risks

Special Issue Reprint

Emerging Topics in Finance and Risk Engineering —In Memory of Peter Carr

Edited by
Dan Pirjol and Lingjiong Zhu

mdpi.com/journal/risks

MDPI

Emerging Topics in Finance and Risk Engineering—In Memory of Peter Carr

Emerging Topics in Finance and Risk Engineering—In Memory of Peter Carr

Dan Pirjol
Lingjiong Zhu

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Editors

Dan Pirjol
School of Business
Stevens Institute of Technology
Hoboken
United States

Lingjiong Zhu
Department of Mathematics
Florida State University
Tallahassee
United States

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Risks* (ISSN 2227-9091) (available at: www.mdpi.com/journal/risks/special_issues/emerging_topics_in_finance_and_risk_engineering_in_memory_of_Peter_Carr).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-2480-9 (Hbk)
ISBN 978-3-7258-2479-3 (PDF)
doi.org/10.3390/books978-3-7258-2479-3

Contents

About the Editors

Dan Pirjol

Dan Pirjol teaches Financial Engineering at the School of Business of the Stevens Institute of Technology in Hoboken, New Jersey.

He received a M.Sc. in Physics from the University of Bucharest and a PhD in Physics from the University of Mainz, Germany. Previously, he worked in the Model Risk and Model Development Sector at JP Morgan, Markit, and Merrill Lynch. His research interests are in applied mathematics, financial engineering, and theoretical physics.

Lingjiong Zhu

Lingjiong Zhu is a Professor at the Department of Mathematics and holds the Thinking Machines Eminent Scholar Chair at Florida State University in Tallahassee, Florida. He received a B.A. from the University of Cambridge in 2008 and a Ph.D. from the NYU's Courant Institute of Mathematical Sciences in 2013. He worked at Morgan Stanley and the University of Minnesota before joining the faculty at Florida State University in 2015. His research interests include applied probability, data science, financial engineering, and operations research.

Preface

This volume includes articles submitted to a Special Issue, which has the same title as the one hosted by *Risks*, and is dedicated to the memory of Peter Carr (1958–2022).

After receiving a Ph.D. in finance from UCLA, Peter Carr worked for eight years as an Assistant Professor of Finance at Cornell University before joining Morgan Stanley as a Vice President in 1996 and later at Bank of America Securities as a Principal in 1999. He was the head of the Quantitative Financial Research Group at Bloomberg from 2003–2010, a Managing Director and Global Head of Market Modeling at Morgan Stanley from 2010 to 2016, before returning to academia in 2016 as the Chair of the Department of Finance and Risk Engineering at NYU's Tandon School of Engineering.

Many of the articles touch on Peter's work. The topics of the articles range from volatility modeling to the use of machine learning for derivatives pricing and are an appropriate reflection of Peter's wide-ranging interests in quantitative finance and risk management. Among the contributors are several former colleagues and collaborators of Peter, and they kindly contributed personal stories related to their interactions with Peter. These were collected into a memorial article, which is included in this reprint.

We are grateful to all the authors who contributed to this volume. Many thanks are due to Mrs. Claire Xiang and Mrs. Sheryl Yin for their tireless editorial efforts, which contributed to the success of this Special Issue. The editors would like to dedicate this volume to the memory of Peter Carr as a teacher, researcher, and friend.

Dan Pirjol and Lingjiong Zhu
Editors

risks

Obituary

In Memory of Peter Carr (1958–2022)

Giuseppe Campolieti [1], Arash Fahim [2], Dan Pirjol [3,*], Harvey Stein [4,†], Tai-Ho Wang [5] and Lingjiong Zhu [2,*]

[1] Department of Mathematics, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON N2L 3C5, Canada; gcampolieti@wlu.ca
[2] Department of Mathematics, Florida State University, Tallahassee, FL 32306, USA; fahim@math.fsu.edu
[3] Stevens Institute of Technology, Hoboken, NJ 07030, USA
[4] Labs Group, Two Sigma, New York, NY 10027, USA; hjstein@gmail.com
[5] Department of Mathematics, Baruch College, The City University of New York, 1 Bernard Baruch Way, New York, NY 10010, USA; tai-ho.wang@baruch.cuny.edu
* Correspondence: dpirjol@stevens.edu (D.P.); zhu@math.fsu.edu (L.Z.)
† Current address: Department of Mathematics, Columbia University, New York, NY 10027, USA.

Abstract: The editors of this special issue and several of the contributing authors have known Peter for a long time. We thought that the special issue will be enriched by adding a few personal notes and recollections about our interactions with Peter.

1. Joe Campolieti—Some Personal Recollections about Peter

I recall the first time I met Peter. It was at an INFORMS Applied Probability conference in NYC in late July of 2001. I was initially introduced to Peter by my colleague, Claudio Albanese who was professor of mathematics at the University of Toronto. At the time, I was teaching at the Masters of Mathematical Finance (MMF) graduate program at the University of Toronto. I taught in that program from 1998 to 2002, before I accepted a tenure-track faculty position at Wilfrid Laurier University as an Associate Professor of Mathematics (and SHARCNET Chair in Financial Mathematics). Prior to meeting Peter, Claudio had mentioned him on a few occasions and always spoke highly of him. Leading up to the 2001 conference, Claudio and I made good progress on the development of some novel so-called solvable models for derivative pricing. We gave talks on this research at the INFORMS 2001 conference. Peter was quite interested in this work as he and Alex Lipton had been working on some ideas related to it. This resulted in a highly referenced paper published in December of 2001 in Risk Magazine, entitled "Black-Scholes goes Hypergeometric" with all four of us as co-authors Albanese et al. (2001). During our stay in NYC, Peter was quick to invite us for dinner at a nice café in the SOHO district. Peter had a great sense of humour and it did not take long to get better acquainted with him. It was very refreshing to see how passionate he was about research and how he was always thinking about new ideas.

I also recall when Peter came to Laurier circa winter 2006. My colleague, Prof. Madhu Kalimipalli, from the Laurier School of Business had invited Peter to give a talk. I made sure to have most of my students in the undergraduate and graduate financial mathematics courses attend his talk. It was an interesting talk that captivated the students. Later that evening, Phelim Boyle (who was also professor at the Laurier School of Business at the time), Madhu and Peter and I had a nice dinner at Sole in Waterloo. Of course, it made for a very interesting conversation. Peter felt at home every time he visited the greater Toronto metropolitan area, thanks to his Canadian roots.

I also recall meeting with Peter at the Bachelier Conference in 2010 in Toronto. It had been a few years since I had seen him. He immediately greeted me with a great smile and enthusiasm. We had a nice chat about what was going on in our lives. He had just rejoined Morgan Stanley at the time. Later in 2011, Laurier began hosting a new series of bi-annual AMMCS conferences. I've been the main co-organizer of the financial mathematics special session within the AMMCS since its inception. In 2013, I invited Peter to give a plenary

check for updates

Citation: Campolieti, Giuseppe, Arash Fahim, Dan Pirjol, Harvey Stein, Tai-Ho Wang, and Lingjiong Zhu. 2024. In Memory of Peter Carr (1958–2022). *Risks* 12: 39. https://doi.org/10.3390/risks12020039

Academic Editor: Steven Haberman

Received: 29 December 2023
Accepted: 18 January 2024
Published: 18 February 2024

talk at the AMMCS and he graciously and enthusiastically accepted. As always, he gave an interesting talk. It was also great to have him participate in the special session. Practically all of the special session speakers joined us for a memorable lunch with Peter. My colleague and collaborator, Prof. Roman Makarov at Laurier, and I had dinner later that evening. We chatted about some new research results. He was impressed with the progress we had made. Later that year Peter generously agreed to write a nice endorsement for a new book I co-authored with Roman ("Financial Mathematics: A Comprehensive Treatment" published in 2014) Campolieti and Makarov (2014).

The last time I met Peter was at the Financial Engineering Conference that took place in early June of 2019. Both of us were given the opportunity to give lengthy back-to-back invited talks in one of the special sessions on mathematical finance. Peter gave a talk on a new model for credit risk. I spoke about spectral expansions for solvable processes with several new results on first passage times, occupation times and other path functionals. This talk allowed me to partly advertise a new book on solvable models that I'm still presently working on completing as it is quite an undertaking. Within that talk I also presented a new structural credit risk model based on a new occupation time hazard model. Peter told me that he really enjoyed my talk and that he was very interested in the general theoretical framework and the applications of the many new analytical formulas I had generated. Who would have thought that this research would soon later be published in Risks within a special issue in memory of Peter in 2022.

He emailed me at the end of December 2020 asking if I had completed my new book and if he could receive a copy of it. Unfortunately, I had not completed it yet. We last corresponded at the very start of January 2021. I mentioned to him that I was taking a six-month sabbatical in the winter of 2021. He said that he would have invited me to NYC if it weren't for the challenges surrounding COVID at the time. A little over a year later, at the start of March 2022, I got the terrible surprising news of his passing. Peter will always be remembered as a great researcher, a great thinker and such a pleasant person to be around. His spirit will live on as he made such a positive impact on many that knew him.

2. Arash Fahim: Peter Carr's Take on Dubins-Schwartz Theorem

For a very long time, I assumed that I know Dubins-Schwartz theorem Dubins and Schwarz (1965). The theorem asserts: Every continuous martingale $M = (M_s)_{s\geq 0}$ can be written as a time-changed Brownian motion $(B_{[M]_s})_{s\geq 0}$, where $[M] = ([M]_s)_{s\geq 0}$ is the (continuous) quadratic variation of M.

One time, when Peter Carr was visiting University of Michigan in Ann Arbor, where I did my postdoc under Erhan Bayraktar's mentorship, I mentioned the theorem during a lunch with Peter and Erhan's group. Peter thought for a minute or two and constructed a bounded function that satisfies Laplace equation as a counter example to the way I describe Dubins-Schwartz theorem. I was feeling embarrassed, and my confidence was shattered. How can a bounded martingale be a time change of Brownian motion which is not bounded. Of course, what Peter did was showing me that the time change can map $[0, \infty)$ into any interval $[0, b)$ that you may imagine. In other words, $[M]$ can be bounded by hitting times of the Brownian motion to some levels, which makes $(B_{[M]_s})_{s\geq 0}$ a bounded martingale.

Apparently, I should have been more careful in learning the theorem, as it is expected from a good mathematician, such as Peter. Thereafter, I challenged any new results I come upon, particularly, by trying to make counter examples. This may have been a normal practice for many, but for me it was an aha moment and a gift from Peter Carr.

3. Dan Pirjol

My first interaction with Peter was sometime in 2011, after giving a talk at Columbia University (on work related to Pirjol (2013)) in the Financial Practitioners seminar ran by Emanuel Derman. I had just moved to JPM and the location in Midtown made it easier to attend seminars, which had started to be held more frequently than during the financial crisis. I did not meet Peter in person at the seminar, but he must have thought that the topic

was sufficiently interesting that he got in touch and told me that he discussed this problem afterwards with the Budapest quant group at Morgan Stanley.

We continued to meet at the NYU seminar organized by Rama Cont and Marco Avellaneda, and at the IAQF and Bloomberg events. I was working on fixed income modeling, and had little overlap with his work which was mostly in equities modeling. I recall attending a talk given by him about the Variance Gamma model, and my ignorance of the topic was such that I believed that the title must refer to one of the Greeks of a variance derivative.

This changed after 2014 when I started working on commodity derivatives. One of the first projects I worked on was a validation study of "Carr randomization", a method he proposed in Carr (1998) for pricing American options on Black-Scholes assets. The bank was considering using this model for pricing and risk management of American options on commodity futures. My job was to find the weak points of the model, so naturally I shot a message to Peter asking what he thinks are the main weaknesses and if possible to suggest possible enhancements. Model developers are usually very reluctant to admit to any shortcomings of their model, but Peter was open about the limitations of his method and also suggested ways to improve on the published version. He also suggested a topic of research—extend his method to the CEV model. However, as we found out later, this had already been done in Wong and Zhao (2010).

We continued to keep in touch after I moved to an academic position. I invited him to give a talk in the Financial Engineering seminar at Stevens, where he talked about one of his last contributions—"Stoptions" and the logistic model for option pricing proposed in Carr and Torricelli (2021). This he delivered with an interesting spin, emphasizing connections to relativity theory, which reflected his wide ranging curiosity and interests about all branches of science. His online seminar broke all records of attendance, with more than 125 viewers, and lasted for one hour and 30 min. He joked after the seminar that "seminars used to be 1.5 h, and I am old school". If a passion for communication, curiosity and insistence on seeing the best in others are "old school" skills, then surely he was the best teacher one could hope for. His lesson will continue to live in all those who knew him.

4. Harvey Stein

In 2001, two years after I moved back to NYC from Tel Aviv, I started attending the Mathematical Finance Seminar at NYU's Courant Institute. It was the premier quantitative finance seminar in NYC, with talks being given by a multitude of famous individuals. Talks were given by Robert Merton, Paul Malliavin, Benoit Mandelbrot, and many other celebrities of the quant world. The research and talks were extremely stimulating and engaging, as were the post-seminar dinners. Of course that was the case, as the seminar was organized by the world renowned quants, Peter Carr and Marco Avellaneda. It was where I met Peter.

Peter had recently left Bank of America and was a visiting professor at NYU. He was an iconic figure at the seminar; always asking questions and engaging the speaker. You didn't need to be familiar with his extensive bibliography to see that he was a leader in the field of quantitative finance; extremely sharp and tremendously knowledgeable. So, in 2003, when I was asked to head and build Bloomberg's global Quantitative Finance R&D group, recruiting Peter to head up my Quantitative Finance Research group was an obvious move. Luckily, Peter was happy to return to industry.

Peter had a talent for viewing everything as an option, and the ability to reason about options economically and mathematically. But, he wasn't just a researcher—he lived quantitative finance. He was constantly in contact with other researchers, discussing their work, inviting them to meetings, holding weekly breakfast seminars, and always very excited about it. He was a whirlwind of activity.

Peter loved applying new mathematical techniques to financial problems. During the Covid pandemic, we corresponded about his work with Doug Costa on viewing optionality as a binary operation. Having an algebraic background, I helped him get a deeper

understanding of the Grothendieck group of a monoid and a canonical representation of that group for the monoid he was working with. In response to a proof of Peter's that used the distributive property of plus over max instead of directly appealing to the definition of max, I told Peter that with proofs like that, he was becoming a real algebraist. That was in December of 2020, to which he responded "That is the nicest thing anyone has said to me this year".

With Peter as the head of the research team, the Quantitative Finance R&D group became a force in quantitative finance. We hired a number of well-known industry leaders, including Pat Hagan, Bruno Dupire, and Bjorn Flesaker. We upgraded all of the Bloomberg option models, making them the industry standard. It was a lot of work, but it was very exciting to be on its forefront, and the group became well-known in the industry. Perhaps it became a little too well-known, as it became common for the banks to poach from the team.

Peter was very excited about building such a strong quantitative research group, and hiring so many high-profile renowned quants. He was especially excited about hiring Bruno. He used to say how great it would be to get Bruno out of retirement. Peter also had great respect for Bruno. In retrospect, given Peter's comments before the interview, I think he might have also been a little intimidated. When interviewing people, to gauge their mathematical depth, I would often ask them to define a Martingale. When they mentioned that the conditional expectation of future values is the current value, I would ask what exactly that meant, and start drilling into details and proofs. Before interviewing Bruno, Peter specifically asked me to ask Bruno these questions, so I did. It didn't go over so well with Bruno. None the less, Bruno and Peter subsequently became close friends.

But all good things ultimately have to end. Senior management couldn't decide whether quant groups should be part of the business unit or within the R&D engineering teams. The group was broken up, and, in 2010, Peter went back to Morgan Stanley. But I will always look back on those times as the halcyon days of quantitative finance at Bloomberg, and reminisce about Peter's (and everyone else's) role in it.

5. Tai-Ho Wang

My first encounter with Peter was a pure coincidence. I came to New York, among one of probably the worst days in history, on 1 September 2001. I was doing my postdoc at NYU's Courant Institute of Mathematical Sciences (CIMS). CIMS has been well-known for years of its tightness in office spaces. I was finally assigned to an office where Peter Laurence (PL) used to work in, who later became one of my main and long term collaborators. PL and I had several joint works on various topics in quantitative finance.

One day in the afternoon while I was working in the office, Peter came to the office, apparently not finding PL but me, and said, "Hi, I am Peter Carr, I am looking for Peter Laurence". Back then not really knowing who he was but vaguely remembered seeing his name somewhere in the building and on some webpage of CIMS, I tried chatting with him a little. I started, "Hi, my name is Tai-Ho Wang. I am from Taiwan". He immediately replied, "Tai-Ho from Taiwan, easy to memorize". My impression on what he said was like "That's right, how come I never thought of that"? Our first conversation basically ended right there since a moment later PL showed up and they went on to other office for discussions. That first impression on him turned out to be the main theme every time I attended Peter's talks or met with him for discussions.

On a fellowship leave, I returned to CIMS in 2006 as a visiting scholar. Up to this point, jointly with PL I gained extensive experiences in applying Lie symmetry to analyze differential equations originated from problems in quantitative finance. Since Peter and PL had been in close contact, knowing from PL that I would be visiting CIMS for a semester, Peter invited me over for a talk at Bloomberg. Peter was very supportive during the talk and pretty interested in the part of results that I presented. In an after-talk conversation, he pointed out to me a relationship between the Girsanov transformation in stochastic analysis and one of the generators of the six dimensional symmetry group for heat equation. When

I finally got down to the bottom of the argument, I was like "That's right, how come I never thought of that"?

Fast forward to 2019, I had joined Baruch College for ten more years. A friend of mine came down to New York and was invited to speak at NYU Tandon. I attended my friend's talk, one the one hand for reunion; on the other hand to meet and say hello to Peter since it had been a while. Peter was so kind to invite me to join their dinner with the speaker and some of his friends and colleagues. In the restaurant, at some point among the multi-threading conversations before the meals were served, Peter came up with this question: Assume under Black-Scholes model with zero interest and dividend rate, what is the distribution for at-the-money Black-Scholes delta at the half-time point prior to expiry? The first reaction that came to my mind was, since Black-Scholes' delta is always between zero and one, a beta distribution. The answer turned out to be yes and no. No, because it is a uniform distribution; yes, because uniform is a special case of beta. This tiny observation was probably not as fancy as his other numerous fascinating works, a good interview question though. However, my impression at that moment was again "Interesting, how come I never thought of that?" It showed Peter had a rare and unique talent in discovering special structures or symmetries hidden behind the models and formulas, which I admire and respect to the highest level.

6. Lingjiong Zhu

I have known Peter for over ten years. I first met him when I was still a PhD student at NYU's Courant Institute of Mathematical Sciences, in the fall of 2012. We met during a reception for the master in financial math program at Courant, where he was an adjunct professor. I got invited to this reception because I was a teaching assistant for Stochastic Calculus, and most of the students of that class were in the financial math program. At the reception, Peter asked me what I was doing for my research and I told him about the Hawkes process papers that I had been writing under the supervision of Prof. S.R.S. Varadhan. He invited me to give a talk to the credit risk group at Morgan Stanley and was so impressed that he asked me to give the same talk to a much larger audience there, which eventually led to my six-month stint at Morgan Stanley in 2013, and gave me a chance to talk to him on an almost daily basis. Peter's enthusiasm for doing research and learning new things had an immense impression on me. He had tremendous energy. He often got up at 4 o'clock in the morning to do research, which was several hours before he came to work!

While at the company, he would continue to work till the evenings. A few times, I would have dinner with him on the 11th floor at the company when everyone else was gone. Peter continued to talk to me on a regular basis and support me academically even after I left Morgan Stanley and moved back to academia, and we started working together on a paper when I was a postdoc at University of Minnesota. I kept close contact with Peter even after I moved to Florida State University in 2015. In February 2017, I had the pleasure of inviting him to visit FSU. We discussed our working paper and got some significant progress during his visit. Both of us gave talks about this work at many conferences and seminars; but it is sad that the paper has not yet been finished. Hopefully I will be able to complete this paper on my own in the future as a way to honor him.

In the summer of 2018, Peter visited China for the first time. He visited Shanghai, my hometown, before flying to Chengdu to visit Southwestern University of Finance and Economics. I happened to be in Shanghai during that summer.

I picked him up at the famous Peace Hotel, and had the pleasure to treat him to lunch at the equally famous, Lv Bo Lang, known for the Shanghai-style dim sum. After lunch, he headed to give a talk at Shanghai Jiao Tong University. The next day, Peter gave a lecture at Fudan University, invited by Jian Sun, a former colleague of ours at Morgan Stanley. I went there as well.

On 25 July, Peter and I met again. We discussed research at a Starbucks near Peace Hotel, before I took him to lunch, this time at Shunfeng Harbour Restaurant. Peter visited

Shanghai again in the summer of 2019. I was also in Shanghai during that summer; but unfortunately we did not meet up this time because Peter got the timing wrong. After all, there was a 12-h time difference!

We used to Skype on Sunday afternoons at 4 pm every other week or sometimes once a month. *Among all the famous people I know in academia, Peter was one of the very few that I could count as a personal friend.* He was not only incredibly smart, but also was super kind and exceedingly generous, especially in terms of helping young colleagues and junior people. Every time I went to New York, I would stop by his office. The first time, I was in his Morgan Stanley's office, and then at his temporary office at Jane Street, and finally at his Brooklyn office at NYU's Tandon School of Engineering. He even invited me to give a talk there once after I finished the visit to Rutgers University. A couple of months before Peter passed, he was invited to give a talk at a machine learning conference targeting quants, and since Peter was not doing much work on machine learning, he instead suggested to the organizers to invite me to give a talk because he knew I had been working on some theoretical machine learning problems. He was such a nice person and always thinking about helping junior people. The last time I spoke to him on Skype was in early January 2022. We agreed to talk on Zoom on 6 February. He didn't show up, and I thought he was too busy and forgot the appointment. When I finally received the bad news from my friends weeks later, it was a complete shock to me. Peter's untimely passing was such a tragedy and a big loss to the community. He will be dearly missed.

Funding: This research received no external funding.

References

Albanese, Claudio, Giuseppe Campolieti, Peter Carr, and Alexander Lipton. 2001. Black-Scholes goes hypergeometric. *Risk* 14: 99–103.

Campolieti, Giuseppe, and Roman N. Makarov. 2014. *Financial Mathematics: A Comprehensive Treatment*. Textbooks in Mathematics. Boca Raton: Chapman and Hall/CRC.

Carr, Peter. 1998. Randomization and the American Put. *The Review of Financial Studies* 11: 597–626. [CrossRef]

Carr, Peter, and Lorenzo Torricelli. 2021. Additive logistic processes in option pricing. *Finance and Stochastics* 25: 689–724. [CrossRef]

Dubins, Lester E., and Gideon Schwarz. 1965. On continuous martingales. *Proceedings of the National Academy of Sciences* 53: 913–16. [CrossRef] [PubMed]

Pirjol, Dan. 2013. Explosive behavior in a log-normal interest rate model. *International Journal on Theoretical and Applied Finance* 16: 1350023. [CrossRef]

Wong, Hoi Ying, and Jing Zhao. 2010. Valuing American options under the CEV model by Laplace-Carson transforms. *Operations Research Letters* 38: 474–81. [CrossRef]

risks

MDPI

Article

Probability Density of Lognormal Fractional SABR Model †

Jiro Akahori [1], Xiaoming Song [2] and Tai-Ho Wang [1,3,*]

[1] Department of Mathematical Sciences, Ritsumeikan University, Noji-Higashi 1-1-1, Kusatsu 525-8577, Shiga, Japan; akahori@se.ritsumei.ac.jp
[2] Department of Mathematics, Drexel University, 32nd and Market Streets, Philadelphia, PA 19096, USA; song@math.drexel.edu
[3] Department of Mathematics, Baruch College, The City University of New York, 1 Bernard Baruch Way, New York, NY 10010, USA
* Correspondence: tai-ho.wang@baruch.cuny.edu
† In memory of Peter Carr, dear friend and inspirational financial engineer.

Abstract: Instantaneous volatility of logarithmic return in the lognormal fractional SABR model is driven by the exponentiation of a correlated fractional Brownian motion. Due to the mixed nature of driving Brownian and fractional Brownian motions, probability density for such a model is less studied in the literature. We show in this paper a bridge representation for the joint density of the lognormal fractional SABR model in a Fourier space. Evaluating the bridge representation along a properly chosen deterministic path yields a small time asymptotic expansion to the leading order for the probability density of the fractional SABR model. A direct generalization of the representation of joint density often leads to a heuristic derivation of the large deviations principle for joint density in a small time. Approximation of implied volatility is readily obtained by applying the Laplace asymptotic formula to the call or put prices and comparing coefficients.

Keywords: asymptotic expansion; lognormal fractional SABR model; mixed fractional Brownian motion; Malliavin calculus; bridge representation

Citation: Akahori, Jiro, Xiaoming Song, and Tai-Ho Wang. 2022. Probability Density of Lognormal Fractional SABR Model. *Risks* 10: 156. https://doi.org/10.3390/risks10080156

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 27 June 2022
Accepted: 20 July 2022
Published: 2 August 2022

1. Introduction

The celebrated Black and Black-Scholes-Merton models have been the benchmark for European options on currency exchange, interest rates, and equities since the inauguration of the trading on financial derivatives. However, empirical evidence has shown that the main drawback of these models is the assumption of constant volatility; the key parameter required in the calculation of option premia under such models. The volatility parameters induced from market data are in fact nonconstant across markets; dubbed as *volatility smile*.

The Stochastic $\alpha\beta\rho$ (SABR hereafter) model, suggested by Hagan, Lesniewski, and Woodward in Hagan et al. (2015), is one of the models, such as local volatility models, stochastic volatility models, and exponential Lévy type of models, etc, that attempts to capture the volatility smile effect. Furthermore, as opposed to local volatility models, in the SABR model the volatility smile moves in the same direction as the underlying with time, see Hagan et al. (2002).

The SABR model is depicted by the following system of stochastic differential equations (SDEs):

$$dF_t = \alpha_t F_t^{\beta} dW_t, \quad F_0 = F, \tag{1}$$

$$d\alpha_t = \nu \alpha_t dZ_t, \quad \alpha_0 = \alpha, \tag{2}$$

with $\beta \in [0,1]$, where F_t denotes the forward price and α_t the instantaneous volatility. W_t and Z_t are correlated Brownian motions with a constant correlation coefficient ρ. The SABR model is at times referred to as lognormal SABR when $\beta = 1$. The SABR formula is an asymptotic expansion for the implied volatilities of call options with various strikes with

small expiry times. For the reader's convenience, we reproduce the SABR formula in the following. Let $\sigma_{BS}(K,\tau)$ be the implied volatility of a vanilla option struck at K and time to expiry τ. The SABR formula states

$$\sigma_{BS}(K,\tau) = \nu \frac{\log(F/K)}{D(\zeta)}\{1 + O(\tau)\} \tag{3}$$

as the time to expiry τ approaches 0. The function D and the parameter ζ involved in (3) are defined respectively as

$$D(\zeta) = \log\left(\frac{\sqrt{1-2\rho\zeta+\zeta^2}+\zeta-\rho}{1-\rho}\right)$$

and

$$\zeta = \begin{cases} \frac{\nu}{\alpha}\frac{F^{1-\beta}-K^{1-\beta}}{1-\beta} & \text{if } \beta \neq 1; \\ \\ \frac{\nu}{\alpha}\log\left(\frac{F}{K}\right) & \text{if } \beta = 1. \end{cases}$$

Generally, the SABR formula is given one order higher, up to order τ. Here we present only the zeroth order for our own purpose.

The geometry of the SABR model is isometrically diffeomorphic to the two-dimensional hyperbolic space, also known as the Poincaré plane. This isometry leads to a derivation of the SABR Formula (3) based on an expression of the heat kernel, known as the McKean kernel, on Poincaré plane. In particular, the lowest order term in (3) has a geometric interpretation. The function D is the geodesic distance from the spot value (F_0, α_0) to the vertical line $F = K$ in the upper half plane $\{(F,\alpha) \in \mathbb{R}^2 : \alpha \geq 0\}$. Hence, the lowest order term in (3) is indeed the ratio between the absolute value of logmoneyness, i.e., $\log(K/F_0)$, and the geodesic distance from (F_0, α_0) to the vertical line $F = K$ in the upper half-plane. We refer readers interested in this topic to Hagan et al. (2015) for more detailed discussions. As expression for heat kernel on hyperbolic space is concerned, Ikeda and Matsumoto in Ikeda and Matsumoto (1999) provided a probabilistic approach and obtained, among other interesting results, a representation for the transition density of hyperbolic Brownian motion, i.e., the heat kernel over the Poincaré plane. See Theorem 2.1 in Ikeda and Matsumoto (1999) for details.

The aforementioned nice isometry between the SABR model and Poincaré plane breaks down if the volatility process, i.e., the α_t process in (1), is driven by a fractional Brownian motion such as the second equation in (6) considered in the paper. Moreover, due to the lack of Markovianity of fractional Brownian motions, thus the nonexistence of the forward and backward Kolmogorov equations, the classical asymptotic expansion approaches, such as the heat kernel or WKB expansion, are no longer applicable. In this regard, the probabilistic approach in Ikeda and Matsumoto (1999) is more applicable and tractable when dealing with processes driven by fractional Brownian motions.

The volatility process is generally conceived as behaving "fractionally" in that the driving noise is a fractional process, e.g., a fractional Brownian motion with a Hurst exponent other than a half. For a far from an exhaustive list, models that attempt to incorporate the fractional feature of volatility include: the ARFIMA model in Granger and Joyeux (1980) and the FIGARCH model Baillie et al. (1996) for discrete-time models; the long memory stochastic volatility model in Comte and Renault (1998) and the affine fractional stochastic volatility model in Comte et al. (2012) for continuous time models. Somewhat on the contrary, in a recent study in Gatheral et al. (2018), the Hurst exponent H is estimated as being less than a half; thereby indicating antipersistency as opposed to the persistency of the volatility process. For a more detailed and in-depth consideration of this issue, we refer interested readers to the discussions in Cont and Das (2022) and Rogers (2019). It is also worth mentioning that generalizations of the Heston model to the fractional version have been considered in El Euch and Rosenbaum (2019) and Guennoun et al. (2018). Heston-related models are usually dealt with via the characteristic or moment-generating

functions. However, in this paper, we take the approach following closely the methodology in Ikeda and Matsumoto (1999). As arbitrage in the modeling is concerned, we remark that, in contrast with the models discussed in for instance Jarrow et al. (2009) and Mishura (2008) within which the underlying prices were assumed driven by fractional Brownian motions, the model considered in the paper is free of arbitrage opportunity since it is the volatility process that is driven by a fractional Brownian motion while the underlying itself is still driven by a (correlated) Brownian motion.

In order to embed the empirically observed fractional feature of the volatility process into the classical SABR model, we suggest in this paper a fractional version of the SABR model as in (6). Modulo a mean-reversion component, this model aligns with the model statistically tested in Gatheral et al. (2018). The main observation in Gatheral et al. (2018) is that in using the square root of the realized variance as a proxy for the instantaneous volatility, the logarithm of the volatility process behaves like a fractional Brownian motion in almost any time scale of frequency. The Hurst exponent H inferred from the time series data is less than a half; indeed, $H \approx 0.1$, see also Cont and Das (2022) and Rogers (2019). This observation of a small Hurst exponent in the volatility process analyzes the model as more technical and challenging from a stochastic analysis point of view. To our knowledge, most of the small time asymptotic expansions for processes driven by fractional Brownian motions have restrictions on the Hurst exponent H of the driving fractional Brownian motion, mostly $H \geq \frac{1}{4}$. One of the advantages of the approach undertaken in the current paper is that it works without restriction on the Hurst exponent H. The key ingredient is a representation in a Fourier space, which we call the bridge representation in Section 2, for the joint density of log spot and volatility, see (9).

A small time asymptotic expansion of the joint density is readily obtained from the bridge representation. The idea is to approximate the conditional expectation in the bridge representation by a judiciously chosen deterministic path since, conditioned on the initial and terminal points, at each point in time a Gaussian process will not wander too far away from its expectation. As long as an asymptotic expansion for the density of the underlying asset is available, obtaining an expansion for implied volatility is almost straightforward by basically comparing the coefficients with a similar expansion obtained by using the lognormal density on the Black or the Black-Scholes-Merton side.

The methodology of deriving the bridge representation (9) can be generalized directly to obtain a bridge representation for the joint density multiple times; hence inducing a representation for finite-dimensional distributions of the fractional SABR model, see Theorem 4. Based on this bridge representation for finite-dimensional distributions, Section 5 is devoted to a heuristic yet appealing derivation of the large deviations principle for the joint density of the fractional SABR model in small time. This large deviations principle in a sense can be regarded as defining a "geodesic distance" over the fractional SABR plane since, as we shall show in Section 5, it recovers the energy functional on the Poincaré plane when $H = \frac{1}{2}$. We leave the rigorous proof of the large deviations principle in future work. An immediate consequence of this large deviation principle is the fractional SABR formula (to the lowest order) (26) which recovers the classical SABR formula when $H = \frac{1}{2}$. The fractional SABR Formula (26) pertains to the guiding principle that the lowest order term in the implied volatility expansion is given by the ratio between the absolute value of the logmoneyness and the geodesic distance to the vertical line $F = K$.

The rest of the paper is organized as follows. The fractional SABR model is specified and the bridge representation for joint density is shown in Section 2. Sections 3 and 4 provide small time asymptotic expansions of the joint density and of the implied volatilities respectively. Section 5 presents the bridge representation for finite-dimensional distributions and the large deviations principle. Finally, the paper concludes in Section 6 with discussions.

2. Model Specification

Throughout the text, $B = \{B_t, t \geq 0\}$ and $W = \{W_t, t \geq 0\}$ denote two independent standard Brownian motions defined on the filtered probability space $(\Omega, \mathcal{F}_t, \mathbb{P})$ satisfying the usual conditions. Let $B^H = \{B_t^H, t \geq 0\}$ be a fractional Brownian motion with Hurst exponent $H \in (0,1)$ generated by B (see Decreusefond and Üstünel 1999), i.e.,

$$B_t^H = \int_0^t K_H(t,s) dB_s,$$

where K_H is the Molchan-Golosov kernel

$$K_H(t,s) = c_H(t-s)^{H-\frac{1}{2}} F\left(H-\frac{1}{2}, \frac{1}{2}-H, H+\frac{1}{2}; 1-\frac{t}{s}\right)\mathbf{1}_{[0,t]}(s), \tag{4}$$

with $c_H = \left[\frac{2H\Gamma\left(\frac{3}{2}-H\right)}{\Gamma(2-2H)\Gamma\left(H+\frac{1}{2}\right)}\right]^{1/2}$ and F is the Gauss hypergeometric function. Also, the autocovariance function of a fractional Brownian motion is denoted by $R(t,s)$ and defined as

$$R(t,s) = \mathbb{E}(B_t^H B_s^H) = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right). \tag{5}$$

Lastly, we assume that all random variables and stochastic processes are defined on $(\Omega, \mathcal{F}_t, \mathbb{P})$.

2.1. The Model

We study the following lognormal fractional SABR (fSABR hereafter) model in risk-neutral probability (for simplicity, interest and dividend rates are both assumed zero):

$$\begin{cases} S_t = s_0 + \int_0^t \alpha_r S_r(\rho dB_r + \bar{\rho} dW_r), \\ \\ \alpha_t = \alpha_0 e^{\nu B_t^H}, \end{cases} \tag{6}$$

where s_0 and α_0 are the given time zero (current observed) values for the processes S and α respectively, $\rho \in (-1,1)$ and $\bar{\rho} = \sqrt{1-\rho^2}$.

In other words, the underlying price S_t follows a stochastic volatility model with the (instantaneous) volatility process α_t, and α_t is given by the exponentiation of a correlated fractional Brownian motion. The main purpose of this section is to derive the bridge representations (9) and (13) for the joint densities of (S_t, α_t). The bridge representation is the crucial starting line in obtaining expansions and approximations of the joint densities to be discussed in Section 3.

By making a change in variables

$$X_t = \ln S_t, \qquad Y_t = \alpha_t,$$

the system (6) can be written more explicitly as

$$\begin{cases} X_t = x_0 + y_0 \int_0^t e^{\nu B_s^H}(\rho dB_s + \bar{\rho} dW_s) - \frac{y_0^2}{2}\int_0^t e^{2\nu B_s^H} ds, \\ \\ Y_t = y_0 e^{\nu B_t^H}, \end{cases} \tag{7}$$

where $x_0 = \ln s_0$ and $y_0 = \alpha_0$.

2.2. Malliavin Calculus with Respect to Brownian Motion

We provide some preliminaries on Malliavin calculus with respect to the two Brownian motions B and W in this subsection. We refer the reader to Hu (2017) and Nualart (2006) for more details.

For any fixed $T > 0$, let $\mathbf{H} = L^2([0,T])$ be the separable Hilbert space of all square-integrable real-valued functions on the interval $[0,T]$ with scalar product denoted by $\langle \cdot, \cdot \rangle_{\mathbf{H}}$. The norm of an element $h \in \mathbf{H}$ will be denoted by $\|h\|_{\mathbf{H}}$. For any $h \in \mathbf{H}$, we put $W(h) = \int_0^T h(t) dW_t$ and $B(h) = \int_0^T h(t) dB_t$.

For any m, $n \in \mathbb{N}$, denote by $C_p^\infty(\mathbb{R}^{m+n})$ the set of all infinitely differentiable functions $g : \mathbb{R}^{m+n} \to \mathbb{R}$ such that g and all of its partial derivatives have polynomial growth. We make use of the notation $\partial_i g = \frac{\partial g}{\partial x_i}$ whenever $g \in C^1(\mathbb{R}^{m+n})$.

Let $\mathcal{S}$ denote the class of smooth and cylindrical random variables such that a random variable $F \in \mathcal{S}$ has the form

$$F = g(W(h_1), \ldots, W(h_m), B(k_1), \ldots, B(k_n)), \tag{8}$$

where g belongs to $C_p^\infty(\mathbb{R}^{m+n})$, $h_1, \ldots, h_m$ and $k_1, \ldots, k_n$ are in $\mathbf{H}$, and $m, n \in \mathbb{N}$.

For a smooth and cylindrical random variable F of the form (8), its Malliavin derivative with respect to W is the $\mathbf{H}$-valued random variable given by

$$D_t^1 F = \sum_{i=1}^m \partial_i g(W(h_1), \ldots, W(h_m), B(k_1), \ldots, B(k_n)) h_i(t),\ t \in [0, T],$$

and respectively its Malliavin derivative with respect to B is given by

$$D_t^2 F = \sum_{i=1}^n \partial_{m+i} g(W(h_1), \ldots, W(h_m), B(k_1), \ldots, B(k_n)) k_i(t),\ t \in [0, T].$$

For any $p \geq 1$, we will denote the domain of D in $L^p(\Omega)$ by $\mathbb{D}^{1,p}$, meaning that $\mathbb{D}^{1,p}$ is the closure of the class of smooth and cylindrical random variables $\mathcal{S}$ with respect to the norm

$$\|F\|_{1,p} = \left(\mathbb{E}|F|^p + \mathbb{E}\Big(\|D^1 F\|_{\mathbf{H}}^2 + \|D^2 F\|_{\mathbf{H}}^2 \Big)^{\frac{p}{2}} \right)^{\frac{1}{p}}.$$

We tailor Theorem 2.1.2 in Nualart (2006) to the following lemma which yields a result on the absolute continuity of the law of a random vector with respect to the Lebesgue measure.

Lemma 1. *Let $F = (F_1, F_2)$ be a random vector in $\mathbb{D}^{1,2}$. If the Malliavin matrix $\gamma := (\langle D^1 F_i, D^1 F_j \rangle_{\mathbf{H}} + \langle D^2 F_i, D^2 F_j \rangle_{\mathbf{H}})_{1 \leq i,j \leq 2}$ of F is invertible a.s. Then the law of F is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. Consequently, the joint density of the random variables (F_1, F_2) exists.*

2.3. Bridge Representation for the Joint Density

In this subsection, we show the existence of the joint density of (X_t, Y_t) for any $t > 0$ by using Malliavin calculus. We also give a bridge representation for the joint density by adapting the methodology introduced in Ikeda and Matsumoto Ikeda and Matsumoto (1999).

Theorem 1. *For any $t > 0$, the law of (X_t, Y_t) satisfying* (7) *is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. Moreover, the joint probability density $p(t; x, y)$ of (X_t, Y_t) has the following bridge representation*

$$\begin{aligned} & p(t; x, y) \\ = \ & \frac{1}{y\sqrt{2\pi\nu^2 t^{2H}}} e^{-\frac{(\ln(y/y_0))^2}{2\nu^2 t^{2H}}} \times \\ & \frac{1}{2\pi} \int_{\mathbb{R}} \mathbb{E}\left[e^{i\left(x - x_0 - \rho y_0 \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2} \right)\xi} e^{-\frac{\bar{\rho}^2 y_0^2 v_t \xi^2}{2}} \,\middle|\, B_t^H = \frac{\ln(y/y_0)}{\nu} \right] d\xi. \end{aligned} \tag{9}$$

where $v_t = \int_0^t e^{2\nu B_s^H} ds$ and $i = \sqrt{-1}$.

Remark 1. *The bridge representation* (9) *can be regarded as a generalization of the well-known McKean kernel, namely, the classical heat kernel over a 2-dimensional hyperbolic space. For reader's reference, the McKean kernel $p_{\mathbb{H}^2}(t; x, y)$ reads*

$$p_{\mathbb{H}^2}(t;x,y) = \frac{\sqrt{2}e^{-t/8}}{(2\pi t)^{3/2}} \int_d^\infty \frac{\xi e^{-\xi^2/2t}}{\sqrt{\cosh\xi - \cosh d}} d\xi,$$

where $d = d(x,y;x_0,y_0)$ *is the geodesic distance from* (x,y) *to* (x_0,y_0)*. The geodesic distance satisfies* $\cosh d(x,y;x_0,y_0) = \frac{(x-x_0)^2+y^2+y_0^2}{2yy_0}$*. Note that the McKean kernel is a density with respect to the Riemannian volume form* $\frac{1}{y^2}dxdy$*. Indeed, in the case where* $H = \frac{1}{2}$*,* $\nu = 1$ *and* $\rho = 0$*, Ikeda-Matsumoto in Ikeda and Matsumoto (1999) showed how to recover the McKean kernel from* (9)*. See also Cheng and Wang (2018) for a different representation in terms of a Bessel bridge for the hyperbolic heat kernel.*

Proof. Notice that we can rewrite (7) as

$$\begin{cases} X_t = x_0 + \int_0^t Y_s(\rho dB_s + \bar{\rho} dW_s) - \frac{1}{2}\int_0^t Y_s^2 ds, \\ \\ Y_t = y_0 e^{\nu B_t^H}. \end{cases}$$

Now we fix any $T \geq t$. Then according to Sections 2.2 and 5.2 in Nualart (2006), the Malliavin derivatives of X_t and Y_t are given as follows

$$\begin{aligned} D_\theta^1 Y_t &= 0, \\ D_\theta^2 Y_t &= y_0 \nu e^{\nu B_t^H} K_H(t,\theta)\mathbf{1}_{[0,t]}(\theta) \end{aligned}$$

and

$$\begin{aligned} D_\theta^1 X_t &= \bar{\rho} Y_\theta \mathbf{1}_{[0,t]}(\theta) = \bar{\rho} y_0 e^{\nu B_\theta^H}\mathbf{1}_{[0,t]}(\theta), \\ D_\theta^2 X_t &= \left(\rho Y_\theta + \int_\theta^t \rho D_\theta^2 Y_s dB_s + \int_\theta^t \bar{\rho} D_\theta^2 Y_s dW_s - \int_\theta^t Y_s D_\theta^2 Y_s ds\right)\mathbf{1}_{[0,t]}(\theta) \\ &= \left(\rho y_0 e^{\nu B_\theta^H} + \rho y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dB_s + \bar{\rho} y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dW_s\right)\mathbf{1}_{[0,t]}(\theta) \\ &\quad - y_0^2 \nu \int_\theta^t e^{2\nu B_s^H} K_H(s,\theta) ds\, \mathbf{1}_{[0,t]}(\theta). \end{aligned}$$

Thus, the Malliavin matrix γ of (X_t, Y_t) is given by

$$\gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix},$$

where

$$\begin{aligned} \gamma_{11} &= \int_0^t (D_\theta^1 X_t)^2 d\theta + \int_0^t (D_\theta^2 X_t)^2 d\theta \\ &= \int_0^t \bar{\rho}^2 y_0^2 e^{2\nu B_\theta^H} d\theta + \int_0^t \Bigg(\rho y_0 e^{\nu B_\theta^H} + \rho y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dB_s \\ &\qquad + \bar{\rho} y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dW_s - y_0^2 \nu \int_\theta^t e^{2\nu B_s^H} K_H(s,\theta) ds\Bigg)^2 d\theta, \end{aligned}$$

$$\begin{aligned} \gamma_{12} = \gamma_{21} &= \int_0^t D_\theta^1 X_t D_\theta^1 Y_t d\theta + \int_0^t D_\theta^2 X_t D_\theta^2 Y_t d\theta \\ &= y_0 \nu e^{\nu B_t^H} \int_0^t K_H(t,\theta)\Bigg(\rho y_0 e^{\nu B_\theta^H} + \rho y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dB_s \\ &\qquad + \bar{\rho} y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dW_s - y_0^2 \nu \int_\theta^t e^{2\nu B_s^H} K_H(s,\theta) ds\Bigg) d\theta, \end{aligned}$$

and

$$\begin{aligned}\gamma_{22} &= \int_0^t (D_\theta^1 Y_t)^2 d\theta + \int_0^t (D_\theta^2 Y_t)^2 d\theta \\ &= y_0^2 \nu^2 e^{2\nu B_t^H} \int_0^t K_H(t,\theta)^2 d\theta.\end{aligned}$$

Then it follows from the Cauchy-Schwarz inequality that almost surely

$$\begin{aligned}\gamma_{12}^2 &< y_0^2 \nu^2 e^{2\nu B_t^H} \int_0^t K_H(t,\theta)^2 d\theta \times \\ &\quad \int_0^t \Big(\rho y_0 e^{\nu B_\theta^H} + \rho y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dB_s \\ &\qquad + \bar{\rho} y_0 \nu \int_\theta^t e^{\nu B_s^H} K_H(s,\theta) dW_s - y_0^2 \nu \int_\theta^t e^{2\nu B_s^H} K_H(s,\theta) ds \Big)^2 d\theta \\ &\leq \gamma_{22} \cdot \gamma_{11},\end{aligned}$$

which implies that the Malliavin matrix γ is invertible a.s. Hence, by Lemma 1 the law of (X_t, Y_t) is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$.

Next, we calculate the joint probability density $p(t;x,y)$ of (X_t, Y_t) as follows. For any bounded and continuous function f defined on $\mathbb{R}^2$, we have

$$\begin{aligned}&\mathbb{E}[f(X_t, Y_t)] \\ &= \mathbb{E}\left[f\left(x_0 + y_0 \int_0^t e^{\nu B_s^H} (\rho dB_s + \bar{\rho} dW_s) - \frac{y_0^2 v_t}{2}, y_0 e^{\nu B_t^H} \right) \right]. \end{aligned} \tag{10}$$

Note that conditioned on $\mathcal{F}_t^B$, $y_0 \bar{\rho} \int_0^t e^{\nu B_s^H} dW_s$ is normally distributed since W and B^H are independent. Moreover,

$$\begin{aligned}&\mathbb{E}\left[y_0 \bar{\rho} \int_0^t e^{\nu B_s^H} dW_s \middle| \mathcal{F}_t^B \right] = 0, \\ &\mathbb{E}\left[\left(y_0 \bar{\rho} \int_0^t e^{\nu B_s^H} dW_s \right)^2 \middle| \mathcal{F}_t^B \right] = y_0^2 \bar{\rho}^2 \int_0^t e^{2\nu B_s^H} ds = y_0^2 \bar{\rho}^2 v_t.\end{aligned}$$

From (10), it follows by conditioning on $\mathcal{F}_t^B$ that

$$\begin{aligned}&\mathbb{E}[f(X_t, Y_t)] \\ &= \mathbb{E}\left[\mathbb{E}\left[f\left(x_0 + y_0 \bar{\rho} \int_0^t e^{\nu B_s^H} dW_s + y_0 \rho \int_0^t e^{\nu B_s^H} dB_s - \frac{y_0^2 v_t}{2}, y_0 e^{\nu B_t^H} \right) \middle| \mathcal{F}_t^B \right] \right] \\ &= \mathbb{E}\left[\int \left\{ f\left(x_0 + \xi + y_0 \rho \int_0^t e^{\nu B_s^H} dB_s - \frac{y_0^2 v_t}{2}, y_0 e^{\nu B_t^H} \right) \frac{e^{-\frac{\xi^2}{2 y_0^2 \bar{\rho}^2 v_t}}}{\sqrt{2\pi y_0^2 \bar{\rho}^2 v_t}} \right\} d\xi \right] \\ &= \mathbb{E}\left[\int \left\{ \frac{1}{\sqrt{2\pi y_0^2 \bar{\rho}^2 v_t}} f\left(x, y_0 e^{\nu B_t^H} \right) e^{-\frac{\left(x - x_0 - y_0 \rho \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2} \right)^2}{2 y_0^2 \bar{\rho}^2 v_t}} \right\} dx \right] \\ &= \int_{\mathbb{R}^2} f(x,y) \mathbb{E}\left[\frac{1}{\sqrt{2\pi y_0^2 \bar{\rho}^2 v_t}} e^{-\frac{\left(x - x_0 - y_0 \rho \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2} \right)^2}{2 y_0^2 \bar{\rho}^2 v_t}} \middle| B_t^H = \frac{\ln(y/y_0)}{\nu} \right] \\ &\qquad \times \frac{1}{y\sqrt{2\pi \nu^2 t^{2H}}} e^{-\frac{(\ln y - \ln y_0)^2}{2\nu^2 t^{2H}}} dx\, dy. \end{aligned} \tag{11}$$

By using the identity

$$e^{-\frac{u^2}{2y_0^2\bar{\rho}^2 v_t}} = \sqrt{\frac{y_0^2\bar{\rho}^2 v_t}{2\pi}} \int_{\mathbb{R}} e^{iu\xi} e^{-\frac{y_0^2\bar{\rho}^2 v_t \xi^2}{2}} d\xi$$

and letting $u = x - x_0 - \rho y_0 \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}$, we have

$$\begin{aligned} & \frac{1}{\sqrt{2\pi y_0^2\bar{\rho}^2 v_t}} e^{-\frac{1}{2y_0^2\bar{\rho}^2 v_t}\left(x - x_0 - y_0\rho \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}\right)^2} \\ = & \frac{1}{2\pi} \int e^{i\left(x - x_0 - \rho y_0 \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}\right)\xi} e^{-\frac{y_0^2\bar{\rho}^2 v_t \xi^2}{2}} d\xi. \end{aligned} \tag{12}$$

Plugging (12) into the right-hand side of (11), we get

$$\begin{aligned} & \mathbb{E}[f(X_t, Y_t)] \\ = & \frac{1}{y\sqrt{2\pi\nu^2 t^{2H}}} \frac{1}{2\pi} \int_{\mathbb{R}^2} f(x,y) e^{-\frac{\left(\ln(y/y_0)\right)^2}{2\nu^2 t^{2H}}} \times \\ & \int_{\mathbb{R}} \mathbb{E}\left[e^{i\left(x - x_0 - \rho y_0 \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}\right)\xi} e^{-\frac{y_0^2\bar{\rho}^2 v_t \xi^2}{2}} \middle| B_t^H = \frac{\ln(y/y_0)}{\nu} \right] d\xi\, dx\, dy. \end{aligned}$$

Finally, we end up with the following bridge representation of the density (9). □

By transforming back to the original variables $(s, a) = (e^x, y)$, we obtain a bridge representation for the joint density $q(t; s, a)$ of (S_t, α_t) in (6).

Corollary 1. *The joint density $q(t; s, a)$ of the lognormal fractional SABR model* (6) *has the following bridge representation*

$$\begin{aligned} q(t; s, a) = & \frac{e^{-\frac{(\ln(a/a_0))^2}{2\nu^2 t^{2H}}}}{a\sqrt{2\pi\nu^2 t^{2H}}} \frac{1}{2\pi s} \\ & \times \int_{\mathbb{R}} \left(\frac{s}{s_0}\right)^{i\xi} \mathbb{E}\left[e^{i\left(-\rho \int_0^t a_0 e^{\nu B_s^H} dB_s + \frac{a_0^2 v_t}{2}\right)\xi} e^{-\frac{\bar{\rho}^2 a_0^2 v_t}{2}\xi^2} \middle| B_t^H = \frac{\ln(a/a_0)}{\nu} \right] d\xi. \end{aligned} \tag{13}$$

3. Expansion Around Deterministic Path

To gain more intuition and, in particular, a more practical form for applications in obtaining approximations of implied volatility, this section is devoted to deriving an expansion to the lowest order of the bridge representation (9) around a properly chosen deterministic path. The expansion will be shown useful in deriving a small time approximation for implied volatility in Section 4.

Recall that the joint density p of (X_t, Y_t) has the representation given in (9) as

$$\begin{aligned} & p(t; x, y) \\ = & \frac{1}{y\sqrt{2\pi\nu^2 t^{2H}}} e^{-\frac{\left(\ln(y/y_0)\right)^2}{2\nu^2 t^{2H}}} \times \\ & \frac{1}{2\pi} \int_{\mathbb{R}} \mathbb{E}\left[e^{i\left(x - x_0 - \rho y_0 \int_0^t e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}\right)\xi} e^{-\frac{\bar{\rho}^2 y_0^2 v_t \xi^2}{2}} \middle| B_t^H = \frac{\ln(y/y_0)}{\nu} \right] d\xi. \end{aligned}$$

Let us start with a few naïve calculations, as follows. We expand the above conditional expectation around the deterministic path m_s, for $0 \le s \le t$, that is determined by the conditional expectation of B_s^H given its terminal point $B_t^H = \frac{\ln(y/y_0)}{\nu}$. Precisely,

$$m_s := \mathbb{E}\left[B_s^H \middle| B_t^H = \frac{\ln(y/y_0)}{\nu} \right] = R\Big(1, \frac{s}{t}\Big) \frac{\ln(y/y_0)}{\nu},$$

where R is defined in (5). By Taylor's expansion, we have, for $n \ge 0$,

$$\begin{aligned} & e^{-i\rho\xi \int_0^t y_0 e^{\nu B_s^H} dB_s} e^{-\frac{1}{2}(\bar{\rho}^2\xi - i)\xi \int_0^t y_0^2 e^{2\nu B_s^H} ds} \\ \approx\ & e^{-i\rho\xi \int_0^t y_0 e^{\nu m_s} dB_s} e^{-\frac{1}{2}(\bar{\rho}^2\xi - i)\xi \int_0^t y_0^2 e^{2\nu m_s} ds} \times \\ & \sum_{k,\ell=0}^{n} \frac{(-i\rho\xi)^k}{k!} \left\{ \int_0^t y_0 \Big(e^{\nu B_s^H} - e^{\nu m_s} \Big) dB_s \right\}^k \times \\ & \frac{1}{\ell!} \left\{ -\frac{1}{2}(\bar{\rho}^2\xi - i)\xi \int_0^t y_0^2 \Big(e^{2\nu B_s^H} - e^{2\nu m_s} \Big) ds \right\}^\ell. \end{aligned}$$

Thus, even for obtaining a naïve expansion, we shall need a systematic way of computing the conditional expectations of the form, for either $k \ge 1$ or $\ell \ge 1$,

$$\mathbb{E}\left[e^{-i\rho\xi \int_0^t y_0 e^{\nu m_s} dB_s} \left\{ \int_0^t \Big(e^{\nu B_s^H} - e^{\nu m_s} \Big) dB_s \right\}^k \left\{ \int_0^t \Big(e^{2\nu B_s^H} - e^{2\nu m_s} \Big) ds \right\}^\ell \middle| B_t^H = \frac{\ln(y/y_0)}{\nu} \right],$$

which is pretty complicated if not impossible. Nevertheless, as far as the leading order is concerned, small-time expansion of the joint density p to the lowest order (i.e., $k = \ell = 0$) is still manageable. The result is summarized in the following theorem.

In the following sequel, for simplification of the notation, we use $\mathbb{E}_{\frac{\eta}{\nu}}[\cdot]$ to denote $\mathbb{E}\big[\cdot | B_t^H = \frac{\eta}{\nu}\big]$, where $\eta = \ln(y/y_0)$. A function g is denoted by $g(t) = O(t^a)$ as $t \to 0^+$ if it satisfies

$$\limsup_{t \to 0^+} \frac{|g(t)|}{t^a} < \infty.$$

Theorem 2. *The joint probability density p of the process (X_t, Y_t) satisfying* (7) *has the following asymptotic to the lowest order*

$$\begin{aligned} & p(t; x, y) \qquad (14) \\ = \ & \frac{1}{2\pi} \frac{1}{y\sqrt{\nu^2 t^{2H}}} e^{-\frac{\eta^2}{2\nu^2 t^{2H}}} \frac{1}{y_0\sqrt{t\psi(\eta)}} e^{-\frac{1}{2y_0^2\psi(\eta)}\left(\frac{x - x_0}{\sqrt{t}} + \frac{y_0^2\sqrt{t}}{2} C_{eR}(\eta) - \rho y_0 t^{-H} C_{RK}(\eta) \frac{\eta}{\nu} \right)^2} \Big(1 + O\Big(\sqrt{t}\Big) \Big), \end{aligned}$$

where

$$\begin{aligned} C_{RK}(\eta) &:= \int_0^1 e^{R(1,u)\frac{\eta}{\nu}} K_H(1,u) du, \\ C_{eR}(\eta) &:= \int_0^1 e^{2R(1,u)\frac{\eta}{\nu}} du, \\ \psi(\eta) &:= C_{eR}(\eta) - \rho^2 C_{RK}^2(\eta). \end{aligned}$$

Proof. To the lowest order, p is given by

$$\begin{aligned} & p(t; x, y) \\ = \ & \frac{e^{-\frac{\eta^2}{2\nu^2 t^{2H}}}}{y\sqrt{2\pi\nu^2 t^{2H}}} \frac{1}{2\pi} \int_{\mathbb{R}} e^{i(x - x_0)\xi} e^{-\frac{1}{2}(\bar{\rho}^2\xi - i)\xi \int_0^t y_0^2 e^{2\nu m_s} ds} \mathbb{E}_{\frac{\eta}{\nu}}\left[e^{-i\rho\xi \int_0^t y_0 e^{\nu m_s} dB_s} \right] d\xi. \qquad (15) \end{aligned}$$

We consider the conditional expectation in the above expression. Note that $\int_0^t e^{\nu m_s} dB_s$ and B_t^H are jointly Gaussian. We apply the following identity to evaluate the conditional expectation: if X and Y are jointly normal with mean 0, we can decompose X as

$$X = \frac{\mathrm{cov}(X,Y)}{V(Y)} Y + \sqrt{\frac{V(X)V(Y) - \mathrm{cov}(X,Y)^2}{V(Y)}} Z,$$

where Y and Z are independent and Z is standard normal. Hence,

$$\mathbb{E}[f(X)|Y=y] = \mathbb{E}\left[f\left(\frac{\mathrm{cov}(X,Y)}{V(Y)} y + \sqrt{\frac{V(X)V(Y) - \mathrm{cov}(X,Y)^2}{V(Y)}} Z \right) \right].$$

In our case, $X = \int_0^t e^{\nu m_s} dB_s$ and $Y = B_t^H$, hence

$$\begin{aligned}
&V(X) = \int_0^t e^{2\nu m_s} ds = t \int_0^1 e^{2R(1,u)\eta} du = C_{eR}(\eta)\, t, \\
&V(Y) = t^{2H}, \\
&\mathrm{cov}(X,Y) = \int_0^t e^{\nu m_s} K_H(t,s) ds = t^{H+\frac{1}{2}} \int_0^1 e^{R(1,u)\eta} K_H(1,u) du = C_{RK}(\eta)\, t^{H+\frac{1}{2}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathbb{E}_{\frac{\eta}{\nu}}\left[e^{-i\rho\xi \int_0^t y_0 e^{\nu m_s} dB_s} \right] \\
= \;& e^{-i\rho\xi y_0 t^{\frac{1}{2}-H} C_{RK}(\eta) \frac{\eta}{\nu}} \mathbb{E}\left[e^{-i\rho\xi y_0 \left\{ \sqrt{t}\sqrt{C_{eR}(\eta) - C_{RK}^2(\eta)} \right\} Z} \right] \\
= \;& \exp\left[-i\rho\xi y_0 t^{\frac{1}{2}-H} C_{RK}(\eta) \frac{\eta}{\nu} - \frac{\rho^2\xi^2 y_0^2 t}{2} \left\{ C_{eR}(\eta) - C_{RK}^2(\eta) \right\} \right].
\end{aligned}$$

Thus, by substituting the above expression into (15), we obtain

$$\begin{aligned}
& p(t;x,y) \\
= \;& \frac{1}{2\pi} \frac{1}{y\sqrt{\nu^2 t^{2H}}} e^{-\frac{\eta^2}{2\nu^2 t^{2H}}} \times \\
& \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i(x-x_0)\xi} e^{-\frac{1}{2}(\bar{\rho}^2\xi - i)\xi t y_0^2 C_{eR}(\eta)} e^{-i\rho\xi y_0 t^{\frac{1}{2}-H} C_{RK}(\eta)\frac{\eta}{\nu} - \frac{\rho^2\xi^2 y_0^2 t}{2}\left\{ C_{eR}(\eta) - C_{RK}^2(\eta) \right\}} d\xi \\
= \;& \frac{1}{2\pi} \frac{1}{y\sqrt{\nu^2 t^{2H}}} e^{-\frac{\eta^2}{2\nu^2 t^{2H}}} \times \\
& \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\left(x - x_0 + \frac{y_0^2 t}{2} C_{eR}(\eta) - \rho y_0 t^{\frac{1}{2}-H} C_{RK}(\eta)\frac{\eta}{\nu} \right)\xi} e^{-\frac{\xi^2 y_0^2 t}{2}\left\{ C_{eR}(\eta) - \rho^2 C_{RK}^2(\eta) \right\}} d\xi \\
= \;& \frac{1}{2\pi} \frac{1}{y\sqrt{\nu^2 t^{2H}}} e^{-\frac{\eta^2}{2\nu^2 t^{2H}}} \frac{1}{y_0\sqrt{t\psi(\eta)}} e^{-\frac{1}{2y_0^2\psi(\eta)}\left(\frac{x-x_0}{\sqrt{t}} + \frac{y_0^2\sqrt{t}}{2} C_{eR}(\eta) - \rho y_0 t^{-H} C_{RK}(\eta)\frac{\eta}{\nu} \right)^2}.
\end{aligned} \tag{16}$$

We postpone the detailed error analysis to Appendix A.1 in the Appendix A. □

Remark 2. *We remark that in the logarithmic scale,* (14) *can be expressed in a more concise form as*

$$\begin{aligned}
& \ln p(t;x,y) \\
= \;& -\frac{1}{2t^{2H}} \left[\frac{\eta^2}{\nu^2} + \frac{1}{y_0^2\psi(\eta)} \left(\frac{x-x_0}{t^{\frac{1}{2}-H}} + \frac{y_0^2 t^{H+\frac{1}{2}}}{2} C_{eR}(\eta) - \rho y_0 C_{RK}(\eta)\frac{\eta}{\nu} \right)^2 \right] + O(\ln t).
\end{aligned}$$

Remark 3. *In the case that* $\nu = 1$, $\rho = 0$, *and* $H = \frac{1}{2}$, *we have*

$$C_{eR}(\eta) = \int_0^1 e^{2R(1,u)\eta} du = \int_0^1 e^{2u\eta} du = \frac{1}{2\eta}(e^{2\eta} - 1) = \frac{y^2 - y_0^2}{2\eta y_0^2}.$$

Then (14) *reduces to*

$$\begin{aligned} & \frac{1}{2\pi} \times \frac{1}{y\sqrt{t}} e^{-\frac{\eta^2}{2t}} \times \frac{1}{y_0\sqrt{C_{eR}(\eta)t}} e^{-\frac{1}{2y_0^2 C_{eR}(\eta)t}\left(x - x_0 + \frac{y_0^2 t}{2} C_{eR}(\eta)\right)^2} \\ = & \frac{1}{2\pi} \frac{1}{y\sqrt{t}} e^{-\frac{\eta^2}{2t}} \frac{1}{y_0\sqrt{C_{eR}(\eta)t}} e^{-\frac{(x-x_0)^2}{2y_0^2 C_{eR}(\eta)t}} e^{-\frac{x-x_0}{2}} (1 + O(t)) \\ = & \frac{1}{2\pi t} e^{-\frac{1}{2t}\left[\eta^2 + \frac{2\eta(x-x_0)^2}{y^2 - y_0^2}\right]} \frac{e^{-\frac{x-x_0}{2}}}{y y_0 \sqrt{C_{eR}(\eta)}} (1 + O(t)). \end{aligned} \tag{17}$$

Notice that in this case (X_t, Y_t) *represents the Brownian motion in the hyperbolic plane whose transition density* $p_{\mathbb{H}}$ *(with respect to the Riemannian area measure) has the leading term in small time asymptotic as*

$$p_{\mathbb{H}}(t; x, y) = \frac{1}{2\pi t} e^{-\frac{d^2(x,y;x_0,y_0)}{2t}} (1 + O(t)),$$

where d *denotes the geodesic distance between* (x, y) *and* (x_0, y_0) *in the hyperbolic plane. For reader's reference, the hyperbolic cosine of the geodesic distance* $d(x, y; x_0, y_0)$ *has the closed form expression*

$$\cosh d(x, y; x_0, y_0) = \frac{(x - x_0)^2 + y_0^2 + y^2}{2y_0 y}.$$

Thus, in a sense the following function in (17)

$$\tilde{d}(x, y; x_0, y_0) := \sqrt{\eta^2 + \frac{2\eta}{y^2 - y_0^2}(x - x_0)^2}$$

can be regarded as an approximation of the hyperbolic geodesic distance. The complete recovery of the hyperbolic geodesic distance is demonstrated in Section 5 below.

4. Small Time Approximation of Option Price and Implied Volatility

We derive in this section the small-time asymptotics of the premium of a call option and its associated implied volatility by applying the small-time asymptotics for the probability density obtained in Section 3 when $H \leq \frac{1}{2}$. It is documented, for example, in Ekström and Lu (2015), that if the underlying asset is governed by an exponential Lévy model, the induced implied volatilities of non-ATM options may explode if jumps exist and the underlying process jumps towards the strike. As we shall see in the following, when $H < \frac{1}{2}$, the small time approximation of implied volatility also explodes; creating a jump-like behavior in the underlying process.

Let $k = \ln K$ be the logmoneyness, t the time to expiry, and recall that $S_t = e^{X_t}$. Though equivalently, we shall be primarily working with the (X_t, Y_t) process as in (7) rather than the (S_t, α_t) process in (6) hereafter. We write the price C of a call as a function of k and t as

$$\begin{aligned} C(k, t) &:= \mathbb{E}\left[(S_t - K)^+\right] = \mathbb{E}\left[(e^{X_t} - e^k)^+\right] \\ &= \iint (e^x - e^k)^+ p(t; x, y) dx\, dy. \end{aligned}$$

To evaluate the last integral, we approximate the joint density p by the small time asymptotics obtained in Theorem 2, then, as $t \to 0^+$, apply Laplace asymptotic formula to the

resulting integral. For the reader's convenience, we provide proof in Appenidx A.2 a variation of the Laplace asymptotic formula that is tailored for our own use.

Lemma 2. *Let $H \leq \frac{1}{2}$. For out-of-money call options, i.e., $k > x_0$, the call price $C(k,t)$ has the following asymptotic as $t \to 0$*

$$\ln C(k,t) \approx -\frac{1}{2t^{2H}} \left\{ \frac{\eta_*^2}{\nu^2} + \frac{1}{y_0^2 \psi(\eta_*)} \left(\frac{k - x_0}{t^{\frac{1}{2}-H}} - \rho y_0 C_{RK}(\eta_*) \frac{\eta_*}{\nu} \right)^2 \right\}, \tag{18}$$

where η_* is the minimizer

$$\eta_* = \operatorname{argmin} \left\{ \eta \in \mathbb{R} : \frac{\eta^2}{\nu^2} + \frac{1}{y_0^2 \psi(\eta)} \left(\frac{k - x_0}{t^{\frac{1}{2}-H}} - \rho y_0 C_{RK}(\eta) \frac{\eta}{\nu} \right)^2 \right\}.$$

Proof. The proof is a straightforward application of the Laplace asymptotic Formula (A12) in Lemma A1. Let $\mathcal{C} = \{(x,\eta) : x \geq k\} \subseteq \mathbb{R}^2$ and $\alpha = \frac{1}{2} - H \geq 0$. By using the asymptotic density (14), consider

$$\begin{aligned}
C(k,t) &= \int_0^\infty \int_k^\infty (e^x - e^k) p(t;x,y) dx dy \\
&= \frac{1}{2\pi} \int_0^\infty \int_k^\infty \left(e^x - e^k \right) \left\{ \frac{1}{y\sqrt{\nu^2 t^{2H}}} e^{-\frac{\eta^2}{2\nu^2 t^{2H}}} \frac{1}{y_0 \sqrt{t\psi(\eta)}} \times \right. \\
&\quad \left. e^{-\frac{1}{2y_0^2 t\psi(\eta)} \left(x - x_0 - \rho y_0 C_{RK}(\eta) \frac{\eta}{\nu} t^{\frac{1}{2}-H} \right)^2} e^{-\frac{x-x_0}{2\psi(\eta)} C_{eR}(\eta)} \left(1 + O\left(\sqrt{t}\right) \right) \right\} dx dy \\
&= \frac{1}{2\pi\nu y_0 t^{H+\frac{1}{2}}} \iint_{\mathcal{C}} \left(\frac{e^x - e^k}{\sqrt{\psi(\eta)}} \right) e^{-\frac{x-x_0}{2\psi(\eta)} C_{eR}(\eta)} \times \\
&\quad e^{-\frac{1}{2t} \left\{ \frac{\eta^2}{\nu^2} t^{2\alpha} + \frac{1}{y_0^2 \psi(\eta)} \left(x - x_0 - \rho y_0 C_{RK}(\eta) \frac{\eta}{\nu} t^\alpha \right)^2 \right\}} \left(1 + O\left(\sqrt{t}\right) \right) dx d\eta.
\end{aligned}$$

Applying the Laplace asymptotic Formula (A12) to the lowest order term in the last expression yields

$$\begin{aligned}
-\ln C(k,t) &\approx \frac{1}{2t} \left\{ \frac{\eta_*^2}{\nu^2} t^{2\alpha} + \frac{1}{y_0^2 \psi(\eta_*)} \left(x_* - x_0 - \rho y_0 C_{RK}(\eta_*) \frac{\eta_*}{\nu} t^\alpha \right)^2 \right\} \\
&= \frac{1}{2t^{2H}} \left\{ \frac{\eta_*^2}{\nu^2} + \frac{1}{y_0^2 \psi(\eta_*)} \left(\frac{x_* - x_0}{t^\alpha} - \rho y_0 C_{RK}(\eta_*) \frac{\eta_*}{\nu} \right)^2 \right\},
\end{aligned}$$

where, for fixed t, (x_*, η_*) is the minimizer of the function

$$\begin{aligned}
(x_*, \eta_*) &= \operatorname{argmin} \left\{ (x,\eta) \in \mathcal{C} : \frac{\eta^2}{\nu^2} t^{2\alpha} + \frac{1}{y_0^2 \psi(\eta)} \left(x - x_0 - \rho y_0 C_{RK}(\eta) \frac{\eta}{\nu} t^\alpha \right)^2 \right\} \\
&= \operatorname{argmin} \left\{ (x,\eta) \in \mathcal{C} : \frac{\eta^2}{\nu^2} + \frac{1}{y_0^2 \psi(\eta)} \left(\frac{x - x_0}{t^\alpha} - \rho y_0 C_{RK}(\eta) \frac{\eta}{\nu} \right)^2 \right\}.
\end{aligned}$$

Since the objective function is continuous in $(x,\eta) \in \mathcal{C}$ and it is a quadratic function in x, it follows that $x_* = k$ when t is small enough, thereby

$$\eta_* = \operatorname{argmin} \left\{ \eta : \frac{\eta^2}{\nu^2} + \frac{1}{y_0^2 \psi(\eta)} \left(\frac{k - x_0}{t^\alpha} - \rho y_0 C_{RK}(\eta) \frac{\eta}{\nu} \right)^2 \right\}.$$

□

Remark 4. *The plots in Figure 1 shows graphically the uniqueness of the minimal point η_* for $H = \frac{1}{4}$ and $H = \frac{3}{4}$. In these particular examples, the contours are convex in the half plane $x > k$, which corresponds to the out-of-money calls. For out-of-money puts, $x < k$, though the contours are not convex, the uniqueness of η_* sustains.*

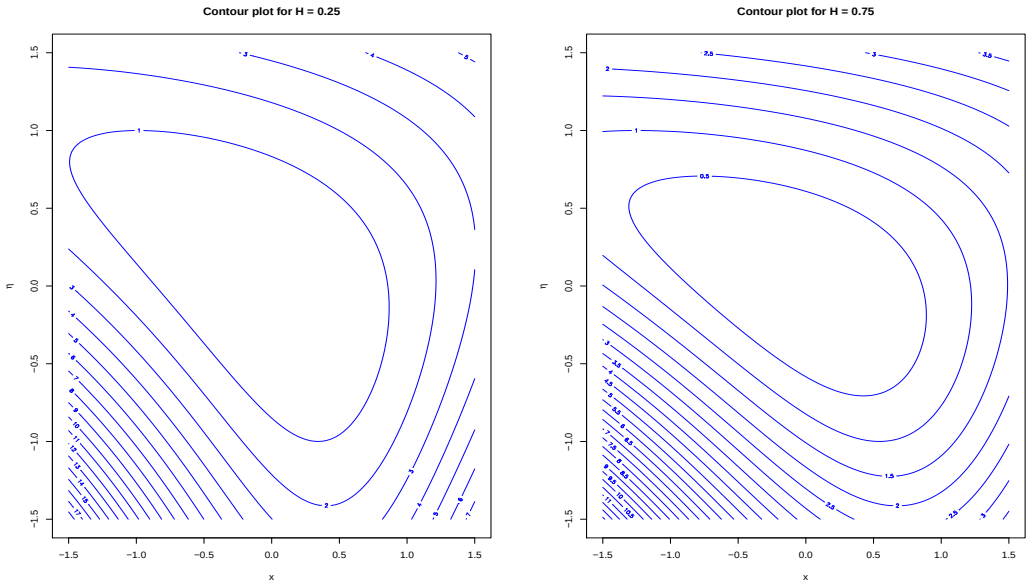


Figure 1. The contour plots. Parameters $\rho = -0.7$, $\nu = 1$, $y_0 = 1$, $t = 0.5$. $H = 0.75$ on the **right**; $H = 0.25$, on the **left**.

So long as we establish an asymptotic for the log price $\ln C(k,t)$ for $k > x_0$, by using the following small time asymptotic for implied volatility in Gao and Lee (2014) or Roper and Rutkowski (2009)

$$\sigma_{\mathrm{BS}}(k,t) = \frac{|k - x_0|}{\sqrt{2t|\ln C(k,t)|}} + O\left(\frac{\ln|\ln C(k,t)|}{\sqrt{t}|\ln C(k,t)|^{3/2}}\right), \tag{19}$$

an asymptotic formula for implied volatility follows immediately. We summarize the result in the following theorem but omitting its proof.

Theorem 3. *Let $H \leq \frac{1}{2}$ and let $k = \ln K$ be the log moneyness and $\alpha = \frac{1}{2} - H$. The implied volatility $\sigma_{\mathrm{BS}}(k,t)$ for out-of-money calls ($k > x_0$) has the following asymptotic in small time to expiry*

$$\sigma^2_{\mathrm{BS}}(k,t) = \sigma^2_{\mathrm{BS}}\left(\frac{k}{t^\alpha}\right) \approx \frac{(k-x_0)^2}{t^{2\alpha}}\left\{\frac{\eta_*^2}{\nu^2} + \frac{1}{y_0^2\psi(\eta_*)}\left(\frac{k-x_0}{t^\alpha} - \rho y_0^2 C_{RK}(\eta_*)\frac{\eta_*}{\nu}\right)^2\right\}^{-1}. \tag{20}$$

The minimal point η_* is given Lemma in 2.

Remark 5. *Note that* (20) *does not recover the SABR formula when $H = \frac{1}{2}$. The derivation of the SABR formula relies heavily on the geometry and symmetry of the underlying SABR plane which is isometric to the Poincaré plane. Figure 2 shows the comparison between the two formulas with time to expiry $t = 1$. Parameters are chosen so as to reproduce the figures in Hagan et al. (2002). In this set of parameters, the maximal difference between the two approximate implied volatility curves is about 1% for logmoneyness $k \in [-1, 1]$.*

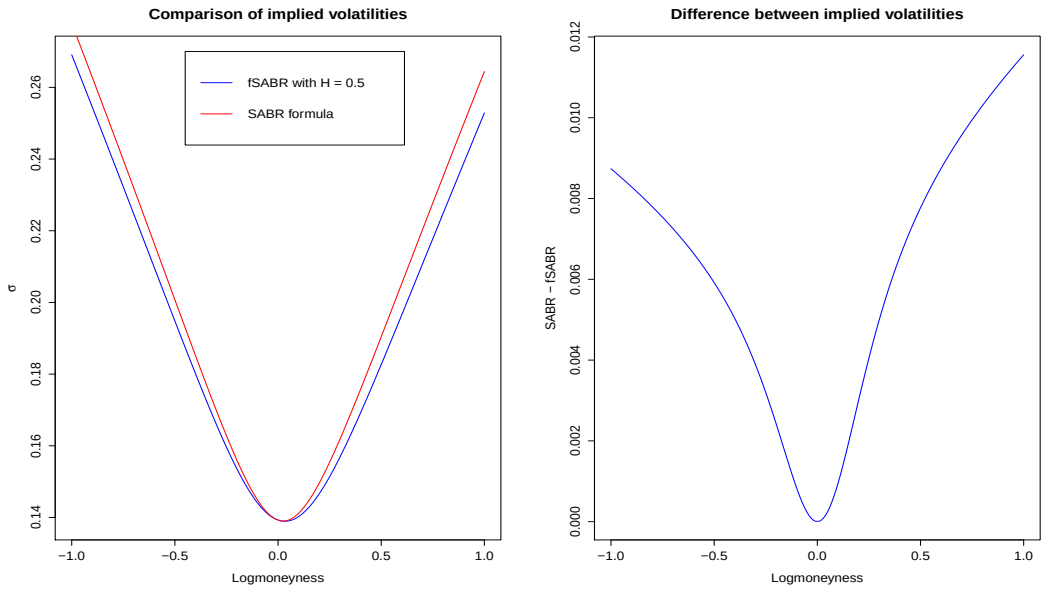


Figure 2. The plot on the **left** shows the approximate implied volatility curves versus logmoneyness with time to expiry $t = 1$ produced by (20) (in blue) the SABR Formula (3) (in red). Parameters are set as $\rho = -0.06867$, $\nu = 0.5778$, $a_0 = 0.13927$. The plot on the **right** shows the difference between the two curves.

We conclude the section by remarking that, as time to expiry t approaches zero, the approximate implied volatility $\sigma_{\mathrm{BS}}(k,t)$ flattens out with $H > \frac{1}{2}$; whereas the whole surface $\sigma_{\mathrm{BS}}(k,t)$ explodes with $H < \frac{1}{2}$ except for the at-the-money option $k = x_0$. Figure 3 shows the plots of approximate implied volatilities σ given in (20) versus logmoneyness k for time to expiry $t = 0.01$ and $t = 1$ respectively, and various Hurst exponents H. As in Figure 2, parameters are chosen as $a_0 = 0.13927$, $\nu = 0.5778$, and $\rho = -0.06867$. The numerical determination of the η_*'s is relatively efficient since it is basically a one-dimensional optimization problem.

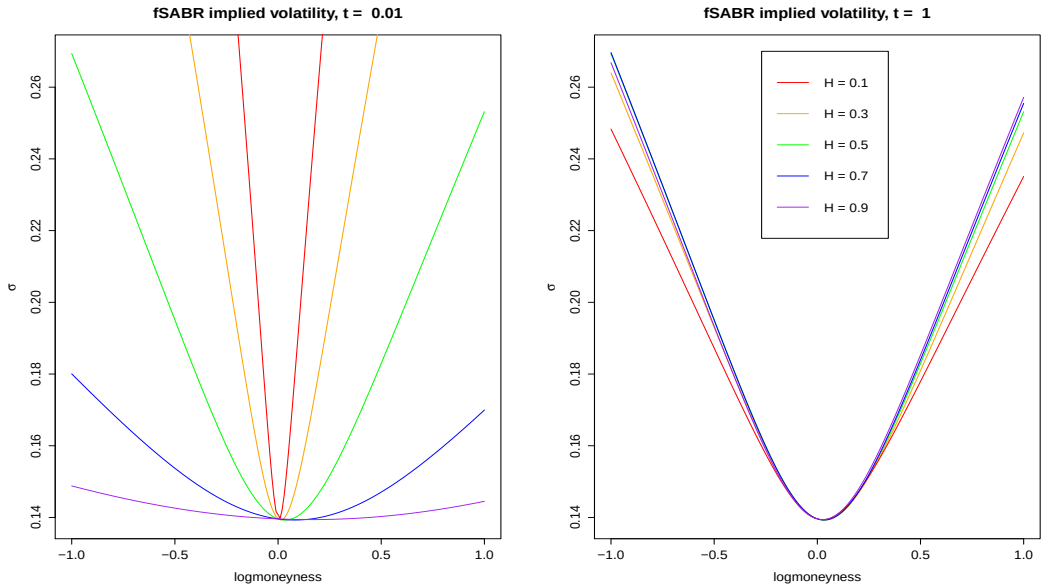


Figure 3. The implied volatility curves for $t = 0.01$ on the **left**, $t = 1$ on the **right**. Parameters are set as $\rho = -0.06867$, $\nu = 0.5778$, $a_0 = 0.13927$. $H = 0.1$ in red, $H = 0.3$ in orange, $H = \frac{1}{2}$ in green, $H = 0.7$ in blue, $H = 0.9$ in purple.

5. A Heuristic Large Deviation Principle

In this section, we provide a heuristic derivation of the sample path large deviation principle for (X_t, Y_t) in small time by bootstrapping the bridge representation to multi-period. For simplicity, we introduce the following vector notations.

$$\begin{aligned}
&\boldsymbol{t} = (t_1, \ldots, t_n), \qquad \boldsymbol{x_t} = (x_{t_1}, \ldots, x_{t_n}), \qquad \boldsymbol{y_t} = (y_{t_1}, \ldots, y_{t_n}),\\
&\boldsymbol{B^H_t} = (B^H_{t_1}, \ldots, B^H_{t_n}), \qquad \boldsymbol{X_t} = (X_{t_1}, \ldots, X_{t_n}), \qquad \boldsymbol{Y_t} = (Y_{t_1}, \ldots, Y_{t_n}),\\
&\boldsymbol{\xi_t} = (\xi_{t_1}, \ldots, \xi_{t_n}), \qquad \boldsymbol{\eta_t} = (\eta_{t_1}, \ldots, \eta_{t_n}), \qquad \boldsymbol{\zeta_t} = (\zeta_{t_1}, \ldots, \zeta_{t_n}).
\end{aligned}$$

Theorem 4. *The multiperiod joint density p of (X_t, Y_t)*

$$p(x_1, y_1, \ldots, x_n, y_n) := \mathbb{P}[(X_{t_1}, Y_{t_1}) = (x_1, y_1), \ldots, (X_{t_n}, Y_{t_n}) = (x_n, y_n)]$$

has the following bridge representation

$$\begin{aligned}
& p(x_1, y_1, \ldots, x_n, y_n) \qquad (21)\\
= \; & \mathbb{E}\left[\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi y_0^2 \bar{\rho}^2 \Delta v_{t_k}}} e^{-\frac{1}{2y_0^2\bar{\rho}^2\Delta v_{t_k}}\left(\Delta x_{t_k} - y_0\rho\int_{t_{k-1}}^{t_k} e^{\nu B^H_s} dB_s + \frac{y_0^2}{2}\Delta v_{t_k}\right)^2} \middle| \nu \boldsymbol{B^H_t} = \boldsymbol{\eta_t}\right] \times\\
& \mathbb{P}\left[y_0 e^{\nu \boldsymbol{B^H_t}} = \boldsymbol{y_t}\right],
\end{aligned}$$

where $\boldsymbol{\eta_t} = \log \boldsymbol{y_t} - \log y_0$, $\Delta x_{t_k} = x_{t_k} - x_{t_{k-1}}$, and $\Delta v_{t_k} = v_{t_k} - v_{t_{k-1}}$ for $k = 1, \ldots, n$. Recall that $v_t = \int_0^t e^{\nu B^H_s} ds$.

Proof. For any bounded measurable function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, consider the expectation

$$\begin{aligned}
& \iint f(\boldsymbol{x_t}, \boldsymbol{y_t}) p(\boldsymbol{x_t}, \boldsymbol{y_t}) d\boldsymbol{x_t} d\boldsymbol{y_t}\\
= \; & \mathbb{E}[f(\boldsymbol{X_t}, \boldsymbol{Y_t})]\\
= \; & \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{X_t}, \boldsymbol{Y_t}) | \mathcal{F}^B_{t_n}\right]\right].
\end{aligned}$$

Let $\xi_{t_i} = \int_0^{t_i} e^{\nu B^H_s} dW_s$, $\zeta_{t_i} = \int_0^{t_i} e^{\nu B^H_s} dB_s$ and thus accordingly $\Delta\xi_{t_i} = \xi_{t_i} - \xi_{t_{i-1}} = \int_{t_{i-1}}^{t_i} e^{\nu B^H_s} dW_s$, $\Delta\zeta_{t_i} = \zeta_{t_i} - \zeta_{t_{i-1}} = \int_{t_{i-1}}^{t_i} e^{\nu B^H_s} dB_s$. Note that, conditioned on $\mathcal{F}^B_{t_n}$, the random variables $\Delta\xi_{t_i}$'s are independent normal with mean 0 and variance Δv_{t_i}. We calculate the conditional expectation as follows.

$$\begin{aligned}
& \mathbb{E}\left[f(\boldsymbol{X_t}, \boldsymbol{Y_t}) | \mathcal{F}^B_{t_n}\right]\\
= \; & \mathbb{E}\left[f\left(x_0 + \rho y_0 \boldsymbol{\zeta_t} - \frac{y_0^2}{2}\boldsymbol{v_t} + \bar{\rho} y_0 \boldsymbol{\xi_t}, y_0 e^{\nu \boldsymbol{B^H_t}}\right) \middle| \mathcal{F}^B_{t_n}\right]\\
= \; & \int f\left(x_0 + \rho y_0 \boldsymbol{\zeta_t} - \frac{y_0^2}{2}\boldsymbol{v_t} + \bar{\rho} y_0 \boldsymbol{\xi_t}, y_0 e^{\nu \boldsymbol{B^H_t}}\right) \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\Delta v_{t_k}}} e^{-\frac{(\Delta\xi_{t_k})^2}{2\Delta v_{t_k}}} d\Delta\boldsymbol{\xi_t}. \qquad (22)
\end{aligned}$$

By applying the change of variables

$$x_{t_k} = x_0 + \rho y_0 \zeta_{t_k} - \frac{y_0^2}{2} v_{t_k} + \bar{\rho} y_0 \xi_{t_k},$$

thus

$$\Delta\xi_{t_k} = \frac{1}{\bar{\rho} y_0}\left(\Delta x_{t_k} - \Delta\zeta_{t_k} - \frac{y_0^2}{2}\Delta v_{t_k}\right).$$

The integral (22) becomes

$$\begin{aligned}
&\int f\Big(\boldsymbol{x_t}, y_0 e^{\nu \boldsymbol{B}_t^H}\Big) \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi \Delta v_{t_k}}} e^{-\frac{(\Delta \xi_{t_k})^2}{2\Delta v_{t_k}}} d\Delta \boldsymbol{\xi_t} \\
= &\int f\Big(\boldsymbol{x_t}, y_0 e^{\nu \boldsymbol{B}_t^H}\Big) \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi \bar{\rho}^2 y_0^2 \Delta v_{t_k}}} e^{-\frac{1}{2\bar{\rho} y_0 \Delta v_{t_k}} \left(\Delta x_{t_k} - \rho y_0 \int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s - \frac{y_0^2}{2} \Delta v_{t_k}\right)^2} d\Delta \boldsymbol{x_t} \\
= &\int f\Big(\boldsymbol{x_t}, y_0 e^{\nu \boldsymbol{B}_t^H}\Big) \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi \bar{\rho}^2 y_0^2 \Delta v_{t_k}}} e^{-\frac{1}{2\bar{\rho} y_0 \Delta v_{t_k}} \left(\Delta x_{t_k} - \rho y_0 \int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s - \frac{y_0^2}{2} \Delta v_{t_k}\right)^2} d\boldsymbol{x_t}
\end{aligned}$$

since the Jacobian between $d\Delta \boldsymbol{x_t}$ and $d\boldsymbol{x_t}$ 1. It follows that

$$\begin{aligned}
&\iint f(\boldsymbol{x_t}, \boldsymbol{y_t}) p(\boldsymbol{x_t}, \boldsymbol{y_t}) d\boldsymbol{x_t} d\boldsymbol{y_t} \\
= &\mathbb{E}\Big[\mathbb{E}[f(\boldsymbol{X_t}, \boldsymbol{Y_t}) | \mathcal{F}_t^B]\Big] \\
= &\mathbb{E}\left[\int f\Big(\boldsymbol{x_t}, y_0 e^{\nu \boldsymbol{B}_t^H}\Big) \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi \bar{\rho}^2 y_0^2 \Delta v_{t_k}}} e^{-\frac{1}{2\bar{\rho} y_0 \Delta v_{t_k}} \left(\Delta x_{t_k} - \rho y_0 \int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s - \frac{y_0^2}{2} \Delta v_{t_k}\right)^2} d\boldsymbol{x_t}\right] \\
= &\iint d\boldsymbol{x_t} d\boldsymbol{y_t} f(\boldsymbol{x_t}, \boldsymbol{y_t}) \times \\
&\mathbb{E}\left[\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi \bar{\rho}^2 y_0^2 \Delta v_{t_k}}} e^{-\frac{1}{2\bar{\rho} y_0 \Delta v_{t_k}} \left(\Delta x_{t_k} - \rho y_0 \int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s - \frac{y_0^2}{2} \Delta v_{t_k}\right)^2} \middle| \nu \boldsymbol{B}_t^H = \boldsymbol{\eta_t}\right] \times \\
&\mathbb{P}\Big[y_0 e^{\nu \boldsymbol{B}_t^H} = \boldsymbol{y_t}\Big].
\end{aligned}$$

This completes the proof of bridge representation (21) since f is arbitrary. □

To move onto a heuristic derivation of the sample path large deviation principle for (X_t, Y_t) in small time, we take logarithm on both sides of (21) and obtain

$$\begin{aligned}
&\log p(x_{t_1}, y_{t_1}, \ldots, x_{t_n}, y_{t_n}) \\
= &\log \mathbb{E}\left[\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi y_0^2 \bar{\rho}^2 \Delta v_{t_k}}} e^{-\frac{1}{2y_0^2 \bar{\rho}^2 \Delta v_{t_k}} \left(\Delta x_{t_k} - y_0 \rho \int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s + \frac{y_0^2}{2} \Delta v_{t_k}\right)^2} \middle| \nu \boldsymbol{B}_t^H = \boldsymbol{\eta}\right] \\
&+ \log \mathbb{P}\Big[\nu \boldsymbol{B}_t^H = \boldsymbol{\eta_t}\Big] - \sum \log y_{t_i}. \qquad (23)
\end{aligned}$$

In the following, we ignore the last term on the right-hand side of (23) and intuitively calculate the limits as $n \to \infty$ of the first two terms. Note that to the leading order we have

$$\log \mathbb{P}\Big[\nu \boldsymbol{B}_t^H = \boldsymbol{\eta_t}\Big] \approx -\frac{1}{2\nu^2} \boldsymbol{\eta}' \boldsymbol{R}^{-1} \boldsymbol{\eta},$$

where $\boldsymbol{R} = [R(t_i, t_j)]$ is the covariance matrix of $\boldsymbol{B}_t^H$. We further discretize the autovariance R of fractional Brownian motion as

$$\begin{aligned}
&R(t_i, t_j) = \mathbb{E}\Big[B_{t_i}^H B_{t_i}^H\Big] = \int_0^{t_i \wedge t_j} K_H(t_i, s) K_H(t_j, s) ds \\
\approx &\sum_{k=0}^{i \wedge j} K_H(t_i, t_k) K_H(t_j, t_k) \Delta t = \boldsymbol{K}' \boldsymbol{K} \Delta t,
\end{aligned}$$

where $\boldsymbol{K}$ denotes the upper triangular matrix

$$\boldsymbol{K}_{ij} = \begin{cases} K_H(t_i, t_j), & \text{if } i \geq j; \\ 0, & \text{otherwise.} \end{cases}$$

Thereby, $\boldsymbol{R}^{-1} = \frac{1}{\Delta t}\boldsymbol{K}^{-1}(\boldsymbol{K}')^{-1}$. Let $\boldsymbol{b} = (b_{t_1}, \dots b_{t_n})$ be the solution to the linear system

$$\frac{\boldsymbol{\eta}}{\nu} = \boldsymbol{K}\boldsymbol{b}\Delta t.$$

It follows that

$$\begin{aligned} & \frac{1}{2\nu^2}\boldsymbol{\eta}'\boldsymbol{R}^{-1}\boldsymbol{\eta} = \frac{1}{2}\Delta t\boldsymbol{b}'\boldsymbol{K}'\boldsymbol{R}^{-1}\boldsymbol{K}\boldsymbol{b}\Delta t \\ = \quad & \frac{1}{2}\boldsymbol{b}'\boldsymbol{b}\Delta t = \frac{1}{2}\sum_{k=1}^{n} b_{t_k}^2 \Delta t \\ \longrightarrow \quad & \frac{1}{2}\int_0^T b_t^2 dt \quad \text{as } n \to \infty. \end{aligned}$$

Also in the limit as $n \to \infty$, we obtain $\eta_t = \nu \int_0^t K_H(t,s) b_s ds$.

On the other hand, for the first term on the right-hand side of (23), we have

$$\begin{aligned} & \log \mathbb{E}\left[\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi y_0^2 \bar{\rho}^2 \Delta v_{t_k}}} e^{-\frac{1}{2y_0^2\bar{\rho}^2\Delta v_{t_k}}\left(\Delta x_{t_k} - y_0\rho\int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s + \frac{y_0^2}{2}\Delta v_{t_k}\right)^2} \middle| \nu \boldsymbol{B}^H = \boldsymbol{\eta}\right] \\ \approx \quad & \sum_{k=1}^{n} \mathbb{E}\left[-\frac{1}{2y_0^2\bar{\rho}^2\Delta v_{t_k}}\left(\Delta x_{t_k} - y_0\rho\int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s\right)^2 \middle| \nu \boldsymbol{B}^H = \boldsymbol{\eta}\right]. \end{aligned}$$

Note that conditioned on $\nu \boldsymbol{B}^H = \boldsymbol{\eta}$, we have

$$\Delta v_{t_k} = \int_{t_{k-1}}^{t_k} e^{2\nu B_s^H} ds \approx e^{2\eta_{t_{k-1}}}\Delta t = e^{2\nu\sum_{j=0}^{k-1} K_H(t_{k-1}, t_j) b_{t_j}\Delta t}\Delta t$$

as well as

$$\begin{aligned} & \Delta x_{t_k} - y_0\rho\int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s \approx \Delta x_{t_k} - y_0\rho e^{\eta_{t_{k-1}}} b_{t_{k-1}}\Delta t \\ = \quad & \left(\frac{\Delta x_{t_k}}{\Delta t} - y_0\rho e^{\nu\sum_{j=0}^{k-1} K_H(t_{k-1}, t_j) b_{t_j}\Delta t} b_{t_{k-1}}\right)\Delta t. \end{aligned}$$

It follows that the first term in (23) has the limit

$$\begin{aligned} & \sum_{k=1}^{n} \mathbb{E}\left[-\frac{1}{2y_0^2\bar{\rho}^2\Delta v_{t_k}}\left(\Delta x_{t_k} - y_0\rho\int_{t_{k-1}}^{t_k} e^{\nu B_s^H} dB_s\right)^2 \middle| \nu \boldsymbol{B}^H = \boldsymbol{\eta}\right] \\ \approx \quad & -\sum_{k=0}^{n} \frac{1}{2y_0^2\bar{\rho}^2 e^{2\nu\sum_{j=0}^{k-1} K_H(t_{k-1}, t_j) b_{t_j}\Delta t}}\left(\frac{\Delta x_{t_k}}{\Delta t} - y_0\rho e^{\nu\sum_{j=0}^{k-1} K_H(t_{k-1}, t_j) b_{t_j}\Delta t} b_{t_{k-1}}\right)^2 \Delta t \\ \longrightarrow \quad & -\frac{1}{2}\int_0^T \frac{1}{y_0^2\bar{\rho}^2 e^{2\nu\int_0^t K_H(t,s) b_s ds}}\left(\dot{x}_t - y_0\rho e^{\nu\int_0^t K_H(t,s) b_s ds} b_t\right)^2 dt \end{aligned}$$

as $n \to \infty$.

Putting the two limits together, we obtain heuristically for $T \approx 0$ that

$$
\begin{aligned}
& -\log \mathbb{P}[X_t = x_t, Y_t = y_t, \text{ for } t \in [0,T]] \\
\approx\ & \frac{1}{2}\int_0^T \frac{1}{y_0^2\bar{\rho}^2 e^{2\nu\int_0^t K_H(t,s)b_s ds}}\left(\dot{x}_t - y_0\rho e^{\nu\int_0^t K_H(t,s)b_s ds} b_t\right)^2 dt + \frac{1}{2}\int_0^T b_t^2 dt \\
=\ & \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2 y_t^2}(\dot{x}_t - \rho y_t b_t)^2 dt + \frac{1}{2}\int_0^T b_t^2 dt \\
=\ & \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2}\left(\frac{\dot{x}_t}{y_t} - \rho b_t\right)^2 dt + \frac{1}{2}\int_0^T b_t^2 dt, \qquad (24)
\end{aligned}
$$

where $b \in L^2[0,T]$ satisfies the integral equation

$$\log y_t - \log y_0 = \nu\int_0^t K_H(t,s)b_s ds$$

for all $t \in [0,T]$. We remark that (24) should serve as the rate function for the sample path large deviation principle in small time for (X_t, Y_t). Moreover, one may define the "geodesic" from the initial point (x_0, y_0) to the terminal point (x_T, y_T) in the fSABR plane as the path (x_t^*, y_t^*) which minimizes the functional (24), i.e.,

$$(x_t^*, y_t^*) := \underset{t\mapsto(x_t,y_t)}{\operatorname{argmin}} \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2}\left(\frac{\dot{x}_t}{y_t} - \rho b_t\right)^2 dt + \frac{1}{2}\int_0^T b_t^2 dt,$$

where again b_t is determined by solving the integral equation

$$\log y_t - \log y_0 = \nu\int_0^t K_H(t,s)b_s ds. \qquad (25)$$

Also, the minimizer can be regarded as the "geodesic" connecting (x_0, y_0) and (x_T, y_T).

Remark 6. *Note that b_t is indeed determined by the inverse operator K_H^{-1} applied to $\log\frac{y_t}{y_0}$. In particular, with $H = \frac{1}{2}$ this inverse operator reduces to the usual derivative. Thus, with $H = \frac{1}{2}$,*

$$b_t = \frac{d}{dt}\left(\log\frac{y_t}{y_0}\right) = \frac{\dot{y}_t}{y_t}.$$

The functional (24) *becomes*

$$
\begin{aligned}
& \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2}\left(\frac{\dot{x}_t}{y_t} - \rho b_t\right)^2 dt + \frac{1}{2}\int_0^T b_t^2 dt \\
=\ & \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2}\left(\frac{\dot{x}_t}{y_t} - \rho\frac{\dot{y}_t}{y_t}\right)^2 dt + \frac{1}{2}\int_0^T \frac{\dot{y}_t^2}{y_t^2} dt \\
=\ & \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2 y_t^2}\left(\dot{x}_t^2 - 2\rho\dot{x}_t\dot{y}_t + \dot{y}_t^2\right) dt.
\end{aligned}
$$

The last expression is the energy functional (up to the constant factor $\frac{1}{2}$) associated with the Riemann metric $ds^2 = \frac{1}{\bar{\rho}^2 y^2}(dx^2 - 2\rho dxdy + dy^2)$. The diffusion process associated with this Riemann metric is governed by the SDEs

$$
\begin{aligned}
dX_t &= Y_t dW_t, \\
dY_t &= Y_t dZ_t,
\end{aligned}
$$

where W_t and Z_t are correlated Brownian motion with constant correlation ρ, which up to a linear transformation is the upper plane model of the Poincaré space. In other words, with $H = \frac{1}{2}$, the functional (24) *recovers the energy functional for the classical Poincaré space, which is isometric to the SABR plane.*

Lastly, with the aid of the sample path large deviation principle (24), it is nearly a common practice, say by applying the Laplace asymptotic formula, to conclude that the log premium of an out-of-money call in small time has the asymptotic as $t \to 0$

$$-\log C(k,t) \approx -\log \mathbb{P}[X_t \geq k] \approx \frac{1}{2}\int_0^T \frac{1}{\bar{\rho}^2}\left(\frac{\dot{x}_t^*}{y_t^*} - \rho b_t^*\right)^2 dt + \frac{1}{2}\int_0^T {b_t^*}^2 dt,$$

where (x_t^*, y_t^*, b_t^*) denotes the optimal path that minimizes the functional (24) subject to the constraint $x_t^* = k$ and y_t^*, b_t^* satisfy the integral Equation (25). Thus, by applying (19), an approximation of implied volatility in small time is readily obtained. We summarize the result in the following proposition which, with $H = \frac{1}{2}$, recovers the SABR Formula (3). However, for $H \neq \frac{1}{2}$, the numerical implementation of (26) is more involved than that of (20) since, as opposed to a one dimensional optimization problem, it is subject to solving a two-dimensional constrained variational problem.

Proposition 1 (fSABR formula)**.** *Let $k = \log\left(\frac{K}{s_0}\right)$ be the log moneyness. The implied volatility $\sigma_{\text{BS}}(k,t)$ in a small time to expiry has the asymptotics*

$$\sigma_{\text{BS}}^2 \approx \frac{k^2}{T}\left(\int_0^T \left\{\frac{1}{\bar{\rho}^2 {y_t^*}^2}(\dot{x}_t^* - \rho y_t^* b_t^*)^2 + {b_t^*}^2\right\} dt\right)^{-1}, \tag{26}$$

where (x_t^*, b^*) is the minimizer of the variational problem

$$(x^*, b^*) = argmin\left\{\dot{x}, b \in L^2[0,T] : \int_0^T \left(\frac{1}{\bar{\rho}^2 y_t^2}(\dot{x}_t - \rho y_t b_t)^2 + b_t^2\right) dt\right\}$$

with $x_T = k$ and y_t^* satisfying

$$\log y_t^* - \log y_0 = \nu \int_0^t K_H(t,s) b_s^* ds$$

for $t \in [0,T]$. Notice that (26) *recovers the SABR Formula* (3) *with $H = \frac{1}{2}$.*

6. Conclusions and Discussion

We showed in this paper a bridge representation in Fourier space and a small time asymptotic for the joint probability of lognormal fractional SABR model for general $\rho \in (-1,1)$. An application of the asymptotics of the joint density is an approximation of the implied volatility in a short time. Due to the different nature of methodologies, the newly obtained approximation of implied volatilities in small time does not recover the celebrated SABR formula for implied volatility (to the zeroth order) when the Hurst exponent H equals a half. To recover the SABR formula, we presented a heuristic derivation of the sample path large deviation principle for the lognormal fractional SABR model by bootstrapping via the multiperiod joint density. We emphasize once again that the same trick is applicable to general fractional SABR models, i.e., to include a local volatility component in the process S_t for an underlying asset. We leave the rigorous proof of the sample path large deviation principle for fractional SABR models in future work. Lastly, the bridge representation methodology is also applicable to the case in which the volatility process is governed by an exponential fractional Ornstein-Uhlenbeck process since a fractional Ornstein-Uhlenbeck process is Gaussian as well. However, as the time to expiry approaches zero, the mean reversion part does not really play a role in the large deviation regime.

Author Contributions: All the authors contributed equally to the article in every aspect. All authors have read and agreed to the published version of the manuscript.

Funding: Jiro Akahori is supported by JSPS KAKENHI Grant Number 23330109, 24340022, 23654056, 25285102, 20K03666, and the project RARE-318984 (an FP7 Marie Curie IRSES). Tai-Ho Wang is partially supported by the PSC-CUNY52 Award and the National Natural Science Foundation of China grant 11971040.

Acknowledgments: We are grateful for helpful discussions with the participants of the conferences: At the Frontiers of Quantitative Finance at International Centre of Mathematical Sciences, Edinburgh, UK, and Mathematics of Quantitative Finance at Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, Germany. We would also like to thank the anonymous referees for their comments and suggestions which helped improve the readability and presentation of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Technical Proofs

In the appendix, we provide a detailed error analysis of the asymptotic expansion for (14) and a version of Laplace's asymptotic formula that is readily applicable to our case.

Appendix A.1. Error Analysis

Let $C_0^\infty(\mathbb{R}^2)$ be the space of smooth functions defined on $\mathbb{R}^2$ with compact support. For a given $f \in C_0^\infty(\mathbb{R}^2)$, recalling $\eta = \ln(y/y_0)$, from (9) we have

$$\begin{aligned}
&\mathbb{E}[f(X_t, Y_t)] = \iint f(x,y)p(t;x,y)dx\,dy \\
= &\frac{1}{2\pi}\iiint f(x,y)\frac{e^{-\frac{\eta^2}{2\nu^2t^{2H}}}}{y\sqrt{2\pi\nu^2t^{2H}}}e^{i(x-x_0)\xi}\mathbb{E}_{\frac{\eta}{\nu}}\left[e^{i\left(-\rho\int_0^t y_0e^{\nu B_s^H}dB_s+\frac{y_0^2v_t}{2}\right)\xi}e^{-\frac{\rho^2y_0^2v_t\xi^2}{2}}\right]d\xi dxdy \\
= &\frac{1}{2\pi}\iint e^{-ix_0\xi}\hat{f}(\xi,y)\frac{e^{-\frac{\eta^2}{2\nu^2t^{2H}}}}{y\sqrt{2\pi\nu^2t^{2H}}}\mathbb{E}_{\frac{\eta}{\nu}}\left[e^{i\left(-\rho\int_0^t y_0e^{\nu B_s^H}dB_s+\frac{y_0^2v_t}{2}\right)\xi}e^{-\frac{\rho^2y_0^2v_t\xi^2}{2}}\right]dyd\xi \\
= &\frac{1}{2\pi}\int e^{-ix_0\xi}\mathbb{E}\left[\hat{f}(\xi,Y_t)e^{i\left(-\rho\int_0^t y_0e^{\nu B_s^H}dB_s+\frac{y_0^2v_t}{2}\right)\xi}e^{-\frac{\rho^2y_0^2v_t\xi^2}{2}}\right]d\xi, \qquad \text{(A1)}
\end{aligned}$$

where

$$\hat{f}(\xi,y) = \int e^{i\xi x}f(x,y)dx$$

is the Fourier transform of f with respect to x.

Note that the right-hand side of (14) equals the right-hand side of (15). We compare (A1) with the following expression obtained by using the approximate joint density in (14) and obtain

$$\begin{aligned}
&\frac{1}{2\pi}\iiint f(x,y)\frac{e^{-\frac{\eta^2}{2\nu^2t^{2H}}}}{y\sqrt{2\pi\nu^2t^{2H}}} \\
&\quad\times e^{i(x-x_0)\xi}e^{-\frac{1}{2}(\rho^2\xi^2-i)\xi\int_0^t y_0^2e^{2\nu m_s}ds}\mathbb{E}_{\frac{\eta}{\nu}}\left[e^{-i\rho\xi\int_0^t y_0e^{\nu m_s}dB_s}\right]d\xi dxdy \\
= &\frac{1}{2\pi}\iint e^{-ix_0\xi}\hat{f}(\xi,y)\frac{e^{-\frac{\eta^2}{2\nu^2t^{2H}}}}{y\sqrt{2\pi\nu^2t^{2H}}} \\
&\quad\times\mathbb{E}_{\frac{\eta}{\nu}}\left[e^{i\left(-\rho\int_0^t y_0e^{\nu m_s}dB_s+\frac{y_0^2}{2}\int_0^t e^{2\nu m_s}ds\right)\xi}e^{-\frac{\rho^2y_0^2\xi^2}{2}\int_0^t e^{2\nu m_s}ds}\right]d\xi dxdy \\
= &\frac{1}{2\pi}\int e^{-ix_0\xi}\mathbb{E}\left[\hat{f}(\xi,Y_t)e^{i\left(-\rho\int_0^t y_0e^{\nu m_s}dB_s+\frac{y_0^2}{2}\int_0^t e^{2\nu m_s}ds\right)\xi}e^{-\frac{\rho^2y_0^2\xi^2}{2}\int_0^t e^{2\nu m_s}ds}\right]d\xi. \qquad \text{(A2)}
\end{aligned}$$

For simplification, denote

$$\lambda_1(t) = e^{i\left(-\rho\int_0^t y_0 e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}\right)\xi} e^{-\frac{\bar{\rho}^2 y_0^2 v_t}{2}\xi^2}$$

and

$$\lambda_2(t) = e^{i\left(-\rho\int_0^t y_0 e^{\nu m_s} dB_s + \frac{y_0^2}{2}\int_0^t e^{2\nu m_s} ds\right)\xi} e^{-\frac{\bar{\rho}^2 y_0^2 \xi^2}{2}\int_0^t e^{2\nu m_s} ds}.$$

Then the modulus of the difference between (A1) and (A2) is equal to

$$\left|\frac{1}{2\pi}\int e^{-ix_0\xi}\mathbb{E}\Big[\hat{f}(\xi, Y_t)(\lambda_1(t) - \lambda_2(t))\Big] d\xi\right|. \tag{A3}$$

The goal is to show that (A3) converges to zero in the order of $t^{\frac{1}{2}}$ as $t \to 0$, for every $f \in C_0^\infty(\mathbb{R}^2)$.

By applying the following inequality, for any $z, w \in \mathbb{C}$,

$$|e^z - e^w| \le \left(e^{\Re(z)} + e^{\Re(w)}\right)|z - w|,$$

where $\Re(z)$ denotes the real part of z, we have

$$\begin{aligned}
& |\lambda_1(t) - \lambda_2(t)| \\
\le\ & \left(e^{-\frac{\bar{\rho}^2 y_0^2 v_t \xi^2}{2}} + e^{-\frac{\bar{\rho}^2 y_0^2 \xi^2}{2}\int_0^t e^{2\nu m_s} ds}\right) \times \\
& \left| i\left(-\rho\int_0^t y_0 e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2}\right)\xi - \frac{\bar{\rho}^2 y_0^2 v_t \xi^2}{2} \right. \\
& \left. -i\left(-\rho\int_0^t y_0 e^{\nu m_s} dB_s + \frac{y_0^2}{2}\int_0^t e^{2\nu m_s} ds\right)\xi + \frac{\bar{\rho}^2 y_0^2 \xi^2}{2}\int_0^t e^{2\nu m_s} ds \right| \\
\le\ & 2|\mathcal{R}_t + i\mathcal{I}_t|
\end{aligned} \tag{A4}$$

since $e^{-\frac{\bar{\rho}^2 y_0^2 v_t}{2}\xi^2} + e^{-\frac{\bar{\rho}^2 y_0^2 \xi^2}{2}\int_0^t e^{2\nu m_s} ds} \le 2$ for all t and ξ. Apparently, $\mathcal{R}_t$ and $\mathcal{I}_t$ are given by

$$\begin{aligned}
\mathcal{R}_t &= \left[-v_t + \int_0^t e^{2\nu m_s} ds\right]\frac{\bar{\rho}^2 y_0^2 \xi^2}{2}, \\
\mathcal{I}_t &= \left(-\rho\int_0^t y_0 e^{\nu B_s^H} dB_s + \frac{y_0^2 v_t}{2} + \rho\int_0^t y_0 e^{\nu m_s} dB_s - \frac{y_0^2}{2}\int_0^t e^{2\nu m_s} ds\right)\xi.
\end{aligned}$$

In the following, K denotes a generic constant whose value may vary in different contexts. Then, by (A3) and (A4) and Hölder's inequality, we have

$$\begin{aligned}
& \left|\frac{1}{2\pi}\int e^{-ix_0\xi}\mathbb{E}\Big[\hat{f}(\xi, Y_t)(\lambda_1(t) - \lambda_2(t))\Big] d\xi\right| \\
\le\ & 2\int \mathbb{E}\Big[|\hat{f}(\xi, Y_t)||\mathcal{R}_t + i\mathcal{I}_t|\Big] d\xi \\
\le\ & 2\left(\mathbb{E}\int |\hat{f}(\xi, Y_t)|^{(1-\epsilon)p} d\xi\right)^{\frac{1}{p}} \left(\int \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q}|\mathcal{R}_t + i\mathcal{I}_t|^q\Big] d\xi\right)^{\frac{1}{q}} \\
\le\ & K\left(\mathbb{E}\int |\hat{f}(\xi, Y_t)|^{(1-\epsilon)p} d\xi\right)^{\frac{1}{p}} \left(\int \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q}(|\mathcal{R}_t|^q + |\mathcal{I}_t|^q)\Big] d\xi\right)^{\frac{1}{q}},
\end{aligned} \tag{A5}$$

for some $\epsilon \in (0, 1)$ and $\frac{1}{p} + \frac{1}{q} = 1$, $p, q > 0$.

Since $f \in C_0^\infty(\mathbb{R}^2)$, it is easy to show the following properties of $\hat{f}$:

(i) for any $r \geq 0$, $\sup_{(\xi,y)\in\mathbb{R}^2} \left|\xi^r \hat{f}(\xi, y)\right| < \infty$;

(ii) for any $r \geq 0$ and $p > 0$, $\int |\xi|^r \sup_{y\in\mathbb{R}} |\hat{f}(\xi, y)|^p d\xi < \infty$.

Note that property (ii) can be easily obtained by property (i).
By the above property (ii), we can show that

$$\limsup_{t\to 0^+} \mathbb{E} \int |\hat{f}(\xi, Y_t)|^{(1-\epsilon)p} d\xi < \infty. \tag{A6}$$

We compute the second term in (A5) separately as follows. By changing variables, we get

$$\begin{aligned}
& \int \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q} |\mathcal{R}_t|^q\Big] d\xi \\
\leq\ & K\bar{\rho}^{2q} y_0^{2q} \int \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q} \left(v_t^q + \left(\int_0^t e^{2\nu m_s} ds\right)^q\right)\right] \xi^{2q} d\xi \\
=\ & K\bar{\rho}^{2q} y_0^{2q} t^q \int \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q} \left(\left(\int_0^1 e^{2\nu B_{tu}^H} du\right)^q + \left(\int_0^t e^{2R(1,u)\eta} du\right)^q\right)\right] \xi^{2q} d\xi \\
=\ & K\bar{\rho}^{2q} y_0^{2q} t^q (L_1 + L_2),
\end{aligned} \tag{A7}$$

where

$$\begin{aligned}
L_1 &:= \int \xi^{2q} \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q} \left(\int_0^1 e^{2\nu B_{tu}^H} du\right)^q\right] d\xi, \\
L_2 &:= \int \xi^{2q} \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q} \left(\int_0^1 e^{2R(1,u)\eta} du\right)^q\right] d\xi.
\end{aligned}$$

By property (ii) of $\hat{f}$, it is easy to see that

$$\limsup_{t\to 0^+} L_2 \leq \left(\int_0^1 e^{2R(1,u)\eta} du\right)^q \int \xi^{2q} \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q}\Big] d\xi < \infty. \tag{A8}$$

For L_1, by Jensen's inequality and Hölder's inequality, we have

$$\begin{aligned}
L_1 &\leq \int \xi^{2q} \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q} \int_0^1 e^{2q\nu B_{tu}^H} du\right] d\xi \\
&\leq \int \xi^{2q} \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\mathbb{E}\left[\left(\int_0^1 e^{2q\nu B_{tu}^H} du\right)^{q_1}\right]\right\}^{\frac{1}{q_1}} d\xi \\
&\leq \int \xi^{2q} \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\int_0^1 \mathbb{E}\Big[e^{2qq_1\nu B_{tu}^H}\Big] du\right\}^{\frac{1}{q_1}} d\xi \\
&= \int \xi^{2q} \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\int_0^1 e^{2(qq_1\nu)^2 (tu)^{2H}} du\right\}^{\frac{1}{q_1}} d\xi.
\end{aligned}$$

where $\frac{1}{p_1} + \frac{1}{q_1} = 1$ with $p_1, q_1 > 0$. Therefore, using property (ii) again, we can easily show

$$\limsup_{t\to 0^+} L_1 \leq \limsup_{t\to 0^+} \int \xi^{2q} \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} d\xi \left\{\int_0^1 e^{2(qq_1\nu)^2 (u)^{2H}} du\right\}^{\frac{1}{q_1}} < \infty. \tag{A9}$$

Thus, it implies from (A7)–(A9) that

$$\int \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q} |\mathcal{R}_t|^q\Big] d\xi = O(t^q), \tag{A10}$$

for any $q > 1$, as $t \to 0^+$.

Similarly, we can write

$$\begin{aligned}
&\int \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q}|\mathcal{I}_t|^q\Big]d\xi \\
\leq\ & K\int |\xi|^q \mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q}\times \\
& \left\{\left|\rho\int_0^t y_0 e^{\nu B_s^H} dB_s\right|^q + \left|\frac{y_0^2 v_t}{2}\right|^q + \left|\rho\int_0^t y_0 e^{\nu m_s} dB_s\right|^q + \left|\frac{y_0^2}{2}\int_0^t e^{2\nu m_s} ds\right|^q\right\}\Big] d\xi \\
=\ & K(J_1 + J_2 + J_3 + J_4),
\end{aligned}$$

where

$$\begin{aligned}
J_1 &:= |\rho|^q \int |\xi|^q \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q}\left|\int_0^t y_0 e^{\nu B_s^H} dB_s\right|^q\right] d\xi, \\
J_2 &:= \int |\xi|^q \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q}\left|\frac{y_0^2 v_t}{2}\right|^q\right] d\xi, \\
J_3 &:= \int |\xi|^q \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q}\left|\rho\int_0^t y_0 e^{\nu m_s} dB_s\right|^q\right] d\xi, \\
J_4 &:= \int |\xi|^q \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q}\left|\frac{y_0^2}{2}\int_0^t e^{2\nu m_s} ds\right|^q\right] d\xi.
\end{aligned}$$

We estimate J_1 through J_4 separately as follows.

- J_1: Choosing $p_1 > 0$ such that $\frac{qq_1}{2} > 1$, by Hölder's inequality, the Burkholder-Davis-Gundy inequality, Jensen's inequality and a change of variables, we obtain Notice that

$$\begin{aligned}
J_1 &\leq |\rho|^q y_0^q \int |\xi|^q \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\mathbb{E}\left[\left|\int_0^t e^{\nu B_s^H} dB_s\right|^{qq_1}\right]\right\}^{\frac{1}{q_1}} d\xi \\
&\leq |\rho|^q y_0^q \int |\xi|^q \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\mathbb{E}\left[\left|\int_0^t e^{2\nu B_s^H} ds\right|^{\frac{qq_1}{2}}\right]\right\}^{\frac{1}{q_1}} d\xi \\
&= |\rho|^q y_0^q t^{\frac{q}{2}} \int |\xi|^q \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\mathbb{E}\left[\left|\int_0^1 e^{2\nu B_{tu}^H} du\right|^{\frac{qq_1}{2}}\right]\right\}^{\frac{1}{q_1}} d\xi \\
&\leq |\rho|^q y_0^q t^{\frac{q}{2}} \int |\xi|^q \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\int_0^1 \mathbb{E}\Big[e^{qq_1\nu B_{tu}^H}\Big] du\right\}^{\frac{1}{q_1}} d\xi \\
&= |\rho|^q y_0^q t^{\frac{q}{2}} \int |\xi|^q \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\int_0^1 e^{\frac{(qq_1\nu)^2}{2}(tu)^{2H}} du\right\}^{\frac{1}{q_1}} d\xi.
\end{aligned}$$

  By property (ii) we have

$$\begin{aligned}
&\limsup_{t\to 0^+} \int |\xi|^q \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} \left\{\int_0^1 e^{\frac{(qq_1\nu)^2}{2}(tu)^{2H}} du\right\}^{\frac{1}{q_1}} d\xi \\
\leq\ & \limsup_{t\to 0^+} \int \xi^{2q} \Big\{\mathbb{E}\Big[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1}\Big]\Big\}^{\frac{1}{p_1}} d\xi \left\{\int_0^1 e^{2(qq_1\nu)^2 (u)^{2H}} du\right\}^{\frac{1}{q_1}} < \infty.
\end{aligned}$$

  Thus, we can see that $J_1 = O(t^{\frac{q}{2}})$ as $t \to 0^+$.
- J_2 and J_4: The asymptotic behavior of J_2 and J_4 is the same as that of $t^q L_1$, and hence, $J_2, J_4 = O(t^q)$ as $t \to 0^+$.
- J_3: By using the same technique to J_1, we have

$$J_3 \leq |\rho|^q y_0^q t^{\frac{q}{2}} \int |\xi|^q \left\{ \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q p_1} \right] \right\}^{\frac{1}{p_1}} \left\{ \mathbb{E}\left[\left| \int_0^1 e^{2R(1,u)\eta} du \right|^{\frac{qq_1}{2}} \right] \right\}^{\frac{1}{q_1}} d\xi$$

and $J_3 = O(t^{\frac{q}{2}})$ as $t \to 0^+$.

Thus, putting all the estimates for the J_i's together we get

$$\int \mathbb{E}\left[|\hat{f}(\xi, Y_t)|^{\epsilon q} |\mathcal{I}_t|^q \right] d\xi = O(t^{\frac{q}{2}}), \tag{A11}$$

for any $q > 1$, as $t \to 0^+$.

Therefore, it implies from (A5), (A6), (A10) and (A11) that

$$\left| \frac{1}{2\pi} \int e^{-ix_0\xi} \mathbb{E}\left[\hat{f}(\xi, Y_t)(\lambda_1(t) - \lambda_2(t)) \right] d\xi \right| = O(t^{\frac{1}{2}}),$$

that is, (A3) converges to zero in the order of $t^{\frac{1}{2}}$ as $t \to 0$, for every $f \in C_0^\infty(\mathbb{R}^2)$.

Appendix A.2. Laplace Asymptotic Formula

We prove the following form of the Laplace asymptotic formula required in the derivation of the small time asymptotic of the price of an out-of-money call.

Lemma A1 (Laplace asymptotic formula)**.** *Let $\mathcal{C}$ be a closed and convex set in $\mathbb{R}^2$ with a nonempty and smooth boundary $\partial\mathcal{C}$. Suppose that $\theta(t,x) := \theta_0(x) + t^\alpha \theta_1(x) + t^{2\alpha}\theta_2(x)$, with $0 \leq 2\alpha < 1$, has continuous second-order partial derivatives in $x \in \mathcal{C}$, and, for every t sufficiently small, the function $\theta(t,x)$ is locally convex in $\mathcal{C}$ and attains its minimum uniquely at $x^*(t) \in \partial\mathcal{C}$. Moreover, there is $\epsilon_0 > 0$ such that for any $0 < \epsilon < \epsilon_0$, there exist t_0 and $\delta > 0$ for which*

$$\theta(t,x) \geq \theta(t, x^*(t)) + \delta, \ \forall (t,x) \in [0,t_0] \times (\mathcal{C} \setminus B_\epsilon(x^*(t))),$$

where $B_\epsilon(x^*(t)) = \{x : |x - x^*(t)| < \epsilon\}$ is the open ball of radius ϵ centered at $x^*(t)$.

Assume that f has continuous second-order partial derivatives in $\mathcal{C}$, is integrable over $\mathcal{C}$ (i.e., $\int_{\mathcal{C}} |f(x)|dx < \infty$) and that f vanishes identically in $\mathcal{C}^c$ and on the boundary $\partial\mathcal{C}$ but the inward normal directional derivative of f at $x^*(t)$ is nonzero.

Then, we have the asymptotic expansion, as $t \to 0^+$,

$$\begin{aligned} &\int_{\mathcal{C}} e^{-\frac{\theta(x,t)}{t}} f(x)dx \\ &= \frac{\sqrt{2\pi}\, t^{\frac{5}{2}}\, e^{-\frac{\theta(t,x^*(t))}{t}}}{\sqrt{\partial^2_{\mathbf{tan}}\theta(t,x^*(t))}|\nabla\theta(t,x^*(t))|} \left[\frac{\nabla f(x^*(t)) \cdot \nabla\theta(t,x^*(t))}{|\nabla\theta(t,x^*(t))|^2} + \frac{1}{2}\frac{\partial^2_{\mathbf{tan}} f(x^*(t))}{\partial^2_{\mathbf{tan}}\theta(t,x^*(t))} + o(1) \right], \end{aligned} \tag{A12}$$

where $\partial^2_{\mathbf{tan}} f(x^*)$ and $\partial^2_{\mathbf{tan}}\theta(t,x^*)$ are the second derivatives of f and θ respectively in the tangential direction to $\mathcal{C}$ at x^*.

Proof. For any $0 < \epsilon < \epsilon_0$, we split the integral on the left side of (A12) into two parts as

$$\int_{\mathcal{C}} e^{-\frac{\theta(t,x)}{t}} f(x)dx = \int_{\mathcal{C} \cap B_\epsilon(x^*(t))} e^{-\frac{\theta(t,x)}{t}} f(x)dx + \int_{\mathcal{C} \setminus B_\epsilon(x^*(t))} e^{-\frac{\theta(t,x)}{t}} f(x)dx. \tag{A13}$$

We treat the two terms on the right-hand side of (A13) individually. For the first term, since the integration region is restricted to a subset of the small ball $B_\epsilon(x^*(t))$, it can be reparametrized by $y = (y^1, y^2)$ so that in the y-coordinates the set $\{y : y^2 = 0\}$ corresponds to $\partial\mathcal{C}$ and the vectors $\{\partial_{y^1}, \partial_{y^2}\}$ form a local orthonormal frame around $x^*(t)$. For simplicity, we further assume that in the y-coordinates $x^*(t)$ is located at the origin. Note that in the y-coordinates the vector ∂_{y^2} is parallel to $\nabla\theta(x^*(t))$ as well as the inward normal vector of $\mathcal{C}$ at $x^*(t)$.

We shall use the convention that repeated indices are summed up over their respective ranges. Denote partial derivatives by subindices, we have for $y \in B_\epsilon(x^*(t))$

$$\theta(t,y) = \theta(t,0) + \theta_2(t,0)y^2 + \frac{1}{2}\theta_{ij}(t,0)y^i y^j + o(|y|^2),$$
$$f(y) = f_i(0)y^i + \frac{1}{2}f_{ij}(0)y^i y^j + o(|y|^2)$$

since $\theta_1(0) = 0$ for θ attains its minimum at the boundary point $x^*(t)$.

Thus, in the y-coordinates the first integral on the right-hand side of (A13) reads

$$\begin{aligned}&\int_{\mathcal{C}\bigcap B_\epsilon(x^*(t))} e^{-\frac{\theta(t,x)}{t}} f(x)dx\\ \approx\ & \int_0^\epsilon \int_{-\epsilon}^\epsilon e^{-\frac{1}{t}\left(\theta(t,0)+\theta_2(t,0)y^2+\frac{1}{2}\theta_{ij}(t,0)y^iy^j\right)}\left[f_i(0)y^i + \frac{1}{2}f_{ij}(0)y^iy^j\right]dy^1dy^2. \quad \text{(A14)}\end{aligned}$$

Now, by a change of variables

$$y^1 = \sqrt{t}z^1, \quad y^2 = tz^2,$$

we can write the above integral on the right-hand side of (A14) as

$$\begin{aligned}&e^{-\frac{\theta(t,0)}{t}}t^{\frac{3}{2}}\int_0^{\frac{\epsilon}{t}}\int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\theta_2(t,0)z^2+\frac{1}{2}\theta_{11}(t,0)(z^1)^2+\theta_{12}(t,0)z^1z^2\sqrt{t}+\frac{1}{2}\theta_{22}(t,0)(z^2)^2t\right)}\times\\ &\left[f_1(0)z^1\sqrt{t} + f_2(0)z^2t + \frac{1}{2}f_{11}(0)(z^1)^2t + f_{12}(0)z^1z^2t^{\frac{3}{2}} + \frac{1}{2}f_{22}(0)(z^2)^2t^2\right]dz^1dz^2. \quad \text{(A15)}\end{aligned}$$

Note that, for any real numbers $a_1, \ldots, a_5$, by dominated convergence theorem, we have

$$\begin{aligned}&\lim_{t\to 0}\int_0^{\frac{\epsilon}{t}}\int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\theta_2(t,0)z^2+\frac{1}{2}\theta_{11}(t,0)(z^1)^2+\theta_{12}(t,0)z^1z^2\sqrt{t}+\frac{1}{2}\theta_{22}(t,0)(z^2)^2t\right)}\times\\ &\left[a_1z^1 + a_2z^2 + a_3(z^1)^2 + a_4(z^2)^2 + a_5z^1z^2\right]dz^1dz^2\\ =\ & \int_0^\infty\int_{-\infty}^\infty e^{-\left(\theta_2(0,0)z^2+\frac{1}{2}\theta_{11}(0,0)(z^1)^2\right)}\times\\ &\left[a_1z^1 + a_2z^2 + a_3(z^1)^2 + a_4(z^2)^2 + a_5z^1z^2\right]dz^1dz^2 \in (-\infty,\infty).\end{aligned}$$

Thus, the quantity in (A15) equals

$$\begin{aligned}&e^{-\frac{\theta(t,0)}{t}}t^{\frac{3}{2}}\left\{\int_0^{\frac{\epsilon}{t}}\int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\theta_2(t,0)z^2+\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)}\times\right.\\ &\left.\left[f_1(0)z^1\sqrt{t} + f_2(0)z^2t + \frac{1}{2}f_{11}(0)(z^1)^2t\right]dz^1dz^2 + O\left(t^{\frac{1}{2}}\right)\right\}\\ =\ & e^{-\frac{\theta(t,0)}{t}}t^{\frac{3}{2}}\left[\sqrt{t}\cdot I + t\cdot II + t\cdot III + O\left(t^{\frac{1}{2}}\right)\right], \quad \text{(A16)}\end{aligned}$$

where

$$\begin{aligned}I &= \int_0^{\frac{\epsilon}{t}}\int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\theta_2(t,0)z^2+\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)}f_1(0)z^1dz^1dz^2,\\ II &= \int_0^{\frac{\epsilon}{t}}\int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\theta_2(t,0)z^2+\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)}f_2(0)z^2dz^1dz^2,\\ III &= \frac{1}{2}\int_0^{\frac{\epsilon}{t}}\int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\theta_2(t,0)z^2+\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)}f_{11}(0)(z^1)^2dz^1dz^2.\end{aligned}$$

As $t \to 0^+$, we calculate each integral individually as follows. For I, since the function in z^1 is an odd function and the integral interval for z^1 is symmetric about the origin, we obtain

$$\begin{aligned} I &= f_1(0) \int_0^{\frac{\epsilon}{t}} e^{-\theta_2(t,0)z^2} dz^2 \times \int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)} z^1 dz^1 \\ &= 0. \end{aligned} \tag{A17}$$

For II and III, notice that $\theta_2(t,0) > 0$ and $\theta_{11}(t,0) > 0$, and hence, we obtain

$$\begin{aligned} II &= f_2(0) \int_0^{\frac{\epsilon}{t}} e^{-\theta_2(t,0)z^2} z^2 dz^2 \times \int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)} dz^1 \\ &\approx f_2(0) \int_0^{\infty} e^{-\theta_2(t,0)z^2} z^2 dz^2 \times \int_{-\infty}^{\infty} e^{-\frac{1}{2}\theta_{11}(t,0)(z^1)^2} dz^1 \\ &= \frac{f_2(0)}{\theta_2^2(t,0)} \times \sqrt{\frac{2\pi}{\theta_{11}(t,0)}}, \end{aligned} \tag{A18}$$

and

$$\begin{aligned} III &= \frac{f_{11}(0)}{2} \int_0^{\frac{\epsilon}{t}} e^{-\theta_2(t,0)z^2} dz^2 \times \int_{-\frac{\epsilon}{\sqrt{t}}}^{\frac{\epsilon}{\sqrt{t}}} e^{-\left(\frac{1}{2}\theta_{11}(t,0)(z^1)^2\right)} (z^1)^2 dz^1 \\ &\approx \frac{f_{11}(0)}{2} \int_0^{\infty} e^{-\theta_2(t,0)z^2} dz^2 \times \int_{-\infty}^{\infty} e^{-\frac{1}{2}\theta_{11}(t,0)(z^1)^2} (z^1)^2 dz^1 \\ &= \frac{f_{11}(0)}{2\theta_2(t,0)} \times \sqrt{\frac{2\pi}{\theta_{11}^3(t,0)}}. \end{aligned} \tag{A19}$$

Therefore, it implies from (A14)–(A19) that, in the y-coordinates,

$$\int_{\mathcal{C} \bigcap B_\epsilon(x^*(t))} e^{-\frac{\theta(t,x)}{t}} f(x) dx \approx e^{-\frac{\theta(t,0)}{t}} t^{\frac{5}{2}} \sqrt{\frac{2\pi}{\theta_{11}(t,0)}} \left[\frac{f_2(0)}{\theta_2^2(t,0)} + \frac{f_{11}(0)}{2\theta_2(t,0)\theta_{11}(t,0)} + o(1) \right]. \tag{A20}$$

For the second term on the right-hand side of (A13), we get

$$\left| \int_{\mathcal{C} \setminus B_\epsilon(x^*)} e^{-\frac{\theta(t,x)}{t}} f(x) dx \right| \leq \int_{\mathcal{C} \setminus B_\epsilon(x^*)} e^{-\frac{\theta(t,x^*)+\delta}{t}} |f(x)| dx \leq e^{-\frac{\delta}{t}} e^{-\frac{\theta(t,x^*)}{t}} \int_{\mathcal{C}} |f(x)| dx. \tag{A21}$$

As a result, the second term is exponentially small (at the rate δ) as $t \to 0^+$ compared to the expansion (A12) obtained for the first term, hence it does not contribute to the asymptotic expansion.

Finally, by (A13), (A20) and (A21) we obtain the Laplace expansion (A12) by rewriting the expressions for the right-hand side of (A20) in the x-coordinates. □

References

Baillie, Richard T., Tim Bollerslev, and Hans Ole Mikkelsen. 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 74: 3–30. [CrossRef]

Cheng, Xue, and Tai-Ho Wang. 2018. Bessel bridge representation for heat kernel in hyperbolic space. *Proceedings of the American Mathematical Society* 146: 1781–92. [CrossRef]

Comte, Fabienne, and Eric Renault. 1998. Long memory in continuous-time stochastic volatility models. *Mathematical Finance* 8: 291–323. [CrossRef]

Comte, Fabienne, Laure Coutin, and Eric Renault. 2012. Affine fractional stochastic volatility models. *Annals of Finance* 8: 337–78. [CrossRef]

Cont, Rama, and Purba Das. 2022. Rough Volatility: Fact or Artifact? *arXiv* arXiv:2203.13820.

Decreusefond, Laurent, and A. Suleyman Üstünel. 1999. Stochastic analysis of the fractional Brownian motion. *Potential Analysis* 10: 177–214. [CrossRef]

Ekström, Erik, and Bing Lu. 2015. Short-time implied volatility in exponential Lévy models. *International Journal of Theoretical and Applied Finance* 18: 1550025. [CrossRef]

El Euch, Omar, and Mathieu Rosenbaum. 2019. The characteristic function of rough Heston models. *Mathematical Finance* 29: 3–38. [CrossRef]
Gao, Kun, and Roger Lee. 2014. Asymptotics of implied volatility to arbitrary order. *Finance and Stochastics* 18: 349–92. [CrossRef]
Gatheral, Jim, Thibault Jaisson, and Mathieu Rosenbaum. 2018. Volatility is rough. *Quantitative Finance* 18: 933–49. [CrossRef]
Granger, Clive W. J., and Roselyne Joyeux. 1980. An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–39. [CrossRef]
Guennoun, Hamza, Antoine Jaquier, Patrick Roome, and Fangwei Shi. 2018. Asymptotic behavior of the fractional Heston model. *SIAM Journal on Financial Mathematics* 9: 337–78. [CrossRef]
Hagan, Patrick, Deep Kumar, Andrew Lesniewski, and Diana Woodward. 2002. Managing smile risk. *Wilmott Magazine* 1: 84–108.
Hagan, Patrick, Andrew Lesniewski, and Diana Woodward. 2015. Probability distribution in the SABR Model of stochastic volatility. In *Large Deviations and Asymptotic Methods in Finance*. Springer Proceedings in Mathematics & Statistics. Cham: Springer, vol. 110, pp. 1–35.
Hu, Yaozhong. 2017. *Analysis on Gaussian Space*. Singapore: World Scientific.
Ikeda, Nobuyuki, and Hiroyuki Matsumoto. 1999. Brownian Motion on the Hyperbolic Plane and Selberg Trace Formula. *Journal of Functional Analysis* 163: 63–100. [CrossRef]
Jarrow, Robert A., Philip Protter, and Hasanjan Sayit. 2009. No arbitrage without semimartingales. *The Annals of Applied Probability* 19: 596–616. [CrossRef]
Mishura, Yuliya. 2008. *Stochastic Calculus for Fractional Brownian Motion and Related Processes*. Lecture Notes in Mathematics. Berlin and Heidelberg: Springer Science & Business Media, vol. 1929.
Nualart, David. 2006. *The Malliavin Calculus and Related Topics*, 2nd ed. Berlin and Heidelberg: Springer.
Rogers, Chris. 2019. Things We Think We Know. Available online: https://www.skokholm.co.uk/wp-content/uploads/2019/11/TWTWKpaper.pdf (accessed on 1 June 2022).
Roper, Michael, and Marek Rutkowski. 2009. A note on the behaviour of the Black-Scholes implied volatility close to expiry. *International Journal of Theoretical and Applied Finance* 12: 427–41. [CrossRef]

risks

MDPI

Article

Modeling Momentum and Reversals

Harvey J. Stein [1,2,*] and Jacob Pozharny [3]

[1] Labs Group, Two Sigma, New York, NY 10013, USA
[2] Department of Mathematics, Columbia University, New York, NY 10027, USA
[3] Bridgeway Capital Management, Houston, Texas 77046, USA
* Correspondence: hjstein@gmail.com

Abstract: Stock prices are well known to exhibit behaviors that are difficult to model mathematically. Individual stocks are observed to exhibit short term price reversals and long term momentum, while their industries only exhibit momentum. Here we show that individual stocks can be modeled by simple mean reverting processes in such a way that these behaviors are captured, the model is arbitrage free, and market informational efficiency is preserved. Simulation shows that in such a market, when mean reversion is sufficiently high, strategies which use reversals would substantially outperform buy and hold strategies.

Keywords: reversals; momentum; mean reversion; market efficiency; investment strategies; no arbitrage

Citation: Stein, Harvey J., and Jacob Pozharny. 2022. Modeling Momentum and Reversals. *Risks* 10: 190. https://doi.org/10.3390/risks10100190

Academic Editors: Dan Pirjol, Lingjiong Zhu and Krzysztof Jajuga

Received: 22 June 2022
Accepted: 28 September 2022
Published: 2 October 2022

1. Introduction

There is a large array of literature that documents reversals and momentum in stock prices. For example, Hameed and Mian (2015) observe reversals in the best and worst subset of monthly performers within industries, whereas the industries themselves only exhibit momentum, as do stocks in the longer term. Figelman (2007) also observes long term stock momentum, albeit on a different time scale.

Many articles try to explain these behaviors by appealing to investor behavior. Daniel et al. (1998) try to explain these features via investor overconfidence and biased self-attribution. Kahneman and Tversky (1979) attribute such features to the certainty and isolation effects. Booth et al. (2016) attributes momentum and reversals to the interaction of price momentum with size related effects.

Here we develop a reduced form model for stock price movements that exhibits these behaviors. We show that reversals and momentum would be observed if stocks were mean reverting to an industry business cycle[1]. We analyze the model analytically and through simulation. Simulation shows that in such a market, strategies which use reversals would outperform buy and hold strategies, often by a factor of 2 or more in high mean reversion environments. We also show that this model is arbitrage free and does not violate market efficiency, thus demonstrating that reversals and momentum can exist in efficient markets.

Mean reversion as a general phenomenon is well known. Mean reversion is commonly posited for short rate models and mean reverting models are used in commodity option pricing. They are useful for modeling pairs trading, as well as analyzing the timing of entering and exiting positions Leung and Li (2015). More recently, mean reverting processes have shown up in the portfolio optimization literature Amédée-Manesme et al. (2019); Kitapbayev and Leung (2018); Zhang et al. (2020). Amédée-Manesme et al. (2019) considers mean reversion in real estate deriving from correlation to mean reverting interest rates. The other references consider assets that mean revert to a fixed constant long term mean. Our work here differs from the above literature in that the mean reversion is to a time dependent function capturing the business cycle rather than to a fixed constant. The models closest to the models we propose here show up in the commodity option pricing literature Geman (2005) and as the Black–Karasinski model in short rate modeling Brigo and Mercurio (2001), but not as equity price processes.

2. Industry and Stock Process Model

We proceed to develop what might be considered a reduced form model for stock price movements.

Consider C_t, the value of an industry at time t. We could model the industry value in a number of ways. If we assume the industry value is deterministic with a time dependent growth rate g_t, then

$$dC_t = g_t C_t dt \tag{1}$$

In this case, the industry value at time t is:

$$C_t = C_0 e^{\int_0^t g_s ds} \tag{2}$$

If firm values follow the industry except for random perturbations away from the industry (noise), it is reasonable to assume not that the stock returns themselves are randomly perturbed, but that the firm value S itself deviates randomly from the industry, but is always equal to the industry on average. We can do that by defining

$$dX_t = -aX_t dt + \sigma dW_t \tag{3}$$

$$X_0 = 0 \tag{4}$$

$$S_t = C_t e^{\alpha_t + X_t} \tag{5}$$

where W is Brownian motion, so that X is an Ornstein-Uhlenbeck process with mean reversion a and instantaneous volatility σ, and α_t is a deterministic drift chosen so that $E[e^{\alpha_t + X_t}] = 1$. Solving the above stochastic differential equation for X yields, for times $t_1 \leq t_2$ that

$$X_{t_2} = X_{t_1} e^{-a(t_2 - t_1)} + e^{-at_2} \int_{t_1}^{t_2} e^{at} \sigma dW_t \tag{6}$$

Thus, X is a Gaussian process with mean and variance given by:

$$E[X_t] = 0 \tag{7}$$

$$\begin{aligned} \text{var}[X_t] &= \int_0^t e^{-2at} e^{2as} \sigma^2 ds \\ &= \frac{\sigma^2}{2a}\left(1 - e^{-2at}\right) \end{aligned} \tag{8}$$

and S_t is lognormal with mean

$$E[S_t] = C_t e^{\alpha_t + \frac{\text{var}[X_t]}{2}}. \tag{9}$$

Choosing

$$\begin{aligned} \alpha_t &= -\frac{\text{var}[X_t]}{2} \\ &= -\frac{\sigma^2}{4a}\left(1 - e^{-2at}\right) \end{aligned} \tag{10}$$

then gives S_t the desired mean.

Note that here we are transitioning from C_t being the overall value of the industry to expressing the optimal value of a company in this industry. For simplicity, we assume all firms revert to this overall level so as to capture behavior between the industry and the economy. The relationship between the sizes of different firms could be captured by instead having each firm revert to the appropriate percentage of the total industry level, and accounting for how these percentages change over time.

Suppose a given industry has a cyclical growth rate. Say its growth rate g_t is cyclical ranging from a low of l to a high of h with a period of p years, so that

$$g_t = \frac{l+h}{2} + \frac{h-l}{2}\sin(2\pi t/p). \tag{11}$$

If its value strictly follows this growth rate, then its value is

$$C_t = C_0 e^{\frac{l+h}{2}t + \frac{(h-l)p}{4\pi}(1-\cos(2\pi t/p))}. \tag{12}$$

This amounts to assuming that there is an underlying deterministic hidden variable driving the value of the industry and that it has cyclic growth.

So, the model we will be considering is n stocks in a given industry, all of which are mean reverting to an overall cyclic industry level:

$$g_t = \frac{l+h}{2} + \frac{h-l}{2}\sin(2\pi t/p) \tag{13}$$

$$dC_t = g_t C_t dt \tag{14}$$

$$dX_{i,t} = -a_i X_{i,t} dt + \sigma_i dW_{i,t} \tag{15}$$

$$S_{i,t} = e^{-\frac{\sigma^2}{4a}\left(1-e^{-2at}\right)+X_t} \tag{16}$$

where the W_i are uncorrelated Brownian motions. In our simulations, we will select particular values of h and l, and use the same mean reversion and volatility for all of the stocks.

3. Model Properties

3.1. Empirical Behavior

To illustrate the model behavior, we consider an industry with 30 companies. The industry follows C_1 with a 5 year cycle, a low rate of return of -1% and a high rate of return of 5%. Each stock has a volatility of 20%. Figures 1 and 2 show the underlying dynamics with a mean reversion of 5 and of 1, respectively. We display the deterministic industry firm level, a sample path for the average stock value, and a sample path for one of the stocks in the industry. The average firm deviates around the average deterministic level, as do individual firms, albeit with much higher volatility. With lower mean reversion, the stock takes much longer to return to the industry level, and the industry does not follow the deterministic level as closely.

3.2. Bounded Variance

For a process Y following geometric Brownian motion, with

$$dY = \mu Y dt + \sigma Y dW \tag{17}$$

for constants μ and σ, we have that

$$d\ln Y = \left(\mu - \frac{\sigma^2}{2}\right)dt + \sigma dW_t, \tag{18}$$

so the variance of $\ln Y$ is $\sigma^2 t$, which grows without bound.

In our case,

$$d\ln S = \left(\frac{C'}{C} + \alpha'\right)dt + dX \tag{19}$$

so the variance of $\ln S$ equals the variance of X, which is bounded.

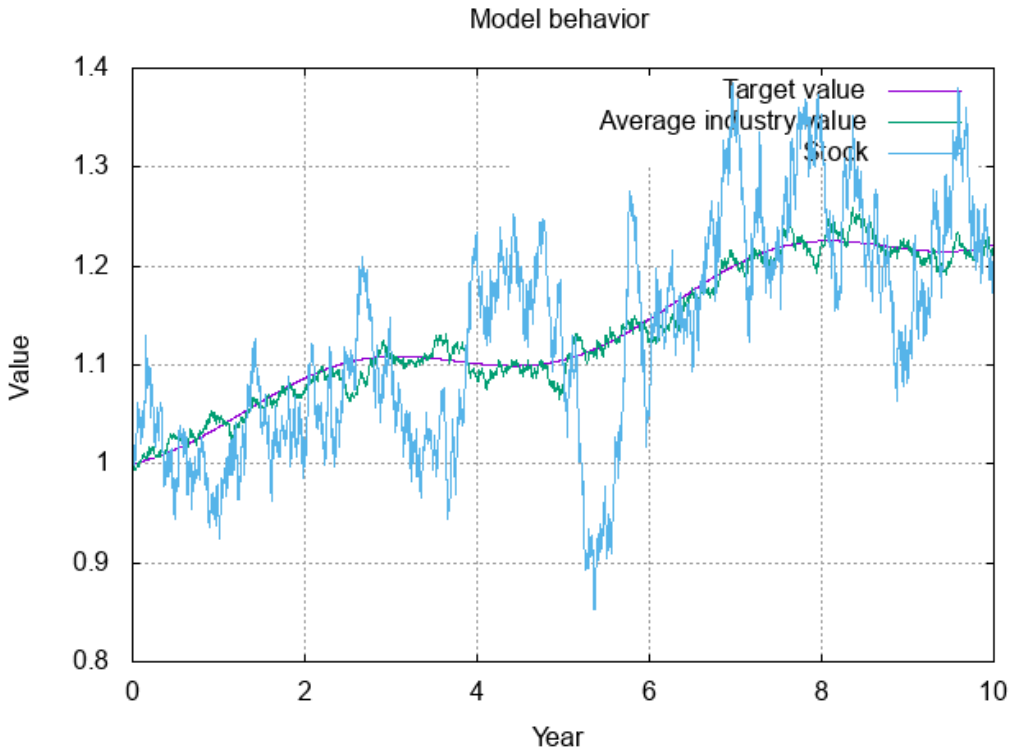


Figure 1. Market and one sample stock, mean reversion 5. The average firm value deviates around the underlying deterministic value, as do individual firms, but with higher volatility.

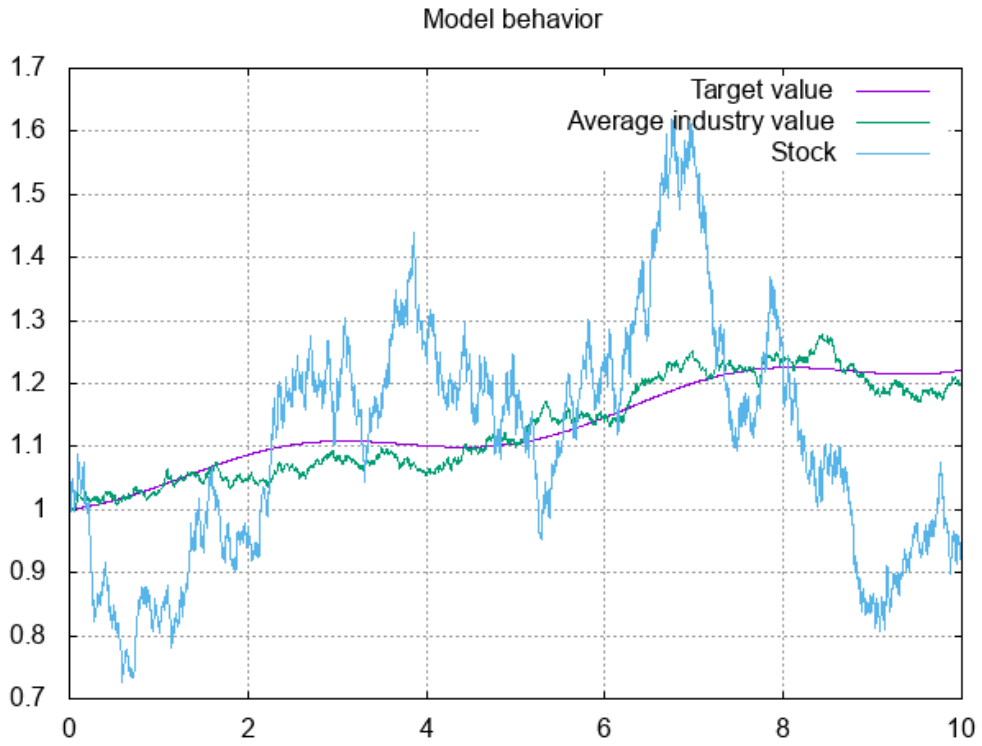


Figure 2. Market and one stock, mean reversion 1. Similar behavior is observed to the mean reversion 5 case, but the stock and market take longer to return to the underlying deterministic level.

One might object to the model on these grounds, as this differs from the common practice of assuming stocks follow geometric Brownian motion, and thus have unbounded

growth of log variance. However, there is little evidence to support such behavior of the long term variance. In practice, we can at best attempt to estimate the variance of $\ln S_t$ from one historically observed path. For time periods greater than a year, we typically have very few samples, or overlapping (correlated) samples, or samples that span over substantial, world changing events. Not to mention the fact that changes in drift will impact sample variances. All of these issues reduce the confidence we have in assuming the variance of $\ln S_t$ grows linearly with t, and thus of holding bounded variance as a point against this model.

3.3. Momentum and Reversals

Because the stock price process is mean reverting to C_t, the drift of S_t is negative when $S_t > C_t$, and is positive when $S_t < C_t$. This constitutes a reversal. Since the rate of reversion is a, the reversal roughly occurs on a time scale of $1/a$. So, strong reversals on a monthly basis would constitute a mean reversion on the order of 6 to 12. Smaller mean reversions will exhibit reversals to a lesser degree.

On the other hand, the industry itself, being the sum of the values of the companies in the industry, will have greatly dampened reversal behavior. It will only exhibit reversal behavior when a large percentage of the companies have inflated values that are not compensated for by substantially deflated values of the remaining companies.

On larger time scales, since the process mean reverts to C_t, when the business cycle is booming, all of the industry stocks will exhibit momentum in that they are mean reverting to C_t which itself exhibits momentum. Similarly, when the industry is in the bust part of the cycle, all of the industry stocks will exhibit poor long term performance.

Since reversals at the industry level are dampened, its momentum will be exhibited both on a small time scale as well as on long time scales.

3.4. Arbitrage and Market Efficiency

There has been much discussion of market efficiency since the ground breaking work of Fama (1970). Discussion has largely been about the extent to which the markets exhibit various degrees of efficiency, and the properties exhibited by efficient markets. For a general summary, see Shiller (2015).

More recently, there has been work on the relationship between market efficiency, martingale properties and the condition of a market being arbitrage free. This was first discovered by Samuelson (1965) and further elaborated on by Jensen (1978). Most recently, Jarrow and Larsson (2020) clarified and formalized the relationship between arbitrage and market efficiency. They showed that a market is informationally efficient in the sense of Fama (1970) if there exists a change of measure making the stock processes over the money market account martingales. The existence of such a change of measure also shows that the market is arbitrage free.

To see that market efficiency holds for this model, it suffices to show that there exists a change of measure making X a martingale. This suffices because it reduces us to the case where S is geometric Brownian motion with deterministic drift. The change of measure making X a martingale would have to be given by changing measure to make W' Brownian motion, where

$$dW' = -\frac{aX}{\sigma}dt + dW \tag{20}$$

We follow the argument given by Kolmo and Eldredge (2012). By Corollary 5.14, Chapter 3 of Karatzas and Shreve (1998), it suffices to show that there exists an unbounded and increasing sequence t_i with $t_0 = 0$ such that

$$E\left[e^{\frac{1}{2}\int_{t_i}^{t_{i+1}} \frac{a^2 X_t^2}{\sigma^2} dt}\right] < \infty \tag{21}$$

for all $i \geq 0$.

Given $S \geq 0$ and $\epsilon > 0$, and applying Jensen's inequality to the above exponential expression, we have that

$$\begin{aligned} e^{\frac{1}{2}\int_S^{S+\epsilon} \frac{a^2 X_t^2}{\sigma^2} dt} &= e^{\frac{1}{\epsilon}\int_S^{S+\epsilon} \frac{\epsilon a^2 X_t^2}{2\sigma^2} dt} \\ &\leq \frac{1}{\epsilon}\int_S^{S+\epsilon} e^{\frac{\epsilon a^2 X_t^2}{2\sigma^2}} dt \end{aligned} \tag{22}$$

By Tonelli's theorem, we can change the order of integration, so

$$E\left[\int_S^{S+\epsilon} e^{\frac{\epsilon a^2 X_t^2}{2\sigma^2}} dt\right] = \int_S^{S+\epsilon} E\left[e^{\frac{\epsilon a^2 X_t^2}{2\sigma^2}}\right] dt. \tag{23}$$

Thus, it suffices to find a constant $\epsilon > 0$ such that for all S, the right hand side of Equation (23) is finite. Then we can choose $t_i = i\epsilon$.

Since $X_t \sim N\left(0, \frac{\sigma^2}{2a}\left(1 - e^{-2at}\right)\right)$, let $\gamma_t = \sqrt{\frac{\sigma^2}{2a}(1 - e^{-2at})}$, and $Z_t = X_t/\gamma_t$. Then $Z_t \sim N(0,1)$, so

$$\begin{aligned} E\left[e^{\frac{\epsilon a^2 X_t^2}{2\sigma^2}}\right] &= E\left[e^{\frac{\epsilon a^2 \gamma_t^2}{2\sigma^2} Z_t^2}\right] \\ &= \frac{1}{\sqrt{1 - \frac{\epsilon a^2 \gamma_t^2}{\sigma^2}}} \end{aligned} \tag{24}$$

provided that

$$1 - \frac{\epsilon a^2 \gamma_t^2}{\sigma^2} > 0 \tag{25}$$

because Z_t^2 is chi-squared with one degree of freedom, so we know its moment generating function.

Solving for ϵ, we see that Equation 25 holds if

$$\epsilon\left(1 - e^{-2at}\right) < \frac{2}{a} \tag{26}$$

Since, for $t \geq 0$, $0 \leq \left(1 - e^{-2at}\right) < 1$, it suffices to select $\epsilon = \frac{1}{a}$. Then the integrand on the right hand side of Equation 23 is continuous and bounded, and hence the integral is finite.

We conclude that the model admits an equivalent martingale measure with respect to the money market account. It then follows that the market is arbitrage free and, from Jarrow and Larsson (2020), that the model proposed here satisfies the efficient market hypothesis. We conclude that momentum and reversals can exist in efficient markets, at least in the above sense.

4. Buy Losers, Sell Winners

Due to the mean reversion, it is likely that good performance will be followed by bad performance and vise-versa. This will occur roughly on a time frame of $1/a$. If one knew the industry level C_t, the mean reversion a, and the volatility σ, then one could determine exactly the drift of each stock, and make an optimal decision as to which stocks to buy and which stocks to short. However, it is difficult to determine these parameters, so we will not know for sure whether the drift of S_t exceeds that of the industry or lags that of the industry.

In practice, reversals are invested in based on the previous period's returns. The strategy is to buy those with poor recent performance (buy the losers) and sell those with high recent performance (sell the winners). If reversals occur in this industry, then such a strategy should outperform the industry average. We call this the "reversals" strategy.

We confirm that reversals consistently occur in our model through simulation. For an industry with 30 stocks, and stock and market parameters as detailed above, we simulate the stock processes and calculate the value of the strategy which each month buys the 9 stocks (30% of the industry) whose previous month's performance was worst, and shorts the 9 stocks with the best previous month's performance. We repeat this experiment 20 times. Figure 3 graphs the value of the strategy over time in each of the 20 experiments over a 10 year period in the case where the mean reversion is 2. Figure 4 does the same in the case where the mean reversion is 1. Table 1 gives the mean, standard deviation and percentiles of the annualized return distribution achieved over the 20 simulations after 10 years.

Table 1. Return statistics of the reversals strategy after 10 years for 20 industry simulations. We run the simulations for mean reversions of 1.0 and 2.0.

Statistic	**MR = 1**	**MR = 2**
mean return	6.5%	12.5%
return std	3.0%	2.5%
min return	1.1%	7.6%
25%ile return	4.1%	10.9%
50%ile return	6.8%	12.8%
75%ile return	8.4%	14.1%
max return	13.0%	16.9%

With a mean reversion of 2, we observe an annualized continuous return between 7.6% and 16.9%, substantially higher than the overall industry average with the above parameters, which is about 2%. With a mean reversion of 1, the improvements are much more modest, with 4 scenarios exhibiting returns less than 4%, the largest return being 13%, and most scenario returns between 4% and 9%. One scenario returns only 1.1%, substantially below the market return of 2%. Improving on that under low mean reversion would require adjusting the strategy.

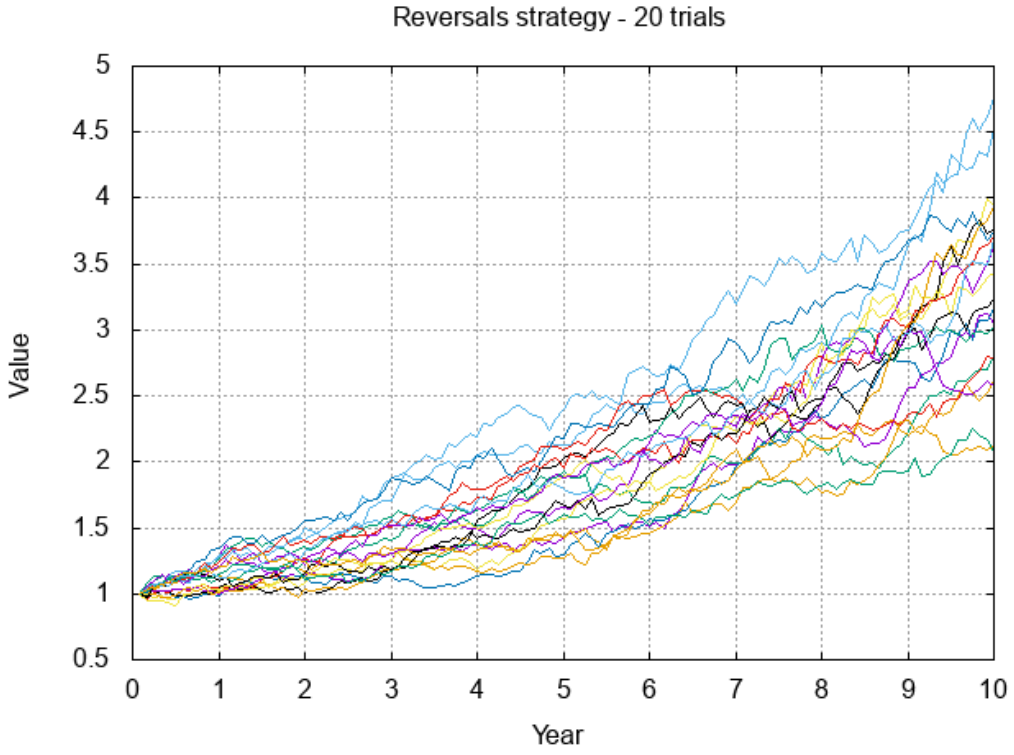


Figure 3. Reversals strategy performance, 20 simulations of 30 stocks, mean reversion 2.

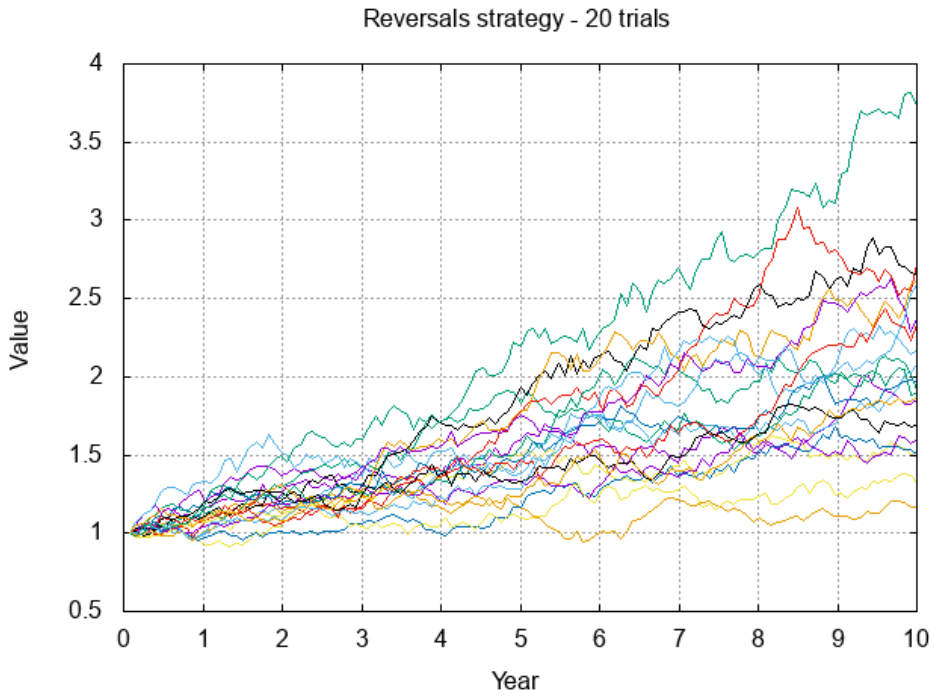


Figure 4. Reversals strategy performance, 20 simulations of 30 stocks, mean reversion 1. With low mean reversion, strategy returns are more modest, and have a higher probability of underperforming the industry.

The behavior as a function of the mean reversion is given in Figure 5. Because the reversals strategy holds fewer stocks, it's return has a higher standard deviation than holding the market. In general, the market return closely follows overall underlying return for C_t, which is about 2%. The performance of the reversals strategy improves on this when the mean reversion is large enough. At a mean reversion of 0.75, the average strategy return is 4.1%, double that of the market, albeit with a standard deviation of 2.8%. At a mean reversion of 1.0, the mean return is 6.5%, with a standard deviation of 3.8%. A mean reversion of 2 yields a mean return of 12% with a standard deviation of 3.1%.

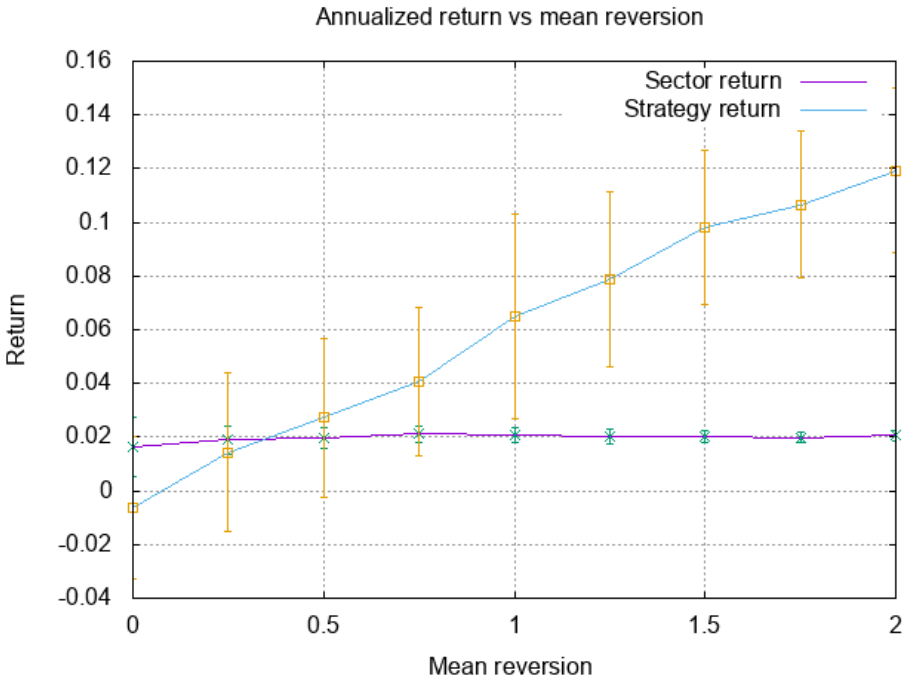


Figure 5. Reversals strategy performance as a function of mean reversion, 10 year horizon, 50 trials for each mean reversion. Error bars show standard deviation of returns.

5. Summary

Motivated by the observation of reversals and momentum in the stock market, we considered a stock process model where each stock in an industry reverts to an underlying deterministic cyclic level representing the industry's business cycle. Mathematical considerations show that in such a model, one would observe short term reversals and long term momentum, while the industry would just exhibit momentum. We showed that such a model is arbitrage free, which also demonstrates that momentum and reversals can exist in efficient markets (as defined by Jarrow and Larsson (2020)).

The above analysis was tested through simulation. In simulations with the industry following a 5 year cycle with returns cycling from -1% to 5%, and stocks with volatility of 20%, we observed that once mean reversion is above about 1.0, a reversal strategy of buying the bottom 30% of the stocks (in terms of the previous month's performance) and selling the top 30% would substantially outperform buy and hold strategies with high probability. At a mean reversion of 1, such a strategy yields a mean return of 6.5% versus a return of about 2% for the market. A mean reversion of 2 yields a mean return of 12.5% with a standard deviation of 2.5%.

Author Contributions: Conceptualization, H.J.S. and J.P.; methodology, H.J.S. and J.P; software, H.J.S.; validation, H.J.S.; formal analysis, H.J.S.; investigation, H.J.S. and J.P.; resources, H.J.S. and J.P.; writing—original draft preparation, H.J.S.; writing—review and editing, H.J.S. and J.P.; visualization, H.J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks to anonymous reviewers for their helpful suggestions, and to Apollo Hogan and Dan Pirjol for helping to incorporate said suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Note

[1] The existence of business cycles is well documented in the literature. See for example Zarnowitz (1992), or, for more recent research on the subject, Fernández-Villaverde and Guerrón-Quintana (2020) and Cerra et al. (2020).

References

Amédée-Manesme, Charles-Olivier, Fabrice Barthélémy, Philippe Bertrand, and Jean-Luc Prigent. 2019. Mixed-asset portfolio allocation under mean-reverting asset returns. *Annals of Operations Research* 281: 65–98. [CrossRef]

Brigo, Damiano, and Fabio Mercurio. 2001. *Interest Rate Models: Theory and Practice*. Berlin: Springer.

Cerra, Valerie, Antonio Fatas, and Sweta Chaman Saxena. 2020. *Hysteresis and Business Cycles*. International Monetary Fund.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. 1998. Investor psychology and security market under- and overreactions. *The Journal of Finance* 53: 1839–85. [CrossRef]

Fama, Eugene F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417. [CrossRef]

Fernández-Villaverde, Jesús, and Pablo A. Guerrón-Quintana. 2020. Uncertainty shocks and business cycle research. *Review of Economic Dynamics* 37: S118–46 [CrossRef] [PubMed]

Figelman, Ilya. 2007. Stock Return Momentum and Reversal. *The Journal of Portfolio Management* 34: 51–67. [CrossRef]

Geman, Hélyette. 2005. *Commodities and Commodity Derivatives: Modeling and Pricing for Agriculturals, Metals and Energy*. Chichester: John Wiley & Sons.

Geoffrey Booth, G., Hung-Gay Fung, and Wai Kin Leung. 2016. A risk-return explanation of the momentum-reversal "anomaly". *Journal of Empirical Finance* 35: 68–77. [CrossRef]

Hameed, Allaudeen, and G. Mujtaba Mian. 2015. Industries and Stock Return Reversals. *Journal of Financial and Quantitative Analysis* 50: 89–117. [CrossRef]

Jarrow, Robert, and Martin Larsson. 2020. Informational efficiency with trading constraints: A characterization. *SIAM Journal on Financial Mathematics* 11: 959–73. [CrossRef]

Jensen, Michael. 1978. Some anomalous evidence regarding market efficiency. *Journal of Financial Economics* 6: 95–101. [CrossRef]

Kahneman, Daniel, and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47: 263–92. [CrossRef]

Karatzas, Ioannis, and Steven Shreve. 1998. *Brownian Motion and Stochastic Calculus*, 2nd ed. New York: Springer Science & Business Media, vol. 113.

Kitapbayev, Yerkin, and Tim Leung. 2018. Mean Reversion Trading with Sequential Deadlines and Transaction Costs. *International Journal of Theoretical and Applied Finance* 21: 1850004. [CrossRef]

Kolmo, and Nate Eldredge. 2012. Can I Apply the Girsanov Theorem to an Ornstein-Uhlenbeck Process? Available online: https://math.stackexchange.com/questions/133691/can-i-apply-the-girsanov-theorem-to-an-ornstein-uhlenbeck-process (accessed on 10 June 2022).

Leung, Tim, and Xin Li. 2015. *Optimal Mean Reversion Trading: Mathematical Analysis and Practical Applications*. Singapore: World Scientific.

Samuelson, Paul. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.

Shiller, Robert J. 2015. *Irrational Exuberance*. Princeton: Princeton University Press.

Zarnowitz, Victor. 1992. Recent Work on Business Cycles in Historical Perspective. In *Business Cycles: Theory, History, Indicators, and Forecasting*. Chicago: National Bureau of Economic Research, University of Chicago Press, pp. 20–76. ISBN 0-226-97890-7.

Zhang, Jize, Tim Leung, and Aleksandr Aravkin. 2020. Sparse mean-reverting portfolios via penalized likelihood optimization. *Automatica* 111: 108651. [CrossRef]

risks

MDPI

Article

Spectral Expansions for Credit Risk Modelling with Occupation Times

Giuseppe Campolieti *, Hiromichi Kato and Roman N. Makarov

Department of Mathematics, Faculty of Science, Wilfrid Laurier University, 75 University Ave. West, Waterloo, ON N2L 3C5, Canada

* Correspondence: gcampolieti@wlu.ca

Abstract: We study two credit risk models with occupation time and liquidation barriers: the structural model and the hybrid model with hazard rate. The defaults within the models are characterized in accordance with Chapter 7 (a liquidation process) and Chapter 11 (a reorganization process) of the U.S. Bankruptcy Code. The models assume that credit events trigger as soon as the occupation time (the cumulative time the firm's value process spends below some threshold level) exceeds the grace period (time allowance). The hazard rate model extends the structural occupation time models and presumes that other random factors may also lead to credit events. Both approaches allow the firm to fulfill its obligations during the grace period. We derive new closed-from pricing formulas for credit derivatives containing the (risk-neutral) probability of defaults and credit default swap (CDS) spreads as special cases, which are derived analytically via a spectral expansion methodology. Our method works for any solvable diffusion, such as the geometric Brownian motion (GBM) and several state-dependent volatility processes, including the constant elasticity of variance (CEV) model. It allows us to write the pricing formulas explicitly as infinite series that converges rapidly. We then calibrate our models (assuming that GBM governs the firm's value) to market CDS spreads from the Total Energy company. Our calibration results show that the computations are fast, and the fit is near-perfect.

Keywords: credit risk models; occupation time; spectral expansions; default probability; credit default spread; hazard rate function; solvable diffusions

Citation: Campolieti, Giuseppe, Hiromichi Kato, and Roman Makarov. 2022. Spectral Expansions for Credit Risk Modelling with Occupation Times. *Risks* 10: 228. https://doi.org/10.3390/risks10120228

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 10 November 2022
Accepted: 25 November 2022
Published: 30 November 2022

1. Introduction

Structural and reduced-form models are the two main mathematical modelling approaches to credit risk. Structural models assume that a credit event triggers based on the current firm's value movement. Such models are linked to the debt-to-equity ratio since the higher the ratio value, the higher the firm's risk. It is reasonable to assume that a default occurs if the firm's value goes (or stays) below some threshold level. The Merton model Merton (1974) is one of the first structural models to analyze defaults. The firm in the Merton model defaults if the firm's value at maturity is less than some threshold level. A significant drawback of the Merton model is that it assumes that default events can only happen at known maturity times. The Black–Cox model Black and Cox (1976) extends the Merton model by allowing the firm to default at any time before or at maturity. The firm defaults once its value hits a specific barrier.

Most of the classical structural models treat default and liquidation as the same event. For example, in the Black–Cox model and in its numerous modifications, default/liquidation occurs when the firm value reaches an absorbing low barrier. According to the U.S. bankruptcy code, a firm that is unable to manage its debt can be given the right to declare bankruptcy under Chapter 11 (a reorganization process) and then reorganize its business. If the reorganization plan fails, Chapter 11 is converted to Chapter 7 (a liquidation process), and the firm is to be liquidated. There are many recent works in

which a distinction between bankruptcy and liquidation is made. See, for instance, the discussions in Moraux (2002); Nardon (2008); Galai et al. (2007); and Broadie et al. (2007). Typically, liquidation time is introduced as the first time the firm's asset value constantly or cumulatively stayed below the bankruptcy level over a certain period of time.

Defaults in the Black–Cox model are characterized in the form of the U.S. Bankruptcy Code Chapter 7. Occupation time models Makarov et al. (2015); Makarov (2016) further extend the Black–Cox model by allowing the firm to stay below a bankruptcy barrier for a specified amount of time as some firms admit a grace period to ensure the firm would be able to fulfill its obligations during the grace period. Therefore, defaults in occupation time models are characterized per both Chapters 7 and 11 of the U.S. Bankruptcy Code. Alternatively, excursion times can be used for a temporal separation between default and liquidation Li et al. (2014). Structural models with Parisian stopping times are related to Parisian options (see Chesney et al. (1997); Haber et al. (1999)).

Reduced-form models are intensity-based models where the likelihood of a default is measured by its hazard rate. Reduced-form models lack the ability to determine defaults endogenously with the firm's value process movements. Alfonsi and Lelong proposed in Alfonsi and Lelong (2012) a hybrid model that unifies the Black–Cox model and reduce-form models, in which default occurs based on hazard rate processes driven by the firm's value and other exogenous factors. However, the Alfonsi–Lelong model only considers the Chapter 7 type defaults.

In this paper, we use both structural and hybrid approaches. The total value of the firm's assets follows some diffusion process, which we call an F-diffusion. We assume that there exists a monotonic mapping that reduces the F-diffusion to a solvable X-diffusion, for which we can derive some fundamental formulas in closed form. In particular, we can obtain the joint probability density of the process value and its occupation time in the form of a spectral series expansion. The simplest example is GBM, which can be mapped to Brownian motion with drift. Although we focus on the GBM case in the numerical study, our methodology applies to a broad class of solvable diffusions.

We propose two occupation time-based models where closed-form pricing formulas for credit derivatives are derived analytically via a spectral expansion methodology. The spectral expansion method works for any solvable diffusions, including such processes as Brownian motion, the squared Bessel (SQB), the Cox–Ingersoll–Ross (CIR) and the Ornstein–Uhlenbeck (OU) processes. It is used to find closed-form credit derivative prices as a discrete expansion form that converges quickly. The first model we consider is an occupation time model in which defaults are characterized in accordance with the U.S. Bankruptcy Code Chapters 7 and 11. One of our occupation time models coincides with the model in Makarov (2016), except that we now employ the spectral expansion method. Furthermore, we derive closed-form pricing formulas for credit derivatives under this model. We also propose a new hybrid hazard rate model, which recovers the Black–Cox model and the Alfonsi–Lelong model as a particular case. Our hybrid model characterizes both Chapters 7 and 11 type defaults and contains closed-form pricing formulas for credit derivatives, leading to fast and accurate calibrations of market CDS spreads.

The paper is organized as follows. Sections 2 and 3 present the concept of occupation time and the main results associated with occupation time processes. Section 4 states our new developments pertaining to the occupation time models that were not exploited in Makarov (2016) and yet will be helpful in later sections. Section 5 describes the hazard rate model, which extends the Black–Cox and Alfonsi–Lelong models. Sections 6 and 7 present default probabilities and implied hazard rate functions and how they relate to one another. Section 8 entails pricing formulas for credit default swap (CDS) spreads. In Section 9, we provide the calibration procedure and results for CDS spreads (for the GBM case).

2. Occupation Time Process for Underlying Diffusion

We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \mathbb{F})$ where $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration for $(\mathbb{P}, \mathbb{F})$-Brownian motion $\{W_t\}_{t\geq 0}$. Let X be a one-dimensional time-homogeneous

regular diffusion process[1] on a state space $\mathcal{I} = (l, r) \subset \mathbb{R}$ with endpoints l, r satisfying $-\infty \leq l < r \leq \infty$. The generator is defined by

$$\mathcal{G}f(x) := \frac{1}{2}\sigma^2(x)f''(x) + \mu(x)f'(x) = \frac{1}{\mathfrak{m}(x)}\left(\frac{f'(x)}{\mathfrak{s}(x)}\right)'; \quad x \in \mathcal{I}, \tag{1}$$

with appropriate boundary condition at the endpoints. The speed $\mathfrak{m}(x)$ and scale $\mathfrak{s}(x)$ densities, where $\mathfrak{s}'(x)$ and $\mathfrak{m}(x)$ are continuous and positive for $x \in \mathcal{I}$ (see Borodin and Salminen (2002)), are defined via the drift, $\mu(x)$, and diffusion, $\sigma(x)$, coefficient functions as follows:

$$\mathfrak{s}(x) := \exp\left(-\int^x \frac{2\mu(z)}{\sigma^2(z)}dz\right), \quad \mathfrak{m}(x) := \frac{2}{\sigma^2(x)\mathfrak{s}(x)}. \tag{2}$$

The diffusion X satisfies the stochastic differential equation (SDE)

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t.$$

An occupation time $\mathcal{A}_t^{\ell,+}$ (or $\mathcal{A}_t^{\ell,-}$) is defined as the cumulative time the diffusion $X \in \mathcal{I}$ stays above (or below) the occupation level $\ell \in \mathcal{I}$ from time 0 to time t:

$$\begin{aligned} \mathcal{A}_t^{\ell,+} &:= \int_0^t \mathbb{I}_{[\ell,r)}(X_s)ds; \quad t \geq 0,\ \ell \in \mathcal{I}, \\ \mathcal{A}_t^{\ell,-} &:= \int_0^t \mathbb{I}_{(l,\ell]}(X_s)ds; \quad t \geq 0,\ \ell \in \mathcal{I}. \end{aligned} \tag{3}$$

The diffusion X with imposed lower and upper (regular) killings at a and b (where $a < b$), respectively, is defined by

$$X_{(a,b),t} := \begin{cases} X_t & t < \mathcal{T}_{(a,b)} \\ \partial^\dagger & t \geq \mathcal{T}_{(a,b)} \end{cases}; \quad X_0 = x \in (a, b), \tag{4}$$

where $\mathcal{T}_{(a,b)} := \inf\{t \geq 0 : X_t \notin (a, b)\}$ is the first exit time from the interval (a, b), and $\partial^\dagger$ denotes the cemetery (killed) state.[2]

Let λ be an instantaneous killing rate defined by:

$$\lambda(X_t) := \alpha_1\mathbb{I}_{[\ell,r)}(X_t) + \alpha_2\mathbb{I}_{(l,\ell]}(X_t); \quad 0 \leq \alpha_1 \leq \alpha_2. \tag{5}$$

Define the following process with instantaneous killing rate λ as:

$$\tilde{X}_{(a,b),t} := \begin{cases} X_{(a,b),t} & \Gamma_t < \xi, \\ \partial^\dagger & \Gamma_t \geq \xi, \end{cases} \tag{6}$$

where

$$\Gamma_t := \int_0^t \lambda(X_s)ds = \alpha_1\mathcal{A}_t^{\ell,+} + \alpha_2\mathcal{A}_t^{\ell,-} \tag{7}$$

is an $\mathbb{F}$-adapted hazard process and $\xi \sim \text{Exp}(1)$ is an $\mathbb{F}$-independent exponential random variable with unit rate. It is enough to consider only the occupation time below ℓ (i.e., $\alpha_1 = 0$) thanks to a simple identity $\mathcal{A}_t^{\ell,+} + \mathcal{A}_t^{\ell,-} = t$. The hazard process in (7) can be simplified to

$$\Gamma_t = \alpha_1 t + (\alpha_2 - \alpha_1)\mathcal{A}_t^{\ell,-}. \tag{8}$$

In the remainder of this section (and the next section) we shall assume $\alpha_1 = 0$ and $\alpha_2 = \alpha \geq 0$. We can define the transition density of the diffusion X with the instantaneous killing rate in (5):[3,4]

$$\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y)dy := \mathbb{P}_x\Big(\tilde{X}_{(a,b),t} \in dy\Big) = \mathbb{E}_x\Big[e^{-\alpha \mathcal{A}^{\ell,-}_t}; X_t \in dy, m_t > a, M_t < b\Big]; \quad (9)$$

for $t > 0, x, y \in (a,b)$, and zero otherwise, where $m_t := \inf_{0\leq u\leq t} X_u$ and $M_t := \sup_{0\leq u\leq t} X_u$.[5] Since both boundaries are NONOSC (non-oscillatory), we are in the Spectral Category I (see, e.g., Campolieti; Campolieti et al. (2013); Linetsky (2004)) and the transition density in (9) admits a discrete spectral expansion form:

$$\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y) = \mathfrak{m}(y) \sum_{n=1}^{\infty} e^{-\tilde{\lambda}_n t} \tilde{\phi}^{\ell,-}_{n,\alpha}(x) \tilde{\phi}^{\ell,-}_{n,\alpha}(y), \quad (10)$$

where $\{\tilde{\phi}^{\ell,-}_{n,\alpha}\}$ are eigenfunctions with eigenvalues $\{\tilde{\lambda}_n\}$ as the set of increasing simple zeros. The explicit formulas for Brownian motion are given in Appendix A. We define the joint density of the occupation time process (below ℓ) and the diffusion with imposed killing at endpoints a and b:

$$f^{(a,b)}_{\mathcal{A}^{\ell,-}_t, X_t}(u,y|\,x)dudy := \mathbb{P}_x\Big(\mathcal{A}^{\ell,-}_t \in du, X_t \in dy, m_t > a, M_t < b\Big), \quad (11)$$

for any $u \in (0,t), x, y \in (a,b)$, and zero otherwise. The joint density is defective at 0 and t where

$$\begin{aligned} \mathbb{P}_x\Big(\mathcal{A}^{\ell,-}_t = 0, X_t \in dy, m_t > a, M_t < b\Big) &= p_{(\ell,b)}(t;x,y)dy, \\ \mathbb{P}_x\Big(\mathcal{A}^{\ell,-}_t = t, X_t \in dy, m_t > a, M_t < b\Big) &= p_{(a,\ell)}(t;x,y)dy. \end{aligned} \quad (12)$$

The joint density can be obtained from the transition density in (9) by Laplace inverting with respect to α: (which can be evaluated numerically via the Gaver–Stehfest algorithm, see Cohen (2007); Gaver (1966); Stehfest (1970))

$$f^{(a,b)}_{\mathcal{A}^{\ell,-}_t, X_t}(u,y|\,x) = \mathcal{L}^{-1}_\alpha\Big\{\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y)\Big\}(u); \quad u \in (0,t),\ x,y \in (a,b). \quad (13)$$

The expectations of a bounded Borel function of the X-diffusion and its occupation time can be evaluated as:[6]

$$\begin{aligned} \mathbb{E}_x\Big[h(X_t)e^{-\alpha \mathcal{A}^{\ell,-}_t}; m_t > a\Big] &= \int_a^r h(y)\tilde{p}^{\ell,-}_{(a,r),\alpha}(t;x,y)dy, \\ \mathbb{E}_x\Big[h\Big(\mathcal{A}^{\ell,-}_t, X_t\Big); m_t > a\Big] &= \int_0^t \int_a^r h(u,y) f^{(a,r)}_{\mathcal{A}^{\ell,-}_t, X_t}(u,y|\,x)dydu \\ &\quad + \begin{cases} \int_a^\ell h(t,y)p_{(a,\ell)}(t;x,y)dy & x \in (a,\ell], \\ \int_\ell^r h(0,y)p^+_\ell(t;x,y)dy & x \in [\ell,r), \end{cases} \end{aligned} \quad (14)$$

where

$$p^+_\ell(t;x,y)dy := \mathbb{P}_x(X_t \in dy, m_t > \ell); \quad x \in [\ell,r), \quad (15)$$

and zero otherwise, is the transition density of X with a lower imposed killing at level ℓ.

Figure 1 shows some graphs of the transition density in (9) using the truncated N-term series in (10) for the case of standard Brownian motion. The top-left graph shows that the N-term series converges with $N = 8$ terms in the series with $a = -5$, $b = 3$, $\ell = -2$, $\alpha = 0.25$, and $t = 1$. If we increase b to 35 in the top-right graph, we can see that it requires more terms ($N \approx 40$) to obtain convergence of the series. If we increase t to 5 (from the top-left to the bottom-left) we can observe that the series requires less terms to demonstrate

converge. The bottom-right graph shows that $\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y) \to p_{(a,b)}(t;x,y)$ as $\alpha \to 0$ and $\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y) \to p_{(\ell,b)}(t;x,y)$ as $\alpha \to \infty$. Figure 2 shows the graph of the joint density in (11) and works well since $\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y)$ is a smooth function of α.

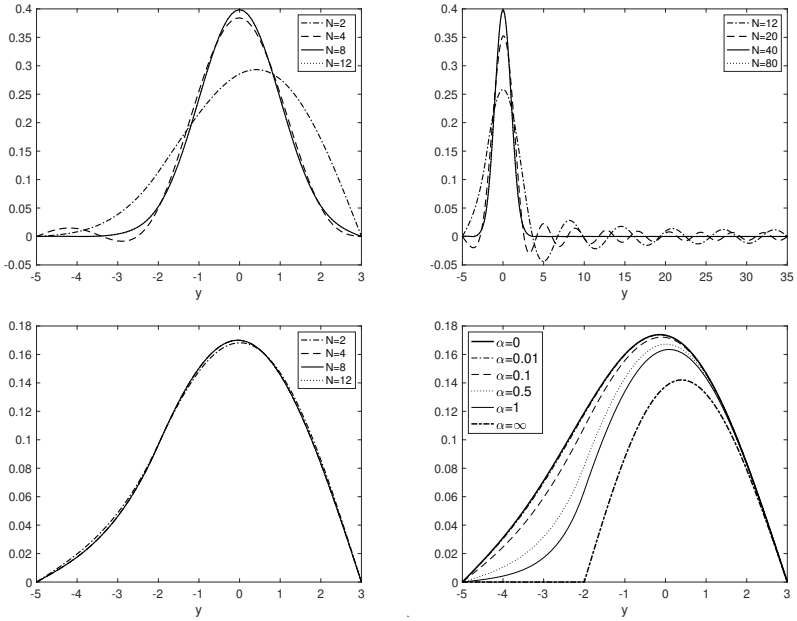


Figure 1. Graphs of the transition density in (9) where $X = W$ is a standard Brownian motion with $a = -5$, $b = 3$ ($b = 35$ for **top-right**), $\ell = -2$, $t = 1$ (**top-left**, bottom-left) $t = 5$ (**top-right**, **bottom-right**), $\alpha = 0.25$ (except **bottom-right**), $N = 30$ (**bottom-right**).

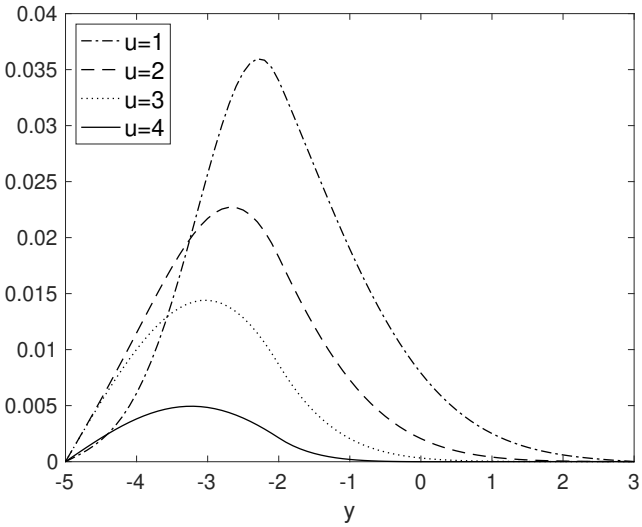


Figure 2. Graphs of the joint density in (11) where $X = W$ is a standard Brownian motion with $a = -5$, $b = 3$, $\ell = -2$, $t = 5$, and $N = 30$. We used 16 terms in the Gaver–Stehfest algorithm.

3. Occupation Time Process for *F*-Diffusion

Consider an F-diffusion $F_t := \mathsf{F}(X_t), t \geq 0$ (starting at $F_0 := \mathsf{F}(x)$, $X_0 = x$), defined in terms of a given diffusion X where $\mathsf{F} : \mathcal{I} \to \mathcal{D} := (F^{(l)}, F^{(r)})$, with $F^{(l)} := \min(\mathsf{F}(l), \mathsf{F}(r))$ and $F^{(r)} := \max(\mathsf{F}(l), \mathsf{F}(r))$, is a smooth monotonic function with unique inverse $\mathsf{X} = \mathsf{F}^{-1}$. Assuming $\mathsf{F}'(x) > 0$ (similar relations apply with a reversal of signs $+/-$ in case $\mathsf{F}'(x) < 0$), the occupation time process below a value $\mathsf{F}(\ell) \in \mathcal{D}$ for the F-diffusion is defined as (and we have a simple relationship between the occupation times of the two diffusions X and F):

$$\mathcal{A}_t^{(F),\mathsf{F}(\ell),-} := \int_0^t \mathbb{I}_{(\mathsf{F}(l),\mathsf{F}(\ell)]}(F_s)ds = \mathcal{A}_t^{\ell,-}. \tag{16}$$

The transition density of the F-diffusion, with the instantaneous killing rate in (5) with $\alpha_1 = 0$ and $\alpha_2 = \alpha$, i.e.,

$$\lambda(F_t) := \alpha \mathbb{I}_{(\mathsf{F}(l),\mathsf{F}(\ell)]}(F_t); \quad \alpha \geq 0, \tag{17}$$

and imposed killings at $\mathsf{F}(a)$ and $\mathsf{F}(b)$, is related to that of the X-diffusion (where $x = \mathsf{X}(F_0)$):

$$\begin{aligned} \tilde{p}^{(F),\mathsf{F}(\ell),-}_{(\mathsf{F}(a),\mathsf{F}(b)),\alpha}(t; F_0, y)dy &:= \mathbb{E}_{F_0}\left[e^{-\alpha \mathcal{A}_t^{(F),\mathsf{F}(\ell),-}}; F_t \in dy, m_t^F > \mathsf{F}(a), M_t^F < \mathsf{F}(b)\right] \\ &= \tilde{p}^{\ell,-}_{(a,b),\alpha}(t; x, \mathsf{X}(y)) \cdot \mathsf{X}'(y)dy, \end{aligned} \tag{18}$$

for $t > 0$ and $F_0, y \in (\mathsf{F}(a), \mathsf{F}(b))$, and zero otherwise. Here, $m_t^F := \inf_{0 \leq u \leq t} F_u$ and $M_t^F := \sup_{0 \leq u \leq t} F_u$. The joint density (and its defective portion) of the F-diffusion, with imposed killings at $\mathsf{F}(a)$ and $\mathsf{F}(b)$, are also related to that of the X-diffusion

$$\begin{aligned} f^{(\mathsf{F}(a),\mathsf{F}(b))}_{\mathcal{A}_t^{(F),\mathsf{F}(\ell),-},F_t}(u, y|\, F_0)dudy &:= \mathbb{P}_{F_0}\left(\mathcal{A}_t^{(F),\mathsf{F}(\ell),-} \in du, F_t \in dy, m_t^F > \mathsf{F}(a), M_t^F < \mathsf{F}(b)\right) \\ &= f^{(a,b)}_{\mathcal{A}_t^{\ell,-},X_t}(u, \mathsf{X}(y)|\, x) \cdot \mathsf{X}'(y)dudy, \end{aligned} \tag{19}$$

for any $u \in (0, t)$ and $F_0, y \in (\mathsf{F}(a), \mathsf{F}(b))$, and zero otherwise. The defective portion of the density at $u = 0$ is:

$$\mathbb{P}_{F_0}\left(\mathcal{A}_t^{(F),\mathsf{F}(\ell),-} = 0, F_t \in dy, m_t^F > \mathsf{F}(a), M_t^F < \mathsf{F}(b)\right) = p_{(\ell,b)}(t; x, \mathsf{X}(y)) \cdot \mathsf{X}'(y)dy, \tag{20}$$

for $F_0, y \geq \mathsf{F}(\ell)$ and zero otherwise. Similarly, the defective portion at $u = t$ is:

$$\mathbb{P}_{F_0}\left(\mathcal{A}_t^{(F),\mathsf{F}(\ell),-} = t, F_t \in dy, m_t^F > \mathsf{F}(a), M_t^F < \mathsf{F}(b)\right) = p_{(a,\ell)}(t; x, \mathsf{X}(y)) \cdot \mathsf{X}'(y)dy, \tag{21}$$

for $F_0, y \leq \mathsf{F}(\ell)$ and zero otherwise. The expectations of a bounded Borel function of the F-diffusion and its occupation time can be expressed in terms of (14):

$$\begin{aligned} \mathbb{E}_{F_0}\left[h(F_t)e^{-\alpha \mathcal{A}_t^{(F),\mathsf{F}(\ell),-}}; m_t^F > \mathsf{F}(a)\right] &= \mathbb{E}_x\left[h(\mathsf{F}(X_t))e^{-\alpha \mathcal{A}_t^{\ell,-}}; m_t > a\right], \\ \mathbb{E}_{F_0}\left[h\left(\mathcal{A}_t^{(F),\mathsf{F}(\ell),-}, F_t\right); m_t^F > \mathsf{F}(a)\right] &= \mathbb{E}_x\left[h\left(\mathcal{A}_t^{\ell,-}, \mathsf{F}(X_t)\right); m_t > a\right]. \end{aligned} \tag{22}$$

In this paper, we consider the F-diffusion as GBM (with state space $\mathcal{D} = (0, \infty)$):

$$F_t = F_0 e^{(\nu - \frac{\sigma^2}{2})t + \sigma W_t}; \quad F_0 > 0, t \geq 0, \tag{23}$$

where $\nu \in \mathbb{R}$ and $\sigma > 0$ are constants. Moreover, we consider an occupation time below a time-dependent occupation level (barrier) $\mathbf{L} := \{L(u) : u \geq 0\}$ where

$$L(t) = L_0 e^{\gamma t}; \quad L_0 \in (0, F_0),\ t \geq 0, \tag{24}$$

with growth rate $\gamma \in \mathbb{R}$, defined by

$$\mathcal{A}_t^{(F),\mathbf{L},-} := \int_0^t \mathbb{I}_{\{F_s \le L(s)\}} ds; \quad t \ge 0. \tag{25}$$

Let $\mu := (\nu - \gamma)/\sigma$ and $\overline{F}_t := e^{-\gamma t} F_t$, $t \ge 0$. Then the occupation time for the GBM process (with the time-dependent occupation level) can be expressed as the occupation time for the new GBM (with a constant occupation level):

$$\mathcal{A}_t^{(F),\mathbf{L},-} = \int_0^t \mathbb{I}_{\{\overline{F}_s \le L_0\}} ds := \mathcal{A}_t^{(\overline{F}),L_0,-}; \quad t \ge 0. \tag{26}$$

We apply a smooth monotonic mapping $\mathsf{F} : \mathcal{I} \to \mathcal{D}$ defined by $\mathsf{F}(x) = F_0 e^{\sigma x}$ with unique inverse $\mathsf{X}(\overline{F}) := \mathsf{F}^{-1}(\overline{F}) = \frac{1}{\sigma} \ln(\overline{F})$, where the underlying process is a Brownian motion with drift μ (starting at $X_0 = x = 0$):

$$X_t = \mu t + W_t; \quad t \ge 0. \tag{27}$$

Then, equation (25) can be expressed as the occupation time for its underlying diffusion (with constant occupation level):

$$\mathcal{A}_t^{(F),\mathbf{L},-} = \int_0^t \mathbb{I}_{(-\infty,\ell]}(X_s) ds := \mathcal{A}_t^{\ell,-}; \quad t \ge 0, \tag{28}$$

where $\ell = \frac{1}{\sigma} \ln(L_0/F_0)$. Moreover, let

$$A(t) = A_0 e^{\gamma t}; \quad A_0 \in (0, L_0),\ t \ge 0, \tag{29}$$

be a time-dependent liquidation barrier. The growth rate γ of the barriers $L(t)$ and $A(t)$ are kept the same, otherwise the Girsanov transformation, that effectively removes the time dependence of the barriers, would fail. The expectations of a bounded Borel function of the GBM and its occupation time can be expressed as integrals over the respective transition densities for the X-diffusion:

$$\begin{aligned}
\mathbb{E}_{F_0}\left[h(e^{-\gamma t}F_t) e^{-\alpha \mathcal{A}_t^{(F),\mathbf{L},-}}; m_t^F > A(t)\right] &= \int_a^\infty h(\mathsf{F}(y)) \tilde{p}^{\ell,-}_{(a,\infty),\alpha}(t;x,y) dy, \\
\mathbb{E}_{F_0}\left[h\left(\mathcal{A}_t^{(F),\mathbf{L},-}, e^{-\gamma t}F_t\right); m_t^F > A(t)\right] &= \int_0^t \int_a^\infty h(u, \mathsf{F}(y)) f^{(a,\infty)}_{\mathcal{A}_t^{\ell,-}, X_t}(u, y \,|\, x) dy du \\
&\quad + \begin{cases} \int_a^\ell h(t, \mathsf{F}(y)) p_{(a,\ell)}(t;x,y) dy & x \in (a, \ell], \\ \int_\ell^\infty h(0, \mathsf{F}(y)) p_\ell^+(t;x,y) dy & x \in [\ell, \infty), \end{cases}
\end{aligned} \tag{30}$$

where $a = \frac{1}{\sigma} \ln(A_0/F_0)$.

4. Occupation Time Model in the Risk-Neutral Measure

We fix a filtered probability space $(\Omega, \mathcal{H}, \widetilde{\mathbb{P}}, \mathbb{H})$[7] with filtration $\mathbb{H} = \{\mathcal{H}_t\}_{t \ge 0}$ where $\mathcal{H}_t$ is a σ-algebra, describing the complete information up to time t. Let $\{F_t\}_{t \ge 0}$ be an almost surely (a.s.) positive $\mathbb{F}$-measurable time-homogeneous Markov process, representing the firm's value process, where $\mathbb{F} \subset \mathbb{H}$ is the natural filtration for $(\widetilde{\mathbb{P}}, \mathbb{F})$-Brownian motion $\{\widetilde{W}_t\}_{t \ge 0}$. A typical example of the firm's value $\{F_t\}_{t \ge 0}$ is GBM:

$$F_t = F_0 e^{(r^f - \frac{\sigma^2}{2})t + \sigma \widetilde{W}_t}; \quad F_0 > 0, t \ge 0, \tag{31}$$

where $r^f \ge 0$ is a constant risk-free rate, $\sigma > 0$ is a constant volatility. The diffusion process $\{F_t\}_{t \ge 0}$ satisfies the SDE $dF_t = r^f F_t dt + \sigma F_t d\widetilde{W}_t$.

For the rest of the paper, we shall assume that the firm's value is a GBM process and presume an occupation time below a time-dependent occupation level (barrier) $\mathbf{L} := \{L(u) = L_0 e^{\gamma u} : u \geq 0\}$ (the parameter $\gamma \in \mathbb{R}$ is the growth rate), defined in (25).

The time dependence of the barrier $\mathbf{L}$ adds more flexibility in the model (particularly for GBM which has a constant log-volatility) by introducing an extra effective drift parameter, while still keeping the discounted value process a martingale in the risk-neutral measure. The extra drift parameter arising from the exponent of the time-dependent barriers allows for a better (more flexible) calibration of the model to default probabilities across the different maturities (short versus long term). The parameter γ can also reflect the compound interest rate on the firm's debt. Additionally, higher values of γ increase the severity of both the grace period and the default over time.

The random variable $\mathcal{A}_t^{(F),\mathbf{L},-}$ measures the cumulative amount of time the process $\{F_t\}_{t\geq 0}$ stays below the occupation level. Let $\tau^{(\vartheta)} := \tau_L^{\vartheta} \wedge \tau_A := \min(\tau_L^{\vartheta}, \tau_A)$ be an $\mathbb{F}$-stopping time where

$$\tau_A := \inf\{t \geq 0 : F_t \leq A(t)\}, \qquad \tau_L^{\vartheta} := \inf\{t \geq 0 : \mathcal{A}_t^{(F),\mathbf{L},-} > \vartheta\}, \tag{32}$$

$A(t) = A_0 e^{\gamma t}, t \geq 0$ is a time-dependent liquidation barrier, ϑ is a nonnegative grace period, and we set $\inf\{\emptyset\} = \infty$ by convention.

In this model, the stopping time $\tau_L := \inf\{t \geq 0 : F_t \leq L(t)\}$ characterizes Chapter 11 type default. The grace period for reorganizing firm's business begins at the moment when the firm's value process hits the occupation barrier $\mathbf{L}$. The default time $\tau^{(\vartheta)}$ characterizes Chapter 7 type default. If the firm's value drops down to the liquidation level $\mathbf{L}$ or if the cumulative amount of time the firm's value process stays below the occupation level exceeds the grace period ϑ, Chapter 11 is converted to Chapter 7, and the firm is to be liquidated.

We may conveniently employ an appropriate transformation to the firm's value process to make both barriers $A(t)$ and $L(t)$ constant. Since $e^{-\gamma t} F_t$ is a monotonic function of X_t, a $(\widetilde{\mathbb{P}}, \mathbb{F})$-Brownian motion with drift $\mu := (r^f - \gamma)/\sigma$, the default time $\tau^{(\vartheta)} = \tau_\ell^{\vartheta} \wedge \tau_a$ can be re-expressed as (where $a = \frac{1}{\sigma}\ln(A_0/F_0)$ and $\ell = \frac{1}{\sigma}\ln(L_0/F_0)$):

$$\tau_A = \inf\{t \geq 0 : X_t \leq a\} := \tau_a, \qquad \tau_L^{\vartheta} = \inf\{t \geq 0 : \mathcal{A}_t^{\ell,-} > \vartheta\} := \tau_\ell^{\vartheta}. \tag{33}$$

Let τ be an a.s. positive $\mathcal{F}$-measurable random variable. We fix a finite time horizon $T > 0$, and let $D_{\tau\wedge T}$ be the (integrable) payoff of a defaultable claim, taking the following form:

$$D_{\tau\wedge T} := h(\tau \wedge T, X_{\tau\wedge T}) = h(T, X_T)\mathbb{I}_{\{\tau\geq T\}} + h(\tau, X_\tau)\mathbb{I}_{\{\tau<T\}}, \tag{34}$$

where $h : [0,T] \times \mathcal{I} \to \mathbb{R}$ is a Borel function. Thus, the no-arbitrage time-t value of the claim is

$$\begin{aligned} D_t :=& B_t \widetilde{\mathbb{E}}\Big[B_{\tau\wedge T}^{-1} D_{\tau\wedge T} \,|\, \mathcal{F}_t\Big] \\ =& B_t\, \widetilde{\mathbb{E}}\Big[B_\tau^{-1} h(\tau, X_\tau)\mathbb{I}_{\{\tau<T\}} \,|\, \mathcal{F}_t\Big] + B_t \widetilde{\mathbb{E}}\Big[B_T^{-1} h(T, X_T)\mathbb{I}_{\{\tau\geq T\}} \,|\, \mathcal{F}_t\Big]; \quad t \in [0,T], \end{aligned} \tag{35}$$

where $B_t := e^{r^f t}$ is the bank account and $\widetilde{\mathbb{E}}$ is the expectation operator under $\widetilde{\mathbb{P}}$ (the risk-neutral expectation). In what follows, we state a new theorem pertaining to general pricing formulas under the occupation time models (and the new results will be used to prove general pricing formulas under the hazard rate models to be described in Section 5). However, at first, we shall state and prove the following lemma, describing the time-homogeneity of the price process.

Lemma 1. *Let $\tau^{(\vartheta)} = \tau_\ell^\vartheta \wedge \tau_a$ be the default time defined in (33), and*[8]

$$D_{t,\mathsf{A},x}^{\vartheta,T} := B_t \widetilde{\mathbb{E}}\left[B^{-1}_{\tau^{(\vartheta)}\wedge T} D_{\tau^{(\vartheta)}\wedge T} | \mathcal{F}_t\right] = B_t \widetilde{\mathbb{E}}_{t,\mathsf{A},x}\left[B^{-1}_{\tau^{(\vartheta)}\wedge T} D_{\tau^{(\vartheta)}\wedge T}\right] \tag{36}$$

be the time-t value of a T-maturity credit derivative with a grace period ϑ. Then, on the set $\{\tau^{(\vartheta)} > t\} = \{\tau_\ell^\vartheta > t, \tau_a > t\}$,

$$D_{t,\mathsf{A},x}^{\vartheta,T} = D_{0,x}^{\vartheta',T-t}, \tag{37}$$

where $\vartheta' := \vartheta - \mathsf{A}$ is the (realized) remaining grace period at time t.

Proof. Define $\hat{\tau}^{(\vartheta)} := \hat{\tau}_\ell^\vartheta \wedge \hat{\tau}_a$ where

$$\hat{\tau}_a := \inf\{s \geq 0 : X_{t+s} \leq a\}, \qquad \hat{\tau}_\ell^\vartheta := \inf\left\{s \geq 0 : \int_t^{t+s} \mathbb{I}_{(-\infty,\ell]}(X_u)du > \vartheta\right\}. \tag{38}$$

Then we can easily show that, on the set $\{\tau^{(\vartheta)} > t\}$,

$$\begin{aligned} \tau_a &= t + \inf\{s \geq 0 : X_{t+s} \leq a\} = t + \hat{\tau}_a, \\ \tau_\ell^\vartheta &= t + \inf\{s \geq 0 : \mathcal{A}_{s+t}^{\ell,-} - \mathcal{A}_t^{\ell,-} > \vartheta - \mathcal{A}_t^{\ell,-}\} = t + \hat{\tau}_\ell^{\vartheta'}, \end{aligned} \tag{39}$$

where $\vartheta' = \vartheta - \mathcal{A}_t^{\ell,-}$. Therefore $\tau^{(\vartheta)} = t + \hat{\tau}^{(\vartheta')}$ (a.s.) and we obtain

$$D_{t,\mathsf{A},x}^{\vartheta,T} = B_t \widetilde{\mathbb{E}}_{t,\mathsf{A},x}\left[B^{-1}_{\tau^{(\vartheta)}\wedge T} D_{\tau^{(\vartheta)}\wedge T}\right] = \widetilde{\mathbb{E}}_x\left[B^{-1}_{\hat{\tau}^{(\vartheta')}\wedge(T-t)} D_{\hat{\tau}^{(\vartheta')}\wedge(T-t)}\right] = D_{0,x}^{\vartheta',T-t}. \tag{40}$$

□

Theorem 1. *The time-0 value of the credit derivative in (36) is given by* [9]*:*

$$\begin{aligned} D_{0,x}^{\vartheta,T} &= B_T^{-1}\left\{\int_0^\vartheta \int_a^\infty h(T,y) f^{(a,\infty)}_{\mathcal{A}_T^{\ell,-},X_T}(u,y|x)dydu + \int_\ell^\infty h(T,y)p_\ell^+(T;x,y)dy\right\} \\ &+ \left\{\int_\vartheta^T B_s^{-1}\int_a^\infty h(s,y) f^{(a,\infty)}_{\mathcal{A}_s^{\ell,-},X_s}(\vartheta,y|x)dyds + B_\vartheta^{-1}\int_a^\ell h(\vartheta,y)p_{(a,\ell)}(\vartheta;x,y)dy\right\} \\ &- \left\{\int_0^\vartheta B_s^{-1}h(s,a)\int_a^\infty \frac{\partial p_a^+}{\partial s}(s;x,y)dyds\right. \\ &+ \int_\vartheta^T B_s^{-1}h(s,a)\left(\int_0^\vartheta\int_a^\infty \frac{\partial f^{(a,\infty)}_{\mathcal{A}_s^{\ell,-},X_s}}{\partial s}(u,y|x)dydu + \int_\ell^\infty \frac{\partial p_\ell^+}{\partial s}(s;x,y)dy\right)ds \\ &\left.+ \int_\vartheta^T B_s^{-1}h(s,a)\int_a^\infty f^{(a,\infty)}_{\mathcal{A}_s^{\ell,-},X_s}(\vartheta,y|x)dyds\right\}, \end{aligned} \tag{41}$$

for $\vartheta \in (0,T)$, and

$$D_{0,x}^{\vartheta,T} = B_T^{-1}\int_a^\infty h(T,y)p_a^+(T;x,y)dy - \int_0^T B_s^{-1}h(s,a)\int_a^\infty \frac{\partial p_a^+}{\partial s}(s;x,y)dyds, \tag{42}$$

for $\vartheta \geq T$.[10]

Proof. Here, we summarize what is needed to compute (41). By the Optional Sampling Theorem, we have

$$\begin{aligned} D_{0,x}^{\vartheta,T} &= B_T^{-1}\widetilde{\mathbb{E}}\left[h(T,X_T)\mathbb{I}_{\left\{\tau_a \geq T, \tau_\ell^\vartheta \geq T\right\}}\right] + \widetilde{\mathbb{E}}_x\left[B^{-1}_{\tau_\ell^\vartheta} h(\tau_\ell^\vartheta, X_{\tau_\ell^\vartheta})\mathbb{I}_{\left\{\tau_\ell^\vartheta < \tau_a, \tau_\ell^\vartheta < T\right\}}\right] \\ &+ \widetilde{\mathbb{E}}_x\left[B^{-1}_{\tau_a} h(\tau_a, X_{\tau_a})\mathbb{I}_{\left\{\tau_a \leq \tau_\ell^\vartheta, \tau_a < T\right\}}\right], \end{aligned} \tag{43}$$

where $X_{\tau_a} = a$ (i.e., the value of the process, as soon as it hits the default barrier, is a). By computing each expectation in (43) (we omit the lengthy proof), we obtain (41). □

Each bracketed term in (41) corresponds to one of the three mathematical expectations in (43). If the firm avoids being liquidated and is solvent at maturity, the time-t value of the claim is given by the first term in (43). If the firm is liquidated prior to maturity due to exceeding the grace period, the claim's value is given by the second term in (43). The last term in (43) is the time-t claim's value corresponding to the scenario when the firm is liquidated early due to reaching the liquidation barrier $A(t)$. For simplicity of presentation, we assume that the same function h describes the payoff value for each scenario.

5. Hazard Rate Model

Suppose now that $(\Omega, \mathcal{H}, \widetilde{\mathbb{P}}, \mathbb{H})$ is a filtered probability space where we set $\mathbb{H} := \mathbb{F} \vee \mathbb{J}$ with filtration $\mathbb{J}$ so that τ is a $\mathbb{J}$-stopping time. Hybrid models unify the structural and reduced-form models by an $\mathbb{F}$-adapted hazard process Γ defined by

$$\Gamma_t := \int_0^t \lambda(F_s)ds; \quad t \geq 0, \tag{44}$$

with an $\mathbb{F}$-adapted hazard rate process λ. For now, let us assume that

$$\lambda(F_s) = \alpha \mathbb{I}_{\{F_s \leq L(s)\}}; \quad \alpha \geq 0, \tag{45}$$

where $L(s)$ is the occupation barrier defined in (24). We define an $\mathbb{H}$-stopping time $\tau^{(\alpha)} := \tau_L^\alpha \wedge \tau_A = \tau_\ell^\alpha \wedge \tau_a$, where

$$\tau_L^\alpha := \inf\{t \geq 0 : \alpha \mathcal{A}_t^{(F),\mathbf{L},-} > \xi\} = \inf\{t \geq 0 : \alpha \mathcal{A}_t^{\ell,-} > \xi\} := \tau_\ell^\alpha; \quad \xi \sim \text{Exp}(1), \tag{46}$$

and $\tau_A = \tau_a$ was defined previously. The default time $\tau^{(\alpha)}$ characterizes both Chapters 7 and 11 type defaults. Usually, it is convenient to rewrite the expression in (46) as

$$\tau_\ell^\alpha = \inf\{t \geq 0 : \mathcal{A}_t^{\ell,-} > \xi^\alpha\}; \quad \xi^\alpha \sim \text{Exp}(\alpha), \tag{47}$$

so that the hazard rate model[11] can be viewed as the occupation time model with an (exogenous) randomization in $\vartheta \sim \text{Exp}(\alpha)$. In the case where the firm's value is the GBM process, it is obvious that when $\alpha = 0$, the hazard rate model reduces to the Black–Cox model with default barrier $A(t)$. Similarly when $\alpha \to \infty$, the hazard rate model reduces to the Black–Cox model with default barrier $L(t)$. We can easily extend to where the hazard rate process is

$$\lambda(F_s) = \alpha_1 \mathbb{I}_{\{F_s \geq L(s)\}} + \alpha_2 \mathbb{I}_{\{F_s \leq L(s)\}}; \quad 0 \leq \alpha_1 \leq \alpha_2. \tag{48}$$

In this case, we define the default time as $\tau^{(\alpha_1,\alpha_2)} := \tau_\ell^{\alpha_1,\alpha_2} \wedge \tau_a$, where

$$\tau_\ell^{\alpha_1,\alpha_2} := \inf\{t \geq 0 : \alpha_1 \mathcal{A}_t^{\ell,+} + \alpha_2 \mathcal{A}_t^{\ell,-} > \xi\}; \quad \xi \sim \text{Exp}(1). \tag{49}$$

For the GBM case, we can recover the Alfonsi–Lelong model by sending the default level $A(t)$ in the F-process to zero, i.e., $a \to -\infty$. We will not employ the same approach used in the Alfonsi–Lelong model, but instead we shall make use of the spectral expansion methodology to obtain closed-form pricing formulas. The rest of this section is devoted to general pricing formulas for credit derivatives under this new framework. Before we state the main theorem, we will state and prove the following lemma. The lemma draws the connections between the occupation time and hazard rate models.

Lemma 2. *Let $\tau^{(\alpha)} = \tau_\ell^\alpha \wedge \tau_a$ be the default time defined in (46), and*

$$D_{t,x}^{\alpha,T} := B_t \widetilde{\mathbb{E}}\Big[B^{-1}_{\tau^{(\alpha)}\wedge T} D_{\tau^{(\alpha)}\wedge T} \big|\, \mathcal{H}_t\Big] = B_t \widetilde{\mathbb{E}}\Big[B^{-1}_{\tau^{(\alpha)}\wedge T} D_{\tau^{(\alpha)}\wedge T} \big|\, X_t = x\Big] \tag{50}$$

be the time-t value of the credit derivative under the hazard rate model (with λ defined in (45)). Then, on the set $\{\tau^{(\alpha)} > t\} = \{\tau_\ell^\alpha > t, \tau_a > t\}$, we have

$$D_{t,x}^{\alpha,T} = \alpha \mathcal{L}_\vartheta \Big\{ D_{0,x}^{\vartheta, T-t} \Big\}(\alpha), \tag{51}$$

and is independent of $\mathcal{A}_t^{\ell,-}$. Moreover, the price process under the occupation time model can be recovered by

$$D_{t,\mathsf{A},x}^{\vartheta,T} = \mathcal{L}_\alpha^{-1} \Big\{ \alpha^{-1} D_{t,x}^{\alpha,T} \Big\}(\vartheta'); \quad \vartheta' := \vartheta - \mathsf{A}. \tag{52}$$

Proof. Let Y be an integrable $\mathcal{H}_\infty$-measurable random variable and Γ_t be an $\mathbb{F}$-adapted hazard process, then by page 145 (5.11) and (5.12) from Bielecki and Rutkowski (2013), we obtain

$$\widetilde{\mathbb{E}}\Big[\mathbb{I}_{\{\tau_\ell^\alpha > t\}} Y \big|\, \mathcal{H}_t\Big] = \mathbb{I}_{\{\tau_\ell^\alpha > t\}} \widetilde{\mathbb{E}}\Big[e^{\Gamma_t} Y \big|\, \mathcal{F}_t\Big]. \tag{53}$$

By substituting $Y = \mathbb{I}_{\{\tau_a > t\}} B^{-1}_{\tau^{(\alpha)}\wedge T} D_{\tau^{(\alpha)}\wedge T}$ and $\Gamma_t = \alpha \mathcal{A}_t^{\ell,-}$ into (53), we obtain (51). □

Theorem 2. *The time-0 value of the credit derivative in (50) is given by*[12]*:*

$$\begin{aligned} D_{0,x}^{\alpha,T} =& B_T^{-1} \int_a^\infty h(T,y) \tilde{p}^{\ell,-}_{(a,\infty),\alpha}(T;x,y) dy \\ &+ \alpha \int_0^T B_s^{-1} \int_a^\infty h(s,y) \Big(\tilde{p}^{\ell,-}_{(a,\infty),\alpha}(s;x,y) - p_\ell^+(s;x,y) \Big) dy ds \\ &- \int_0^T B_s^{-1} h(s,a) \int_a^\infty \left(\frac{\partial \tilde{p}^{\ell,-}_{(a,\infty),\alpha}}{\partial s}(s;x,y) + \alpha [\tilde{p}^{\ell,-}_{(a,\infty),\alpha}(s;x,y) - p_\ell^+(s;x,y)] \right) dy ds. \end{aligned} \tag{54}$$

Proof. By Lemma 2, we obtain (54) from (41) and (42) by direct computations. □

Comparing (42) with (54), we notice that the r.h.s. of (42) is the limit of the r.h.s. of (54), as $\alpha \to 0$ or as $\ell \to a$. That is, the credit derivative value in (54) converges to the value under the Black–Cox model with default barrier $A(t)$.

6. Probability of Default

A survival probability is a credit derivative of the form in (34) with payoff $D_{\tau\wedge T} = \mathbb{I}_{\{\tau \geq T\}}$ and $B_t := 1$. By Theorem 1, the (unconditional) survival probability at time t under the occupation time model is

$$\widetilde{\mathbb{P}}_x(\tau^{(\vartheta)} > T) = \int_0^\vartheta \int_a^\infty f^{(a,\infty)}_{\mathcal{A}_T^{\ell,-}, X_T}(u,y|\,x) dy du + \int_\ell^\infty p_\ell^+(T;x,y) dy. \tag{55}$$

(Note: The survival probability is clearly zero for $x \leq a$.) By Theorem 2, the (unconditional) survival probability at time T under the hazard rate model (with λ defined in (45)) is[13]

$$\widetilde{\mathbb{P}}_x(\tau^{(\alpha)} > T) = \int_a^\infty \tilde{p}^{\ell,-}_{(a,\infty),\alpha}(T;x,y) dy. \tag{56}$$

For any regular diffusion, Equation (56) can be calculated using the spectral expansion method (assuming the spectral series can be integrated term-by-term), where $b \to \infty$:

$$\int_a^b \tilde{p}^{\ell,-}_{(a,b),\alpha}(T;x,y)dy = \sum_{n=1}^{\infty} e^{-\tilde{\lambda}_n T} \int_a^b \mathfrak{m}(y)\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)dy. \tag{57}$$

By Lemma 2, the (unconditional) survival probability in (55) under the occupation time model can be obtained by (which can be evaluated numerically via the Gaver–Stehfest algorithm)

$$\widetilde{\mathbb{P}}_x(\tau^{(\vartheta)} > T) = \mathcal{L}_\alpha^{-1}\Big\{\alpha^{-1}\widetilde{\mathbb{P}}_x(\tau^{(\alpha)} > T)\Big\}(\vartheta). \tag{58}$$

Some graphs of the probability of default for the GBM process are shown in Figure 3. We can observe that the larger the α value, the greater the probability of default is. This makes sense because as α increases, the firm has to carry more risks by allowing the counterparty to penalize more for the firm's value staying below the occupation level. Similarly, we can say that the smaller the ϑ value, the greater the probability of default is, as the firm is required to default immediately once the occupation time exceeds the grace period. The difference between the models is that default probability values under the hazard rate model are more flexible across all maturity times, but the occupation time model does not correct the short-time behaviour whose values overlap with that in the Black–Cox model with default barrier $A(t)$.

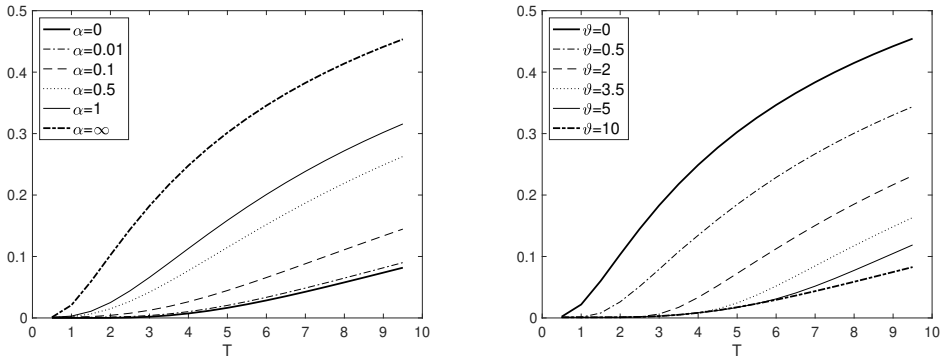


Figure 3. Graphs of default probabilities for the hazard rate (**left**) and occupation time models (**right**). The underlying firm's value is GBM with $F_0 = 100$, $L_0 = 50$, $A_0 = 20$, $r^f = 5\%$, $\sigma = 30\%$, $\gamma = 5\%$, and $N = 30$. For the right graph, we used 16 terms in the Gaver–Stehfest algorithm.

7. Implied Hazard Rate Function

A reduced-form model is a model where a default event is characterized by a hazard rate (ordinary) function $\lambda(t)$. The hazard rate function is defined so that $\lambda(t)\Delta t$ is the probability of defaulting between time t and $t + \Delta t$ conditional on no default until time t (where τ is a default time):

$$\lambda(t) := \lim_{\Delta t \to 0} \frac{\widetilde{\mathbb{P}}(t \le \tau \le t + \Delta t | \tau > t)}{\Delta t} = \frac{-\frac{d}{dt}\widetilde{\mathbb{P}}(\tau > t)}{\widetilde{\mathbb{P}}(\tau > t)}. \tag{59}$$

If τ is the default time pre-specified by a model, then the hazard rate function implied by the model, denoted by $\lambda^*(t)$, can be calculated using (59). Implied hazard rate functions provide a uniform way of comparing default behaviours across different models. Under the hazard rate model (with λ defined in (45)), the implied hazard rate function can be

computed analytically via the spectral expansion method (assuming that the series in (57) can be differentiated term-by-term):[14]

$$\lambda^*(t) = \frac{\sum_{n=1}^{\infty} \tilde{\lambda}_n e^{-\tilde{\lambda}_n t} \int_a^b \mathfrak{m}(y) \tilde{\phi}_{n,\alpha}^{\ell,-}(x) \tilde{\phi}_{n,\alpha}^{\ell,-}(y) dy}{\sum_{n=1}^{\infty} e^{-\tilde{\lambda}_n t} \int_a^b \mathfrak{m}(y) \tilde{\phi}_{n,\alpha}^{\ell,-}(x) \tilde{\phi}_{n,\alpha}^{\ell,-}(y) dy}. \tag{60}$$

Some graphs of the implied hazard rate functions are given in Figure 4. We can see that the implied hazard rate functions are generally increasing in time. For large α (small ϑ), implied hazard rate functions increase sharply up to certain times, and decrease gradually (and end up converging to a certain value). The hazard rate functions under the occupation time model are relatively lower (especially for short times) compared to that under the hazard rate model. This observation makes sense from the shape of the probability of defaults described in Section 6.

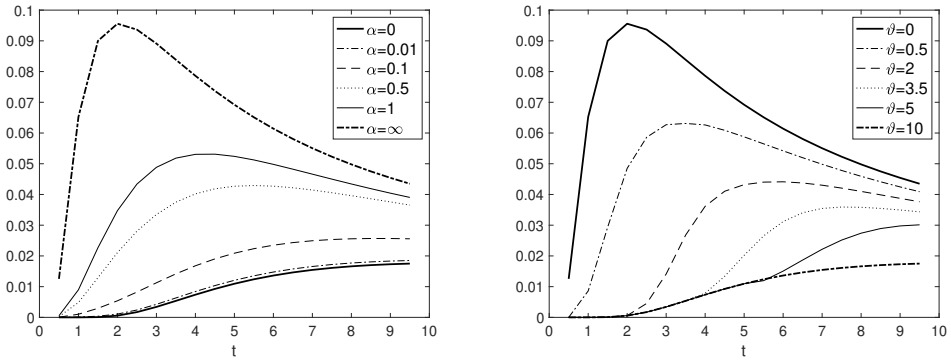


Figure 4. Graphs of implied hazard rate functions for the hazard rate model (**left**) and occupation time model (**right**). The underlying firm's value is GBM with $F_0 = 100$, $L_0 = 50$, $A_0 = 20$, $r^f = 5\%$, $\sigma = 30\%$, $\gamma = 5\%$, and $N = 30$. For the right graph, we used 16 terms in the Gaver–Stehfest algorithm.

8. Credit Default Swap (CDS) Spreads

We assume that there are p regular (time-proportional) payments on the time grid $\{T_1, T_2, \ldots, T_p\}$ with $0 < T_1 < \cdots < T_p = T$ until the default event and one last payment at time τ. The fair price of the CDS spread is given by

$$R(0,T) := \frac{DL(0,T)}{PL(0,T)} = \frac{LGD[e^{-r^f T}\widetilde{\mathbb{P}}(\tau \le T) + \int_0^T r^f e^{-r^f u}\widetilde{\mathbb{P}}(\tau \le u)du]}{\int_0^T e^{-r^f u}\widetilde{\mathbb{P}}(\tau > u)du - \int_0^T r^f e^{-r^f u}(u - T_{\beta(u)-1})\widetilde{\mathbb{P}}(\tau > u)du}. \tag{61}$$

where $LGD \in [0,1]$ is the Loss Given Default, which is assumed to be deterministic, and $\beta(t) \in \{1, \ldots, p\}$ is the index of the next payment date satisfying $T_{\beta(t)-1} \le t < T_{\beta(t)}$. For our convenience, we shall rewrite the Equation (61) in terms of the survival probabilities:

$$R(0,T) = \frac{1 - e^{-r^f T}\widetilde{\mathbb{P}}(\tau > T) - r^f \mathfrak{f}(T)}{\mathfrak{f}(T) - r^f(\mathfrak{g}(T) - [T_{p-1}\mathfrak{f}(T) - \sum_{i=1}^{p}(T_i - T_{i-1})\mathfrak{f}(T_i)])} \tag{62}$$

where

$$\mathfrak{f}(t) := \int_0^t e^{-r^f u}\widetilde{\mathbb{P}}(\tau > u)du, \qquad \mathfrak{g}(t) := \int_0^t u e^{-r^f u}\widetilde{\mathbb{P}}(\tau > u)du. \tag{63}$$

We use the following proposition to calculate the fair value of the CDS spreads under the hazard rate model.

Proposition 1. *Under the hazard rate model (with λ defined in (45)), the time-0 value of a CDS spread is given in (62), where $\mathfrak{f}$ and $\mathfrak{g}$ are defined in (63), admit spectral expansion forms:*[15,16]

$$\begin{aligned}\mathfrak{f}(t) &= \int_a^b \widetilde{G}_{\alpha}^{\ell,-}(r^f;x,y)dy - \sum_{n=1}^{\infty} \frac{e^{-(\tilde{\lambda}_n+r^f)t}}{\tilde{\lambda}_n+r^f} \int_a^b \mathfrak{m}(y)\tilde{\phi}_{n,\alpha}^{\ell,-}(x)\tilde{\phi}_{n,\alpha}^{\ell,-}(y)dy,\\ \mathfrak{g}(t) &= -\int_a^b \frac{\partial}{\partial\lambda}\widetilde{G}_{\alpha}^{\ell,-}(\lambda;x,y)dy\Big|_{\lambda=r^f}\\ &\quad - \sum_{n=1}^{\infty} \frac{1+(\tilde{\lambda}_n+r^f)t}{(\tilde{\lambda}_n+r^f)^2} e^{-(\tilde{\lambda}_n+r^f)t} \int_a^b \mathfrak{m}(y)\tilde{\phi}_{n,\alpha}^{\ell,-}(x)\tilde{\phi}_{n,\alpha}^{\ell,-}(y)dy,\end{aligned} \tag{64}$$

where $\widetilde{G}^{\ell,-}$ is the Green function defined by $\widetilde{G}_{\alpha}^{\ell,-}(\lambda;x,y) := \mathcal{L}_t\{\tilde{p}_{\alpha}^{\ell,-}(t;x,y)\}(\lambda)$.

Proof. For $\mathfrak{f}$, we can integrate term by term to obtain

$$\mathfrak{f}(t) = \int_a^b \mathfrak{m}(y)\left(\sum_{n=1}^{\infty} \frac{\tilde{\phi}_{n,\alpha}^{\ell,-}(x)\tilde{\phi}_{n,\alpha}^{\ell,-}(y)}{\tilde{\lambda}_n+r^f} - \sum_{n=1}^{\infty} \frac{e^{-(\tilde{\lambda}_n+r^f)t}}{\tilde{\lambda}_n+r^f}\tilde{\phi}_{n,\alpha}^{\ell,-}(x)\tilde{\phi}_{n,\alpha}^{\ell,-}(y)\right)dy. \tag{65}$$

We know, by definition, that the first series in (65) is simply the Green function

$$\widetilde{G}_{\alpha}^{\ell,-}(\lambda;x,y) = \mathfrak{m}(y)\sum_{n=1}^{\infty} \frac{\tilde{\phi}_{n,\alpha}^{\ell,-}(x)\tilde{\phi}_{n,\alpha}^{\ell,-}(y)}{\tilde{\lambda}_n+\lambda}. \tag{66}$$

For $\mathfrak{g}$ we can employ the integration by parts formula to get

$$\mathfrak{g}(t) = \sum_{n=1}^{\infty}\left(\frac{1}{(\tilde{\lambda}_n+r^f)^2} - \frac{1+(\tilde{\lambda}_n+r^f)t}{(\tilde{\lambda}_n+r^f)^2}e^{-(\tilde{\lambda}_n+r^f)t}\right)\int_a^b \mathfrak{m}(y)\tilde{\phi}_{n,\alpha}^{\ell,-}(x)\tilde{\phi}_{n,\alpha}^{\ell,\pm}(y)dy, \tag{67}$$

where the first series in (67) can be expressed in terms of the Green function, and hence we obtain (64). □

Both series in (64) converge rapidly since the denominator $\tilde{\lambda}_n + r^f$ tends to ∞ as $n \to \infty$.

9. Calibration of CDS Spreads: GBM Case

We calibrate our models to some market data where the underlying firm's value is assumed to be GBM. We extract the market CDS spreads data for the Total Energies company (TTE) with spot time on 1 June 2022. The market data contain eight sample market data points at the following tenor values: $T = 0.5, 1, 2, 3, 4, 5, 7, 10$ years. Our calibration approach is find optimal parameter values that (locally) minimizes the gap between the market and model prices through a loss function. Here, we use the (unweighted) root mean squared error (RMSE) as the loss function (where Θ is the set of parameters in a given model):

$$L(\Theta) = \sqrt{\frac{\sum_{i=1}^{8}\left(R_{T_i}^{Mdl} - R_{T_i}^{Mkt}\right)^2}{8}}$$

where $R_{T_i}^{Mkt}$ is the i^{th} observed market CDS spread and $R_{T_i}^{Mdl}$ is the CDS spread implied by the model at T_i. We compare the following four models:

- The Black–Cox model Black and Cox (1976)
- The occupation time model Makarov (2016)
- The Alfonsi–Lelong model Alfonsi and Lelong (2012)
- The new hazard rate model (with λ defined in (48))

The summary of the market data used, as well as the calibration results, can be found in Tables 1–3. We employ an iterative scheme for our calibration procedure, that is, we

calibrate parameters in the Black–Cox model, and then use the calibrated values as initial guesses to calibrate other parameters in the occupation time and hazard rate models. We do not use the iterative scheme for the Alfonsi–Lelong model since that model does not contain the default barrier $A(t)$. Figure 5 shows that the Black–Cox and occupation time models fitted poorly for small tenors. The hazard rate model significantly improve the result and led to a near-perfect calibration. The Alfonsi–Lelong model still outperforms the Black–Cox model, but is not as accurate as the hazard rate model. We use calibrated values from Table 3 to calculate the risk-neutral probabilities implied from the market CDS spreads under the hazard rate model. Based on Figure 6, we see that the Total Energies company is currently at a low default risk.

Table 1. Initial Information about the market.

Variable Name	Description	Value
F_0	initial firm's value	55.59
r^f	constant risk-free rate	5%
σ	constant volatility	28%
LGD	constant Loss Given Default	0.6

Table 2. Market Data of Tenor (year) and CDS spread (bps).

Tenor	0.5	1	2	3	4	5	7	10
CDS	11.86	15.13	21.29	28.79	37.21	45.83	60.03	73.17

Table 3. Calibration Results for model parameters (the underlying process is GBM). NA means not applicable.

Variable	Description	Black–Cox	Occupation	A–L	Hazard Rate
A_0	default barrier (at time 0)	18.96	13.47	NA	12.83
γ	growth rate of barrier	-3.51%	-0.19%	11.22%	-0.32%
L_0	occupation barrier (at time 0)	NA	23.29	28.78	38.42
θ/T	grace period relative to T	NA	0.2368	NA	NA
α_1	killing rate (above L)	NA	NA	0.27%	0.18%
α_2	killing rate (below L)	NA	NA	3.42%	2.33%
	loss function value	5.70×10^{-6}	5.16×10^{-6}	3.10×10^{-8}	5.77×10^{-10}

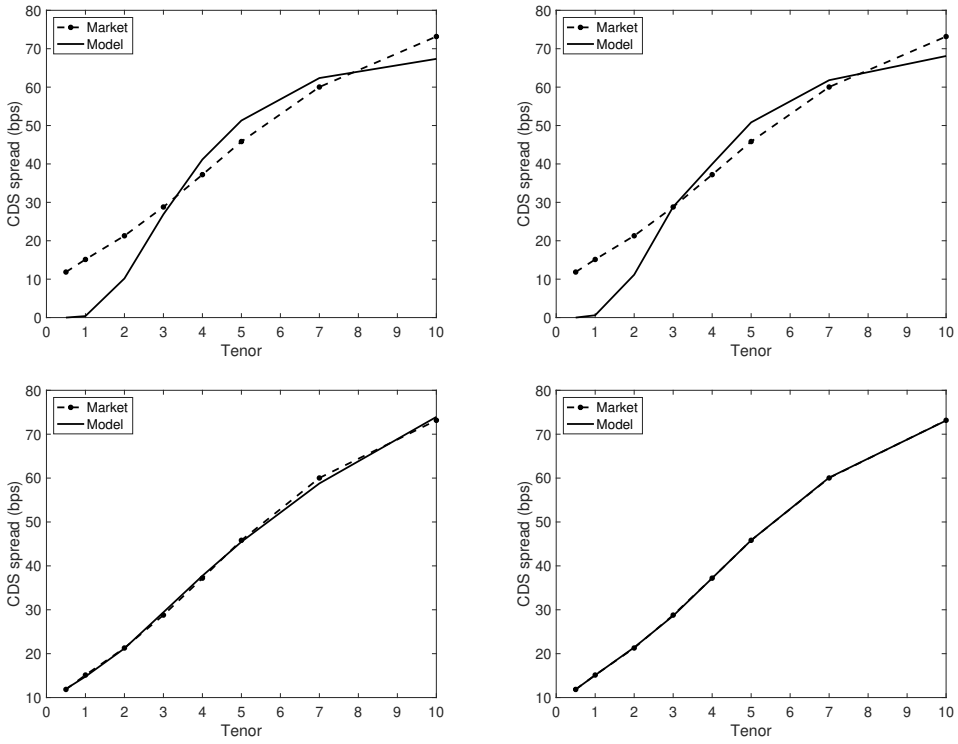


Figure 5. Graphs of CDS Spreads for four models: Black–Cox (**top-left**), occupation time (**top-right**), Alfonsi–Lelong (**bottom-left**), and hazard rate (**bottom-right**). The underlying firm's value is GBM.

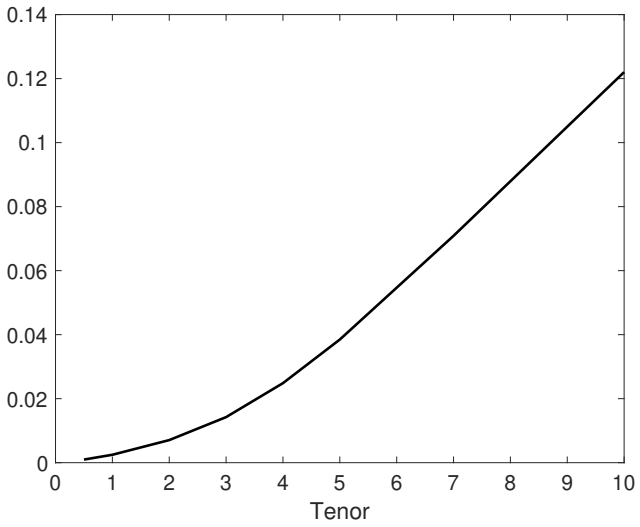


Figure 6. Implied risk-neutral probability for the hazard rate model (the underlying firm's value is GBM).

10. Conclusions

This study led to the innovation of two new credit risk models, namely, the occupation time and hazard rate models. They captured both Chapters 7 and 11 type defaults (the occupation time model was considered in Makarov et al. (2015); Makarov (2016) but the papers did not use the spectral expansion method). We derived closed-form pricing formulas for credit derivatives. Moreover, the hazard rate model prices can be expressed explicitly for diffusions, such as a geometric Brownian motion, and many other solvable processes. The pricing formulas under the occupation time model can be obtained by a Laplace inverse transformation of the hazard rate model. The Laplace inversion is

performed numerically using the Gaver–Stehfest algorithm. The hazard rate model can capture versatile default probability values, which, in the occupation time models, are immovable at short maturity times. Our models are calibrated to the market CDS spreads from the Total Energies company. Our calibration results show that the computations are fast and lead to near-perfect calibrations to typical market CDS spread data.

Our main future work is to employ explicit expressions for alternative solvable diffusions. We can then consider pricing credit derivatives under such alternative models including the Constant Elasticity of Variance (CEV) model and other nonlinear local volatility models. Additionally, we can construct new structural models of credit risk based on the last passage time. In this paper, we have only considered senior debts, but we may also look into junior (subordinated) debts as well. One example is a contingent convertible (CoCo) bond which is a bond that converts into equity once the debt-to-equity ratio falls to a certain threshold level. CoCo bonds are popular among firms since firms can avoid default events, to a certain extent, by converting the CoCo bonds into equity once a catastrophic event triggers. We may also consider the pricing and calibrations of the new models to standard equity options, and thereby study the interplay between equity and credit markets.

Author Contributions: Conceptualization, G.C. and R.N.M.; methodology, G.C. and H.K.; software, G.C., H.K. and R.N.M.; validation, H.K.; formal analysis, G.C.; investigation, G.C., H.K. and R.N.M.; resources, H.K.; data curation, H.K.; writing—original draft preparation, H.K.; writing—review and editing, G.C. and R.N.M.; visualization, H.K.; supervision, G.C. and R.N.M.; project administration, G.C.; funding acquisition, G.C. and R.N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant numbers 2020-04782 and 2018-06176.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Explicit Expressions for a Drifted Brownian Motion

We provide explicit expressions on default probabilities and the CDS spread valuations for a Brownian motion with drift $\mu \in \mathbb{R}$, starting at $x \in (a, b)$ with imposed killings at levels a, b where $a < x < b$.

Appendix A.1. Default Probabilities

In what follows, we provide explicit formulas for the integral in (57). Let $\{\tilde{\lambda}_n\}$ be the eigenvalues (refer to Equation (10)), satisfying the following eigenvalue equation:[17]

$$\begin{cases} \sqrt{2(\tilde{\lambda}_n-\alpha)}\cos\Big(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a)\Big)\sin\Big(\sqrt{2\tilde{\lambda}_n}(b-\ell)\Big) & \\ +\sqrt{2\tilde{\lambda}_n}\cos\Big(\sqrt{2\tilde{\lambda}_n}(b-\ell)\Big)\sin\Big(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a)\Big)=0, & \tilde{\lambda}_n > \alpha, \\ \sqrt{2(\alpha-\tilde{\lambda}_n)}\sin\Big(\sqrt{2\tilde{\lambda}_n}(b-\ell)\Big)\cosh\Big(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a)\Big) & \\ +\sqrt{2\tilde{\lambda}_n}\cos(\sqrt{2\tilde{\lambda}_n}(b-\ell))\sinh\Big(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a)\Big)=0, & \tilde{\lambda}_n \in (0,\alpha), \end{cases} \tag{A1}$$

and let $\{\tilde{\phi}^{\ell,-}_{n,\alpha}\}$ be the eigenfunctions given as follows. We only provide formulas for $x \geq \ell$ (i.e., the firm's value starts above the liquidation level) since the other case is rarely used in practice.

- If $x \in [\ell, b), y \in (a, \ell], \tilde{\lambda}_n \in (0, \alpha)$,

$$\begin{aligned}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)=&\Big\{\sin(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[(b-a)\sinh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a)+\tfrac{\cosh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))}{\sqrt{2(\alpha-\tilde{\lambda}_n)}}\Big]\\&-\cos(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[\tfrac{\sinh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))}{\sqrt{2\tilde{\lambda}_n}}+\Big(\sqrt{\tfrac{\alpha-\tilde{\lambda}_n}{\tilde{\lambda}_n}}(b-\ell)-\sqrt{\tfrac{\tilde{\lambda}_n}{\alpha-\tilde{\lambda}_n}}(\ell-a)\Big)\cosh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))\Big]\Big\}^{-1}\\&\times\sin(\sqrt{2\tilde{\lambda}_n}(b-x))\sinh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(y-a)).\end{aligned}\tag{A2}$$

- If $x\in[\ell,b), y\in(a,\ell], \tilde{\lambda}_n>\alpha$,

$$\begin{aligned}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)=&\Big\{\sin(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[(b-a)\sin(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a)-\tfrac{\cos(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))}{\sqrt{2(\tilde{\lambda}_n-\alpha)}}\Big]\\&-\cos(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[\tfrac{\sin(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))}{\sqrt{2\tilde{\lambda}_n}}+\Big(\sqrt{\tfrac{\tilde{\lambda}_n-\alpha}{\tilde{\lambda}_n}}(b-\ell)+\sqrt{\tfrac{\tilde{\lambda}_n}{\tilde{\lambda}_n-\alpha}}(\ell-a)\Big)\cos(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))\Big]\Big\}^{-1}\\&\times\sin(\sqrt{2\tilde{\lambda}_n}(b-x))\sin(\sqrt{2(\tilde{\lambda}_n-\alpha)}(y-a)).\end{aligned}\tag{A3}$$

- If $x\in[\ell,b), y\in[\ell,b), \tilde{\lambda}_n\in(0,\alpha)$,

$$\begin{aligned}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)=&\tfrac{\sqrt{2(\alpha-\tilde{\lambda}_n)}\cosh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))\sin(\sqrt{2\tilde{\lambda}}(\ell-a))-\sqrt{2\tilde{\lambda}_n}\cos(\sqrt{2\tilde{\lambda}_n}(\ell-a))\sinh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a)}{\sqrt{2\tilde{\lambda}_n}\sin(\sqrt{2\tilde{\lambda}_n}(b-a))}\\&\times\Big\{\cos(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[\tfrac{\sinh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))}{\sqrt{2\tilde{\lambda}_n}}+\Big(\sqrt{\tfrac{\alpha-\tilde{\lambda}_n}{\tilde{\lambda}_n}}(b-\ell)-\sqrt{\tfrac{\tilde{\lambda}_n}{\alpha-\tilde{\lambda}_n}}(\ell-a)\Big)\cosh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))\Big]\\&-\sin(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[(b-a)\sinh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a)+\tfrac{\cosh(\sqrt{2(\alpha-\tilde{\lambda}_n)}(\ell-a))}{\sqrt{2(\alpha-\tilde{\lambda}_n)}}\Big]\Big\}^{-1}\\&\times\sin(\sqrt{2\tilde{\lambda}_n}(b-x))\sin(\sqrt{2\tilde{\lambda}_n}(b-y)).\end{aligned}\tag{A4}$$

- If $x\in[\ell,b), y\in[\ell,b), \tilde{\lambda}_n>\alpha$,

$$\begin{aligned}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)=&\tfrac{\sqrt{2(\tilde{\lambda}_n-\alpha)}\cos(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))\sin(\sqrt{2\tilde{\lambda}}(\ell-a))-\sqrt{2\tilde{\lambda}_n}\cos(\sqrt{2\tilde{\lambda}_n}(\ell-a))\sin(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a)}{\sqrt{2\tilde{\lambda}_n}\sin(\sqrt{2\tilde{\lambda}_n}(b-a))}\\&\times\Big\{\cos(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[\tfrac{\sin(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))}{\sqrt{2\tilde{\lambda}_n}}+\Big(\sqrt{\tfrac{\tilde{\lambda}_n-\alpha}{\tilde{\lambda}_n}}(b-\ell)+\sqrt{\tfrac{\tilde{\lambda}_n}{\tilde{\lambda}_n-\alpha}}(\ell-a)\Big)\cos(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))\Big]\\&-\sin(\sqrt{2\tilde{\lambda}_n}(b-\ell))\Big[(b-a)\sin(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a)-\tfrac{\cos(\sqrt{2(\tilde{\lambda}_n-\alpha)}(\ell-a))}{\sqrt{2(\tilde{\lambda}_n-\alpha)}}\Big]\Big\}^{-1}\\&\times\sin(\sqrt{2\tilde{\lambda}_n}(b-x))\sin(\sqrt{2\tilde{\lambda}_n}(b-y)).\end{aligned}\tag{A5}$$

Then, by the Girsanov theorem (or via a Doob h-transform), the transition density of the drifted Brownian motion follows as

$$\tilde{p}^{\ell,-}_{(a,b),\alpha}(T;x,y)=e^{\mu(y-x)-\frac{\mu^2T}{2}}\sum_{n=1}^{\infty}2e^{-\tilde{\lambda}_nT}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y),\tag{A6}$$

where $\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)$ is given by (A2)–(A5). To implement (57), the integrals

$$\int_a^b e^{\mu(y-x)}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)dy,\ n\geq1$$

can be obtained explicitly by direct computations.

Appendix A.2. CDS Spread

Let $\{\tilde{\lambda}_n\}$ be the eigenvalues and let $\{\tilde{\phi}^{\ell,-}_{n,\alpha}\}$ be the eigenfunctions defined in (A1)–(A5). Then we have explicit formulas for $\mathfrak{f}$ and $\mathfrak{g}$, defined in Proposition 1:

$$\begin{aligned}\mathfrak{f}(t) &= \int_a^b e^{\mu(y-x)}\widetilde{G}^{\ell,-}_{\alpha}(r^f+\mu^2/2;x,y)dy \\ &\quad - \sum_{n=1}^{\infty}\frac{e^{-(\tilde{\lambda}_n+r^f+\mu^2/2)t}}{\tilde{\lambda}_n+r+\mu^2/2}\int_a^b 2e^{\mu(y-x)}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)dy, \\ \mathfrak{g}(t) &= -\int_a^b e^{\mu(y-x)}\frac{\partial}{\partial\lambda}\widetilde{G}^{\ell,-}_{\alpha}(\lambda;x,y)dy\Big|_{\lambda=r^f+\mu^2/2} \\ &\quad - \sum_{n=1}^{\infty}\frac{1+(\tilde{\lambda}_n+r^f+\mu^2/2)t}{(\tilde{\lambda}_n+r^f+\mu^2/2)^2}e^{-(\tilde{\lambda}_n+r^f+\mu^2/2)t}\int_a^b 2e^{\mu(y-x)}\tilde{\phi}^{\ell,-}_{n,\alpha}(x)\tilde{\phi}^{\ell,-}_{n,\alpha}(y)dy,\end{aligned} \tag{A7}$$

where $\widetilde{G}^{\ell,-}_{\alpha}(\lambda;x,y)$ is the Green function with explicit expressions given below (we only provide for $x \geq \ell$):

- If $y \in (a,\ell]$,

$$\widetilde{G}^{\ell,-}_{(a,b),\alpha}(\lambda;x,y)=\frac{2\sinh(\sqrt{2\lambda}(b-x))\sinh(\sqrt{2\lambda+2\alpha}(y-a))}{\sqrt{2\lambda}\cosh(\sqrt{2\lambda}(b-\ell))\sinh(\sqrt{2\lambda+2\alpha}(\ell-a))+\sqrt{2\lambda+2\alpha}\cosh(\sqrt{2\lambda+2\alpha}(\ell-a))\sinh(\sqrt{2\lambda}(b-\ell))},$$

- If $y \in [\ell,b)$,

$$\begin{aligned}\widetilde{G}^{\ell,-}_{(a,b),\alpha}(\lambda;x,y)=&\frac{2\sinh(\sqrt{2\lambda}(x\wedge y-a))\sinh(\sqrt{2\lambda}(b-x\vee y))}{\sqrt{2\lambda}\sinh(\sqrt{2\lambda}(b-a))}+\frac{2\sinh(\sqrt{2\lambda}(b-x))\sinh(\sqrt{2\lambda}(b-y))}{\sqrt{2\lambda}\sinh(\sqrt{2\lambda}(b-a))}\\ &\times\frac{\sqrt{2\lambda}\cosh(\sqrt{2\lambda}(\ell-a))\sinh(\sqrt{2\lambda+2\alpha}(\ell-a))-\sqrt{2\lambda+2\alpha}\cosh(\sqrt{2\lambda+2\alpha}(\ell-a))\sinh(\sqrt{2\lambda}(\ell-a))}{\sqrt{2\lambda}\cosh(\sqrt{2\lambda}(b-\ell))\sinh(\sqrt{2\lambda+2\alpha}(\ell-a))+\sqrt{2\lambda+2\alpha}\cosh(\sqrt{2\lambda+2\alpha}(\ell-a))\sinh(\sqrt{2\lambda}(b-\ell))},\end{aligned}$$

and the integral

$$\int_a^b e^{\mu(y-x)}\widetilde{G}^{\ell,-}_{\alpha}(\lambda;x,y)dy$$

can be obtained explicitly as well. In practice, due to the lengthy expressions of the Green function, we may obtain

$$\int_a^b e^{\mu(y-x)}\frac{\partial}{\partial\lambda}\widetilde{G}^{\ell,-}_{\alpha}(\lambda;x,y)dy = \frac{\partial}{\partial\lambda}\int_a^b e^{\mu(y-x)}\widetilde{G}^{\ell,-}_{\alpha}(\lambda;x,y)dy$$

by numerical differentiations (since it is a smooth function in λ).

Notes

1 Here we distinguish the X-diffusion from the F-diffusion (i.e., the firm's value process) $F := \mathsf{F}(X)$ which is obtained through a smooth monotonic mapping $\mathsf{F}: \mathcal{I} \to \mathcal{D}$ (where $\mathcal{D}$ is the state space for the F-diffusion) with unique inverse $\mathsf{X} = \mathsf{F}^{-1}$.

2 The cemetery state $\partial^\dagger$ is not included in the interval $\mathcal{I}$. When the process is killed and immediately sent to the cemetery state, it stays there indefinitely.

3 $\mathbb{E}_x[X;A] := \mathbb{E}_x[X\mathbb{I}_A]$ for any random variable X and event A.

4 $\mathbb{P}_x\Big(\tilde{X}_{(a,b),t} \in dy\Big) := \mathbb{P}\Big(\tilde{X}_{(a,b),t} \in dy \big| X_0 = x\Big)$.

5 When $\alpha = 0$, we obtain the regular transition density of $X_{(a,b)}$ without an instantaneous killing:

$$p_{(a,b)}(t;x,y)dy := \mathbb{P}_x(X_t \in dy, m_t > a, M_t < b); \quad t > 0,\ x,y \in (a,b),$$

and zero otherwise.

6 The transition and joint densities are pointwise convergent as $b \to r$:

$$\tilde{p}^{\ell,-}_{(a,r),\alpha}(t;x,y) := \lim_{b\to r}\tilde{p}^{\ell,-}_{(a,b),\alpha}(t;x,y), \qquad f^{(a,r)}_{\mathcal{A}^{\ell,-}_t,X_t}(u,y|\,x) := \lim_{b\to r} f^{(a,b)}_{\mathcal{A}^{\ell,-}_t,X_t}(u,y|\,x).$$

[7] We assume there exists a risk-neutral probability $\widetilde{\mathbb{P}}$ equivalent to $\mathbb{P}$ so that the discounted price process of the defaultable claim is a Doob–Levy $(\widetilde{\mathbb{P}}, \mathbb{F})$-martingale. In the GBM model, we clearly have the discounted firm's value process $e^{-r^f t}F_t = F_0 e^{-\frac{\sigma^2}{2}t+\sigma\widetilde{W}_t}, t \geq 0$, as a $(\widetilde{\mathbb{P}}, \mathbb{F})$-martingale.

[8] $\widetilde{\mathbb{E}}_{t,\mathsf{A},x}\big[h(\tau \wedge T, \tilde{X}_{\tau\wedge T})\big] := \widetilde{\mathbb{E}}\Big[h(\tau \wedge T, \tilde{X}_{\tau\wedge T})|\, \mathcal{A}_t^{\ell,-} = \mathsf{A}, X_t = x\Big]$.

[9] For $0 \leq t < T$, Theorem 1 extends to the time-t value of the credit derivative since $D^{\vartheta,T}_{t,\mathsf{A},x} = D^{\vartheta-\mathsf{A},T-t}_{0,x}$, thanks to Lemma 1.

[10] $\tau^{(\vartheta)} \wedge T = \tau_a \wedge T$ (a.s.), for $\vartheta \geq T$, where the T-maturity credit derivative price $D^{\vartheta,T}_{0,x}$ corresponds to that in the Black–Cox model with default barrier $A(t)$.

[11] The name "hazard rate model" comes from the fact that we are employing a hazard rate process to model default probabilities.

[12] For $0 \leq t < T$, Theorem 2 extends to the time-t value of the credit derivative: $D^{\alpha,T}_{t,x} = D^{\alpha,T-t}_{0,x}$.

[13] We can easily extend it to the hazard rate model with λ defined in (48), since

$$\widetilde{\mathbb{P}}_x(\tau^{(\alpha_1,\alpha_2)} > T) = e^{-\alpha_1 T}\int_a^\infty \tilde{p}^{\ell,-}_{(a,\infty),\alpha_2-\alpha_1}(T; x, y).$$

[14] Under the occupation time model, the implied hazard rate function can be computed by Laplace inverting, with respect to α, of the numerator and denominator in (60) separately.

[15] We can easily extend it to the hazard rate model with λ defined in (48), by sending $r^f \to r^f + \alpha_1$ and $\alpha \to \alpha_2 - \alpha_1$.

[16] Under the occupation time model, the implied hazard rate function can be computed by Laplace inverting, with respect to α, of $\mathfrak{f}$ and $\mathfrak{g}$ in (64).

[17] The eigenvalues $\{\tilde{\lambda}_n\}_{n\geq 1}$ can be obtained numerically by the bisection or Newton-Raphson methods.

References

Alfonsi, Aurélien, and Jérôme Lelong. 2012. A closed-form extension to the Black-Cox model. *International Journal of Theoretical and Applied Finance* 15: 1250053. [CrossRef]

Bielecki, Tomasz R., and Marek Rutkowski. 2013. *Credit Risk: Modeling, Valuation and Hedging*. Berlin/Heidelberg: Springer.

Black, Fischer, and John C. Cox. 1976. Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance* 31: 351–67. [CrossRef]

Borodin, Andrei N., and Paavo Salminen. 2002. *Handbook of Brownian Motion—Facts and Formulae*, 2nd ed. Probability and Its Applications. Basel: Birkhäuser.

Broadie, Mark, Mikhail Chernov, and Suresh Sundaresan. 2007. Optimal debt and equity values in the presence of Chapter 7 and Chapter 11. *Journal of Finance* 62: 1341–77. [CrossRef]

Campolieti, Giuseppe. Solvable Diffusions in Financial Mathematics. *manuscript in preparation, to be published.*

Campolieti, Giuseppe, Roman N. Makarov, and Karl Wouterloot. 2013. Pricing step options under the CEV and other solvable diffusion models. *International Journal of Theoretical and Applied Finance* 16: 1350027. [CrossRef]

Chesney, Marc, Monique Jeanblanc-Picqué, and Marc Yor. 1997. Brownian excursions and Parisian barrier options. *Advances in Applied Probability* 29: 165–84. [CrossRef]

Cohen, Alan M. 2007. *Numerical Methods for Laplace Transform Inversion*. New York: Springer, vol. 5.

Galai, Dan, Alon Raviv, and Zvi Wiener. 2007. Liquidation triggers and the valuation of equity and debt. *Journal of Banking and Finance* 31: 3604–20. [CrossRef]

Gaver, Donald P., Jr. 1966. Observing stochastic processes, and approximate transform inversion. *Operations Research* 14: 444–59. [CrossRef]

Haber, Richard J., Phillip J. Schönbucher, and Paul Wilmott. 1999. Pricing Parisan Options. *The Journal of Derivatives* 6: 71–79. [CrossRef]

Li, Bin, Qihe Tang, Lihe Wang, and Xiaowen Zhou. 2014. Liquidation risk in the presence of Chapters 7 and 11 of the US bankruptcy code. *Journal of Financial Engineering* 1: 1450023. [CrossRef]

Linetsky, Vadim. 2004. The spectral decomposition of the option value. *International Journal of Theoretical and Applied Finance* 7: 337–84. [CrossRef]

Makarov, Roman N., Adam Metzler, and Zi Ni. 2015. Modelling default risk with occupation times. *Finance Research Letters* 13: 54–65. [CrossRef]

Makarov, Roman N. 2016. Modeling liquidation risk with occupation times. *International Journal of Financial Engineering* 3: 1650028. [CrossRef]

Merton, Robert C. 1974. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance* 29: 449–70.

Moraux, Franck. 2002. Valuing Corporate Liabilities When the Default Threshold is Not an Absorbing Barrier. Available online: https://www.worldscientific.com/doi/10.1142/9789814759595_0014 (accessed on 12 December 2021).

Nardon, Martina. 2008. First passage and excursion time models for valuing defaultable bonds: a review with some insights. *Frontiers in Finance and Economics* 5: 1–25.

Stehfest, Harald. 1970. Algorithm 368: Numerical inversion of Laplace transforms. *Communications of the ACM* 13: 47–49. [CrossRef]

risks

MDPI

Article

Optimal Investment in a Dual Risk Model

Arash Fahim † and Lingjiong Zhu *,†

Department of Mathematics, Florida State University, Tallahassee, FL 32306, USA
* Correspondence: zhu@math.fsu.edu
† These authors contributed equally to this work.

Abstract: Dual risk models are popular for modeling a venture capital or high-tech company, for which the running cost is deterministic and the profits arrive stochastically over time. Most of the existing literature on dual risk models concentrates on the optimal dividend strategies. In this paper, we propose to study the optimal investment strategy on research and development for the dual risk models to minimize the ruin probability of the underlying company. We will also study the optimization problem when, in addition, the investment in a risky asset is allowed.

Keywords: dual risk model; minimizing ruin probability; optimal investment

1. Introduction

The classical Cramér–Lundberg model, or the classical compound Poisson risk model, assumes that the surplus process of an insurance company follows the dynamics:

$$dX_t = \rho dt - dJ_t, \qquad X_0 = x > 0, \tag{1}$$

where $\rho > 0$ is the premium rate and $J_t = \sum_{i=1}^{N_t} Y_i$ is a compound Poisson process, where N_t is a Poisson process with intensity $\lambda > 0$ and claim sizes Y_i are i.i.d. positive random variables independent of the Poisson process with $\mathbb{E}[Y_1] < \infty$. One central question in the ruin theory is to study the ruin probability $\mathbb{P}(\tau < \infty)$, where $\tau := \inf\{t > 0 : X_t < 0\}$.

In recent years, there have been a lot of studies in the insurance and finance literature on the so-called dual risk model, see, e.g., (Afonso et al. 2013; Avanzi et al. 2013; Avanzi et al. 2007; Bayraktar and Egami 2008; Cheung 2012; Cheung and Drekic 2008; Ng 2009, 2010; Rodríguez-Martínez et al. 2015; Yang and Sendova 2014),

with wealth process following the dynamics:

$$dX_t = -\rho dt + dJ_t, \qquad X_0 = x > 0, \tag{2}$$

where $\rho > 0$ is the cost of running the company and $J_t = \sum_{i=1}^{N_t} Y_i$, is the stream of profits, where N_t is a Poisson process with intensity $\lambda > 0$ and Y_i are i.i.d. $\mathbb{R}^+$ valued random variables with common probability density function $p(y)$, $y > 0$, independent of the Poisson process. The dual risk model is used to model the wealth of a venture capital, whose profits depend on the research and development. The classical risk model (1) is most often interpreted as the surplus of an insurance company. On the other hand, the dual risk model (2) can be understood as the wealth of a venture capital or high-tech company. The analogue of the premium in the classical model is the running cost in the dual model, and the claims become the future profits of the company. The ruin probability and the Laplace transform of the ruin time have been well studied for the dual risk model; see, e.g., Afonso et al. (2013). When there is a random delay for the innovations turned to profits, the dual risk model becomes time-inhomogeneous and the ruin probabilities and the distribution of the ruin times are studied in Zhu (2017).

One of the most fundamental questions in the dual risk model is the optimal dividend strategy. Avanzi et al. (2007) worked on optimal dividends in the dual risk model where

Citation: Fahim, Arash, and Lingjiong Zhu. 2023. Optimal Investment in a Dual Risk Model. *Risks* 11: 41. https://doi.org/10.3390/risks11020041

Academic Editor: Mogens Steffensen

Received: 29 December 2022
Revised: 2 February 2023
Accepted: 6 February 2023
Published: 9 February 2023

the optimal strategy is a barrier strategy. Avanzi et al. (2013) studied a dividend barrier strategy for the dual risk model, whereby dividend decisions are made only periodically, but still allow ruin to occur at any time. A dual model with a threshold dividend strategy with exponential interclaim times was studied in Ng (2009). Afonso et al. (2013) also worked on dividend problem in the dual risk model, assuming exponential interclaim times. A new approach for the calculation of expected discounted dividends was presented and ruin and dividend probabilities, number of dividends, time to a dividend, and the distribution for the amount of single dividends were studied. Dividend moments in the dual risk model were considered in Cheung and Drekic (2008). They derived integro-differential equations for the moments of the total discounted dividends which can be solved explicitly assuming the jump size distribution has a rational Laplace transform. The expected discounted dividends assuming the profits follow a Phase Type distribution were studied in Rodríguez-Martínez et al. (2015). The Laplace transform of the ruin time, expected discounted dividends for the Sparre–Andersen dual model were derived in Yang and Sendova (2014). More recently, Yang et al. (2020) obtained an explicit expression of the expected discounted discounted dividends in a dual risk model with the threshold dividend strategy and the optimal threshold level were derived. Avanzi et al. (2020) considered the optimal periodic dividend strategies for a general class of dual risk models with fixed transaction costs. In Fahim and Zhu (2022), they obtained the asymptotic analysis for optimal dividends in the dual risk model. Liu et al. (2023) studied the optimal dividend strategy for the dual model with surplus-dependent expense.

So far, the optimization problems studied in the literature on dual risk models have been almost exclusively devoted to the optimal dividend strategy. In this paper, we consider a different type of optimization problem. For a venture capital, or a high-tech company, the investment strategy on research and development (R&D) is crucial. A decision to increase the investment on research and development will increase the running cost of the company, but that will also boost the possibility of the future profits. Therefore, we believe that it is of fundamental interest to understand the optimal investment strategy to strengthen the position of the company.

It is well known that research and development is a basic engine of economic and social growth. It is a considerable amount of spending among many leading corporations in the world. A 2014 FORTUNE article listed the top ten biggest R&D spenders worldwide in the year 2013, including Volkswagen, Samsung, Intel, Microsoft, Roche, Novartis, Toyota, Johnson & Johnson, Google and Merck, with Intel spending as much as 20.1% of their revenue on R&D, see Casey and Hackett (2014). Many technology giants increase their R&D spending consistently, year over year, see, e.g., Table 1 for the R&D and percentage of the revenues of Alphabet, Amazon, Tesla in the years 2018–2021[1]. Notice that in the case of Alphabet, even though the R&D expenditure increases year by year, it increases in line with the increase of the total revenues so that as the percentage of revenues, the number does not change much. The same can be said about Amazon. For some companies, both the absolute R&D expenditure amount and the percentage as the revenues remain reasonably stable, see, e.g., Table 1 for Merck in the years 2018–2021, with the year of 2020 being the only exception which witnessed an unusually high R&D expenditure. For some companies, both the absolute R&D expenditure amount and the revenues can change dramatically, see, e.g., Table 1 for Alphabet, Amazon, Tesla in the years 2018–2021. The case of Tesla is exceptional but not unusual for a new high-tech company in the sense that the total revenues has astronomical growth and the R&D expenditure as the percentage of revenues actually declines during this period even though it had a spectacular increase in R&D expenditure in the year of 2021. Another company that has enjoyed similar phenomenal growth as Tesla is Amazon, see Table 1. However, Amazon's overall growth is not as fast as Tesla.

Since it is expensed rather than capitalized, cuts on research and development increase profit in the short term, but they can hurt the strength of a company in the long run, even if the detrimental impact of the cuts may not be felt for a few years. In the most

recent recession, firms with revenues greater than 100 million USD reduced their research and development intensity (divided by revenue) by 5.6%, even though the advertising intensity actually increased 3.4%, see Marie Knott (2012). In the long run, the research and development does help the company grow and increase the value of a company. Using a measure of the so-called research quotient, a study over all publicly traded US companies from 1981 through 2006 suggested that a 10% increase in research quotient results in an increase in market value of 1.1%, see Marie Knott (2012). Indeed, the US government also encourages the research and development activities. The Research & Experimentation Tax Credit is a general business tax credit passed by the Congress in 1981, as a response to the concerns that research spending declines had adversely affected the country's economic growth, productivity gains, and competitiveness within the global marketplace. According to a study by Ernst & Young, in the year 2005, 17,700 US corporations claimed 6.6 billion USD R&D tax credits on their tax returns[2].

Table 1. R&D spending by Alphabet, Amazon, Tesla and Merck during 2018–2021.

Alphabet	**2018**	**2019**	**2020**	**2021**
R&D (millions)	$21,419	$26,018	$27,573	$31,562
Revenues (millions)	$136,819	$161,857	$182,527	$257,637
As % of Revenues	15.7%	16.1%	15.1%	12.3%
Amazon	**2018**	**2019**	**2020**	**2021**
R&D (millions)	$28,837	$35,931	$42,740	$56,052
Revenues (millions)	$232,887	$280,522	$386,064	$469,822
As % of Revenues	12.4%	12.8%	11.1%	11.9%
Tesla	**2018**	**2019**	**2020**	**2021**
R&D (millions)	$1460	$1343	$1491	$2593
Revenues (millions)	$21,461	$24,578	$31,536	$58,823
As % of Revenues	6.8%	5.5%	4.7%	4.4%
Merck	**2018**	**2019**	**2020**	**2021**
R&D (millions)	$9,752	$9,724	$13,397	$12,245
Revenues (millions)	$42,294	$39,121	$41,518	$48,704
As % of Revenues	23.1%	24.9%	32.3%	25.1%

Optimal investment problems have a long history in finance and related fields. For example, Merton (1969, 1971) formulated and studied the problem of optimal allocation between risky assets and a risk-free asset to maximize expected utility; Fleming and Zariphopoulou (1991) considered the optimal investment and consumption problem where short-selling is not allowed but borrowing is allowed. Davis (1990), and Shreve and Soner (1994) studied optimal investment and consumption with proportional transaction costs and Morton and Pliska (1990) considered optimal portfolio management with fixed transaction costs. Grossman and Zhou (1993) studied optimal investment strategies for controlling drawdowns. Fleming and Sheu (2000) studied the optimal investment problem to maximize the long-term growth rate of expected utility of wealth. Hipp and Plum (2000) studied the optimal investment for insurers. Carr et al. (2001) considered the problem of optimal investment in a risky asset, and in derivatives written on the price process of this asset. Finally, there are also a limited number of works on the optimal venture capital investments, see, e.g., Bayraktar and Egami (2008). However, to the best of our knowledge, the optimal investment in research and development for the dual risk model has never been studied in the previous literature, and our paper is the first one that considers this problem.

We propose to study the optimal investment strategy on research and development for the dual risk models to minimize the ruin probability of the underlying company. In addition to the investment in research and development, we will also allow the investment in a risky asset, e.g., a market index. The possibility that an insurer can invest part of the

surplus into a risky asset to minimize the ruin probability was studied by Browne (1995) for the case that the insurance business is modeled by a Brownian motion with constant drift and the risky asset is modeled as a geometric Brownian motion. Later, Hipp and Plum (2000) studied the optimal investment in a market index for insurers in the classical compound Poisson risk model. We will study the optimal investment problem when both investment in research and development and investment in a risky asset are allowed. Unlike the problem of minimizing the ruin probability for an insurer in the classical risk model Hipp and Plum (2000), we will obtain closed-form formulas in the dual risk model.

Since the works of Browne (1995) and Hipp and Plum (2000), the optimal investment in the market for the classical risk model and related models have been extensively studied. In Liu and Yang (2004), they generalized the works by Hipp and Plum (2000) by including a risk-free asset. In Schmidli (2002), the optimization problem of minimizing the ruin probability for the classical risk model is studied when investment in a risky assent and proportional reinsurance are both allowed. The asymptotic ruin probability for the classical risk model under the optimal investment in a risky asset is obtained by Gaier et al. (2003) for large initial wealth. The asymptotics for small claim sizes were obtained in Hipp and Schmidli (2004). In Yang and Zhang (2005), they studied the optimal investment for an insurer when the risk process is compound Poisson process perturbed by a standard Brownian motion and the insurer can invest in the money market and in a risky asset. In Gaier and Grandits (2002), the case when the claim sizes are of regularly varying tails were studied. The results were then extended to include interest rates in Gaier and Grandits (2004). The case for subexponential claims was investigated in Schmidli (2005). In Promislow and Young (2005), they studied the problem of minimizing the probability of ruin of an insurer when the claim process is modeled by a Brownian motion with drift optimizing over the investment in a risky asset and purchasing quota-share reinsurance. In Wang et al. (2007), they adopted the martingale approach to study the optimal investment problem for an insurer when the insurer's risk process is modeled by a Lévy process with possible investment in a security market described by the standard Black–Scholes model. When the underlying investor is an individual rather than an insurance company, the optimal investment problem of minimizing the ruin probability was studied in, e.g., Bayraktar and Young (2007). In Azcue and Muler (2009), they studied the minimization of the ruin probability for the classical risk model with possible investment in a risky asset that follows a geometric Brownian motion under the borrowing constraints. There have been many other works in this area. For a survey, we refer to Paulsen (2008) and the references therein.

This paper is organized as follows. We first introduce a state-dependent dual risk model that generalizes the classical dual risk model (Section 2). When the size of a company increases, the cost usually also increases, while the resource of income will also increase in general, which makes it natural to study a state-dependent dual risk model. Then, we study the optimal investment strategy on research and development to minimize the ruin probability of the company (Section 3), with a further discussion of a state-dependent example in Section 3.1. As a special case, the state-independent model is discussed in Section 3.2, with a further discussion of a state-independent example in Section 3.3. Next, we study the joint investment in research and development and a market index to minimize the ruin probability in Section 4. Finally, we provide some numerical studies in Section 5 to better understand how the minimized ruin probability and the optimal strategy depend on the parameters in the model.

2. A State-Dependent Dual Risk Model

We introduce a state-dependent dual risk model with the wealth process being defined as follows:

$$dX_t = -\rho(X_t)dt + dJ_t, \qquad X_0 > 0, \tag{3}$$

where $J_t = \sum_{i=1}^{N_t} Y_i$, where N_t is a simple point process with intensity $\lambda(X_{t-})$ at time t, and Y_i are i.i.d. positive random variables with finite mean and independent of $\mathcal{F}_{\tau_i-}$, where

$\mathcal{F}_t$ is the natural filtration generated by X_t process, τ_i is the i-th arrival time of N_t and we further assume that $\rho(\cdot), \lambda(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ are increasing functions. The state-dependent dual risk model (3) was first introduced in Zhu (2015b), in which ruin probability and the Laplace transform of the ruin time were studied.

The motivation of introducing state dependence for the dual risk model is the following. First, the cost of a company usually increases as the size of the company increases. For example, the running cost of a small business and a Fortune 500 company are vastly different. Second, as the size of a company increases, the arrival intensity of the future profits might increase. It may be due to the fact that the larger a company gets, the more resources for income it will obtain. It is also well known in the finance literature that as a company gets larger and stronger, it can enjoy more benefits, e.g., net present value (NPV), which for example might be due to the opportunities brought by franchising. As we can see from Table 1, the R&D expenditure may be far from being constant as the size of the company and the revenue of the company change. More realistically, the R&D expenditure and other costs of running the company should be state-dependent.

Let $\tau := \inf\{t > 0 : X_t \leq 0\}$ be the ruin time of X_t process. The eventual ruin probability is defined as the function $\psi(x) := \mathbb{P}(\tau < \infty | X_0 = x)$ to emphasize the dependence on the initial wealth x. Note that for the state-independent dual risk model, $\lambda(\cdot) \equiv \lambda$ and $\rho(\cdot) \equiv \rho$, under the assumption $\lambda\mathbb{E}[Y_1] > \rho$, the ruin probability $\psi(x)$ is less than 1. Indeed, $\psi(x) = e^{-\alpha x}$, where $\alpha > 0$ is the unique solution to the equation; see, e.g., Afonso et al. (2013):

$$\rho\alpha + \lambda \int_0^\infty [e^{-\alpha y} - 1] p(y) dy = 0. \tag{4}$$

For the state-dependent dual risk model, there is no simple closed-form formula for the ruin probability. Nevertheless, for the special case when the jump sizes Y_i are i.i.d. exponentially distributed, there is a closed-form expression for the ruin probability; see Theorem 1 in Zhu (2015b).

Finally, we notice that the X_t process in (3) is an extension of the (nonlinear) marked Hawkes process with exponential kernel (see, e.g., Brémaud and Massoulié 1996; Gao and Zhu 2018a, 2018b; Hawkes 1971; Zhu 2015a),

that is, N_t is a simple point process with intensity $\lambda(X_t)$, where

$$X_t := X_0 e^{-\beta t} + \sum_{i:\tau_i < t} Y_i e^{-\beta(t-\tau_i)}, \tag{5}$$

where τ_i is the i-th arrival time of N_t, and Y_i are i.i.d. positive random variables independent of $\mathcal{F}_{\tau_i -}$ with finite mean and $X_0, \beta > 0$ are given constants, where X_t in (5) satisfies the dynamics (3) with $\rho(x) := \beta x$. When $\lambda(\cdot)$ is linear, it is called linear Hawkes process, named after Hawkes (1971). When $\lambda(\cdot)$ is nonlinear, the Hawkes process is said to be nonlinear which was first introduced by Brémaud and Massoulié (1996). Hawkes processes have wide applications in finance, neuroscience, social networks, criminology, seismology, and many other fields; see Gao and Zhu (2021) and the references therein. Since the X_t process in (3) is an extension of the (nonlinear) marked Hawkes process with exponential kernel, our paper also contributes to the literature on the Hawkes process.

3. Minimizing the Ruin Probability

In this section, we study the optimization control problem of minimizing the ruin probability for the dual risk model. The management of the underlying company can decide whether or not to increase the capital spending on research and development to boost the future profits. Our goal is to find the optimal expenditure on research and development to minimize the probability that the company is eventually ruined.

Before we proceed, we introduce the investment on research and development $C \in \mathcal{C}$, where $\mathcal{C}$ is the set of all admissible strategies, defined as

$$\mathcal{C} := \{C : [0, \infty) \times \Omega \to \mathbb{R}_{\geq 0} : C \text{ is progressively measurable, bounded and predictable}\}. \tag{6}$$

Given the control $C \in \mathcal{C}$, the wealth process has the dynamics

$$dX_t^C = -(\rho(X_t) + C_t)dt + dJ_t^C, \tag{7}$$

where $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing and $J_t = \sum_{i=1}^{N_t} Y_i$, where Y_i are defined same as before and N_t is a simple point process with intensity $F(X_{t-}, C_{t-})$ at time t, where $F(x, c) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is measurable in (x, c) and increasing in both x and c and $F(x, 0) = \lambda(x)$ for every $x \in \mathbb{R}_+$, where $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing.

We define τ^C as the ruin time of the X^C process under the control $C \in \mathcal{C}$ by $\tau^C := \inf\{t \geq 0 : X_t^C \leq 0\}$. We are interested in studying the optimization problem:

$$V(x) := \min_{C \in \mathcal{C}} \mathbb{P}\Big(\tau^C < \infty | X_0^C = x\Big). \tag{8}$$

From the optimal control point of view, it is also interesting to study the state-dependent case, which adds a technical contribution to the literature of stochastic optimal control theory. We will show that the optimal strategy is in general state-dependent when the underlying dual risk model is state-dependent, and it exhibits a closed-form expression.

Theorem 1. *The optimal strategy C^* is given by*

$$C_t^* = C^*(X_t) \in \arg\min_{C \geq 0} \frac{\rho(X_t) + C}{F(X_t, C)}, \tag{9}$$

provided that the minimum exists.

Proof of Theorem 1. For any control $C \in \mathcal{C}$, we have

$$dX_t^C = -(\rho(X_t) + C_t)dt + dJ_t^C, \tag{10}$$

where $J_t^C = \sum_{i=1}^{N_t^C} Y_i$, where N_t^C is a simple point process with intensity $F(X_{t-}, C_{t-})$ at time t and Y_i are i.i.d. with probability density function $p(y)$ defined as before.

Let us introduce a random time change and define the random time $T(t)$ via:

$$\int_0^{T(t)} F(X_{s-}, C_{s-})ds = t. \tag{11}$$

Then, it is easy to see that $T(0) = 0$ and $T(t) \to \infty$ as $t \to \infty$ since $C \in \mathcal{C}$ is bounded. It follows from (10) that

$$dX_{T(t)} = -(\rho(X_{T(t)}) + C_{T(t)})dT(t) + dJ_{T(t)}^C. \tag{12}$$

Under the random time change (11), we have

$$\frac{dT(t)}{dt} = \frac{1}{F(X_t, C_t)},$$

and $J_{T(t)}^C$ is distributed as $\overline{J}_t := \sum_{i=1}^{\overline{N}_t} Y_i$, where $\overline{N}_t$ is a standard Poisson process with intensity 1; see, e.g., Meyer (1971) for the random time change for simple point processes. Therefore, we obtain

$$dX_{T(t)} = -\frac{\rho(X_{T(t)}) + C_{T(t)}}{F(X_t, C_t)}dt + d\overline{J}_t. \tag{13}$$

Let us also notice that $\mathbb{P}(X_t \text{ ever gets ruined}) = \mathbb{P}(X_{T(t)} \text{ ever gets ruined})$. Therefore, the optimal strategy is given by (9) provided that the minimum exists. This completes the proof. □

In Theorem 1, we obtain the closed-form expression of the optimal strategy C^*. However, we do not have a closed form for the minimized ruin probability $\mathbb{P}(\tau^{C^*} < \infty | X_0^{C^*} = x)$. Next, we will show that we can obtain a closed form for the ruin probability in the special case when the jump sizes Y_i follow exponential distributions. We first recall the following result from Zhu (2015b), which states that the ruin probability for a state-dependent dual risk model with the exponentially distributed Y_i has a closed-form expression.

Theorem 2 (Theorem 1 in Zhu (2015b))**.** *Consider the dual risk model: $dX_t = -\rho(X_t)dt + dJ_t$, where $X_0 = x > 0$, $J_t = \sum_{i=1}^{N_t} Y_i$, where Y_i are exponential random variables with the probability density function $p(y) = \nu e^{-\nu y}$, $\nu > 0$, and N_t is a simple point process with intensity $\lambda(X_{t-})$ at time t, where $\rho(\cdot), \lambda(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ are increasing functions. Then,*

$$\mathbb{P}(\tau < \infty | X_0 = x) = \frac{\int_x^\infty \frac{\lambda(y)}{\rho(y)} e^{\nu y - \int_0^y \frac{\lambda(w)}{\rho(w)} dw} dy}{\int_0^\infty \frac{\lambda(y)}{\rho(y)} e^{\nu y - \int_0^y \frac{\lambda(w)}{\rho(w)} dw} dy}. \tag{14}$$

As a corollary of Theorems 1 and 2, we obtain the closed form for the minimized ruin probability when the jump sizes Y_i are i.i.d. exponentially distributed.

Proposition 1. *Assume $p(y) = \nu e^{-\nu y}$, where $\nu > 0$. Assume also the integral*

$$\int_0^\infty \frac{F(y, C^*(y))}{\rho(y) + C^*(y)} e^{\nu y - \int_0^y \frac{F(w, C^*(w))}{\rho(w) + C^*(w)} dw} dy$$

exists and is finite. Then,

$$\min_{C \in \mathcal{C}} \mathbb{P}\Big(\tau^C < \infty | X_0^C = x\Big) = \frac{\int_x^\infty \frac{F(y, C^*(y))}{\rho(y) + C^*(y)} e^{\nu y - \int_0^y \frac{F(w, C^*(w))}{\rho(w) + C^*(w)} dw} dy}{\int_0^\infty \frac{F(y, C^*(y))}{\rho(y) + C^*(y)} e^{\nu y - \int_0^y \frac{F(w, C^*(w))}{\rho(w) + C^*(w)} dw} dy}. \tag{15}$$

Proof of Proposition 1. The proposition follows immediately from Theorems 1 and 2. □

3.1. A State-Dependent Example

In this section, we study a state-dependent example in detail. We assume that

$$F(x, c) = \lambda(x) + \delta(x) c^\gamma, \tag{16}$$

where $\delta(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing, and $\gamma > 0$. We recall that $\lambda(\cdot)$ is increasing and thus, $\lambda(\cdot) \geq \lambda(0) > 0$. Let us also assume that $\rho(\cdot) \leq \rho(\infty) < \infty$. Under our assumptions, $F(x, c)$ is increasing in both x and c, and $F(x, 0) = \lambda(x)$.

Notice when $\gamma > 1$, for any constant strategy $C_t \equiv C$, where $C > 0$ is sufficiently large, the ruin probability is bounded above by the ruin probability of the following process:

$$dX_t = -(\rho(\infty) + C)dt + dJ_t, \tag{17}$$

where $J_t = \sum_{i=1}^{N_t} Y_i$ is compound Poisson with N_t being the Poisson process with intensity $\lambda(0) + \delta(0)C^\gamma$.

By the ruin probability for state-independent dual risk model (see, e.g., Afonso et al. 2013), the ruin probability of the X_t process defined in (17) is given by $e^{-\alpha_C x}$, where α_C is the unique positive solution to the equation:

$$(\rho(\infty) + C)\alpha_C + (\lambda(0) + \delta(0)C^\gamma) \int_0^\infty [e^{-\alpha_C y} - 1] p(y) dy = 0. \tag{18}$$

We can rewrite this equation as:

$$\frac{\rho(\infty)+C}{\lambda(0)+\delta(0)C^{\gamma}}\alpha_C = \int_0^{\infty}[1-e^{-\alpha_C y}]p(y)dy. \tag{19}$$

The right-hand side of the above equation is bounded between 0 and 1. In the left-hand side of the above equation, $\lim_{C\to\infty}\frac{\rho(\infty)+C}{\delta(0)C^{\gamma}} = 0$, which implies that $\alpha_C \to \infty$ as $C \to \infty$. Hence, $V(x) \leq \inf_{C>0} e^{-\alpha_C x} = 0$ and the minimized ruin probability is trivially zero.

Therefore, in the rest of this section, we only consider two cases: (i) $0 < \gamma < 1$; (ii) $\gamma = 1$.

3.1.1. The $0 < \gamma < 1$ Case

Under the assumption that $0 < \gamma < 1$, it is easy to see from Theorem 1 that the optimal strategy $C_{T(t)}$ is the strategy that minimizes the drift:

$$\frac{\rho(X_{T(t)})+C_{T(t)}}{\lambda(X_{T(t)})+\delta(X_{T(t)})C_{T(t)}^{\gamma}}. \tag{20}$$

It is easy to compute from (20) that the optimal strategy satisfies

$$\lambda(X_{T(t)})+\delta(X_{T(t)})(1-\gamma)C_{T(t)}^{\gamma} = \rho(X_{T(t)})\delta(X_{T(t)})\gamma C_{T(t)}^{\gamma-1}. \tag{21}$$

Therefore, for any $t > 0$, the optimal strategy C_t satisfies

$$\lambda(X_t)+\delta(X_t)(1-\gamma)C_t^{\gamma} = \rho(X_t)\delta(X_t)\gamma C_t^{\gamma-1}. \tag{22}$$

It is clear that the optimal strategy C_t is a function of X_t and we denote it as $C^*(X_t)$. Then, under the optimal strategy,

$$dX_t = -(\rho(X_t)+C^*(X_t))dt + dJ_t, \tag{23}$$

where $J_t = \sum_{i=1}^{N_t} Y_i$, where N_t has intensity $\lambda(X_{t-})+\delta(X_{t-})C^*(X_{t-})^{\gamma}$ at time t.

When the probability density function $p(y) = \nu e^{-\nu y}$ of jump sizes Y_i is exponential, it follows from Proposition 1 that we have the following result:

Proposition 2. *Assume $p(y) = \nu e^{-\nu y}$, where $\nu > 0$. Assume also the integral*

$$\int_0^{\infty}\frac{\lambda(y)+\delta(y)C^*(y)^{\gamma}}{\rho(y)+C^*(y)}e^{\nu y-\int_0^y\frac{\lambda(w)+\delta(w)C^*(w)^{\gamma}}{\rho(w)+C^*(w)}dw}dy$$

exists and is finite. Then,

$$V(x) = \frac{\int_x^{\infty}\frac{\lambda(y)+\delta(y)C^*(y)^{\gamma}}{\rho(y)+C^*(y)}e^{\nu y-\int_0^y\frac{\lambda(w)+\delta(w)C^*(w)^{\gamma}}{\rho(w)+C^*(w)}dw}dy}{\int_0^{\infty}\frac{\lambda(y)+\delta(y)C^*(y)^{\gamma}}{\rho(y)+C^*(y)}e^{\nu y-\int_0^y\frac{\lambda(w)+\delta(w)C^*(w)^{\gamma}}{\rho(w)+C^*(w)}dw}dy}. \tag{24}$$

Proof of Proposition 2. The proposition follows immediately from Proposition 1. □

Next, in the following example, we show that with particular model specifications, the optimal C^* and the minimized ruin probability $V(x)$ in (24) admit a simpler closed-form formula.

Example 1. *Let $\rho(x) = \rho_0$, $\lambda(x) = \lambda_0(c_1x + c_2)$, and $\delta(x) = \delta_0(c_1x + c_2)$, where $\rho_0, \lambda_0, \delta_0, c_1$, c_2 are positive constants. Then, the optimal investment rate $C^*(x)$ is a constant $C^*(x) \equiv C_0$, where C_0 is the unique positive solution to the equation:*

$$\lambda_0 + \delta_0(1-\gamma)C_0^\gamma = \rho_0\delta_0\gamma C_0^{\gamma-1}. \tag{25}$$

Hence, the minimized ruin probability in (24) *can be computed as:*

$$\begin{aligned} V(x) &= \frac{\int_x^\infty \frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}(c_1y+c_2)e^{\nu y-\int_0^y \frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}(c_1w+c_2)dw}dy}{\int_0^\infty \frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}(c_1y+c_2)e^{\nu y-\int_0^y \frac{\lambda_0+\delta(w) C_0^\gamma}{\rho_0+C_0}(c_1w+c_2)dw}dy} \\ &= \frac{\int_x^\infty (c_1y+c_2)e^{\left(\nu-\frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}c_2\right)y-\frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}\frac{c_1}{2}y^2}dy}{\int_0^\infty (c_1y+c_2)e^{\left(\nu-\frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}c_2\right)y-\frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}\frac{c_1}{2}y^2}dy} \\ &= \frac{\frac{1}{4d^{3/2}}e^{-dy^2}\left[\sqrt{\pi}e^{\frac{c^2}{4d}+dy^2}(ac+2bd)erf(\frac{2dy-c}{2\sqrt{d}})-2a\sqrt{d}e^{cy}\right]\Big|_{y=x}^{\infty}}{\frac{1}{4d^{3/2}}e^{-dy^2}\left[\sqrt{\pi}e^{\frac{c^2}{4d}+dy^2}(ac+2bd)erf(\frac{2dy-c}{2\sqrt{d}})-2a\sqrt{d}e^{cy}\right]\Big|_{y=0}^{\infty}} \\ &= \frac{2a\sqrt{d}e^{cx-dx^2}+\sqrt{\pi}e^{\frac{c^2}{4d}}(ac+2bd)erfc(\frac{2dx-c}{2\sqrt{d}})}{2a\sqrt{d}+\sqrt{\pi}e^{\frac{c^2}{4d}}(ac+2bd)erfc(\frac{-c}{2\sqrt{d}})}, \end{aligned} \tag{26}$$

where $erf(x) := \frac{2}{\sqrt{2\pi}}\int_0^x e^{-t^2}dt$ is the error function and $erfc(x) := 1 - erf(x)$ is the complementary error function and $a := c_1$, $b := c_2$, and

$$c := \nu - \frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}c_2, \qquad d := \frac{\lambda_0+\delta_0 C_0^\gamma}{\rho_0+C_0}\frac{c_1}{2}. \tag{27}$$

3.1.2. The $\gamma = 1$ Case

When $\gamma = 1$, it follows from Theorem 1 that the optimal $C^*(x)$ satisfies $C^*(x) = 0$ in the region where $\delta(x) \le \frac{\lambda(x)}{\rho(x)}$ and the "optimal" $C^*(x) = \infty$ in the region where $\delta(x) > \frac{\lambda(x)}{\rho(x)}$.

Remark 1. *If we impose a research and development budget constraint by $M \in (0, \infty)$, the maximum capacity, then the admissible set of controls is given by $\mathcal{C}_M := \{C \in \mathcal{C} : \sup_{t\ge 0} C_t \le M\}$. Then, the above analysis implies that $C^*(x) = 0$ in the region $\delta(x) \le \frac{\lambda(x)}{\rho(x)}$ and $C^*(x) = M$ in the region $\delta(x) > \frac{\lambda(x)}{\rho(x)}$.*

Next, in the following example, we show that with particular model specifications, the optimal C^* the minimized ruin probability $V(x)$ admit simpler closed-form formulas.

Example 2. *Let $\rho(x) = \rho_0(c_1x + c_2)$, $\lambda(x) = \left(\nu + \frac{\lambda_0}{1+x}\right)\rho(x)$, and $\delta(x) = \delta_0$, where ρ_0, c_1, c_2, λ_0, δ_0 are positive constants. We further assume that $\nu < \delta_0 < \nu + \lambda_0$. Then, the optimal C^* is given by:*

$$C^*(x) = \begin{cases} 0 & \text{if } x \le \frac{\lambda_0-\delta_0+\nu}{\delta_0-\nu}, \\ +\infty & \text{if } x > \frac{\lambda_0-\delta_0+\nu}{\delta_0-\nu}. \end{cases} \tag{28}$$

Let us define:

$$x^* := \frac{\lambda_0-\delta_0+\nu}{\delta_0-\nu}. \tag{29}$$

Then, we can compute that for any $y \le x^*$,

$$\int_0^y \frac{\lambda(w)+\delta(w)C^*(w)}{\rho(w)+C^*(w)}dw = \int_0^y \left(\nu + \frac{\lambda_0}{1+w}\right)dw = \nu y + \lambda_0 \log(1+y), \tag{30}$$

and for any $y > x^*$,

$$\int_0^y \frac{\lambda(w)+\delta(w)C^*(w)}{\rho(w)+C^*(w)}dw = \nu x^* + \lambda_0 \log(1+x^*) + \delta_0(y-x^*). \tag{31}$$

Therefore, for $x > x^*$, we have

$$\begin{aligned} &\int_x^\infty \frac{\lambda(y)+\delta(y)C^*(y)}{\rho(y)+C^*(y)} e^{\nu y - \int_0^y \frac{\lambda(w)+\delta(w)C^*(w)}{\rho(w)+C^*(w)}dw} dy \\ &= \int_x^\infty \delta_0 e^{\nu y - \nu x^* - \lambda_0 \log(1+x^*) - \delta_0(y-x^*)} dy = \frac{e^{-\nu x^* + \delta_0 x^*}}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu} e^{-(\delta_0-\nu)x}, \end{aligned} \tag{32}$$

and for $x \le x^*$, we have

$$\begin{aligned} &\int_x^\infty \frac{\lambda(y)+\delta(y)C^*(y)}{\rho(y)+C^*(y)} e^{\nu y - \int_0^y \frac{\lambda(w)+\delta(w)C^*(w)}{\rho(w)+C^*(w)}dw} dy \\ &= \int_x^{x^*} \left(\nu + \frac{\lambda_0}{1+y}\right) e^{\nu y - \nu y - \lambda_0 \log(1+y)} dy + \frac{1}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu} \\ &= \frac{\nu}{1-\lambda_0}\left[(1+x^*)^{-\lambda_0+1} - (1+x)^{-\lambda_0+1}\right] + (1+x)^{-\lambda_0} - (1+x^*)^{-\lambda_0} + \frac{1}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu}. \end{aligned} \tag{33}$$

Hence, we conclude that for $x > x^*$, we have

$$V(x) = \frac{\frac{e^{-\nu x^* + \delta_0 x^*}}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu} e^{-(\delta_0-\nu)x}}{\frac{\nu}{1-\lambda_0}\left[(1+x^*)^{-\lambda_0+1} - 1\right] + 1 - (1+x^*)^{-\lambda_0} + \frac{1}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu}}, \tag{34}$$

and for $x \le x^*$, we have

$$V(x) = \frac{\frac{\nu}{1-\lambda_0}\left[(1+x^*)^{-\lambda_0+1} - (1+x)^{-\lambda_0+1}\right] + (1+x)^{-\lambda_0} - (1+x^*)^{-\lambda_0} + \frac{1}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu}}{\frac{\nu}{1-\lambda_0}\left[(1+x^*)^{-\lambda_0+1} - 1\right] + 1 - (1+x^*)^{-\lambda_0} + \frac{1}{(1+x^*)^{\lambda_0}} \frac{\delta_0}{\delta_0 - \nu}}. \tag{35}$$

3.2. The State-Independent Case

In this section, we consider the state-independent case, that is,

$$\rho(\cdot) \equiv \rho, \qquad \lambda(\cdot) \equiv \lambda, \tag{36}$$

and

$$F(\cdot, c) \equiv F(c), \tag{37}$$

where $\rho, \lambda > 0$ and $F : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing. Under the assumptions (36) and (37), we have the following result, which is a corollary of Theorem 1 and the ruin probability for the state-independent dual risk model (Equation (4)).

Theorem 3. *The optimal strategy C^* is constant, given by*

$$C^* = \arg\min_{C \ge 0} \frac{\rho + C}{F(C)}, \tag{38}$$

provided that the minimum exists and the minimized ruin probability is $V(x) = e^{-\beta x}$, *where*

$$(\rho + C^*)\beta + F(C^*)\int_0^\infty [e^{-\beta y} - 1]p(y)dy = 0. \tag{39}$$

Proof of Theorem 3. Under the assumptions (36) and (37), it follows from Theorem 1 that the optimal strategy C^* is constant, which is given by $C^* = \arg\min_{C\geq 0} \frac{\rho+C}{F(C)}$. With the optimal C^*, we have

$$dX_t = -(\rho + C^*)dt + dJ_t, \tag{40}$$

where $J_t = \sum_{i=1}^{N_t} Y_i$ is compound Poisson, where N_t is Poisson with intensity $F(C^*)$.

By the formula for the ruin probability for the state-independent dual risk model, see, e.g., Equation (4), we have $V(x) = e^{-\beta x}$, where β satisfies the equation (39). This completes the proof. □

3.3. A State-Independent Example

In this section, we consider a state-independent example, that is,

$$\rho(\cdot) \equiv \rho, \qquad \lambda(\cdot) \equiv \lambda, \tag{41}$$

and

$$F(x, c) = \lambda + \delta c^\gamma, \qquad \delta, \gamma > 0. \tag{42}$$

In this special case, by Theorem 3, the optimal strategy C^* is constant and given by

$$C^* = \arg\min_{C\geq 0} \frac{\rho + C}{\lambda + \delta C^\gamma}. \tag{43}$$

By following the discussions in the more general state-dependent case in Section 3.1, the case $\gamma \geq 1$ is trivial and in the rest we only consider the cases $0 < \gamma < 1$ and $\gamma = 1$.

3.3.1. The $0 < \gamma < 1$ Case

We first consider the case that $0 < \gamma < 1$. In this case, the intensity $F(X_t, C_t) = \lambda + \delta C_t^\gamma$ is a concave and increasing function of C_t. What this indicates is that the initial investment of research and development can boost the prospect of future profits, but the margin decreases with the increase of the investment.

When it is allowed to invest in research and development, we will see later that the condition

$$(\rho - \lambda\mathbb{E}[Y_1]) - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)(\mathbb{E}[Y_1])^{\frac{1}{1-\gamma}} < 0 \tag{44}$$

is sufficient to guarantee that $V(x) < 1$. Note that this is weaker than the usual condition $\rho - \lambda\mathbb{E}[Y_1] < 0$ for the dual risk model. We have the following result.

Proposition 3. *Under the assumption* (44),

$$V(x) = \min_{C\in\mathcal{C}} \mathbb{P}\Big(\tau^C < \infty | X_0^C = x\Big) = e^{-\beta x}, \tag{45}$$

where β *is the unique positive value that satisfies the equation:*

$$\begin{aligned} &\beta\left[\rho + \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{\beta}{1 - \int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}\right] \\ &\quad - \left[\lambda + \delta\left(\frac{1}{\delta\gamma}\right)^{\frac{\gamma}{\gamma-1}}\left(\frac{\beta}{1 - \int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{\gamma}{\gamma-1}}\right]\left(1 - \int_0^\infty e^{-\beta y}p(y)dy\right) = 0, \end{aligned} \tag{46}$$

and the optimal strategy is given by

$$C^* = \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}} \left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}, \tag{47}$$

which also satisfies the following equation:

$$\lambda + (1-\gamma)\delta(C^*)^\gamma = \rho\delta\gamma(C^*)^{\gamma-1}. \tag{48}$$

Proof of Proposition 3. It follows from Theorem 3 that the optimal strategy is given by

$$C^* = \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}} \left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}, \tag{49}$$

and the minimized ruin probability $V(x)$ satisfies Equation (46).

To show that (46) has a unique positive solution, it is equivalent to show that $F(\beta) = 0$ has a unique positive solution where

$$F(\beta) := \beta\left[\rho - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma}-1\right)[g(\beta)]^{\frac{1}{1-\gamma}} - \lambda g(\beta)\right], \tag{50}$$

and

$$g(\beta) := \frac{1-\int_0^\infty e^{-\beta y}p(y)dy}{\beta}. \tag{51}$$

It is easy to compute that for $\beta > 0$,

$$g'(\beta) = \frac{1}{\beta^2}\int_0^\infty \left[\beta y e^{-\beta y} - 1 + e^{-\beta y}\right]p(y)dy. \tag{52}$$

Let $h(x) := xe^{-x} - 1 + e^{-x}$, $x \geq 0$. Then, $h(0) = 0$ and $h(x) \to -1$ as $x \to \infty$. Moreover, $h'(x) = -xe^{-x} < 0$ for $x > 0$. Thus, $h(x) \leq 0$ for any $x \geq 0$ and therefore, $g'(\beta) \leq 0$ for any $\beta > 0$ and $g(\beta)$ is a decreasing function of β.

Note that $F(\beta) = 0$ for $\beta > 0$ if and only if $G(\beta) = 0$ for $\beta > 0$, where

$$G(\beta) := \rho - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma}-1\right)[g(\beta)]^{\frac{1}{1-\gamma}} - \lambda g(\beta). \tag{53}$$

Note that by L'Hôpital's rule, $\lim_{\beta\to 0^+} g(\beta) = \mathbb{E}[Y_1]$. Therefore,

$$\lim_{\beta\to 0^+} G(\beta) = (\rho - \lambda\mathbb{E}[Y_1]) - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma}-1\right)(\mathbb{E}[Y_1])^{\frac{1}{1-\gamma}} < 0. \tag{54}$$

On the other hand, $g(\beta) \to 0$ as $\beta \to \infty$; therefore, $G(\beta) \to \rho > 0$ as $\beta \to \infty$. Since $g(\beta)$ is a decreasing function in β and $0 < \gamma < 1$, it follows that $G(\beta)$ is increasing in β. Hence, we conclude that $G(\beta) = 0$ has a unique positive solution. This completes the proof. □

In the following example, we show that when Y_i are exponentially distributed, we are able to compute out β and C^* in simple closed forms.

Example 3. *When $p(y) = \nu e^{-\nu y}$, $\nu > 0$, β satisfies*

$$\beta\left[\rho + \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}(\beta+\nu)^{\frac{1}{\gamma-1}}\right] = \left[\lambda + \delta\left(\frac{1}{\delta\gamma}\right)^{\frac{\gamma}{\gamma-1}}(\beta+\nu)^{\frac{\gamma}{\gamma-1}}\right]\frac{\beta}{\beta+\nu}, \tag{55}$$

which implies that

$$\rho(\beta+\nu)=\lambda+\left(\frac{1}{\gamma}-1\right)\left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}(\beta+\nu)^{\frac{\gamma}{\gamma-1}}. \tag{56}$$

In particular, when $\gamma=\frac{1}{2}$, *we obtain* $\rho(\beta+\nu)^2=\lambda(\beta+\nu)+\frac{\delta^2}{4}$, *which implies* $\beta=\frac{\lambda+\sqrt{\lambda^2+\rho\delta^2}}{2\rho}-\nu$, *and thus, the optimal* C^* *is given by*

$$C^*=\frac{\delta^2\rho^2}{(\lambda+\sqrt{\lambda^2+\rho\delta^2})^2}. \tag{57}$$

Remark 2. *We have already shown in Proposition 3 that* $V(x)=e^{-\beta x}$, *where* β *is the unique positive solution to Equation* (46) *and that it is equivalent to*

$$\rho-(\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma}-1\right)[g(\beta)]^{\frac{1}{1-\gamma}}-\lambda g(\beta)=0, \tag{58}$$

where $g(\beta)$ *is defined in* (51). *Now, let us discuss how the value* β *(and hence the value function* $V(x)=e^{-\beta x}$*) and the optimal investment rate* C^* *depend on the parameters* ρ, λ *and* δ. *By* (58), *we have the following observations:*

(i) As ρ *increases,* $g(\beta)$ *increases. Since* $g(\beta)$ *is decreasing in* β, *we conclude that* β *decreases as* ρ *increases. Intuitively, this means that as the fixed running cost for research and investment increases, the ruin probability increases. Asymptotically, as* $\rho\to 0$, $g(\beta)\to 0$. *When* $g(\beta)\to 0$, *since* $0<\gamma<1$, *we must have* $[g(\beta)]^{\frac{1}{1-\gamma}}\ll g(\beta)$. *Therefore, by* (58), *as* $\rho\to 0$, *we have* $g(\beta)\sim\frac{\rho}{\lambda}$. *From the definition of* $g(\beta)$, *we have* $g(\beta)\sim\frac{1}{\beta}$ *as* $\beta\to\infty$. *Hence, we conclude that* $\beta\sim\frac{\lambda}{\rho}$, *as* $\rho\to 0$. *Therefore, the optimal* C^* *satisfies*

$$C^*\sim(\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{\rho}{\lambda}\right)^{\frac{1}{1-\gamma}},\qquad \text{as } \rho\to 0. \tag{59}$$

(ii) As δ *increases,* $g(\beta)$ *decreases. Since* $g(\beta)$ *is decreasing in* β, *we conclude that* β *increases as* δ *increases. Intuitively, this indicates that if the prospect of future profits given the investment in research and development increases, then the ruin probability decreases. Asymptotically, as* $\delta\to\infty$, *we have* $g(\beta)\to 0$, *and thus,* $(\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma}-1\right)[g(\beta)]^{\frac{1}{1-\gamma}}\to\rho$, *which implies that as* $\delta\to\infty$, *we have* $g(\beta)\sim\frac{\rho^{1-\gamma}}{\gamma\delta}\left(\frac{1}{\gamma}-1\right)^{\gamma-1}$. *Since* $g(\beta)\sim\frac{1}{\beta}$ *as* $\beta\to\infty$, *we conclude that* $\beta\sim\frac{\gamma\delta}{\rho^{1-\gamma}}\left(\frac{1}{\gamma}-1\right)^{1-\gamma}$, *as* $\delta\to\infty$. *Moreover, the optimal* C^* *satisfies:*

$$C^*\to\frac{\rho}{\frac{1}{\gamma}-1},\qquad \text{as } \delta\to\infty. \tag{60}$$

Now, if $\delta\to 0$, *then* $g(\beta)\to\frac{\rho}{\lambda}$. *Therefore, as* $\delta\to 0$, $\beta\to\alpha$, *where we recall that* α *is the unique positive value so that* $1-\int_0^\infty e^{-\alpha y}p(y)dy=\alpha\frac{\rho}{\lambda}$, *which is the same as defined in* (4). *Moreover, the optimal* C^* *satisfies*

$$C^*\sim(\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{\rho}{\lambda}\right)^{\frac{1}{1-\gamma}},\qquad \text{as } \delta\to 0. \tag{61}$$

Intuitively, it says that as $\delta\to 0$, *there is no value investing in research and development.*

(iii) Similarly, as λ *increases,* β *increases, and the ruin probability decreases. As* $\lambda\to\infty$, *we have* $g(\beta)\to 0$. *Thus,* $\lambda g(\beta)\to\rho$, *and* $g(\beta)\sim\frac{\rho}{\lambda}$. *Since* $g(\beta)\sim\frac{1}{\beta}$ *as* $\beta\to\infty$, *we conclude that* $\beta\sim\frac{\lambda}{\rho}$, *as* $\lambda\to\infty$. *Moreover, the optimal* C^* *satisfies:*

$$C^*\sim(\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{\rho}{\lambda}\right)^{\frac{1}{1-\gamma}},\qquad \text{as } \lambda\to\infty. \tag{62}$$

(iv) Assume that the parameters are chosen so that

$$(\rho - \lambda\mathbb{E}[Y_1]) - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)(\mathbb{E}[Y_1])^{\frac{1}{1-\gamma}} \to 0. \tag{63}$$

Then, it follows that $g(\beta) \to \mathbb{E}[Y_1]$ *and* $\beta \to 0$. *More precisely, as* $\beta \to 0$, $g(\beta) \sim \mathbb{E}[Y_1] - \frac{\beta}{2}\mathbb{E}[Y_1^2]$ *if* $\mathbb{E}[Y_1^2] < \infty$, *and* (58) *becomes*

$$\rho - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)\left(\mathbb{E}[Y_1] - \frac{\beta}{2}\mathbb{E}\left[Y_1^2\right]\right)^{\frac{1}{1-\gamma}} - \lambda\left(\mathbb{E}[Y_1] - \frac{\beta}{2}\mathbb{E}\left[Y_1^2\right]\right) = O(\beta^2), \tag{64}$$

as $\beta \to 0$. *Then, it follows that*

$$\begin{aligned}\rho - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)\left(\mathbb{E}[Y_1]^{\frac{1}{1-\gamma}} - \frac{1}{2(1-\gamma)}(\mathbb{E}[Y_1])^{\frac{\gamma}{1-\gamma}}\mathbb{E}\left[Y_1^2\right]\beta\right) \\ - \lambda\left(\mathbb{E}[Y_1] - \frac{\beta}{2}\mathbb{E}\left[Y_1^2\right]\right) = O(\beta^2),\end{aligned} \tag{65}$$

as $\beta \to 0$. *Hence, we conclude that*

$$\beta \sim \frac{-(\rho - \lambda\mathbb{E}[Y_1]) + (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)(\mathbb{E}[Y_1])^{\frac{1}{1-\gamma}}}{(\delta\gamma)^{\frac{1}{1-\gamma}}\frac{1}{2\gamma}(\mathbb{E}[Y_1])^{\frac{\gamma}{1-\gamma}}\mathbb{E}[Y_1^2] + \frac{\lambda}{2}\mathbb{E}[Y_1^2]}. \tag{66}$$

Moreover, the optimal C^* *satisfies:*

$$C^* \sim (\delta\gamma)^{\frac{1}{1-\gamma}}(\mathbb{E}[Y_1])^{\frac{1}{1-\gamma}}. \tag{67}$$

Remark 3. *The value function* $V(x) = e^{-\beta x}$ *and the optimal investment rate* C^* *also depend on the parameter* γ. *We will study the* $\gamma = 1$ *case in details later. For the moment, let us try to understand the asymptotic behavior of the value function and the optimal investment rate as* $\gamma \to 1^-$. *We will also obtain the asymptotics as* $\gamma \to 0^+$. *Let us recall that the optimal* C^* *satisfies the equation:*

$$\lambda + (1-\gamma)\delta(C^*)^\gamma = \rho\delta\gamma(C^*)^{\gamma-1}. \tag{68}$$

Thus, we have $(1-\gamma)\delta(C^*)^\gamma \le \rho\delta\gamma(C^*)^{\gamma-1}$ *which implies that* $C^* \le \frac{\rho\gamma}{1-\gamma}$. *Thus,* $C^* \to 0$ *as* $\gamma \to 0$. *Note that* $\lim_{\gamma\to 0^+}\gamma^\gamma = 1$. *Therefore, we can check that*

$$C^* \sim \frac{\rho\delta}{\lambda+\delta}\gamma, \qquad \text{as } \gamma \to 0^+. \tag{69}$$

Now, let us consider the $\gamma \to 1^-$ *limit. Let us rewrite that Equation* (68) *as*

$$\frac{\lambda}{(1-\gamma)^{1-\gamma}} + \delta D^\gamma = \frac{\rho\delta\gamma}{D^{1-\gamma}}, \tag{70}$$

where $D = (1-\gamma)C^*$. *Let us first consider the case* $\rho\delta > \lambda$. *Notice first that* $\lim_{\gamma\to 1^-}(1-\gamma)^{1-\gamma} = 1$. *First, D cannot go to 0 as* $\gamma \to 1^-$, *because otherwise, the left-hand side of* (70) *goes to* λ *and as D goes to 0,* $D < 1$ *and* $D^{1-\gamma} \le 1$, *so the right-hand side of* (70) *is greater than* $\rho\delta\gamma$. *Then, in the limit as* $\gamma \to 1^-$, *we obtain* $\lambda \ge \rho\delta$, *which is a contradiction. Second, D cannot go to* ∞ *as* $\gamma \to 1^-$. *To see this, notice that as* $D \to \infty$, *the left hand side of* (70) *goes to* ∞ *and in the right-hand side of* (70), *for large D,* $D > 1$ *and* $D^{1-\gamma} \ge 1$ *and hence, the right-hand side is less than* $\rho\delta$, *which is a contradiction.*

Therefore, if $\rho\delta > \lambda$, D converges to a positive constant, which from (70) *we can see that the limit is $\frac{\rho\delta-\lambda}{\delta}$, and we have*

$$C^* \sim \frac{\rho\delta - \lambda}{\delta}\frac{1}{1-\gamma}, \qquad \text{as } \gamma \to 1^-. \tag{71}$$

If $\rho\delta < \lambda$, then the optimal $C^* \to 0$ as $\gamma \to 1^-$. To see this, notice that if $\limsup_{\gamma\to 1^-} C^* \in (0,\infty)$, then in (68)*, we have $\limsup_{\gamma\to 1^-} \rho\delta\gamma(C^*)^{\gamma-1} = \rho\delta$ and $\limsup_{\gamma\to 1^-}[\lambda + (1-\gamma)\delta(C^*)^\gamma] = \lambda$, which is a contradiction since $\rho\delta < \lambda$. If $\limsup_{\gamma\to 1^-} C^* = \infty$, then for $C^* > 1$, we have from* (68) *that $\lambda < \lambda + (1-\gamma)\delta(C^*)^\gamma = \rho\delta\gamma(C^*)^{\gamma-1} < \rho\delta$, which is again a contraction. Hence, we must have $C^* \to 0$.*

Since $C^* \to 0$, $(1-\gamma)\delta(C^*)^\gamma \ll \rho\delta\gamma(C^*)^{\gamma-1}$, and thus

$$C^* \sim \left(\frac{\lambda}{\rho\delta\gamma}\right)^{\frac{1}{\gamma-1}} \sim \frac{1}{e}\left(\frac{\rho\delta}{\lambda}\right)^{\frac{1}{1-\gamma}}, \qquad \text{as } \gamma \to 1^-. \tag{72}$$

If $\rho\delta = \lambda$, the optimal C^* satisfies the equation:

$$\lambda = \frac{(1-\gamma)\delta(C^*)^\gamma}{\gamma(C^*)^{\gamma-1} - 1}. \tag{73}$$

Assume that $C^* > 0$ is fixed, then by L'Hôpital's rule,

$$\lim_{\gamma\to 1^-} \frac{(1-\gamma)\delta(C^*)^\gamma}{\gamma(C^*)^{\gamma-1}-1} = \lim_{\gamma\to 1^-} \frac{-\delta(C^*)^\gamma + (1-\gamma)\delta(C^*)^\gamma \log C^*}{(C^*)^{\gamma-1} + \gamma(C^*)^{\gamma-1}\log C^*} = \frac{-\delta C^*}{1+\log C^*}. \tag{74}$$

Therefore, as $\gamma \to 1^-$, C^* converges to the unique positive solution to the equation: $\delta x + \lambda(1+\log x) = 0$.

3.3.2. The $\gamma = 1$ Case

When $\gamma = 1$, it follows from Theorem 3 that the optimal strategy C^* is constant and it is given by

$$C^* = \arg\min_{C\geq 0} \frac{\rho + C}{\lambda + \delta C}. \tag{75}$$

When $\frac{\rho}{\lambda} < \frac{1}{\delta}$, then $\inf_{C\geq 0}\frac{\rho+C}{\lambda+\delta C} = \frac{\rho}{\lambda}$ and the optimal strategy is $C_t \equiv 0$. In this case, the value function $V(x) = e^{-\beta x}$, where

$$\rho\beta + \lambda\int_0^\infty [e^{-\beta y} - 1]p(y)dy = 0. \tag{76}$$

When $\frac{\rho}{\lambda} > \frac{1}{\delta}$, then $\inf_{C\geq 0}\frac{\rho+C}{\lambda+\delta C} = \frac{1}{\delta}$. Additionally, for any $C \in \mathcal{C}$ and $\overline{C} := \|C\|_\infty$, the strategy $\overline{C}$ is more optimal than C. The "optimal strategy" is $C_t \equiv \infty$. Let us also assume that $\delta\mathbb{E}[Y_1] > 1$. In this case, the value function $V(x) = e^{-\beta x}$, where

$$\beta + \delta\int_0^\infty [e^{-\beta y} - 1]p(y)dy = 0. \tag{77}$$

When $\frac{\rho}{\lambda} = \frac{1}{\delta}$, in terms of ruin probability, it does not make a difference whether the company decides to invest in research and development or not.

Remark 4. *When $\frac{\rho}{\lambda} \geq \frac{1}{\delta}$, $V(x) = e^{-\beta x}$, where β satisfies* (77) *that is independent of ρ and λ. Asymptotically, when $\frac{\rho}{\lambda} \to 0$, it is easy to see that $\beta \sim \frac{\lambda}{\rho}$.*

Example 4. *In the special case that* $p(y) = \nu e^{-\nu y}$, *when* $\frac{\rho}{\lambda} < \frac{1}{\delta}$, *then the optimal* $C \equiv 0$ *and* $V(x) = e^{-(\frac{\lambda}{\rho} - \nu)x}$, *and when* $\frac{\rho}{\lambda} > \frac{1}{\delta}$ *and* $\frac{\delta}{\nu} > 1$, *then the optimal* $C \equiv \infty$ *and* $V(x) = e^{-(\delta - \nu)x}$.

4. Investing in a Market Index

We have already studied the optimal investment in research and development for a venture capital or high-tech company in the dual risk model in Section 3, and now, let us also add the possibility of the alternative investment in a risky asset in the market, which is a capital market index modeled by a geometric Brownian motion.

For simplicity, we restrict our discussions to the state-independent case as in Section 3.3:

$$\rho(\cdot) \equiv \rho, \qquad \lambda(\cdot) \equiv \lambda, \tag{78}$$

where $\rho, \lambda > 0$ and

$$F(x, c) = \lambda + \delta c^{\gamma}, \qquad \delta, \gamma > 0. \tag{79}$$

Let us assume that the market index S_t follows a geometric Brownian motion:

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \tag{80}$$

where $\mu, \sigma > 0$ and W_t is a standard Brownian motion.

Assume that at time t, the company can invest θ_t shares of the market index S_t and C_t in research and development. Thus, the wealth process of the company satisfies the dynamics:

$$dX_t = -(\rho + C_t)dt + dJ_t^C + \theta_t dS_t, \qquad X_0 = x > 0 \tag{81}$$

The invested amount in the market index is $A_t = \theta_t S_t$ at time t.

We are interested in finding optimal investment strategies to minimize the probability of ruin:

$$V(x) := \inf_{C \in \mathcal{C}, A \in \mathcal{A}} \mathbb{P}(\tau < \infty | X_0 = x), \tag{82}$$

where $\mathcal{C}$ is the same as defined before and $\mathcal{A}$ is the admissible strategies for investment in the market index, defined as:

$$\mathcal{A} := \Big\{ A : [0, \infty) \times \Omega \to \mathbb{R} : A \text{ is progressively measurable} \\ \text{and for any } t > 0, \mathbb{E}\left[\int_0^t A_s^2 ds\right] < \infty. \Big\}. \tag{83}$$

For any given $C \in \mathcal{C}$ and $A \in \mathcal{A}$, we write $X^{C,A} = X$ to emphasize the dependence on C and A.

With additional investment in a market index, the random time change argument in the analysis in Section 3 no longer applies. Instead, we rely on the stochastic optimal control theory (see, e.g., Fleming and Soner 1993), which suggests that the Hamilton–Jacobi–Bellman equation for $V(x)$ is given by

$$\inf_{C \geq 0, A \in \mathbb{R}} \Big\{ -(\rho + C)V'(x) + (\lambda + \delta C^{\gamma}) \int_0^{\infty} [V(x + y) - V(x)] p(y) dy \\ + A\mu V'(x) + \frac{1}{2} A^2 \sigma^2 V''(x) \Big\} = 0, \tag{84}$$

with boundary condition $V(0) = 1$.

Similar to the case in Section 3, the case $\gamma \geq 1$ leads to triviality and for the rest, we consider two cases: $0 < \gamma < 1$ and $\gamma = 1$.

4.1. The $0 < \gamma < 1$ Case

In this section, we consider the $0 < \gamma < 1$ case. We start with the following technical lemma.

Lemma 1. *$V(x) = e^{-\beta x}$ is a solution to the Hamilton–Jacobi–Bellman Equation* (84)*, where $\beta > 0$ is the unique solution to the equation:*

$$\beta\left[\rho + \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}\right] - \left[\lambda + \delta\left(\frac{1}{\delta\gamma}\right)^{\frac{\gamma}{\gamma-1}}\left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{\gamma}{\gamma-1}}\right]\left(1-\int_0^\infty e^{-\beta y}p(y)dy\right) - \frac{1}{2}\frac{\mu^2}{\sigma^2} = 0. \tag{85}$$

Given $V(x) = e^{-\beta x}$ and let

$$(C^*, A^*) \in \operatorname{argmin}\Big\{ -(\rho + C)V'(x) + (\lambda + \delta C^\gamma)\int_0^\infty [V(x+y) - V(x)]p(y)dy + A\mu V'(x) + \frac{1}{2}A^2\sigma^2 V''(x)\Big\}. \tag{86}$$

Then, we have

$$C^* = \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}, \qquad A^* = \frac{\mu}{\sigma^2\beta}. \tag{87}$$

Proof of Lemma 1. Assume that $V'(x) < 0$ and $V''(x) > 0$, then the optimal C and A are given respectively by

$$C = \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{V'(x)}{\int_0^\infty [V(x+y) - V(x)]p(y)dy}\right)^{\frac{1}{\gamma-1}}, \qquad A = -\frac{\mu V'(x)}{\sigma^2 V''(x)}, \tag{88}$$

and the Hamilton–Jacobi–Bellman equation becomes

$$-\left[\rho + \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{V'(x)}{\int_0^\infty [V(x+y) - V(x)]p(y)dy}\right)^{\frac{1}{\gamma-1}}\right]V'(x) + \left[\lambda + \delta\left(\frac{1}{\delta\gamma}\right)^{\frac{\gamma}{\gamma-1}}\left(\frac{V'(x)}{\int_0^\infty [V(x+y) - V(x)]p(y)dy}\right)^{\frac{\gamma}{\gamma-1}}\right] \cdot \int_0^\infty [V(x+y) - V(x)]p(y)dy - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{(V'(x))^2}{V''(x)} = 0. \tag{89}$$

We can see that $V(x) = e^{-\beta x}$, where $\beta > 0$ is the unique solution to the equation:

$$\beta\left[\rho + \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}\right] - \left[\lambda + \delta\left(\frac{1}{\delta\gamma}\right)^{\frac{\gamma}{\gamma-1}}\left(\frac{\beta}{1-\int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{\gamma}{\gamma-1}}\right]\left(1-\int_0^\infty e^{-\beta y}p(y)dy\right) - \frac{1}{2}\frac{\mu^2}{\sigma^2} = 0. \tag{90}$$

Recall the definition $g(\beta) = \frac{1}{\beta}\left[1 - \int_0^\infty e^{-\beta y} p(y)dy\right]$ and we want to show that the equation

$$H(\beta) := \rho - (\delta\gamma)^{\frac{1}{1-\gamma}} \left(\frac{1}{\gamma - 1}\right) [g(\beta)]^{\frac{1}{1-\gamma}} - \lambda g(\beta) - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{1}{\beta} = 0 \tag{91}$$

has a unique positive solution. It is easy to see that $\lim_{\beta\to 0^+} g(\beta) = \mathbb{E}[Y_1]$ and $\lim_{\beta\to\infty} g(\beta) = 0$. Thus, $H(\beta) \sim -\frac{1}{2}\frac{\mu^2}{\sigma^2\beta} < 0$ as $\beta \to 0^+$ and $H(\beta) \to \rho$ as $\beta \to \infty$. We have already proved that $g(\beta)$ is decreasing in β. Moreover, $\frac{1}{\beta}$ is also decreasing in β. Therefore, $H(\beta)$ is increasing in β and hence, there exists a unique positive value β so that $H(\beta) = 0$.

Finally, we can compute that the optimal C^* and A^* are given by (87). This completes the proof. □

A Verification Theorem

Let us recall from (84) that the Hamilton–Jacobi–Bellman equation is given by

$$\begin{aligned} 0 = \inf_{C>0, A\in\mathbb{R}} \Big\{ &- (\rho + C)V'(x) + (\lambda + \delta C^\gamma)\int_0^\infty [V(x+y) - V(x)]p(y)dy \\ &+ A\mu V'(x) + \frac{1}{2}A^2\sigma^2 V''(x)\Big\}, \end{aligned} \tag{92}$$

with boundary condition $V(0) = 1$.

Theorem 4 (Vertification)**.** *If $w \in \mathrm{C}_\mathrm{b}^2$ is a solution of* (92) *with $w(0) = 1$, such that for any $C \in \mathcal{C}$ and $A \in \mathcal{A}$*

$$\lim_{K\to\infty} w(K) = 0, \tag{93}$$

then, $w \le V$. In addition, if

$$C^*(x) := \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}} \left(\frac{w'(x)}{\int_0^\infty [w(x+y) - w(x)]p(y)dy}\right)^{\frac{1}{\gamma-1}} \text{ and } A^*(x) = -\frac{\mu w'(x)}{\sigma^2 w''(x)},$$

are such that

$$dX_t^* = -(\rho + C^*(X_t^*))dt + dJ_t^{C^*(X_{t-}^*)} + A^*(X_t^*)dS_t$$

has a solution and $C_\cdot^* := C^*(X_\cdot^*) \in \mathcal{C}$ and $A_\cdot^* := A^*(X_\cdot^*) \in \mathcal{A}$, then $w = V$.

Proof of Theorem 4. We follow the supermartingale argument presented in (Rogers 2013, Theorem 1.1). Since w is bounded and continuously differentiable with bounded derivative, by Itô lemma for jump processes, we have

$$\begin{aligned} \mathbb{E}\left[w\left(X_t^{C,A}\right)\middle|\mathcal{F}_s\right] = w\left(X_s^{C,A}\right) + \mathbb{E}\Big[\int_s^t \Big(&- (\rho + C_u)w'\left(X_u^{C,A}\right) \\ &+ (\lambda + \delta C_u^\gamma)\int_0^\infty \left[w\left(X_u^{C,A} + y\right) - w\left(X_u^{C,A}\right)\right]p(y)dy \\ &+ A_u\mu w'\left(X_u^{C,A}\right) + \frac{1}{2}A_u^2\sigma^2 w''\left(X_u^{C,A}\right)\Big)du\Big|\mathcal{F}_s\Big] \ge w\left(X_s^{C,A}\right), \end{aligned} \tag{94}$$

for any $C \in \mathcal{C}$ and $A \in \mathcal{A}$. Therefore, $w(X_t^{C,A})$ is a submartingale. Let τ_K be the first time that the $X_t^{C,A}$ process hits $K > 0$. Since w is uniformly bounded, by optional stopping theorem,

$$w(x) \le \mathbb{E}\left[w\left(X_{\tau_K\wedge\tau}^{C,A}\right)\right] = \mathbb{E}\left[w\left(X_{\tau_K}^{C,A}\right)1_{\{\tau_K<\tau\}} + 1_{\{\tau_K\ge\tau\}}\right] = w(K)\mathbb{P}(\tau_K < \tau) + \mathbb{P}(\tau < \tau_K).$$

It follows from (93) and monotone convergence theorem that the right-hand side above converges to $\mathbb{P}(\tau < \infty)$ as $K \to \infty$ and thus,

$$w(x) \leq \mathbb{P}(\tau < \infty).$$

By taking infimum over $C \in \mathcal{C}$ and $A \in \mathcal{A}$, we obtain $w \leq V$. All the above inequalities change to equality for $C_t = C^*(X^*_{t-})$ and $A_t = A^*(X^*_{t-})$. This completes the proof. □

Corollary 1. *$w(x) = e^{-\beta x}$ with β defined in* (85) *satisfies* (93) *and thus, $w = V$.*

Proof of Corollary 1. We already showed, in Lemma 1, that w is a classical solution of the boundary value problem (92). Moreover, since C^* and A^* defined by (86) are admissible controls (constants). By Theorem 4 and because (93) trivially holds, we have $V(x) = w(x) = e^{-\beta x}$. The proof is complete. □

Next, we provide some asymptotic analysis.

Remark 5. *As in Remark 2, let us discuss the dependence of C^*, β and hence, $V(x) = e^{-\beta x}$ on the parameters ρ, λ and δ. Since the results are similar to Remark 2, we omit the details and only summarize the results here. Note that β satisfies*

$$\rho - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)[g(\beta)]^{\frac{1}{1-\gamma}} - \lambda g(\beta) - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{1}{\beta} = 0, \tag{95}$$

where $g(\beta)$ is defined in (51).

(i) As $\rho \to 0^+$, we have $\beta \sim \frac{\lambda + \frac{1}{2}\frac{\mu^2}{\sigma^2}}{\rho}$, and $C^* \sim (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{\rho}{\lambda + \frac{1}{2}\frac{\mu^2}{\sigma^2}}\right)^{\frac{1}{1-\gamma}}$.

(ii) As $\delta \to \infty$, we have $\beta \sim \frac{\gamma}{\rho^{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)^{1-\gamma}\delta$, and $C^* \to \frac{\rho}{\frac{1}{\gamma} - 1}$. As $\delta \to 0$, we have $\beta \to \alpha$, where α is the unique positive value so that

$$\rho\alpha + \lambda\int_0^\infty [e^{-\alpha y} - 1]p(y)dy - \frac{1}{2}\frac{\mu^2}{\sigma^2} = 0. \tag{96}$$

Moreover, as $\delta \to 0$, we have $C^* \sim (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\lambda}\left(\rho - \frac{1}{2\alpha}\frac{\mu^2}{\sigma^2}\right)\right)^{\frac{1}{1-\gamma}}$.

(iii) As $\lambda \to \infty$, we have $\beta \sim \frac{\lambda}{\rho}$, and $C^* \sim (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{\rho}{\lambda}\right)^{\frac{1}{1-\gamma}}$.

Remark 6. *Here, we investigate the asymptotic behavior of the value function and the optimal investment rate as $\gamma \to 1^-$ and $\gamma \to 0^+$. Note that the optimal C^* and β satisfy*

$$\rho - \left(\frac{1}{\gamma} - 1\right)C^* - \frac{\lambda}{\delta\gamma}(C^*)^{1-\gamma} - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{1}{\beta} = 0, \tag{97}$$

and

$$C^* = \left(\frac{1}{\delta\gamma}\right)^{\frac{1}{\gamma-1}}\left(\frac{\beta}{1 - \int_0^\infty e^{-\beta y}p(y)dy}\right)^{\frac{1}{\gamma-1}}. \tag{98}$$

(i) As $\gamma \to 0^+$, $C^* \sim \eta\gamma$ for some $\eta > 0$ and $\beta \to \iota$ for some $\iota > 0$. It is easy to check that $\eta, \iota > 0$ satisfy $\eta = \frac{1 - \int_0^\infty e^{-\iota y}p(y)dy}{\iota}$ and $\rho - \eta - \frac{\lambda}{\delta}\eta - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{1}{\iota} = 0$. Thus

$$\rho - \left(1 + \frac{\lambda}{\delta}\right)\frac{1 - \int_0^\infty e^{-\iota y}p(y)dy}{\iota} - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{1}{\iota} = 0. \tag{99}$$

(ii) Next, let us consider $\gamma \to 1^-$.

If $\delta\mathbb{E}[Y_1] > 1$, *then there exists a unique value* $\iota > 0$ *such that* $\delta = \frac{\iota}{1-\int_0^\infty e^{-\iota y}p(y)dy}$. *Assume further that* $\rho - \frac{\lambda}{\delta} - \frac{1}{2}\frac{\mu^2}{\sigma^2\iota} > 0$. *Then, we have* $C^* \sim \frac{\eta}{1-\gamma}$ *and* $\beta \to \iota$ *as* $\gamma \to 1^-$, *where* $\eta = \rho - \frac{\lambda}{\delta} - \frac{1}{2}\frac{\mu^2}{\sigma^2\iota}$.

If $\rho - \frac{\lambda}{\delta} - \frac{1}{2}\frac{\mu^2}{\sigma^2\iota} < 0$, *the optimal* $C^* \to 0$ *as* $\gamma \to 1^-$ *and* $C^* \sim \left(\frac{\delta\gamma}{\lambda}\left(\rho - \frac{1}{2}\frac{\mu^2}{\sigma^2}\frac{1}{\beta}\right)\right)^{\frac{1}{1-\gamma}}$ *and* $\beta \to \iota$ *as* $\gamma \to 1^-$. *We can check that* η, ι *satisfy the equations* $\eta = \frac{\lambda}{\delta\left(\rho - \frac{1}{2}\frac{\mu^2}{\sigma^2\iota}\right)}$ *and* $\frac{\iota}{\delta} = 1 - \int_0^\infty e^{-\iota y}p(y)dy$. *As* $\gamma \to 1^-$, *we have* $C^* \sim \frac{1}{e}\left(\frac{\delta}{\lambda}\left(\rho - \frac{1}{2}\frac{\mu^2}{\sigma^2\iota}\right)\right)^{\frac{1}{1-\gamma}}$.

If $\rho - \frac{\lambda}{\delta} - \frac{1}{2}\frac{\mu^2}{\sigma^2\iota} = 0$, *then, as* $\gamma \to 1^-$, *we have that* C^* *converges to the unique positive solution to the equation* $\delta x + \lambda(1 + \log x) = 0$.

4.2. The $\gamma = 1$ *Case*

Consider the case where $\gamma = 1$, i.e., for $x > 0$. Then, we have a singular control problem on $C \in \mathcal{C}$ (see, e.g., Fleming and Soner 1993) and the value function $V(x)$ satisfies the Hamilton–Jacobi–Bellman equation:

$$0 = \min\left\{ -\rho V'(x) + \lambda\int_0^\infty [V(x+y) - V(x)]p(y)dy + \inf_{A\in\mathbb{R}}\left\{A\mu V'(x) + \frac{1}{2}A^2\sigma^2V''(x)\right\}, \right. \\ \left. \delta\int_0^\infty [V(x+y) - V(x)]p(y)dy - V'(x)\right\}, \tag{100}$$

with boundary condition $V(0) = 1$. Optimizing over A, it reduces to the following equation:

$$0 = \min\left\{ -\rho V'(x) + \lambda\int_0^\infty [V(x+y) - V(x)]p(y)dy - \frac{\mu^2(V')^2}{2\sigma^2V''}, \right. \\ \left. \delta\int_0^\infty [V(x+y) - V(x)]p(y)dy - V'(x)\right\}, \tag{101}$$

with boundary condition $V(0) = 1$.

For $w \in \mathrm{C}_\mathrm{b}^2$, we define

$$\mathcal{P} := \left\{x \in \mathbb{R}_+ \ : \ \delta\int_0^\infty [w(x+y) - w(x)]p(y)dy - w'(x) > 0\right\}.$$

According to (Fleming and Soner 1993, Chapter 8), w is a classical solution of (101) if
(i) On $\mathcal{P}$, w satisfies

$$0 = -\rho w'(x) + \lambda\int_0^\infty [w(x+y) - w(x)]p(y)dy - \frac{\mu^2(w')^2}{2\sigma^2w''}.$$

(ii) On $\mathbb{R}_+$, w satisfies

$$0 \le -\rho w'(x) + \lambda\int_0^\infty [w(x+y) - w(x)]p(y)dy - \frac{\mu^2(w')^2}{2\sigma^2w''}, \\ 0 \le \delta\int_0^\infty [w(x+y) - w(x)]p(y)dy - w'(x). \tag{102}$$

(iii) $w(0) = 1$.

Lemma 2. *$w(x) = e^{-(\beta_1 \vee \beta_2)x}$ is a classical solution of* (101) *where β_1 is the unique positive solutions of $F(\beta) = 0$ and β_2 is the unique positive solution of $G(\beta) = 0$ if it exists or zero otherwise. Here, F and G are given by*

$$F(\beta) := \rho\beta + \lambda \int_0^\infty [e^{-\beta y} - 1]p(y)dy - \frac{1}{2}\frac{\mu^2}{\sigma^2},$$
$$G(\beta) := \beta + \delta \int_0^\infty [e^{-\beta y} - 1]p(y)dy.$$

Proof of Lemma 2. If $G'(0) = 1 - \delta\mathbb{E}[Y_1] \geq 0$, then $\beta_2 = 0$ and $G(\beta_1) > 0$. This implies that $\mathcal{P} = \mathbb{R}_+$. By straightforward calculations,

$$-\rho w'(x) + \lambda \int_0^\infty [w(x+y) - w(x)]p(y)dy - \frac{\mu^2 (w')^2}{2\sigma^2 w''} = wF(\beta_1) = 0,$$
$$\delta \int_0^\infty [w(x+y) - w(x)]p(y)dy - w'(x) = wG(\beta_1) > 0.$$

If $G'(0) = 1 - \delta\mathbb{E}[Y_1] < 0$ and $\beta_1 > \beta_2$, then $G(\beta_1) > 0$ and we have $\mathcal{P} = \mathbb{R}_+$. Similar to the previous paragraph, we obtain that w is a classical solution. If $G'(0) = 1 - \delta\mathbb{E}[Y_1] < 0$ and $\beta_1 \leq \beta_2$, then $F(\beta_2) \geq 0$ and we have $\mathcal{P} = \emptyset$. Thus,

$$-\rho w'(x) + \lambda \int_0^\infty [w(x+y) - w(x)]p(y)dy - \frac{\mu^2 (w')^2}{2\sigma^2 w''} = wF(\beta_2) \geq 0,$$
$$\delta \int_0^\infty [w(x+y) - w(x)]p(y)dy - w'(x) = wG(\beta_2) = 0.$$

The proof is complete. □

A Verification Theorem

Theorem 5 (Verification)**.** *Let $w \in \mathrm{C}_\mathrm{b}^2$ be a decreasing classical solution of problem* (101) *such that condition* (93) *holds. Then, $w(x) \leq V(x)$, where $V(x)$ is the value function of the ruin probability minimization problem with investment.*
In addition, if $\mathcal{P} = \mathbb{R}_+$, then $w(x) = V(x)$.

Proof of Theorem 5. Let $A = \{A_s\}_{s\geq 0}$ be an admissible strategy and $C := \{C_t\}_{t\geq 0}$ be a non-decreasing singular function, i.e., $C_t := \int_0^t dc_s$ where c_s is a non-negative measure. Then,

$$X_t^{C,A} = x - \rho t - C_t + J_t^C + \int_0^t A_s dS_s\,,$$

where $J_t^C = \sum_{i=1}^{N_t^C} Y_i$ where N_t^C is a simple point process with compensator $\lambda t + \delta C_t$. Then, by Itô's formula for C_b^2 functions, we have

$$\begin{aligned}\mathbb{E}\Big[w\Big(X_t^{C,A}\Big)\Big|\mathcal{F}_s\Big] = w\Big(X_s^{C,A}\Big) + \mathbb{E}\Bigg[\int_s^t \Bigg(&-\rho w' + \lambda \int_0^\infty [w(x+y) - w(x)]p(y)dy \\ &+ A_u\mu w' + \frac{1}{2}A_u^2\sigma^2 w''\Bigg)\Big(X_u^{C,A}\Big)du \\ &+ \int_s^t \Bigg(- w' + \delta \int_0^\infty [w(x+y) - w(x)]p(y)dy\Bigg)\Big(X_u^{C,A}\Big)dC_u^0 \\ &+ \sum_{s\leq u\leq t}\Big(w(X_u^{C,A} - \Delta C_u) - w\Big(X_u^{C,A}\Big)\Big)\Bigg].\end{aligned}$$

Here, $C_u = C_u^0 + \Delta C_u$ where C_u^0 is the continuous part of C and ΔC_u is the pure jump part of C_u. Notice that by the definition of classical solution, (102) holds and therefore,

the first two terms inside the expectation above are non-negative. In addition, since w is non-increasing, we have $w(X_u^{C,A} - \Delta C_u) - w(X_u^{C,A}) \geq 0$. Thus, $\mathbb{E}[w(X_t^{C,A})|\mathcal{F}_s] \geq w(X_s^{C,A})$ and $w(X_t^{C,A})$ is a submartingale. Similar to the arguments in the proof of Theorem 4, (93) implies that $w(x) \leq \mathbb{P}(\tau < \infty)$. By taking the infimum over (C, A), we obtain $w \leq V$.

Now, assume that $\mathcal{P} = \mathbb{R}_+$ and set $C \equiv 0$. It follows from the definition of A^* and Itô's formula that

$$\mathbb{E}[w(X^*_{t\wedge\tau})] = w(x) + \mathbb{E}\Bigg[\int_0^{t\wedge\tau}\Big(-\rho w' + \lambda\int_0^\infty [w(x+y) - w(x)]p(y)dy + A^*\mu w' + \frac{1}{2}(A^*)^2\sigma^2 w''\Big)(X^*_s)ds\Bigg] = w(x),$$

In the above, X^* satisfies $X^*_t = x - \rho t + J^\lambda_t + \int_0^t A^*(X^*_s)dW_s$. If we let $t \to \infty$, we obtain $w(x) = \mathbb{P}(\tau^* < \infty) \geq V(x)$ where τ^* is the ruin time for process X^*. The proof is complete. □

Corollary 2. *The classical solution $w(x) = e^{-(\beta_1 \vee \beta_2)x}$ of boundary value problem* (92) *satisfies the assumption of the verification and thus, $w = V$.*

Proof of Corollary 2. First, the condition (93) trivially holds. Therefore, if $\beta_1 > \beta_2$, then $\mathcal{P} = \mathbb{R}_+$ and $w = V$ is followed by Theorem 5. It remains to show the result for the case that when $\beta_1 \leq \beta_2$, i.e., $\mathcal{P} = \emptyset$. For $c > 0$, let $w_c(x) = \mathbb{P}(\tau_c < \infty)$ with $X_t = x - (\rho + c)t + J^c_t + \int_0^t A^* dW_s$ with $A^* = \frac{\mu}{\sigma^2\beta_2}$. Then, immediately, we obtain $w_c \geq V$. We want to show that $w_c(x) \to w(x) = e^{-\beta_2 x}$ as $c \to \infty$. Notice that w_c satisfies the equation

$$0 = -(\rho + c)w_c'(x) + (\lambda + \delta c)\int_0^\infty [w_c(x+y) - w_c(x)]p(y)dy - \frac{\mu^2(w_c')^2}{2\sigma^2 w_c''},$$

with the boundary condition $w_c(0) = 1$. The unique bounded solution of the above equation is given by $w_c(x) = e^{-\beta(c)x}$ where $\beta(c)$ satisfies

$$-(\rho + c)\beta(c) + (\lambda + \delta c)\int_0^\infty [e^{-\beta(c)y} - 1]p(y)dy - \frac{\mu^2}{2\sigma^2} = 0. \tag{103}$$

Notice that for any $c > 0$, $\beta(c)$ is uniquely determined and is continuous on c. In addition, straightforward calculations show that $\beta(c)$ is increasing, i.e.,

$$\beta'(c) = \frac{1}{c}\frac{\rho + \lambda\int_0^\infty [1 - e^{-\beta(c)y}]p(y)dy + \frac{\mu^2}{2\sigma^2}}{\rho + c + (\lambda + \delta c)\int_0^\infty e^{-\beta(c)y}yp(y)dy} > 0.$$

Thus, $\bar{\beta} := \lim_{c\to\infty}\beta(c)$ exists and $\bar{\beta} > 0$ and after dividing (103) by c and taking limit when $c \to \infty$, we obtain

$$G(\bar{\beta}) = -\bar{\beta} + \lambda\int_0^\infty \Big[e^{-\bar{\beta}y} - 1\Big]p(y)dy = 0.$$

Since G has a unique positive solution, we must have $\bar{\beta} = \beta_2$ and therefore, we obtain $V(x) \leq \lim_{c\to\infty} w_c(x) = e^{-\beta_2 x}$. This completes the proof. □

5. Numerical Studies

In this section, we carry out numerical studies to illustrate and understand better how the minimized ruin probability and the optimal investment rate depend on the parameters in the dual risk model.

5.1. State-Independent Ruin Probability with Optimal Investment

In this section, we assume that the dual risk model is state-independent, and in particular, we assume that $\rho(\cdot) \equiv \rho$, $\lambda(\cdot) \equiv \lambda$, and $F(\cdot, c) \equiv \lambda + \delta c^{\gamma}$. We also assume that Y_i are i.i.d. exponentially distributed so that $p(y) = \nu e^{-\nu y}$ for some $\nu > 0$. We also assume that $\lambda \mathbb{E}[Y_1] = \frac{\lambda}{\nu} > \rho$ so that the ruin probability is less than 1 without any investment in research and development. Indeed, the ruin probability is given by $e^{-\alpha x}$, where, according to (4), α satisfies the equation:

$$\rho\alpha + \lambda \int_0^{\infty} [e^{-\alpha y} - 1]\nu e^{-\nu y} dy = \rho\alpha - \lambda \frac{\alpha}{\nu + \alpha} = 0, \tag{104}$$

which implies that $\alpha = \frac{\lambda}{\rho} - \nu$.

In Figure 1, we compare the ruin probability without any investment, the minimized ruin probability with investment in research and development, and the minimized ruin probability when investment in both research and development and a market index are allowed. For simplicity, we assume that $\gamma = \frac{1}{2}$ so that as in Example 3, the minimized ruin probability is $V(x) = e^{-\beta x}$, where $\beta = \frac{\lambda + \sqrt{\lambda^2 + \rho\delta^2}}{2\rho} - \nu$, and by investing in research and development, it reduces the ruin probability. Now, if additional investment in a risky asset, e.g., a market index is allowed, then the ruin probability can be further reduced and the minimized ruin probability becomes $V(x) = e^{-\beta x}$, where by letting $p(y) = \nu e^{-\nu y}$ and $\gamma = \frac{1}{2}$ in (85), we deduce that $\beta > 0$ is the unique solution to the equation

$$\beta\rho - \frac{\beta\delta^2}{4} \frac{1}{(\nu + \beta)^2} - \frac{\lambda\beta}{\nu + \beta} - \frac{1}{2} \frac{\mu^2}{\sigma^2} = 0. \tag{105}$$

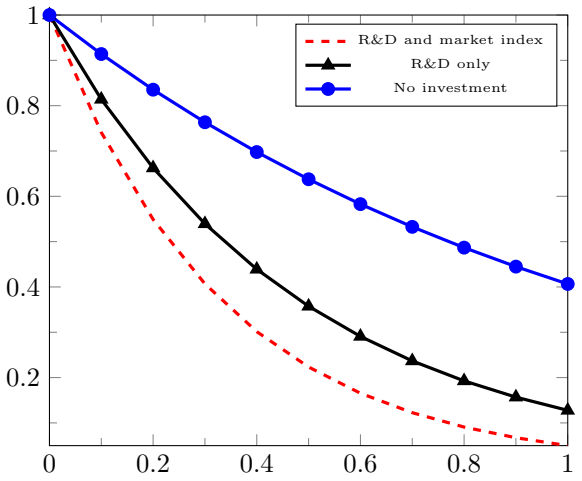


Figure 1. Illustration of the ruin probability without any investment (blue curve with circle markers), the minimized ruin probability with investment in research and development (black curve with triangle markers), and the minimized ruin probability when investment in both research and development and a market index are allowed (red dashed curve). The x-axis denotes the initial wealth of the underlying company and the y-axis denotes the (minimized) ruin probability. Here, we take $\gamma = \frac{1}{2}$, $\rho = 0.1$, $\nu = 0.1$, $\lambda = 0.1$, $\delta = 1$, $\mu = 0.1$ and $\sigma = 0.2$.

In Figure 2, we investigate the dependence of the optimal C^* on the parameters γ and δ given $\rho = 2$, $\nu = 2$ and $\lambda = 0.1$. Let us recall that when investment in research and development is allowed, the optimal investment rate C^* is the unique positive solution to the following equation:

$$\lambda + (1 - \gamma)\delta(C^*)^{\gamma} = \rho\delta\gamma(C^*)^{\gamma - 1}. \tag{106}$$

When additional investment in a market index is allowed, the optimal investment rate C^* for the investment in research and development remains the same. Notice that from (106), the optimal C^* is independent of the distribution of Y_i. Therefore, the definition of C^* is independent of the condition (44) under which the minimized ruin probability is less than 1. Intuitively, that is because C^* optimizes over the drift term by the random time change technique, but when the condition (44) is violated, even the optimal C^* still gives the ruin probability equal to 1. In Figure 2, we give the heat map plot of the optimal C^* as function of γ and δ. Note that for $p(y) = \nu e^{-\nu y}$ the condition (44) is equivalent to

$$\rho - \frac{\lambda}{\nu} - (\delta\gamma)^{\frac{1}{1-\gamma}}\left(\frac{1}{\gamma} - 1\right)\frac{1}{\nu^{\frac{1}{1-\gamma}}} < 0. \tag{107}$$

When this condition is violated, then it corresponds to the darker region in the bottom half of the plot in Figure 2. The boundary is achieved when the left-hand side of (107) is zero. In this region, the ruin probability is always 1 regardless of the investment in research and development. When the condition (107) is satisfied, it corresponds to the upper half of the plot in Figure 2. In this region, it is easy to observe that as δ increases, C^* increases. For the plot in Figure 2, the optimal C^* is less sensitive to the change of the parameter γ.

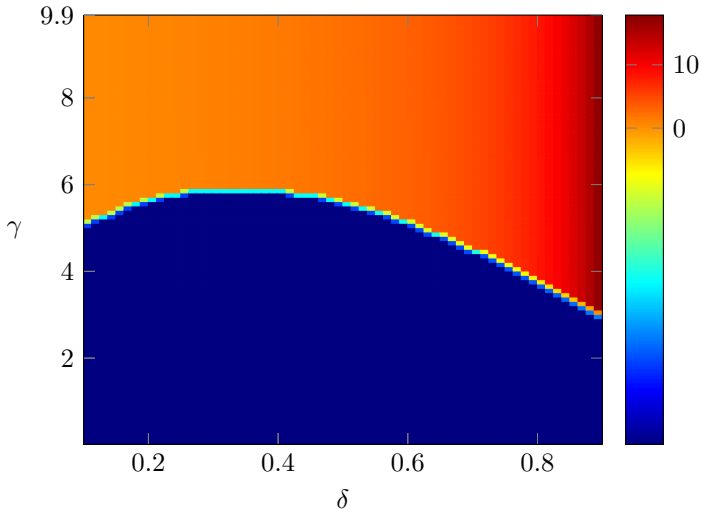


Figure 2. This shows C^* as a function of γ and δ. In the darker region in the bottom half of the plot, this is where ruin probability is always 1 regardless of the investment. In the upper half of the plot, the minimized ruin probability is less than 1 and it shows the heat map. Here, we take $\rho = 2$, $\nu = 2$ and $\lambda = 0.1$.

In Figure 3, we investigate the dependence of the optimal C^* on the parameters ρ and λ given $\delta = 1$, $\nu = 0.1$ and $\gamma = \frac{1}{2}$. For $\gamma = \frac{1}{2}$, we showed in Example 3 that the optimal C^* is given by

$$C^* = \frac{\delta^2\rho^2}{(\lambda + \sqrt{\lambda^2 + \rho\delta^2})^2}. \tag{108}$$

When $p(y) = \nu e^{-\nu y}$ and $\gamma = \frac{1}{2}$, the condition (44) reduces to $\rho - \frac{\lambda}{\nu} - \frac{\delta^2}{4\nu^2} < 0$. When this condition is violated, the ruin probability is always 1 regardless of the investment and it corresponds to the dark region in the bottom right corner of the plot in Figure 3. When this condition is satisfied, the heat map plot of the optimal C^* as a function of ρ and λ is illustrated in Figure 3. We can see that as ρ increases, the optimal C^* increases, and as λ increases, the optimal C^* decreases.

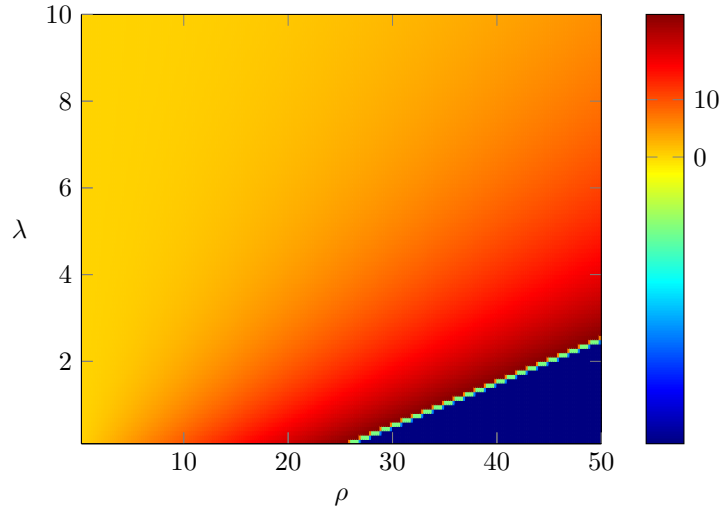


Figure 3. This shows C^* as a function of ρ and λ. In the darker region in the bottom right corner of the plot, this is where ruin probability is always 1 regardless of the investment. In the rest of the plot, the minimized ruin probability is less than 1. Here, we take $\nu = 0.1$, $\gamma = 0.5$ and $\delta = 1$.

5.2. State-Dependent Ruin Probability with Optimal Investment

In this section, we assume that the dual risk model is state-dependent, and in particular, we assume that $F(x, c) = \lambda(x) + \delta(x)c^\gamma$. We also assume that Y_i are i.i.d. exponentially distributed so that $p(y) = \nu e^{-\nu y}$ for some $\nu > 0$.

First, let us consider a special example in the case of $0 < \gamma < 1$. Let us consider the model in Example 1. For simplicity, let us assume that $\gamma = \frac{1}{2}$. Recall that in Example 1, $\rho(x) = \rho_0$, $\lambda(x) = \lambda_0(c_1 x + c_2)$, and $\delta(x) = \delta_0(c_1 x + c_2)$. The optimal investment rate $C^*(x) \equiv C_0$ is a constant and is given by:

$$C_0 = \frac{\delta_0^2 \rho_0^2}{(\lambda_0 + \sqrt{\lambda_0^2 + \rho_0 \delta_0^2})^2}. \tag{109}$$

The minimized ruin probability is given by

$$\frac{2a\sqrt{d}e^{cx-dx^2} + \sqrt{\pi}e^{\frac{c^2}{4d}}(ac + 2bd)\text{erfc}(\frac{2dx-c}{2\sqrt{d}})}{2a\sqrt{d} + \sqrt{\pi}e^{\frac{c^2}{4d}}(ac + 2bd)\text{erfc}(\frac{-c}{2\sqrt{d}})}, \tag{110}$$

where x is the initial wealth, $a := c_1$, $b := c_2$, $c := \nu - \frac{\lambda_0 + \delta_0 C_0^{1/2}}{\rho_0 + C_0} c_2$, and $d := \frac{\lambda_0 + \delta_0 C_0^{1/2}}{\rho_0 + C_0} \frac{c_1}{2}$. By setting $C_0 = 0$ in (110), we obtain the ruin probability without any investment in research and development.

In Figure 4, the blue curve with circle markers stands for the ruin probability without investment and the red dashed curve stands for the minimized ruin probability with investment. These two curves differ from exponential decays, which is due to the flexibility of the state-dependent model. As observed in Zhu (2015b), for the state-dependent dual risk model, the ruin probability can have subexponential, exponential and superexponential decays in terms of the initial wealth. Additionally, for the state-dependent dual risk model, the ruin probability may not be convex in the initial wealth (as we can see from the blue curve with circle markers in Figure 4).

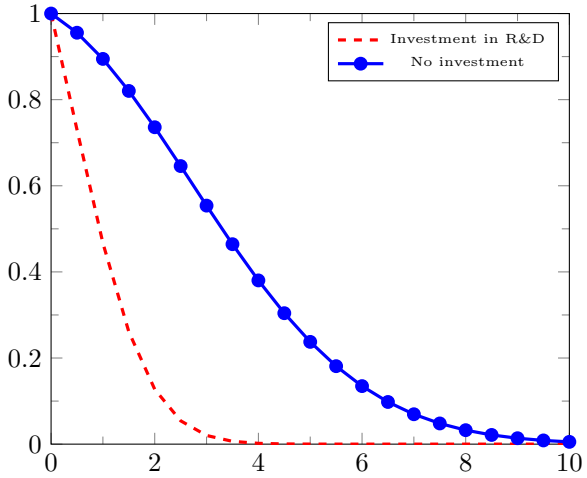


Figure 4. Illustration of the ruin probability without any investment (blue curve with circle markers), the minimized ruin probability with investment in research and development (red dashed curve). The x-axis denotes the initial wealth of the underlying company and the y-axis denotes the (minimized) ruin probability. Here, we take $\gamma = 0.5$, $\rho_0 = 1$, $\nu = 0.1$, $\lambda_0 = 0.1$, $\delta = 1$, $c_1 = 1$ and $c_2 = 1$.

Next, let us consider an example for $\gamma = 1$ for the state-dependent dual risk model. Let us recall that in Example 2, $\rho(x) = \rho_0(c_1 x + c_2)$, $\lambda(x) = \left(\nu + \frac{\lambda_0}{1+x}\right)\rho(x)$, and $\delta(x) = \delta_0$, and under the assumption that $\nu < \delta_0 < \nu + \lambda_0$, the optimal C^* is given by $C^* = 0$ if $x \leq x^*$ and $C^* = \infty$ if $x > x^*$, where $x^* := \frac{\lambda_0 - \delta_0 + \nu}{\delta_0 - \nu}$. From Example 2, with optimal investment, the minimized ruin probability is given by $V(x)$ in (34) if $x > x^*$ and the minimized ruin probability is given by $V(x)$ in (35) if $x \leq x^*$, where x is the initial wealth. Without any investment, as in Theorem 2, under the assumption that $\lambda_0 > 1$, we can compute that the ruin probability is given by

$$V(x) = \frac{\int_x^\infty \left(\nu + \frac{\lambda_0}{1+y}\right) \frac{1}{(1+y)^{\lambda_0}} dy}{\int_0^\infty \left(\nu + \frac{\lambda_0}{1+y}\right) \frac{1}{(1+y)^{\lambda_0}} dy} = \frac{\nu(1+x)^{-\lambda_0+1} + (\lambda_0 - 1)(1+x)^{-\lambda_0}}{\lambda_0 + \nu - 1}, \tag{111}$$

which is strictly between 0 and 1.

In Figure 5, we plot the curve of the ruin probability as a function of the initial wealth without investment (blue curve with circle markers) and the minimized ruin probability as a function of the initial wealth with the optimal investment in research and development (red dashed curve) as in the example of the state-dependent dual risk model we described above. In Figure 5, the critical threshold for the optimal investment strategy is $x^* = 3$ in the plot. When the wealth process is below this threshold x^*, the optimal strategy for investment in R&D is not to invest, and when the wealth process is above this threshold x^*, the optimal strategy for investment in R&D is to invest as aggressively as possible. When $x < x^*$, from (35), we can see that $V(x)$ decays polynomially in x, and when $x > x^*$, from (34), we can see that $V(x)$ decays exponentially in x.

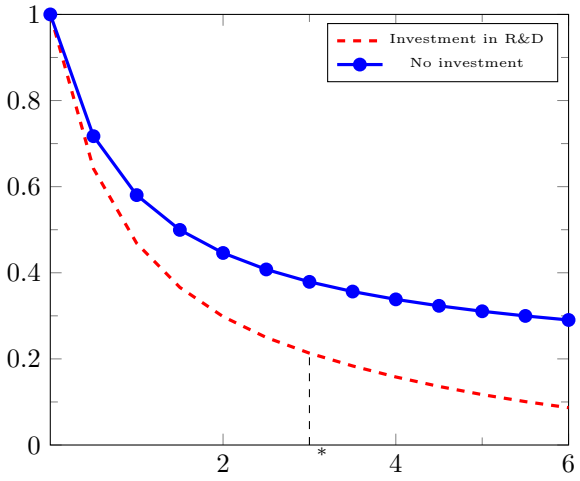


Figure 5. Illustration of the ruin probability without any investment (blue curve with circle markers), the minimized ruin probability with investment in research and development (red dashed curve). The x-axis denotes the initial wealth of the underlying company and the y-axis denotes the (minimized) ruin probability. x^* on the x-axis is the critical threshold above which the optimal strategy is to invest as much as possible in R&D, and below which the optimal strategy is not to invest at all in R&D. Here, we take $\rho_0 = 1$ (irrelevant), $\nu = 0.1$, $\lambda_0 = 1.2$, $\delta_0 = 0.4$ and $c_1 = c_2 = 1$ (irrelevant) and $\gamma = 1$.

Author Contributions: Conceptualization, A.F. and L.Z.; methodology, A.F. and L.Z.; formal analysis, A.F. and L.Z.; investigation, A.F. and L.Z.; resources, A.F. and L.Z.; writing—original draft preparation, A.F. and L.Z.; writing—review and editing, A.F. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Arash Fahim gratefully acknowledges support from the National Science Foundation via the award NSF-DMS-1447067. Lingjiong Zhu is grateful to the support from the National Science Foundation via the awards NSF-DMS-1613164, NSF-DMS-2053454, NSF-DMS-2208303.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to two anonymous referees for their helpful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 Available at https://www.macrotrends.net/ (accessed on 8 February 2023).

2 See Supporting innovation and economic growth: The broad impact of the R&D credit in 2005. Prepared by Ernst & Young LLP for the R&D Coalition. April 2008. Available at https://www.scribd.com/document/207312025/e-y-RatiosR-DTaxCreditStudy2008final (accessed on 28 December 2022).

References

Afonso, Lourdes B., Rui M. R. Cardoso, and Alfredo D. Egídio dos Reis. 2013. Dividend problems in the dual risk model. *Insurance: Mathematics and Economics* 53: 906–18. [CrossRef]

Avanzi, Benjamin, Eric C. K. Cheung, Bernard Wong, and Jae-Kyung Woo. 2013. On a periodic dividend barrier strategy in the dual model with continuous monitoring of solvency. *Insurance: Mathematics and Economics* 52: 98–113. [CrossRef]

Avanzi, Benjamin, Hans U. Gerber, and Elias S. W. Shiu. 2007. Optimal dividends in the dual model. *Insurance: Mathematics and Economics* 41: 111–23. [CrossRef]

Avanzi, Benjamin, Hayden Lau, and Bernard Wong. 2020. Optimal periodic dividend strategies for spectrally positive Lévy risk processes with fixed transaction costs. *Insurance: Mathematics and Economics* 93: 315–32.

Azcue, Pablo, and Nora Muler. 2009. Optimal investment strategy to minimize the ruin probability of an insurance company under borrowing constraints. *Insurance: Mathematics and Economics* 44: 26–34. [CrossRef]
Bayraktar, Erhan, and Masahiko Egami. 2008. Optimizing venture capital investments in a jump diffusion model. *Mathematical Methods of Operations Research* 67: 21–42. [CrossRef]
Bayraktar, Erhan, and Virginia R. Young. 2007. Minimizing the probability of lifetime ruin under borrowing constraints. *Insurance: Mathematics and Economics* 41: 196–221. [CrossRef]
Brémaud, Pierre, and Laurent Massoulié. 1996. Stability of nonlinear Hawkes processes. *The Annals of Probability* 24: 1563–88. [CrossRef]
Browne, Sid. 1995. Optimal investment policies for a firm with a random risk process: exponential utility and minimizing the probability of ruin. *Mathematics of Operations Research* 20: 937–58. [CrossRef]
Carr, Peter, Xing Jin, and Dilip B. Madan. 2001. Optimal investment in derivative securities. *Finance and Stochastics* 5: 33–59.
Casey, Michael, and Robert Hackett. 2014. The 10 Biggest R&D Spenders Worldwide. Retrieved from Fortune. Available online: http://fortune.com/2014/11/17/top-10-research-development (accessed on 8 February 2023).
Cheung, Eric C. K. 2012. A unifying approach to the analysis of business with random gains. *Scandinavian Actuarial Journal* 2012: 153–82. . [CrossRef]
Cheung, Eric C. K., and Steve Drekic. 2008. Dividend moments in the dual risk model: Exact and approximate approaches. *Astin Bulletin: The Journal of the IAA* 38: 399–422. [CrossRef]
Davis, Mark H. A. 1990. Portfolio selection with transaction costs. *Mathematics of Operations Research* 15: 676–713.
Fahim, Arash, and Lingjiong Zhu. 2022. Asymptotic analysis for optimal dividends in a dual risk model. *Stochastic Models* 38: 605–37. [CrossRef]
Fleming, Wendell H., and H. Mete Soner. 1993. *Controlled Markov Processes and Viscosity Solutions*. New York: Springer.
Fleming, Wendell H., and Shuenn-Jyi Sheu . 2000. Risk-sensitive control and an optimal investment model. *Mathematical Finance* 10: 197–213. [CrossRef]
Fleming, Wendell H., and Thalia Zariphopoulou. 1991. An optimal investment/consumption model with borrowing. *Mathematics of Operations Research* 16: 671–891. [CrossRef]
Gaier, Johanna, and Peter Grandits. 2002. Ruin probabilities in the presence of regularly varying tails and optimal investment. *Insurance: Mathematics and Economics* 30: 211–17. [CrossRef]
Gaier, Johanna, and Peter Grandits. 2004. Ruin probabilities and investment under interest force in the presence of regularly varying tails. *Scandinavian Actuarial Journal* 2004: 256–78. [CrossRef]
Gaier, Johanna, Peter Grandits, and Walter Schachermayer. 2003. Asymptotic ruin probabilities and optimal investment. *The Annals of Applied Probability* 13: 1054–76. [CrossRef]
Gao, Fuqing, and Lingjiong Zhu. 2021. Precise deviations for Hawkes processes. *Bernoulli* 27: 221–48. [CrossRef]
Gao, Xuefeng, and Lingjiong Zhu. 2018a. Large deviations and applications for Markovian Hawkes processes with a large initial intensity. *Bernoulli* 24: 2875–905. [CrossRef]
Gao, Xuefeng, and Lingjiong Zhu. 2018b. Limit theorems for Markovian Hawkes processes with a large initial intensity. *Stochastic Processes and Their Applications* 128: 3807–39. [CrossRef]
Grossman, Sanford J., and Zhongquan Zhou. 1993. Optimal investment strategies for controlling drawdowns. *Mathematical Finance* 3: 241–76. [CrossRef]
Hawkes, Alan G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika 58*: 83–90. . 1093/biomet/58.1.83. [CrossRef]
Hipp, Christian, and Hanspeter Schmidli. 2004. Asymptotics of ruin probabilities for controlled risk processes in the small claims case. *Scandinavian Actuarial Journal* 2004: 321–35. [CrossRef]
Hipp, Christian, and Michael Plum. 2000. Optimal investment for insurers. *Insurance: Mathematics and Economics* 27: 215–28. [CrossRef]
Liu, Chi Sang, and Hailiang Yang. 2004. Optimal investment for an insurer to minimize its probability of ruin. *North American Actuarial Journal* 8: 11–31. [CrossRef]
Liu, Shanshan, Zhaoyang Liu, and Guoxin Liu. 2023. Optimal dividend strategy for the dual model with surplus-dependent expense. *Communications in Statistics-Theory and Methods* 52: 543–66. [CrossRef]
Marie Knott, Anne. 2012. The trillion-dollar R&D fix. *Harvard Business Review* 76.
Merton, Robert C. 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics* 51: 247–57.
Merton, Robert C. 1971. Optimal consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* 3: 373–413.
Meyer, Paul-André. 1971. Démonstration simplifiée d'un théorème de Knight. In *Séminaire de Probabilités, V (Univ. Strasbourg, Année Universitaire 1969–1970)*. Berlin: Springer, vol. 191, pp. 191–95.
Morton, Andrew J., and Stanley R. Pliska. 1990. Optimal portfolio management with fixed transaction costs. *Mathematical Finance* 5: 337–56. [CrossRef]
Ng, Andrew C. Y. 2009. On a dual model with a dividend threshold. *Insurance: Mathematics and Economics* 44: 315–24. [CrossRef]
Ng, Andrew C. Y. 2010. On the upcrossing and downcrossing probabilities of a dual risk model with phase-type gains. *Astin Bulletin* 40: 281–306. [CrossRef]
Paulsen, Jostein. 2008. Ruin models with investment income. *Probability Surveys* 5: 416–34. [CrossRef]

Promislow, S. David, and Virginia R. Young. 2005. Minimizing the probability of ruin when claims follow Brownian motion with drift. *North American Actuarial Journal* 9: 109–28. [CrossRef]

Rodríguez-Martínez, Eugenio V., Rui M. R. Cardoso, and Alfredo D. Egídio dos Reis. 2015. Some advances on the Erlang(n) dual risk model. *Astin Bulletin* 45: 127–50. [CrossRef]

Rogers, Leonard C. G. 2013. *Optimal Investment*. Springer Briefs in Quantitative Finance. Heidelberg: Springer. . 978-3-642-35202-7. [CrossRef]

Schmidli, Hanspeter. 2002. On minimizing the ruin probability by investment and reinsurance. *The Annals of Applied Probability* 12: 890–907. [CrossRef]

Schmidli, Hanspeter. 2005. On optimal investment and subexponential claims. *Insurance: Mathematics and Economics* 36: 25–35. [CrossRef]

Shreve, Steven E., and H. Mete Soner. 1994. Optimal investment and consumption with transaction costs. *Annals of Applied Probability* 4: 609–92.

Wang, Zengwu, Jianming Xia, and Lihong Zhang. 2007. Optimal investment for an insurer: The martingale approach. *Insurance: Mathematics and Economics* 40: 322–34. [CrossRef]

Yang, Chen, and Kristina P. Sendova. 2014. The ruin time under the Sparre-Andersen dual model. *Insurance: Mathematics and Economics* 54: 28–40. [CrossRef]

Yang, Chen, Kristina P. Sendova, and Zhong Li. 2020. Parisian ruin with a threshold dividend strategy under the dual Lévy risk model. *Insurance: Mathematics and Economics* 90: 135–50.

Yang, Hailiang, and Lihong Zhang. 2005. Optimal investment for insurer with jump-diffusion risk process. *Insurance: Mathematics and Economics* 37: 615–34. [CrossRef]

Zhu, Lingjiong. 2015a. Large deviations for Markovian nonlinear Hawkes processes. *Annals of Applied Probability* 25: 548–81. [CrossRef]

Zhu, Lingjiong. 2015b. A state-dependent dual risk model. *arXiv*, arXiv:1510.03920.

Zhu, Lingjiong. 2017. A delayed dual risk model. *Stochastic Models* 33: 149–70. [CrossRef]

risks

MDPI

Article

Backward Deep BSDE Methods and Applications to Nonlinear Problems

Yajie Yu, Narayan Ganesan and Bernhard Hientzsch *

Corporate Model Risk, Wells Fargo, New York, NY 10017, USA; jessica.yu@wellsfargo.com (Y.Y.); narayan.ganesan.8@gmail.com (N.G.)
* Correspondence: bernhard.hientzsch@wellsfargo.com

Abstract: We present a pathwise deep Backward Stochastic Differential Equation (BSDE) method for Forward Backward Stochastic Differential Equations with terminal conditions that time-steps the BSDE backwards and apply it to the differential rates problem as a prototypical nonlinear problem of independent financial interest. The nonlinear equation for the backward time-step is solved exactly or by a Taylor-based approximation. This is the first application of such a pathwise backward time-stepping deep BSDE approach for problems with nonlinear generators. We extend the method to the case when the initial value of the forward components X can be a parameter rather than fixed and similarly to also learn values at intermediate times. We present numerical results for a call combination and for a straddle, the latter comparing well to those obtained by Forsyth and Labahn with a specialized PDE solver.

Keywords: differential rates; FBSDEs; nonlinear pricing; deep learning for pricing

Citation: Yu, Yajie, Narayan Ganesan, and Bernhard Hientzsch. 2023. Backward Deep BSDE Methods and Applications to Nonlinear Problems. *Risks* 11: 61. https://doi.org/10.3390/risks11030061

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 2 November 2022
Revised: 24 February 2023
Accepted: 10 March 2023
Published: 16 March 2023

1. Introduction

As proposed in Han and Jentzen (2017), deep learning (DL) and deep neural networks (DNN) can be used to solve high-dimensional nonlinear PDEs by converting them to Forward Backward Stochastic Differential Equations (FBSDE) and building neural networks to learn the control and initial value of the corresponding stochastic control problem. They use their method on, for instance, a differential rates problem, as studied in Mercurio (2015), for a combination of two call options. Hientzsch (2021) also gives an overview of pricing different instruments in quantitative finance via deep BSDE and FBSDE.

To summarize, deep BSDE methods rewrite the FBSDE problem into a stochastic control problem where one searches for a control (ultimately approximating the gradient of the solution) and an initial value or initial value function which minimize certain loss functions. The minimization typically occurs through stochastic gradient descent approaches and variants (such as Adam), where the stochastic gradient with respect to the parameters of the control and of the initial value function is obtained from a mini-batch of realizations of the underlying dynamics (X) and corresponding realizations of the solution of the BSDE (Y) given the current control Π. The forward deep BSDE methods (first described in Han et al. (2018)) and the backward deep BSDE methods described here differ by how the realization of Y is computed given the current control and the realization of X and also by the loss function to be minimized. These pathwise deep BSDE methods have the advantage that one only needs to implement a discretization of the forward dynamics X and a pathwise computation of the backward dynamics for Y and the loss function generically in a deep learning framework such as TensorFlow and can use standard deep learning techniques. They also can be implemented for very high-dimensional problems avoiding or at least mitigating the curse of dimensionality.

Han et al. (2018) propose time-stepping both forward and backward SDE forward in time and transform the final value problem to a stochastic control problem in which the objective function measures how well the given final value has been approximated. We

call their method the "forward deep BSDE" method since it time-steps the BSDE forward. Wang et al. (2018) consider a BSDE with zero drift term which can be trivially time-stepped backwards and propose and demonstrate forward and backward methods with fixed X_0, describing the first pathwise backward deep BSDE method. Liang et al. (2021) solve BSDEs with linear generators with both forward and backward methods. They indicate a Taylor expansion approach for nonlinear generators, without giving details or results for nonlinear problems. We describe the general approach and the application to the differential rates setting for two variants of this pathwise backward deep BSDE method, which is, to the best of our knowledge, the first application of this pathwise backward method to nonlinear problems.

The backward method starts with the given final value at maturity and then time-steps the BSDE backwards until a given initial time t_0, which is assumed to be 0 without loss of generality in this paper. In continuous time and complete markets, a trading strategy can completely eliminate randomness; thus, all realizations of Y_{t_0} (initial value of derivative) for the same initial risk factors X_{t_0} should have the same value. Thus, if we minimize a measure of the range of Y_{t_0}, we should obtain a minimum of 0 and a risk-eliminating trading strategy. In the time-discrete and/or incomplete setting, the randomness can no longer be eliminated, but its impact can be minimized. In the pathwise backward methods, the variance of Y_{t_0} serves as that measure—either with respect to a Monte Carlo mean or a to-be-learned parameter. Similarly, for random X_{t_0}, we minimize the square distance from an also to-be-determined function $\mathsf{Yinit}(\mathsf{X}_{\mathsf{t}_0})$ represented by a DNN. This extension to random X_{t_0} is new.

We consider the differential rates problem together with Black–Scholes dynamics for European options. To force nonlinear behavior, one can consider, for example, a linear combination of calls with coefficients with opposite signs or a straddle. Differential rates mean that positive cash balances in the trading strategy accrue interest at a lower lending rate, while negative cash balances (debts/loans) accrue interest at a higher borrowing rate.

For the differential rates problem, (Han and Jentzen 2017, Section 4.4) mention a nonlinear PDE which can be solved by appropriate nonlinear PDE solvers in small dimensions (see, for instance, Forsyth and Labahn (2007)). For a more general setting, Mercurio (2015) presents PDEs and proposes PDE solution or binomial tree methods. None of these methods work in higher dimensions due to the curse of dimensionality. All these methods require problem-specific implementations of nonlinear PDE or tree solvers.

There are other approaches to such nonlinear problems, including in high dimensions, that also rely on the equivalent FBSDE formulation. As mentioned above, Han and Jentzen (2017) solve a nonlinear differential rates problem with their pathwise forward deep BSDE method. Huré et al. (2020) solves nonlinear problems by a BSDE rollback method, where the solution and its gradient at each time-step are sequentially learned by minimizing the residual of the time-discretized BSDE. Warin (2018) solve nonlinear problems by repeatedly nested Monte Carlo. While somewhat straightforward to implement, this only works for shorter maturities and requires substantial computational resources. Raissi (2018) solves the BSDE by adding loss terms for the residuals of all time-steps to the usual final loss term. Training for such complex loss functions can be quite challenging and unreliable and might not lead to solutions that satisfy all constraints and loss functions equally well, and standard stochastic optimization methods often struggle to optimize well. None of these works except Han and Jentzen (2017) solve the differential rates problem.

In this paper, we first introduce FBSDE for general nonlinear problems, with particular details for the differential rates problem, time-discretize them, and then derive exact solutions and Taylor approximations for the backward step problem. We then quickly describe the forward and backward deep BSDE approaches that we consider—both the batch-variance variant already described in the literature but also the novel initial variable and network versions, the last one for varying or random X_{t_0}, together with a computational graph for the network version. Then, we apply these methods to the differential rates problem for the call combination case from (Han and Jentzen 2017, Section 4.4) and for the

straddle case from Forsyth and Labahn (2007). We compare the results for a case with fixed X_{t_0} and for a case with varying X_{t_0} with the results from Forsyth and Labahn (2007) and see that they agree well. Finally, we conclude.

2. FBSDE for Nonlinear Problems

We are interested in solving a nonlinear PDE for a single function u that depends on time t and an n-dimensional state vector x of the following general form:

$$u_t(t,x) + \mathcal{L}_t u(t,x) + f(t,x,u(t,x),\nabla_x u(t,x)) = 0, \tag{1}$$

with

$$\mathcal{L}_t u(t,x) := \frac{1}{2}\mathrm{Tr}\Big((\sigma\sigma^T)(t,x)(\mathrm{Hess}_x u)(t,x)\Big) + \mu(t,x)\nabla_x u(t,x), \tag{2}$$

where $\nabla_x u$ is the gradient vector with respect to the x and $\mathrm{Hess}_x u$ is the Hessian matrix with respect to the x, with $\mu(t,x) \in \mathcal{R}^n$ and $\sigma(t,x) \in \mathcal{R}^{n,n}$ being appropriate vector and matrix functions of their arguments, together with terminal condition at maturity T given as

$$u(T,x) = g(x). \tag{3}$$

A nonlinear Feyman–Kac theorem[1] shows that the solution of the above PDE also satisfies the following FBSDE system under appropriate assumptions:

The forward SDE (FSDE) for the vector $X_t \in \mathcal{R}^n$:

$$dX_t = \mu(t,X_t)dt + \sigma(t,X_t)dW_t, \tag{4}$$

and the backward SDE (BSDE) in terms of strategy $\Pi_t \in \mathcal{R}^n$:

$$-dY_t = f(t,X_t,Y_t,\Pi_t)dt\ - \Pi_t^T\sigma(t,X_t)dW_t, \tag{5}$$

with terminal condition[2]

$$Y_T = g(X_T), \tag{6}$$

where

$$Y_t = u(t,X_t), \Pi_t = \nabla_X u(t,X_t). \tag{7}$$

In terms of pricing applications in finance, X_t is the vector of values of the underlying assets, and $g(X_T)$ is the final payoff of the European option that one tries to replicate with a self-financing portfolio in the underlying asset(s) and a remaining cash position. That portfolio will contain $\pi_j(t)$ worth of the jth underlying asset (corresponding to index j in the vector X_t), and Π_t is the vector of the $\pi_j(t)$.The portfolio (including cash position) is worth Y_t at time t.

Now, Y_t or equivalently $u(t,X_t)$ represent the needed wealth at t to exactly or approximately replicate the payoff when starting at X_t at time t. One can thus define price (by replication) $\mathsf{price}(t,X_t;X_T \mapsto g(X_T))$ as the solution of the FBSDE and/or the nonlinear PDE. In linear pricing, one has

$$\mathsf{price}(t,X_t;X_T \mapsto g(X_T)) = -\mathsf{price}(t,X_t;X_T \mapsto -g(X_T)). \tag{8}$$

In nonlinear pricing, these two prices are no longer necessarily the same but will give an upper and a lower price.

Using the Euler–Maruyama method to discretize time direction forward for both X_t and Y_t, we have

$$X_{t_{i+1}} = X_{t_i} + \mu(t_i,X_{t_i})\Delta t_i + \sigma(t_i,X_{t_i})\Delta W^i \tag{9}$$

and

$$Y_{t_{i+1}} = Y_{t_i} - f(t_i,X_{t_i},Y_{t_i},\Pi_{t_i})\Delta t_i + \Pi_{t_i}^T\sigma(t_i,X_{t_i})\Delta W^i. \tag{10}$$

2.1. Backward Time-Stepping

As discussed in the introduction, in deep BSDE methods, we compute pathwise realizations of Y given pathwise realizations of X. Since the FBSDE that we are considering are decoupled, a realization of X can be computed independently and ahead of the realization of Y. In the pathwise backward deep BSDE methods, we compute the realization of Y backwards, starting from a given final value $Y_T = g(X_T)$ and using the realizations of ΔW^i from X. In terms of filtration, this means that we are operating under a filtration that has all the information about W_t until $t = T$ and are measurable with respect to information about X and W up to time T (and not time t as in the forward method). Since we are using realizations of X and corresponding realizations of Y given the current strategy Π. to compute gradients with respect to strategy parameters, the strategy can be assumed known. Thus, the formulas in this subsection all contain realizations, not random variables.

2.1.1. Exact Backward Time-Stepping

To backward step in time direction, we rewrite (10) as

$$Y_{t_i} - f(t_i, X_{t_i}, Y_{t_i}, \Pi_{t_i})\Delta t_i = Y_{t_{i+1}} - \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Delta W^i \tag{11}$$

and solve for Y_{t_i}.

For a differential rates setup in a risk-neutral measure, the f generator function in the BSDE is

$$f(t, X_t, Y_t, \Pi_t) = -r^l(t)Y_t + (r^b(t) - r^l(t))\left(\sum_{j=1}^{n} \pi_j(t) - Y_t\right)^+. \tag{12}$$

This driver expresses that all assets $X_j(t)$ and positive cash balances grow at a risk-neutral rate $r^l(t)$ unless the cash position $Y_t - \sum_{j=1}^n \pi_j(t)$ is negative, and that negative cash balance will grow at a rate $r^b(t)$ corresponding to the higher borrowing rate as compared with the lower or equal lending rate.

There are two cases for Equation (12):

(1). If $\sum_{j=1}^n \pi_j(t) > Y(t)$:

$$f(t, X_t, Y_t, \Pi_t) = -r^l(t)Y_t + (r^b(t) - r^l(t))\left(\sum_{j=1}^{n} \pi_j(t) - Y_t\right). \tag{13}$$

Inserting this into Equation (11) and solving, we obtain

$$Y_{t_i} = \frac{Y_{t_{i+1}} + \left(r^b(t_i) - r^l(t_i)\right)\left(\sum_{j=1}^n \pi_j(t_i)\right)\Delta t_i - \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Delta W^i}{1 + r^b(t_i)\Delta t_i}. \tag{14}$$

(2). If $\sum_{j=1}^n \pi_j(t) \leq Y(t)$:

$$f(t, X_t, Y_t, \Pi_t) = -r^l(t)Y_t. \tag{15}$$

Inserting this into Equation (11) and solving, we obtain

$$Y_{t_i} = \frac{Y_{t_{i+1}} - \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Delta W^i}{1 + r^l(t_i)\Delta t_i}. \tag{16}$$

However, we do not know Y_{t_i} before solving the nonlinear Equation (11) for it. From (14) and (16) and the conditions involving Y_{t_i}, we obtain that the condition $Y_{t_i} < \sum_{j=1}^n \pi_j(t_i)$ is equivalent to

$$Y_{t_i+1} < \left(1 + r^l(t_i))\Delta t_i\right) \sum_{j=1}^{n} \pi_j(t_i) + \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Delta W^i \tag{17}$$

and the same for the relation with $\geq$. Thus, if (17) is satisfied, we use (14), otherwise (16).

2.1.2. Time-Stepping from Taylor Expansion

By a first-order Taylor expansion, we have

$$\begin{aligned} f\Big(t_i, X_{t_i}, Y_{t_i}, \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Big) \quad &\approx \quad f\Big(t_i, X_{t_i}, Y_{t_{i+1}}, \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Big) \\ &\quad - \frac{\partial f}{\partial Y}\Big(t_i, X_{t_i}, Y_{t_{i+1}}, \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Big)\big(Y_{t_{i+1}} - Y_{t_i}\big). \end{aligned} \tag{18}$$

Inserting this into Equation (11) and solving for Y_{t_i}, we have the following:

$$Y_{t_i} = Y_{t_{i+1}} + \frac{f\Big(t_i, X_{t_i}, Y_{t_{i+1}}, \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Big)\Delta t_i - \Pi_{t_i}^T \sigma(t_i, X_{t_i}) \Delta W^i}{1 - \frac{\partial f}{\partial Y}\Big(t_i, X_{t_i}, Y_{t_{i+1}}, \Pi_{t_i}^T \sigma(t_i, X_{t_i})\Big)\Delta t}. \tag{19}$$

Note that f and $\frac{\partial f}{\partial u}$ are evaluated at $Y_{t_{i+1}}$.

With the same setup for the differential rates problem, it is clear that there are only two possible forms for f:

(1). If $\sum_{j=1}^n \pi_j(t_i) > Y_{t_{i+1}}$:

$$f(t_i, X_{t_i}, Y_{t_{i+1}}, \Pi_{t_i}) = -r^l(t_i) Y(t_i) + (r^b(t_i) - r^l(t_i))\left(\sum_{j=1}^n \pi_j(t_i) - Y_{t_{i+1}}\right) \tag{20}$$

and

$$\frac{\partial f}{\partial Y} = -r^b(t_i). \tag{21}$$

Inserting this into Equation (19), we obtain the same (14).

(2). If $\sum_{j=1}^n \pi_j(t_i) \leq Y_{t_{i+1}}$:

$$f(t_i, X_{t_i}, Y_{t_{i+1}}, \Pi_{t_i}) = -r^l(t_i) Y_{t_{i+1}} \tag{22}$$

and

$$\frac{\partial f}{\partial Y} = -r^l(t_i). \tag{23}$$

Inserting this into Equation (19), we again obtain (16).

Notice that both the exact and Taylor backward steps have the same form (14) and (16); the difference is in the conditions when they are applied.

3. Deep BSDE Approach

3.1. Forward Approach

As introduced in Han and Jentzen (2017), with forward time-stepped Equations (9) and (10), one minimizes the loss function

$$E(||Y_T^{F,\Pi} - g(X_N)||^2) \tag{24}$$

over parameters of the strategies $\Pi_\cdot$ and over initial value Y_0 , where $Y_T^{F,\Pi}$ is the result of forward stepping (10) with strategy vector Π.

The initial portfolio value Y_0 is a parameter of the minimization problem, as are all the parameters of the DNN functions $\pi_i(t_i, X_{t_i})$ treated as functions of X_{t_i} (that give the stochastic vector process Π_t as value of the holdings of the risky underlying securities in the portfolio). Since X_0 is fixed, instead of learning a function $\pi_0(X_0)$, one learns a parameter π_0. Alternatively, one can learn a single function $\pi(t_i, X_{t_i})$ as function of t_i and X_{t_i}, which means that all the parts of the computational graph that represent the evaluation of $\pi(t, x)$ share the same DNN parameters.[3] The minimization problem is then solved with standard

deep learning approaches. For the case of random X_0, one also learns the initial value of Y_0 as a function $\mathsf{Yinit}(X_0)$ of X_0 using the same loss function.

The minimization is typically implemented through mini-batch stochastic gradient descent and similar approaches, such as Adam. The gradients of the expected value with respect to the trainable parameters are approximated by the gradients of the empirical sum over the loss function as evaluated on a number of trajectories. In this work, we generate new trajectories for each mini-batch for each epoch/stochastic gradient descent step, and we tested both on fixed testing batches as well as freshly generated batches. In general, strategies and initial value functions will be somewhat noisy, since the approximation of the expectation respective its gradient will depend on the mini-batch size. Such noise can be reduced by increasing the mini-batch size or by various kinds of postprocessing. Given a strategy Π and keeping it fixed, one can also compute a refined $\mathsf{Yinit}(X_0)$ by performing a separate optimization only on the parameters of Yinit.

3.2. Backward Approach

In the backward approach, one time-steps Equation (9) forward but time-steps Equation (10) backward, starting from $Y_T = g(X_T)$. As discussed in the previous section, one can use an analytical solution of (11) or some Taylor expansion approach. Using either approach, one will obtain an expression or implementation

$$Y_{t_i} = \mathsf{ybackstep}(t_i, Y_{t_{i+1}}, X_{t_i}, \Pi_{t_i}, \Delta W^i). \tag{25}$$

For the differential rates setup, the backward step ybackstep is given by (14) and (16) depending on whether $Y_{t_{i+1}}$ satisfies (17) or not (for the exact step), or whether $\sum_{j=1}^{n} \pi_j(t_i) \leq Y_{t_{i+1}}$ or not (for the Taylor step). In general, ybackstep can be any exact or approximate solution of (11).

As discussed before, the time-stepping of both (9) (forward) and (10) (backward) occurs in a single realization—one generates a realization ΔW^i for all i, simulates X_{t_i} for all i from (9) and then performs the time-stepping for this specific realization for $\mathcal{Y}$. A little bit more formally, let us define recursively backwards for any given current strategy function $\pi(t, x)$ the following random variables $\mathcal{Y}^\pi$ that are measurable as of time T: $\mathcal{Y}^\pi_T = \mathcal{Y}^\pi_{t_N} = g(X_T)$ and

$$\mathcal{Y}^\pi_{t_i}(\omega) = \mathsf{ybackstep}(t_i, \mathcal{Y}^\pi_{t_{i+1}}(\omega), X_{t_i}(\omega), \pi(t_i, X_{t_i}(\omega)), \Delta W^i(\omega)) \tag{26}$$

realization by realization. This is a well-defined sequence of random variables, all measurable as of T.

For fixed X_0, the loss function used in Wang et al. (2018) and Liang et al. (2021) is

$$\mathsf{var}(\mathcal{Y}^\pi_0) \tag{27}$$

and one optimizes over functions π to find the strategy that minimizes this initial variance. (In the limit for vanishing time-steps, the gradient of the PDE solution will lead to zero variance).

For the stochastic gradient descent approaches, this variance will be approximated by the variance over the current mini-batch. Thus, for the mini-batch stochastic gradient step, the loss function will be the mini-batch variance

$$E(||\mathcal{Y}^\pi_0 - \bar{Y}_0||^2), \tag{28}$$

where $\bar{Y}_0$ will be the mean over the mini-batch. Similarly to before, approximating the expectation with a finite sum over the mini-batch will introduce noise into the optimization and into the computation of $\bar{Y}_0$, which depends on the size of the mini-batch.

Thus, for estimates of the initial value of the instrument, one does not necessarily have to use only the last mini-batch mean, one can compute the mean of $\mathcal{Y}^\pi_0$ over a larger sample

of paths or batches generated with a fixing trading strategy. Instead of using the mini-batch mean in the loss function, one can learn $\bar{Y}_0$ as a parameter/variable (resulting in the same loss function but with different meaning of $\bar{Y}_0 = \mathsf{Yinit}$).

Once X_0 is random, one can no longer use batch variance in a straightforward way. Instead (and inspired by the parameter version just discussed), one uses a loss function

$$E(||\mathcal{Y}_0^\pi - \mathsf{Yinit}(X_0)||^2), \tag{29}$$

where the $\mathsf{Yinit}(X_0)$ is a function represented by a DNN which is learned as part of the DL approaches.

Restating more formally again, we can define

$$Y_0^\pi = \mathsf{Yinit}(X_0; \pi) = E[\mathcal{Y}_0^\pi | X_0] \tag{30}$$

and define a loss function[4]

$$\mathcal{L}(\pi) = E(||\mathcal{Y}_0^\pi - \mathsf{Yinit}(X_0; \pi)||^2). \tag{31}$$

We now are solving a (time-discrete but continuous in space) stochastic control problem

$$\pi^* = \operatorname*{argmin}_\pi \mathcal{L}(\pi) \tag{32}$$

and we define $u_0^*(X_0) = \mathsf{Yinit}(X_0; \pi^*)$. We expect that π^* will be an approximation (up to time-discretization error) for the gradient of the solution of the FBSDE (and thus PDE) in $t = 0 \dots T$ and u_0^* will be an approximation of the solution at time $t = 0$. Proceeding similarly for arbitrary t, we can similarly obtain approximations of the solution at any time t in $0 \dots T$.

We solve (32) by gradient-based optimization methods to iteratively improve strategies until we obtain a minimum. The involved expectations cannot be analytically computed as functions of π, so one needs to approximate them by sampling. As discussed above, one uses mini-batch stochastic gradient descent and methods based on such (such as Adam). Moreover, instead of exactly determining the Yinit function as a conditional expectation given a strategy π, we update parameters for both π and Yinit at the same time.

As an illustration, the computational graph to compute a single sample for the empirical loss function for the backward method with random X_0 is shown in Figure 1. (For the initial variable version for fixed X_0, both Y_{init} and Π_0 would be variables independent of X_0, rather than networks depending on X_0, and X_0 would be an input.) First, one simulates X forward, starting at the first time-step, proceeding through intermediate time-steps, and reaching the final time-step, according to (9). At the final time-step, $\mathcal{Y}_T(\omega)$ is set to $g(X_T)$, and backward steps **ybstep** are taken (as discussed in Section 2.1), proceeding throughintermediate steps, until one reaches the first time-step again. At the first time-step, one computes $\mathcal{Y}_0(\omega) - \mathsf{Yinit}(X_0(\omega))$ ("Mismatch" shown in green), and the empirical loss function is defined as an average over the square of the mismatch. Unfilled boxes are given implementations/operations that do not change, pink boxes are networks to be trained, and blue circles are randomly generated each time.

Algorithm 1 shows an entire pathwise backward deep BSDE method as pseudocode. The computational graph shown in Figure 1 represents lines 5–9 in the pseudocode, with L being the sum of squares of the mismatch.

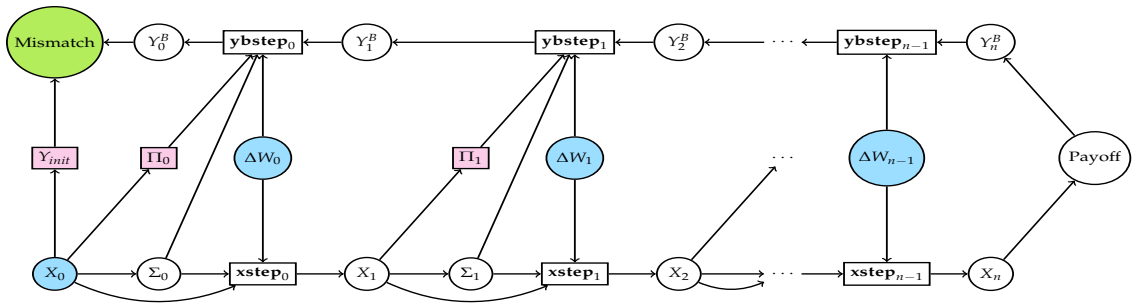


Figure 1. Computational graph for the entire method with initial network.

Algorithm 1 The pathwise Backward deep BSDE Method

procedure PATHWISEBACKWARDDEEP BSDE(*batchsize*)
▷ Initialization
  Initialize DNNs $\mathsf{Yinit}(X;\theta_u)$ and $\pi(t,X;\theta_\pi)$ with random parameters θ_u and θ_π
  repeat
    for $nbatch \leftarrow 1, batchsize$ **do**
      Generate trajectory $X_{t_i}^{nbatch}$
      Generate corresponding backward trajectory $Y_{t_i}^{nbatch}$ with current π
    end for
    $L \leftarrow \sum_{nbatch=1}^{batchsize}\left(Y_0^{nbatch} - \mathsf{Yinit}(X_0^{nbatch})\right)^2$
    Update θ_u by $\nabla_{\theta_u} L$ with SGD, Adam, or similar
    Update θ_π by $\nabla_{\theta_\pi} L$ with SGD, Adam, or similar
  until stopping criterion satisfied
    ▷ $\mathsf{Yinit}(x;\theta_u)$ approximates $u(0,x)$ and $\pi(t,x;\theta_\pi)$ approximates $\nabla_x u(t,x)$
end procedure

Similarly, one can introduce additional terms

$$E(||Y_{t_i} - \mathsf{Ylearned}_i(X_{t_i})||^2) \tag{33}$$

at some (or all) intermediate times t_i to learn some approximations for the solution function $\mathsf{Ylearned}_i(X_{t_i})$ as a function represented by a DNN, which is learned as part of the minimization of the combined loss function (with initial and intermediate time terms). The DNN thus learned will be an approximation of $u(t_i, X_{t_i}) = Y_{t_i}^\pi = E[\mathcal{Y}_{t_i}^\pi | X_{t_i}]$. Alternatively, one could first learn the control from maturity to the last intermediate time using the loss function for the intermediate time and then learn the control for the interval to the previous intermediate times piece by piece until one reaches the initial time.

Considering the strategy Π fixed, one can thus obtain functions Yinit and $\mathsf{Ylearned}_i$ as solutions of least square problems, and one can use standard approaches to compute them. In the method described above, strategies are judged by variance against these functions, and we are looking for strategies that minimize variance. One could alternatively look for strategies that optimize other risk measures, such as quantiles, expected shortfall, or unequally weighted or one-sided variances (to only or predominantly optimize over trajectories where not enough initial capital was provided in hindsight), as long as one can compute these loss functions appropriately on mini-batches and optimize them well. Given any such strategies, one can then determine an appropriate initial value or price also in different ways, not only as conditional expectations as above but possibly such that the probability that initial capital was not enough is at most a given value, or that the expected initial mismatch is bounded by a certain number, even in the case that initial capital was not enough.

The setting proposed here with a variance against the conditional expectation most closely fits with the setting of the forward pathwise deep BSDE and the underlying PDE, and we thus use it here exclusively. We intend to study the other settings in future work.

All the pathwise backward methods except the one using batch variance are novel, to the best of our knowledge, and we are the first to apply them to nonlinear generators f, in particular nonlinear with respect to y.

4. Results

We present results on two financial derivatives treated in the literature. The two financial derivatives are a call combination (long one call on the maximum across assets with a strike of 120 and short two calls on the maximum with a strike of 150, with a maturity of 0.5 years) as in[5] Han and Jentzen (2017) and a straddle on the maximum (long both a put and a call, with a strike of 100.0 with a maturity of 1 year) as in Forsyth and Labahn (2007). These two instruments correspond to the payoff g in (6) as $g(M_T) = (M_T - K_1)^+ - 2(M_T - K_2)^+$, with $K_1 = 120$ and $K_2 = 150$ for the call combination and $g(M_T) = (M_T - K)^+ + (K - M_T)^+$ with $K = 100$ for the straddle, both with $M_t = \max_{i=1}^n X_t^i$ as the maximum across assets, which in one dimension simplifies to $M_t = X_t$. While we are presenting results for the one-dimensional case to compare with the results of Forsyth and Labahn (2007), the same method can be applied to high dimensions. We refer to Ganesan et al. (2022), who show the application of the forward pathwise deep BSDE methods to high-dimensional boundary value/barrier option problems and leave high-dimensional examples for the backward pathwise deep BSDE methods introduced here for future work.

Both examples use constant-coefficient Black–Scholes dynamics for the underlying X_t, where $\mu(t, X_t) = \mu$ and $\sigma(t, X_t) = \sigma$ in (4), but with different constant values, as listed in the subsections. Both examples are differential rates problems, where the BSDE (5) has the generator $f(t, X_t, Y_t, \Pi_t)$, as given in (12), once again with different parameters for the two examples.

4.1. Call Combination

For the example from Han and Jentzen (2017), we picked $\sigma = 0.2$, $\mu = 0.06$, $r_l = 0.04$, and $r_b = 0.06$. We used 50 time-steps. For the fixed X_0 case, we picked $X_0 = 120$. For the random/varying X_0 case, we picked a uniform distribution within the range $[70, 170]$. We use various batch sizes, prescaling, Adam with default parameters and exponentially decaying learning rate, two hidden layers with $dim + 10 = 11$ neurons, and activation function Softplus for the first two layers and then identity on the output layer.

The loss function behaves similarly for all methods, and we show an example in Figure 2.

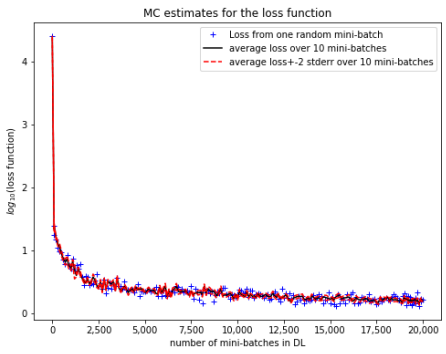
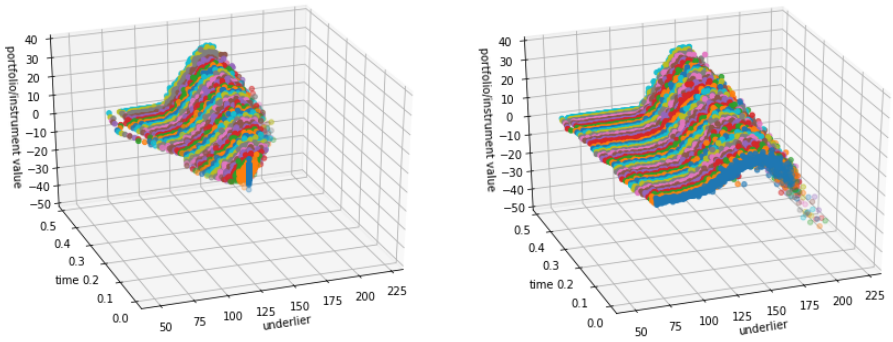


Figure 2. Loss function over 20,000 mini-batches for long call combination for batch-variance method with exact backward step. Loss functions for the other variants look very similar. Batch size 512.

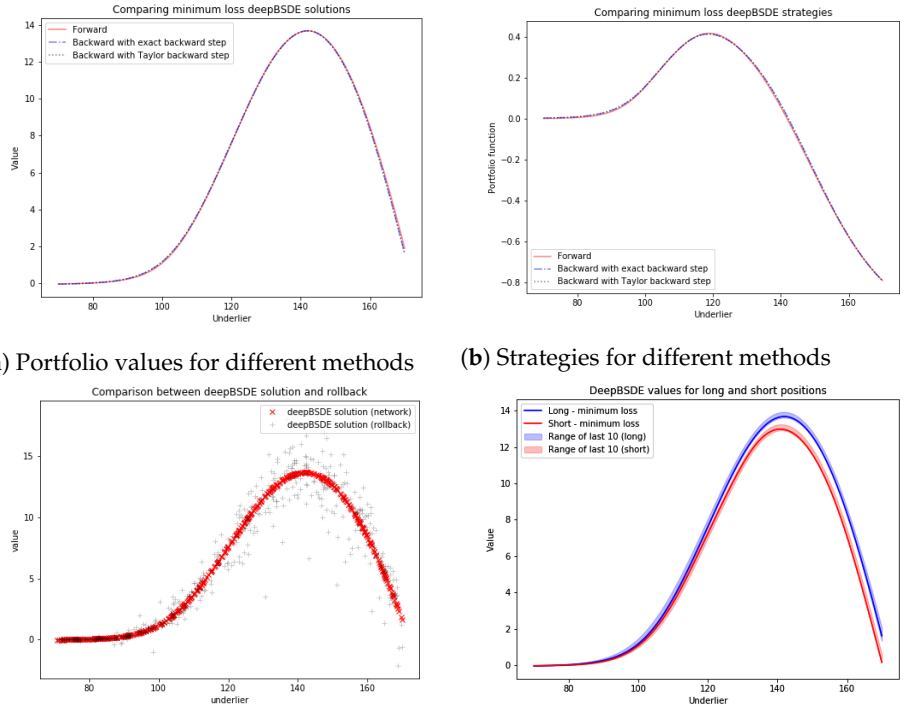
Figure 3 shows the Y path values for fixed X_0 (on the left) and for uniform random X_0 (on the right) for the long call combination. Notice that the random X_0 variant covers much more of the solution surface. Figure 4 shows initial Yinit network results vs. rollback, minimal loss solution and the range of the 10 last validation solutions for long and short, and solution and strategy from different methods. We see that the solutions from different

deep BSDE (including Taylor vs. exact step) are close to each other, and the strategies are also rather similar, with the forward strategy slightly different but close.



(**a**) Fixed X_0. (**b**) Random X_0.

Figure 3. Y path values at 20,000 mini-batches for the batch-variance version with exact backward step (other variants look very similar) for long call combination with fixed X_0 on the left and initial network version with random X_0 on the right. Notice the much smaller coverage for fixed X_0. Batch size 512.



(**a**) Portfolio values for different methods (**b**) Strategies for different methods

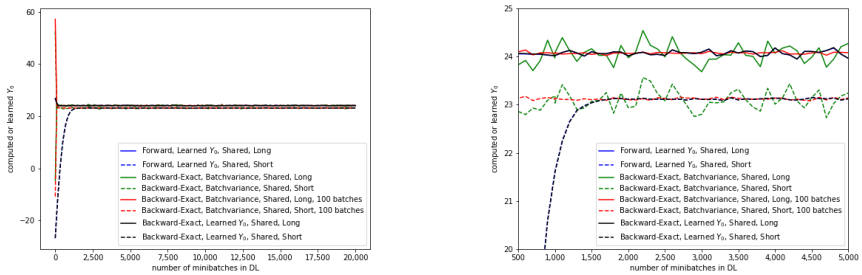
(**c**) Backward exact—Yinit vs. rollback (**d**) Backward exact—long and short

Figure 4. Call combination: results for random X_0. Batch size 512. Panels (**a**–**c**) show results for a long position. (**d**) shows results for both long and short positions.

4.2. Straddle

(Forsyth and Labahn 2007, Table 1 on page 28 in hjb.pdf) picked $\sigma = 0.3$, $\mu = r_b = 0.05$, and $r_l = 0.03$. We used 100 time-steps (one of the numbers of time-steps for which results are given in tables in Forsyth and Labahn (2007)). The strike for the straddle is 100. We used various batch sizes, prescaling, Adam with default parameters and exponentially decaying learning rate, two hidden layers with $dim + 10 = 11$ neurons, and activation function Softplus for the first two layers and then identity on the output layer, as in the call combination case.

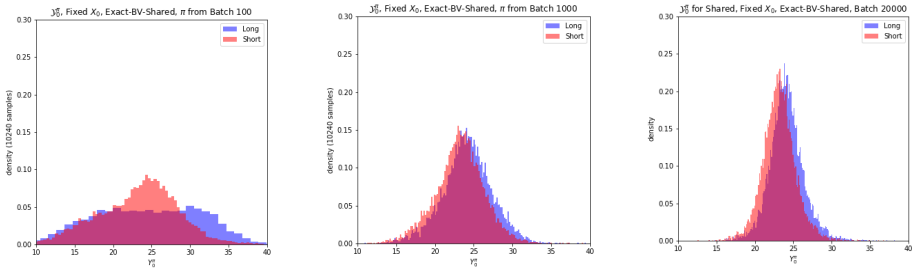
We first consider the fixed X_0 case. Like Forsyth and Labahn (2007), we pick X_0 to be 100. In Figure 5, we plot Y_0 estimates and parameter for different backward and forward methods for a certain range of mini-batches, showing the behavior once the methods have converged to a region where the mini-batch size limits how well the loss function is approximated (and thus the results vary within a range). We see that the method that learns the Y_0 parameter initially converges more slowly (which the lower price converging even more slowly than the upper price) than the batch-variance methods. However, once close to true value, its convergence is smoother and better than the batch-variance methods—and it varies less. Computing the mean over 100 mini-batches rather than over one leads to a faster and smoother convergence of the initial value for the batch-variance variants.



(**a**) All mini-batches. (**b**) Detail.

Figure 5. Y_0 estimates or parameters for the straddle case—all 20,000 mini-batches (**a**) and detail (**b**). Exact backward step. (Batch size 256). The results for Taylor backward step look very similar.

We use this example to visualize the details of the pathwise backward methods. We set π to the strategy obtained after optimizing over 100, 1000, and 20,000 mini-batches using the exact backward step with the batch-variance variant for fixed X_0. Using these strategies, we generate many samples from the rollback random variable $\mathcal{Y}_0^\pi$ and show them in Figure 6. The optimization starts with a π DNN with random weights and biases. In the beginning of the optimization, the strategy cannot control the distribution of the rollback well yet, but it slowly reduces the distribution's variance, making it (approximately) unimodal. After batch 100, the strategy for the short position already creates a unimodal distribution for the rollback, while the strategy for the long position still results in a distribution that is bimodal and more spread out. After batch 1000, both strategies result in unimodal and narrowing distributions. After batch 20,000, the strategies are even narrower and more peaked, with clearer separation between short and long.



(**a**) With π from batch 100 (**b**) With π from batch 1000 (**c**) With π from batch 20,000

Figure 6. Distribution of the rollback $\mathcal{Y}_0^\pi$ for strategies of certain batches in the optimization. Straddle. Backward method with exact backward step. Single DNN for π. Batch size 512.

In Figure 7, we show the distribution of batch means for freshly generated mini-batches, now only for the policy from the 20,000th mini-batch. This distribution is much narrower and clearly separated between long and short. For comparison, we also plotted the mean over all generated mini-batches.

In Figure 8, we show the distribution of batch losses for freshly generated mini-batches, again only for the policy from the 20,000th mini-batch. For comparison, we also plotted the average loss over all generated mini-batches. We can see that the losses are similarly distributed for short and long positions. To estimate the loss very well from a single mini-batch, one would need a batch size that is 10 or 100 times larger. However, using stochastic gradient descent approaches has the advantage of implicit regularization and avoidance of local minima.

Figure 9 shows loss curves over five independent runs started from different seeds. Crosses indicate losses approximated by single mini-batches; black shows loss approximations averaged over ten mini-batches, with red showing ranges implied by ten mini-batches. While the behavior under different seeds is different in the beginning, it becomes very similar and lies in the same range once the optimization through stochastic gradient descent or Adam progresses more and more. The curves all go through the same phases—an initial fast decrease which subsequently slows down and then slowly decreases and stabilizes once the loss reaches the magnitude of estimation and stochastic gradient noise.

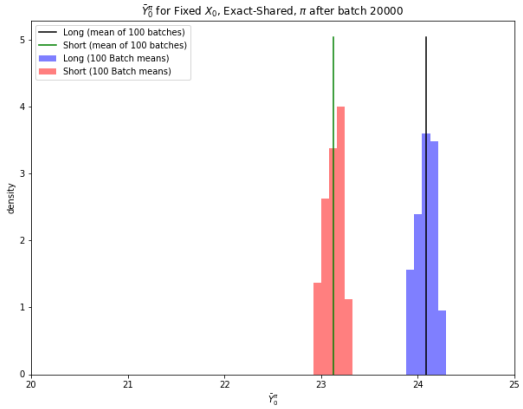


Figure 7. Distribution of batch means of the rollback $\mathcal{Y}_0^\pi$ for the strategies after the 20,000th batch in the optimization. Straddle. Backward method with exact backward step. Single DNN for π. Batch size 512.

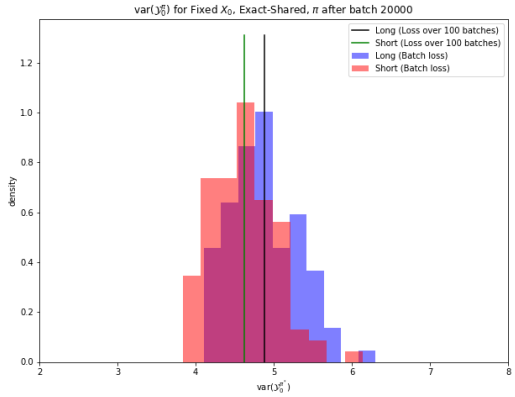


Figure 8. Distribution of batch losses for the rollback $\mathcal{Y}_0^\pi$ for the strategies after the 20,000th batch in the optimization. Straddle. Backward method with exact backward step. Single DNN for π. Batch size 512.

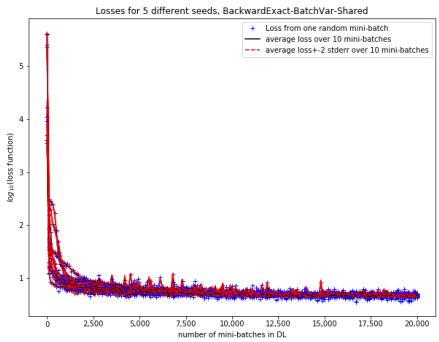


Figure 9. Loss function over 20,000 mini-batches for the short and long straddle for batch-variance method with exact backward step with a single DNN for π for five independent realizations. Behavior is very similar for other variants.

One can similarly extract, visualize, and study the distribution of the rollbacks and their means and losses defined on them (with respect to batch mean, with respect to a learned mean, or with respect to an externally given reference mean). In the noisy region where the loss has stabilized up to noise, the distribution of the rollbacks, means, and losses during the stochastic gradient descent method or Adam method starts to resemble the distribution under one (randomized or deterministic) policy but with much larger sample size. This actually means that if we work with the distribution of rollbacks, means, or losses over a sequence of mini-batches from the optimization from that noise region, we will obtain results corresponding to methods with larger samples and mini-batch sizes, allowing us to obtain quite accurate results despite the relatively small mini-batch size (such as 128, 256, or 512). Since the loss estimate from a single mini-batch is not very accurate, selecting the strategy to be from the mini-batch with the smallest loss estimate does not necessarily result in picking the strategy with the smallest actual loss. However, one can pick several candidate strategies and estimate the loss more accurately and then select the one with the smallest accurate loss or use an ensemble from some with small accurate losses. Since our results (mean or range over mini-batches during optimization) agree well with the range of results given by Forsyth and Labahn (2007), we do not do so here, but we plan to do so in future work.

We compare our lower and upper price against the results given in (Forsyth and Labahn 2007, Tables 2 and 3) in Table 1. The results are very close to each other. Since Forsyth and Labahn use a PDE method, they report results for a certain number of space steps, while our method does not discretize space. They report results for higher numbers of space steps (and an equally higher number of time-steps) which are even closer to our results, but the different number of time-steps does not allow definite conclusions.

In Table 2, we report the results over five different random seeds. One can see that the price across seeds varies within an interval around the prices given in Forsyth and Labahn. The particular seed does not impact the price a lot. The range of prices over mini-batches in the optimization in the noise region provides estimates that are consistent with estimates across different random seeds.

For the random X_0 case, we pick X_0 uniformly within the range $[50, 150]$ but plot results within the range $[80, 120]$. We saved and discretized Figure 1 from (Forsyth and Labahn 2007, Figure 1) (the hjb PDF version), extracted relative coordinates for the points on the curve and converted them to values, and plotted them as curves in the figures (curves shown in black).

Figure 10 shows comparisons of the backward exact deep BSDE method (both minimum loss solution and range of last 10) against the curves from (Forsyth and Labahn 2007, Figure 1), while Figure 11 shows comparisons of different deep BSDE methods against each other. It can be seen that the curves from (Forsyth and Labahn 2007, Figure 1) are within the range for both batch sizes and that the different methods mostly agree, although

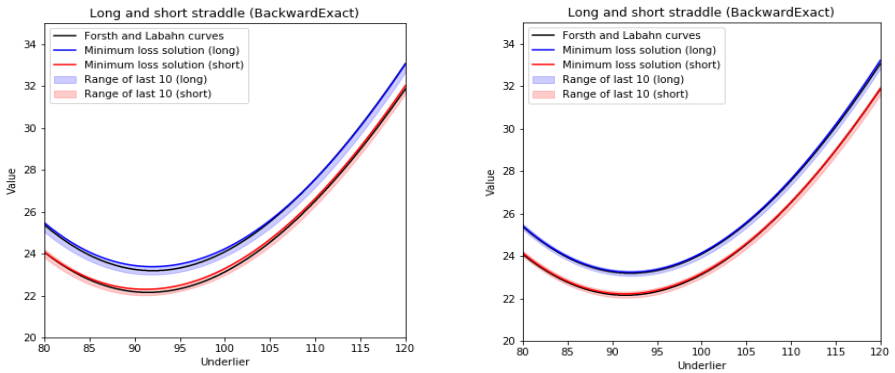
somewhat more so for batch size 512.[6] Similarly to before, exact vs. Taylor step only has minimal impact.

Table 1. Different pricers. * Means that Taylor is different by 0.01. Fixed X_0. Y_0 range over every 100 in the last 1000 in parentheses. Middle of range as price.

Method	Upper Price	Lower Price
Results from Forsyth and Labahn-101 nodes		
Fully Implicit HJB PDE (implicit control)	24.02	23.06
Crank–Nicolson HJB PDE (implicit control)	24.05	23.09
Fully Implicit HJB PDE (pwc policy)	24.01	23.07
Crank–Nicolson HJB PDE (pwc policy)	24.07	23.09
Forward deep BSDE—20,000 batches, size 256		
Learned Y_0 (shared)	24.06 (23.99–24.14)	23.10 (23.02–23.17)
Learned Y_0 (separate)	24.07 (24.01–24.12)	23.10 (23.06–23.15)
Backward deep BSDE—20,000 batches, size 256		
Batch variance/1 (shared)	24.14 (23.98–24.30)	23.19 (23.00–23.37)
Batch variance/100 (shared)	24.08 (24.06–24.09)	23.13 (23.09–23.16)
Batch variance/1 (separate)	24.06 (23.94–24.19 *)	23.10 (22.95–23.25)
Batch variance/100 (separate)	24.07 (24.06–24.09 *)	23.12 (23.10–23.13)
Learned Y_0 (shared)	24.06 (23.99–24.14)	23.10 (23.02–23.17)
Learned Y_0 (separate)	24.06 (24.01–24.11 *)	23.10 (23.06–23.15)

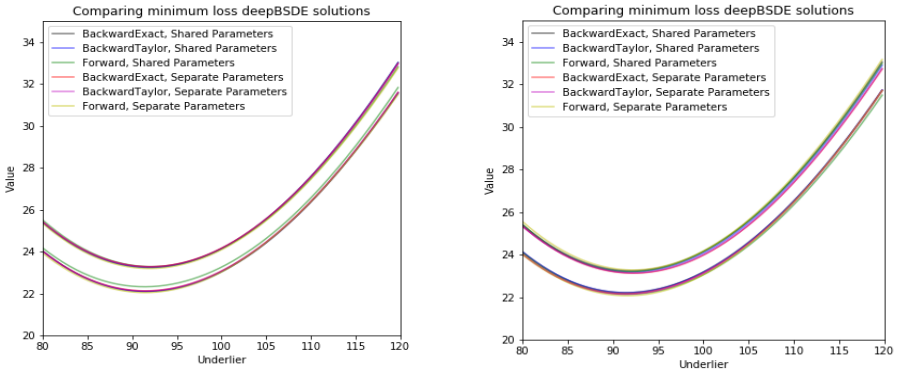
Table 2. Impact of random seed. Fixed X_0. Exact backward step. Range of price over seeds. Y_0 range over every 100 in the last 1000 in parentheses.

Method	Upper Price	Lower Price
Results from Forsyth and Labahn—101 nodes		
Fully Implicit (implicit control)	24.02	23.06
Crank–Nicolson (implicit control)	24.05	23.09
Fully Implicit (pwc policy)	24.01	23.07
Crank–Nicolson (pwc policy)	24.07	23.09
Backward deep BSDE—20,000 batches, size 512, five seeds		
Batch variance/1 (shared)	24.02–24.08 (23.87–24.29)	23.06–23.12 (22.91–23.33)
Batch variance/100 (shared)	24.06–24.07 (24.03–24.09)	23.11–23.12 (23.09–23.15)



(**a**) Batch size 128 (**b**) Batch size 512

Figure 10. $\mathsf{Yinit}(X_0)$ for two batch sizes and backward exact step plotted against Forsyth and Labahn curves from (Forsyth and Labahn 2007, Figure 1).



(**a**) Batch size 256 (**b**) Batch size 512

Figure 11. Yinit(X_0) for various methods for two batch sizes.

5. Conclusions

We first introduced FBSDE for general nonlinear problems, with particular details for the differential rates problem, time-discretized them, and then derived exact solutions and Taylor approximations for the backward step equation. We then quickly described the pathwise forward and backward deep BSDE approaches that we consider—both the batch-variance variant already described in the literature and also the novel initial variable and network versions; the last one is for random X_0. Then, we applied these methods for the differential rates problem for the call combination case from Han et al. (2018) and for the straddle case from Forsyth and Labahn (2007). We compared the results for a case with fixed X_0 and for a case with varying X_0 with the results from Forsyth and Labahn (2007) and saw that they agree well with Forsyth and Labahn (2007) and each other.

The deep BSDE methods described in this paper use a very different approach from the PDE methods in Forsyth and Labahn (2007), but they give results that agree well with those published there. This makes us confident that these methods can be used to generically and efficiently approximate solutions to such nonlinear pricing problems, using relatively small batch sizes such as 128, 256, or 512.

6. Disclaimer

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Wells Fargo Bank, N.A., its parent company, affiliates, and subsidiaries.

Author Contributions: Conceptualization, Y.Y., N.G. and B.H.; methodology, Y.Y., N.G. and B.H.; software, Y.Y., N.G. and B.H.; validation, Y.Y., N.G. and B.H.; formal analysis, Y.Y., N.G. and B.H.; investigation, Y.Y., N.G. and B.H.; writing—original draft preparation, Y.Y., N.G. and B.H.; writing—review and editing, B.H.; visualization, Y.Y., N.G. and B.H.; supervision, B.H.; project administration, B.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable

Acknowledgments: The authors would like to thank Orcan Ogetbil, Daniel Weingard, and Xin Wang for proof reading drafts and giving helpful feedback; Vijayan Nair for discussion regarding methods, presentation, and results, as well as for reviewing the paper; and Agus Sudjianto for supporting this research. They would also like to thank Dan Pirjol for suggestions and recommendations that helped improve the readability and presentation of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 Set $Z_t = \Pi_t^T \sigma(t, X_t)$ and $\tilde{f}(t, X, Y, Z) = f(t, X, Y, \sigma(t, X_t)^{-T} Z_t^T)$. Then, (Pardoux 1998, Theorem 2.2) with this Z_t and $\tilde{f}$ shows that one can construct a solution Y_t and Z_t of the BSDE from a classical $C^{2,1}$ solution of a corresponding PDE and a solution X_t. Rewriting the PDE and BSDE in terms of Π_t^T instead of Z_t and using function f rather than $\tilde{f}$ gives the form reported below. For the opposite direction, (Pardoux 1998, Theorem 2.4) shows how a solution Y_t of the BSDE (corresponding to a X that starts at x at time t) gives a continuous function $u(t, x)$, which is a viscosity solution of the corresponding PDE.

2 In general, the terminal condition could be given as a random variate G_T that is measurable with respect to the information available of time T (i.e., the sigma algebra generated by X_t with $t \leq T$). The FBSDE approach then will be more general than the PDE approach. If there is an exact (or approximate) Markovianization with a Markov state M_t, the strategy Π_t and the solution Y_t would in general be functions $\pi(t, M_t)$ and $u(t, M_t)$ of that Markov state. We only treat the usual final value case here.

3 There are many introductions into DL, DNN, and common forms of DNN. For a minimal one geared towards deep BSDE, see Hientzsch (2021).

4 This is actually the expectation of the conditional variance $\mathsf{var}(\mathcal{Y}_0^\pi | X_0)$ over the distribution of X_0.

5 Here, we consider the 1-dimensional case, while Han and Jentzen (2017) consider the 100-dimensional case.

6 Notice that (Forsyth and Labahn 2007, Figure 1) do not give the number of space or time-steps used for their plot.

References

Forsyth, Peter A., and George Labahn. 2007. Numerical methods for controlled Hamilton-Jacobi-Bellman PDEs in finance. *Journal of Computational Finance* 11: 1–44. Available online: https://cs.uwaterloo.ca/~paforsyt/hjb.pdf (accessed on 15 March 2023). [CrossRef]

Ganesan, Narayan, Yajie Yu, and Bernhard Hientzsch. 2022. Pricing barrier options with deep backward stochastic differential equation methods. *Journal of Computational Finance* 25. Available online: https://ssrn.com/abstract=3607626 (accessed on 15 March 2023). [CrossRef]

Han, Jiequn, and Arnulf Jentzen. 2017. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics* 5: 349–80.

Han, Jiequn, Arnulf Jentzen, and Weinan E. 2018. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences* 115: 8505–10. [CrossRef] [PubMed]

Hientzsch, Bernhard. 2021. Deep learning to solve forward-backward stochastic differential equations. *Risk Magazine*, February. Available online: https://ssrn.com/abstract=3494359 (accessed on 15 March 2023).

Huré, Côme, Huyên Pham, and Xavier Warin. 2020. Deep backward schemes for high-dimensional nonlinear PDEs. *Mathematics of Computation* 89: 1547–79. [CrossRef]

Liang, Jian, Zhe Xu, and Peter Li. 2021. Deep learning-based least squares forward-backward stochastic differential equation solver for high-dimensional derivative pricing. *Quantitative Finance* 21: 1309–23. Available online: https://ssrn.com/abstract=3381794 (accessed on 15 March 2023). [CrossRef]

Mercurio, Fabio. 2015. Bergman, Piterbarg, and beyond: Pricing derivatives under collateralization and differential rates. In *Actuarial Sciences and Quantitative Finance*. Berlin: Springer, pp. 65–95. Available online: https://ssrn.com/abstract=2326581 (accessed on 15 March 2023).

Pardoux, Étienne. 1998. Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order. In *Stochastic Analysis and Related Topics VI*. Berlin: Springer, pp. 79–127.

Raissi, Maziar. 2018. Forward-backward stochastic neural networks: Deep learning of high-dimensional partial differential equations. *arXiv* arXiv:1804.07010.

Wang, Haojie, Han Chen, Agus Sudjianto, Richard Liu, and Qi Shen. 2018. Deep learning-based BSDE solver for LIBOR market model with application to bermudan swaption pricing and hedging. *arXiv* arXiv:1807.06622. Available online: https://ssrn.com/abstract=3214596 (accessed on 15 March 2023).

Warin, Xavier. 2018. Nesting Monte Carlo for high-dimensional non-linear PDEs. *Monte Carlo Methods and Applications* 24: 225–47. [CrossRef]

risks

MDPI

Article

Sparse Modeling Approach to the Arbitrage-Free Interpolation of Plain-Vanilla Option Prices and Implied Volatilities

Daniel Guterding

Technische Hochschule Brandenburg, Magdeburger Straße 50, 14770 Brandenburg an der Havel, Germany; daniel.guterding@th-brandenburg.de

Abstract: We present a method for the arbitrage-free interpolation of plain-vanilla option prices and implied volatilities, which is based on a system of integral equations that relates terminal density and option prices. Using a discretization of the terminal density, we write these integral equations as a system of linear equations. We show that the kernel matrix of this system is, in general, ill-conditioned, so that it cannot be solved for the discretized density using a naive approach. Instead, we construct a sparse model for the kernel matrix using singular value decomposition (SVD), which allows us not only to systematically improve the condition number of the kernel matrix, but also determines the computational effort and accuracy of our method. In order to allow for the treatment of realistic inputs that may contain arbitrage, we reformulate the system of linear equations as an optimization problem, in which the SVD-transformed density minimizes the error between the input prices and the arbitrage-free prices generated by our method. To further stabilize the method in the presence of noisy input prices or arbitrage, we apply an L_1-regularization to the SVD-transformed density. Our approach, which is inspired by recent progress in theoretical physics, offers a flexible and efficient framework for the arbitrage-free interpolation of plain-vanilla option prices and implied volatilities, without the need to explicitly specify a stochastic process, expansion basis functions or any other kind of model. We demonstrate the capabilities of our method in a number of artificial and realistic test cases.

Keywords: option pricing; plain-vanilla options; volatility interpolation; arbitrage; inverse problem; optimization; regularization; sparse modeling

Citation: Guterding, Daniel. 2023. Sparse Modeling Approach to the Arbitrage-Free Interpolation of Plain-Vanilla Option Prices and Implied Volatilities. *Risks* 11: 83. https://doi.org/10.3390/risks11050083

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 24 March 2023
Revised: 21 April 2023
Accepted: 25 April 2023
Published: 28 April 2023

1. Introduction

Since the dawn of modern quantitative finance, academics and practitioners have tried to understand the dynamics of asset price fluctuations, which, in financial markets, are called volatility. Understanding volatility is not only necessary for the risk management of financial products in general, but also, specifically, for the pricing of option contracts, since their value depends on the size of expected price fluctuations for the respective underlying assets. These studies have produced various books and reviews (Carr and Lee 2009; Clark 2010; Derman and Miller 2016; Gatheral 2006), as well as many focused studies for which we can cite only a few examples here: (Aït-Sahalia et al. 2001; Baker et al. 2004; Egloff et al. 2010; Hagan et al. 2001; Jiang 2020; Lipton 2002; Mixon 2002; Xing et al. 2010). Over time, many option pricing models have emerged, which assume specific dynamics for the underlying asset and its volatility. Among them are the seminal Black and Scholes (1973) model and the, similarly influential, Heston (1993) model, which specify an explicit stochastic process for the price of the underlying asset, and also for the volatility, in the case of the Heston model.

Most option pricing models require knowledge of various parameters which reflect the state of the market, such as interest rate, dividends or implied volatilities. If, for example, the ubiquitous Black–Scholes model (Black and Scholes 1973) is applied to realistic settings, it requires an implied volatility for each strike for which a price is intended. This implied

volatility characterizes the size of price fluctuations for the option's underlying asset, which is *implied* by the price of the option, i.e., only when this value of volatility is inserted into the pricing formula does one obtain the same price as is observed on the market.

Plain-vanilla European options for many underlyings are traded on exchanges, but for a fixed discrete set of strikes. To price a plain-vanilla option with a strike not contained in this set of market-quoted instruments, it is necessary to calculate the volatility for this missing strike out of the available option prices. Often, this is done not only by implying the volatility for single strikes, but by constructing a continuous representation of the implied volatility out of the discrete set of quoted option prices. Other use cases for such a continuous representation of implied volatility include the construction of a local volatility (LV) (Derman and Kani 1994; Dupire 1994) for use in the popular class of local stochastic volatility (LSV) models (Lipton 2002; Lipton et al. 2014), or the pricing of exotic derivatives. (Carr and Lee 2009).

There are many methods for constructing a continuous representation of option prices or implied volatilities, starting from simple spline interpolation, directly applied to the implied volatility, and moving to sophisticated arbitrage-free schemes in either volatility or option prices. Kahale (2004) uses a C^2 interpolating function to produce an arbitrage-free interpolation of option prices, but requires that the inputs also be arbitrage-free. Andreasen and Huge (2011) developed a one-step finite difference scheme to calibrate a piecewise constant local volatility model to quoted option prices. Jäckel (2014) formulates another method to construct C^2 interpolants for option prices and presents an algorithm to remove arbitrage from these interpolants. Another approach, developed by Le Floc'h and Osterlee (2019a, 2019b) directly relates option prices and terminal density using stochastic collocation to various basis functions. Such a relation exists, since the fair price of any financial instrument is the present value of its expected payoff. This expected payoff can be calculated from the payoff function of the instrument and the risk-neutral probability distribution. For plain-vanilla options, which are investigated herein, the payoff is only determined at maturity, so that the terminal density is sufficient to calculate the expected payoff (Breeden and Litzenberger 1978).

A totally different route to a continuous representation of option prices or implied volatility are stochastic volatility models, such as the Heston (1993), or SABR (Hagan et al. 2015) models and related approximate formulae for the volatility smile (Gatheral and Jacquier 2014; Lorig et al. 2017). These models are usually based on, or inspired by, a stochastic process with few parameters. Therefore, the forms of the volatility smiles in these models are somewhat inflexible, and it can be hard to fit them to market-quoted prices. Nevertheless, these models describe the whole dynamics of an underlying asset, instead of just the terminal density at a given point in time, which enables them to also price path-dependent options (Carr et al. 2022; Guterding and Boenkost 2018; Tian et al. 2014) and other exotic derivatives. (Guterding 2021; Zhu and Lian 2012).

Our aim here was to find a method for which it is not necessary to specify a stochastic process and which should also not require specifying a specific functional form for either the volatility or the terminal density. Such a method for the interpolation of implied volatilities would be versatile and applicable to a broad range of realistic situations. Inspired by recent progress on inverse problems in theoretical physics (Otsuki et al. 2020), we present a new method that uses the singular value decomposition (SVD) and L_1-regularization to obtain a sparse model, i.e., one containing only a few non-zero parameters of the relation between plain-vanilla European option prices and the terminal density of the underlying asset. Using a constrained and L_1-regularized optimization, our method is able to extract arbitrage-free representations of both option prices and implied volatilities, even from inputs that contain severe arbitrage, without any de-arbitraging or pre-processing steps. Since we are directly working with the density, our method is also able to extrapolate implied volatilites in an arbitrage-free manner beyond the quoted strike range.

We first review the relation between option prices and terminal density and describe how it can be discretized and written in matrix form. We show why naive attempts to

invert this system of equations fail, in general, and propose a solution that involves transforming the problem into a better-conditioned one using the SVD. Then, we reformulate the procedure for finding the terminal density from a matrix inversion into a constrained optimization problem that also avoids arbitrage. We test our method on several classic examples, such as normal and log-normal densities. Furthermore, we show that our method can easily handle multimodal densities, arbitrage and volatility smiles with kinks, all of these being areas in which stochastic volatility models struggle (Le Floc'h and Osterlee 2019a). We conclude with a summary of the advantages and disadvantages of our method.

2. Methodology

2.1. Relation between Terminal Density and Option Price

Suppose we know the probability distribution or density $\phi(x)$ for the price of an asset at time T. Then, we may calculate the price of a European option on this underlying asset at time t, with strike K and expiry at time T, from the following integral: (Breeden and Litzenberger 1978)

$$\Pr(K) = e^{-r\tau} \int_{-\infty}^{\infty} dx\, \psi(K,x)\phi(x)\,. \tag{1}$$

The time to expiry is given by $\tau = T - t$ and r is a risk-free interest rate. The kernel $\psi(K,x)$ depends on the type of option we want to price. For a European Call option we use the following kernel:

$$\psi_C(K,x) = \max(0, x-K)\,. \tag{2}$$

For a European Put option we use another kernel:

$$\psi_P(K,x) = \max(0, K-x)\,. \tag{3}$$

In cases when the price of the underlying asset is, for example, non-negative, such as a stock, or obeys some other restriction, this may be taken into account not only by picking a suitable probability distribution $\phi(x)$, but also by adjusting the integral boundaries.

For numerical calculations, it is useful to discretize $\phi(x)$. If we pick a relevant interval $[x_{\min} : x_{\max}]$, divide it into L, not necessarily uniform, sub-intervals and apply the trapezoidal rule, we obtain a discrete representation of Equation (1). Using the abbreviation $f(x) = f_{K,r,\tau}(x) = e^{-r\tau}\psi(K,x)\phi(x)$, it reads:

$$\Pr(K) \approx \frac{1}{2}\sum_{i=1}^{L}\Big(f(x_i) + f(x_{i-1})\Big)\Big(x_i - x_{i-1}\Big)\,. \tag{4}$$

If we pick a uniform discretization of the interval $[x_{\min} : x_{\max}]$, we obtain the following simplified approximation:

$$\Pr(K) \approx \Big(\frac{1}{2}\big[f(x_0) + f(x_L)\big] + \sum_{i=1}^{L-1} f(x_i)\Big)\Delta x\,. \tag{5}$$

As we refine the interval into smaller sub-intervals, the calculated price converges to the true price.

2.2. Matrix Representation of the Relation between Option Price and Terminal Density

From Equation (1), it is clear that the same density, $\phi(x)$, should be used for Call and Put options with the same underlying and expiry at T. Different strikes are taken into account through the kernel $\psi(K,x)$. In practice, European Call and Put options are often traded on exchanges, such that quoted prices for a set of discrete strikes are publicly available.

Based on these prices, we would like to find an approximation to the density $\phi(x)$. We write down a matrix equation that relates all M available Call and Put prices to the uniformly discretized density with N points (see Equation (5)):

$$\mathrm{Pr} = \begin{pmatrix} \mathrm{Pr}_1 \\ \vdots \\ \mathrm{Pr}_M \end{pmatrix} = \begin{pmatrix} \frac{1}{2}g_1(x_1) & g_1(x_2) & \dots & g_1(x_{N-1}) & \frac{1}{2}g_1(x_N) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2}g_M(x_1) & g_M(x_2) & \dots & g_M(x_{N-1}) & \frac{1}{2}g_M(x_N) \end{pmatrix} \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_N) \end{pmatrix} = G\phi\,. \tag{6}$$

Here, we absorbed Δx into the function $g_i(x)$. If the option with index i is a Call option with strike K_i, we use:

$$g_i(x) = \Delta x \cdot e^{-r\tau}\psi_C(K_i, x) = \Delta x \cdot e^{-r\tau}\max(0, x - K_i)\,. \tag{7}$$

If the option with index i is a Put option, we use:

$$g_i(x) = \Delta x \cdot e^{-r\tau}\psi_P(K_i, x) = \Delta x \cdot e^{-r\tau}\max(0, K_i - x)\,. \tag{8}$$

2.3. The Difficulty in Implying the Terminal Density from Option Prices

In general, we are interested in a finely resolved density $\phi(x)$ with N discrete points in the interval $[x_{\min} : x_{\max}]$, while only a limited number of option prices M is available. Hence, G in Equation (6) is, in general, not a square, but an $(M \times N)$ matrix, where $M \leq N$ and often even $M \ll N$. Therefore, G is ill-conditioned and cannot, in general, be inverted to find the vector of $\phi(x_i)$ on the right-hand side.

A way to circumvent this difficulty is provided by the singular value decomposition (SVD) of a matrix. The singular value decomposition of G reads:

$$G = USV^T\,. \tag{9}$$

Here, T denotes the transpose of a matrix. Matrices U and V are orthogonal matrices of sizes $(M \times M)$ and $(N \times N)$. S is a matrix of size $(M \times N)$, which contains, on its diagonal, the singular values s_i, where $i = 1, \dots, \min(M, N)$. The singular values are non-negative real numbers in descending order.

This means we can write Equation (6) as:

$$\mathrm{Pr} = G\phi = USV^T\phi\,. \tag{10}$$

Since U and V are orthogonal matrices ($U^T = U^{-1}$), we can multiply from the left by U^T and obtain:

$$U^T\mathrm{Pr} = U^TUSV^T\phi = U^{-1}USV^T\phi = SV^T\phi\,. \tag{11}$$

We now introduce the abbreviations $\mathrm{Pr}' = U^T\mathrm{Pr}$ and $\phi' = V^T\phi$, which we call transformed quantities from now on. Prices Pr' are the SVD-transformed prices, while ϕ' is the SVD-transformed density. With these transformed quantities we can write Equation (11) in the following form:

$$\mathrm{Pr}' = U^T\mathrm{Pr} = SV^T\phi = S\phi'\,. \tag{12}$$

Since the matrix of singular values S in Equation (12) is diagonal, we conclude that an element-wise equation also holds:

$$\mathrm{Pr}'_i = S_{ii}\phi'_i = s_i\phi'_i\,. \tag{13}$$

This shows that the transformation from ϕ to Pr via G can be decomposed into three steps:

1. application of a basis transformation from ϕ to ϕ' via $\phi' = V^T\phi$
2. weighting the elements of ϕ' with the singular values S to get Pr' via $\mathrm{Pr}' = S\phi'$
3. application of a basis transformation from Pr' to Pr via $\mathrm{Pr} = U\mathrm{Pr}'$

Whether such a transformation is easily invertible, is characterized by the decay of singular values and, in particular, by the condition number $C = s_{\max}/s_{\min}$. While the best-conditioned system of equations has $C = 1$, we are dealing here with the kernel matrix G (see Equation (6)), for which $C \gg 1$.

To show this, we analyze the kernel G for an equidistant discretization of the interval $[x_{\min} : x_{\max}]$ with $N = 10{,}000$ points and a variable number of strikes in the same interval. For simplicity, we only take into account Call options. The condition number as a function of the number of option strikes M in the problem is shown in Figure 1.

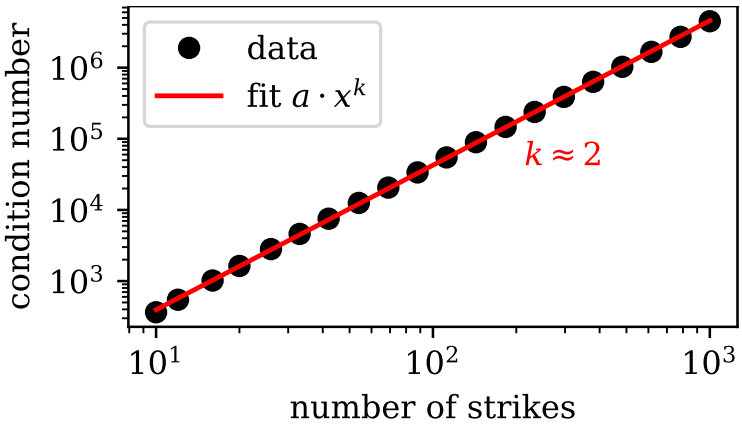


Figure 1. Log–log plot of the condition number of the kernel matrix G defined in Equation (6) as a function of the number of strikes. The fit with $f(x) = a \cdot x^k$ clearly shows that the growth of the condition number follows such a power law with exponent $k \approx 2$.

We fit the condition number of the matrix G with a power law of the form $f(x) = a \cdot x^k$, where we take x to be the number of option strikes. The fit clearly reveals that $k \approx 2$, which means that the condition number increases quadratically with the number of options taken into account. Importantly, even if only as few as ten strikes are considered, we already have $C \gg 1$, i.e., the system is very ill-conditioned and any naive attempt at solving Equation (6) for ϕ will not succeed.

2.4. Rapid Decay of the Kernel Matrix Singular Values

The fact that G is, in general, ill-conditioned leads to problems in treating Equation (6) numerically. In the previous section, we discussed how the condition number increases rapidly as we consider a larger number of options. This indicates that the additional singular values associated with additional market quotes, i.e., additional linear equations, decay rapidly.

Here, we show that the singular values also decay rapidly for a fixed matrix G, i.e., for a fixed number of considered options M and for a fixed number of discretization points N within $[x_{\min} : x_{\max}]$. We chose $M = 25$ equidistant strikes in a fixed interval and varied the number of discretization points N. As before, we only took into account Call options. The normalized singular values s_i/s_1 are shown in Figure 2. We attempted to fit the decay of singular values with an inverse power law of the form $f(x) = x^{-k}$. The initial decay seemed to follow such a law with roughly $k \approx 2.7$ and slowed down a little for the tail of singular values.

Recall from the previous section (see Equation (13)) that the singular values s_i play the role of weights for the transformed density ϕ' to obtain the transformed prices Pr'. From Figure 2, it is obvious that these weights may differ by several orders of magnitude.

In any direct inversion method this would cause severe numerical problems, essentially because we would calculate $\phi'_i = \mathrm{Pr}'_i/s_i$, wherein we would have to divide by the quickly decaying singular values s_i.

However, the role of s_i as weights also shows that most of the relevant information must be contained in the first few basis vectors associated with the largest singular values.

Therefore, we may consider only a limited number of these singular values and basis vectors to reconstitute a better-conditioned approximate version of G.

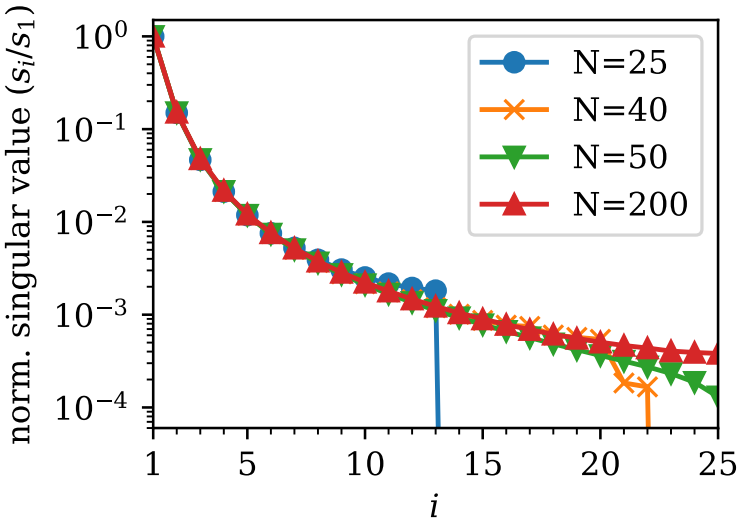


Figure 2. Logarithmic plot of the normalized singular values s_i/s_1 for various numbers of discretization points N. The number of strikes is fixed to $M = 25$. The normalized singular values decay with an inverse power law of the form $f(x) = x^{-k}$ with $k \approx 2.7$.

Let us fix the number of considered singular values to Q, where $1 \leq Q \leq \min(M, N)$. We take the Q largest singular values and form the new diagonal matrix $\tilde{S}$. We also reduce the dimensionality of U with $(M \times M)$ to $\tilde{U}$ with $(M \times Q)$ and of V^T with $(N \times N)$ to $\tilde{V}^T$ with $(Q \times N)$, by keeping only the first Q columns for $\tilde{U}$ or rows for $\tilde{V}^T$, respectively. Thus, the new kernel matrix reads:

$$\tilde{G} = \tilde{U}\tilde{S}\tilde{V}^T\,. \tag{14}$$

The matrix $\tilde{G}$ is still of size $(M \times N)$, but is better conditioned than the initial matrix G, which it aims to approximate. We achieved this by cutting off the tail of singular values s_j with $Q < j \leq \min(M, N)$, for which we know that $s_j \leq s_Q$. Thus, the condition number of $\tilde{G}$ is $\tilde{C} = s_1/s_Q$, for which we know, in general, that $\tilde{C} \leq C$. However, since we have already shown that singular values for our kernel matrix G decay with a power law (see Figure 2), we can safely assume that $\tilde{C} \ll C$ if the chosen Q is sufficiently small, i.e., the tail of small singular values is discarded.

This means we can systematically reduce the condition number of our kernel to obtain a better-conditioned kernel matrix $\tilde{G}$ by retaining only the largest singular values and associated orthogonal vectors, which is accomplished by lowering Q.

The prices Pr and the density ϕ are now related by the new kernel matrix $\tilde{G}$, which gives us a new relation similar to Equation (10):

$$\mathrm{Pr} \approx \tilde{G}\phi = \tilde{U}\tilde{S}\tilde{V}^T\phi\,. \tag{15}$$

How good the approximation of G by $\tilde{G}$ is, obviously depends on how many singular values we retain, i.e., how we pick Q.

We also note that Pr and ϕ have M and N entries, respectively, while the transformed quantities $\mathrm{Pr}' \approx \tilde{U}^T\mathrm{Pr}$ and $\phi' \approx \tilde{V}^T\phi$ both only have Q entries. Since we are often interested in cases where $N \gg M \geq Q$, this can make a large difference computationally.

In this sense, the SVD allows us to describe the relation between density and prices with only a few parameters and related orthogonal vectors. Therefore, we may say that the SVD gives us a sparse model of the original relation defined by Equation (6).

This, however, does not mean that our method is only useful for sparse *inputs*. Rather, our method generates a sparse representation of the relation between *any* number of input prices and *any* number of density discretization points.

2.5. Optimization Problem for Finding the Density

So far, we have discussed how to treat the kernel matrix G of Equation (6) that relates prices Pr and densities ϕ. Remember that, in practice, the prices Pr are known, and, thus, we are interested in the density ϕ. This means we now attempt to solve Equation (6) approximately, by actually solving Equation (15).

This could work in cases in which the input prices are reachable with a non-negative density, i.e., they contain no arbitrage. In many other methods this is solved by filtering the input prices or applying some other form of de-arbitraging. (Jäckel 2014; Kahale 2004; Le Floc'h and Osterlee 2019a).

However, we can resolve the need for de-arbitraging by reformulating Equation (15) in the form of a constrained optimization problem. This optimization problem should give us the non-negative density ϕ with N entries. In our sparse model, ϕ is, however, directly related to $\phi' \approx \tilde{V}^T\phi$, which has only Q entries. Therefore, the most efficient way to find ϕ is to actually find ϕ' via optimization.

To this end, we minimize the squared error in the transformed quantities:

$$\chi^2(\phi'|\mathrm{Pr}') = \frac{1}{2}\|\mathrm{Pr}' - \tilde{S}\phi'\|_2^2 \,. \tag{16}$$

To further enhance the sparsity in the transformed domain, we add another term for L_1-regularization with a free parameter λ:

$$F(\phi'|\mathrm{Pr}', \lambda) = \frac{1}{2}\|\mathrm{Pr}' - \tilde{S}\phi'\|_2^2 + \lambda\|\phi'\|_1 \,. \tag{17}$$

The same optimization may also be carried out based on the deviation from the non-transformed prices Pr:

$$F(\phi'|\mathrm{Pr}, \lambda) = \frac{1}{2}\|\mathrm{Pr} - \tilde{U}\tilde{S}\phi'\|_2^2 + \lambda\|\phi'\|_1 \,. \tag{18}$$

We still cannot guarantee that the density ϕ is non-negative, as it should be. Therefore, we constrain the optimization with the following additional conditions:

$$\phi_i = \left(\tilde{V}\phi'\right)_i \geq 0 \quad \forall\, i \,. \tag{19}$$

Furthermore, the integral of the density, expressed using the trapezoidal rule as before, should be equal to one:

$$\begin{aligned} 1 =& \Big(\frac{1}{2}(\phi_1 + \phi_N) + \sum_{i=2}^{N-1}\phi_i\Big)\Delta x \\ =& \Big(\frac{1}{2}\Big[(\tilde{V}\phi')_1 + (\tilde{V}\phi')_N\Big] + \sum_{i=2}^{N-1}(\tilde{V}\phi')_i\Big)\Delta x \,. \end{aligned} \tag{20}$$

Our goal is now to find the transformed density ϕ' that minimizes Equation (18) under the constraints defined by Equations (19) and (20). The true density ϕ can then be calculated from the solution ϕ' of the optimization problem via $\phi = \tilde{V}\phi'$.

In Appendix A we discuss further details of how to calculate the error in prices and when it may be reasonable to convert between Call and Put prices.

In cases in which not only the mid-price is known, but also the Bid–Ask spread for each option is known, one could reformulate the optimization problem in Equation (18) to penalize calculated prices that lie outside the spread. We leave this problem for further studies.

2.6. Finding a Solution to the Optimization Problem

We implemented the system of equations defined by Equation (18) under the constraints defined by Equations (19) and (20) using the domain-specific language `CVXPY` (Agrawal et al. 2018; Diamond and Boyd 2016), available as a package for the Python programming language.

The transformed density ϕ' that minimizes the squared error with additional L_1-regularization on ϕ' (see Equation (18)) can be found using various optimization algorithms. While some authors recommend using their own implementation of the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2010), we have found that the open-source solvers `ECOS` (Domahidi et al. 2013) and `SCS` (O'Donoghue et al. 2016) both deliver excellent performances, especially with systems, such as the system we attempt to solve, which usually have only a few degrees of freedom (remember that ϕ' has only Q entries). Therefore, we refer readers who are interested in details of the implementation of these solvers to the respective papers. In our case, `ECOS` seemed to be a good choice for the numerical solver.

For further considerations on how to efficiently solve the optimization problem, please see Appendix B. For readers interested in the computer code of our reference implementation, we provide a full working example on Github.[1]

2.7. A Measure for the Similarity of Probability Distributions

For test cases with known probability distributions we wanted to quantify the degree to which our method recovered the known density. A suitable measure for the similarity of probability distributions is the Bhattacharyya (1943) distance d_B, which, for two probability distributions, $p(x)$ and $q(x)$, is defined as:

$$d_B(p,q) = -\ln\left[\int_{-\infty}^{\infty} dx\,\sqrt{p(x)q(x)}\right]. \tag{21}$$

If the probability distributions $p(x)$ and $q(x)$ are identical, i.e., the overlap is maximal, the integral under the logarithm is equal to one and the Bhattacharyya distance is zero. For all other cases, the overlap calculated from the integral is between zero and one, or exactly zero when there is no overlap. The Bhattacharyya distance is a non-negative number which approaches $+\infty$ in cases where there are no overlaps.

In practice, we sample the probability distributions that are exactly known on the same grid for which we know our implied density $\phi(x)$ and calculate the integral in Equation (21), using the trapezoidal rule.

3. Examples

In this section, we calculate option prices for a number of known densities and show that our method is able to accurately recover the terminal density only using the supplied prices. Furthermore, we demonstrate that our method is able to reconstruct densities from realistic input prices without any de-arbitraging, filtering or other pre-processing steps.

Since our implied density is non-negative, we can safely apply linear interpolation to the density and calculate option prices with strikes between the available input prices. This enables us to also interpolate implied volatilities by calculating these from the interpolated option prices.

3.1. Normal Density

A normal density corresponds to the Bachelier model, which, having mostly been ignored for a long time, has regained attention in the context of negative interest rates and negative prices for oil-futures. (Bachelier 1901; Choi et al. 2022).

Using the forward price F of the underlying asset, the strike of the option K, the normal volatility σ and the time to expiry τ, we define the moneyness m of an option in the following way:

$$m = m(F, K, \sigma, \tau) = \frac{F - K}{\sigma\sqrt{\tau}}. \tag{22}$$

Denoting the normal density function as ϕ_N and the cumulative normal density function as Φ_N, the pricing formula of the Bachelier model can be expressed as follows for a Call option:

$$\mathrm{Pr_C} = e^{-r\tau}\left[(F - K)\Phi_N(m) + \sigma\sqrt{\tau}\phi_N(m)\right]. \tag{23}$$

For the Put option it reads:

$$\mathrm{Pr_P} = e^{-r\tau}\left[(K - F)\Phi_N(-m) + \sigma\sqrt{\tau}\phi_N(m)\right]. \tag{24}$$

We fixed the initial underlying price to $S_0 = 0.1$, the normal volatility to $\sigma = 0.1$, the interest rate to $r = 0.05$ and the time to expiry to $\tau = 1$. We then calculated the prices of 200 Call and Put options for a uniformly discretized grid of strikes between $K_{\min} = -0.7$ and $K_{\max} = 0.7$.

We used our method with the option prices so-calculated to imply the density $\phi(x)$ on a uniformly discretized grid with $x_{\min} = -0.9$, $x_{\max} = 0.9$ and $N = 1000$. We retained $Q = 150$ singular values.

The normal distribution, upon which the Bachelier model is based, was recovered to a high degree of accuracy (see Figure 3). Further evidence is shown in Figure 4, which depicts the error in prices χ^2 calculated from Equation (16) and the Bhattacharyya distance of $\phi(x)$ with respect to the normal distribution calculated via Equation (21).

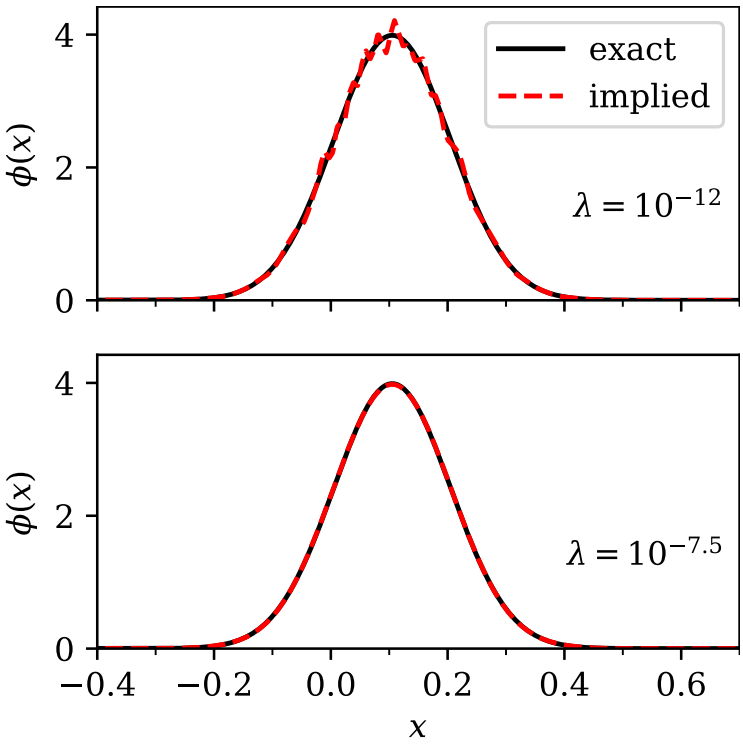


Figure 3. Comparison of the exact (bold line) and implied (dashed line) densities $\phi(x)$ for two different values of the regularization parameter λ. The exact density was normal with $\sigma = 0.1$ and shifted to $\mu = S_0 \exp(r\tau) \approx 0.105$. The top panel shows an under-regularized implied density ($\lambda = 10^{-12}$), while the bottom panel shows a close to optimal implied density with $\lambda = 10^{-7.5}$.

For $\lambda < 10^{-8}$ the error in prices was almost independent of λ. That meant that many different solutions to the optimization problem existed, which produced a basically perfect fit to the prices. This was possible since we retained a large number of singular values. In the density $\phi(x)$ this manifested in the form of tiny oscillations, as can be seen in Figure 3. In other words, without regularization the input prices did not contain enough information for the optimization to yield a smooth density at the high resolution we selected for $\phi(x)$. As expected, all densities, irrespective of the value of λ, were non-negative.

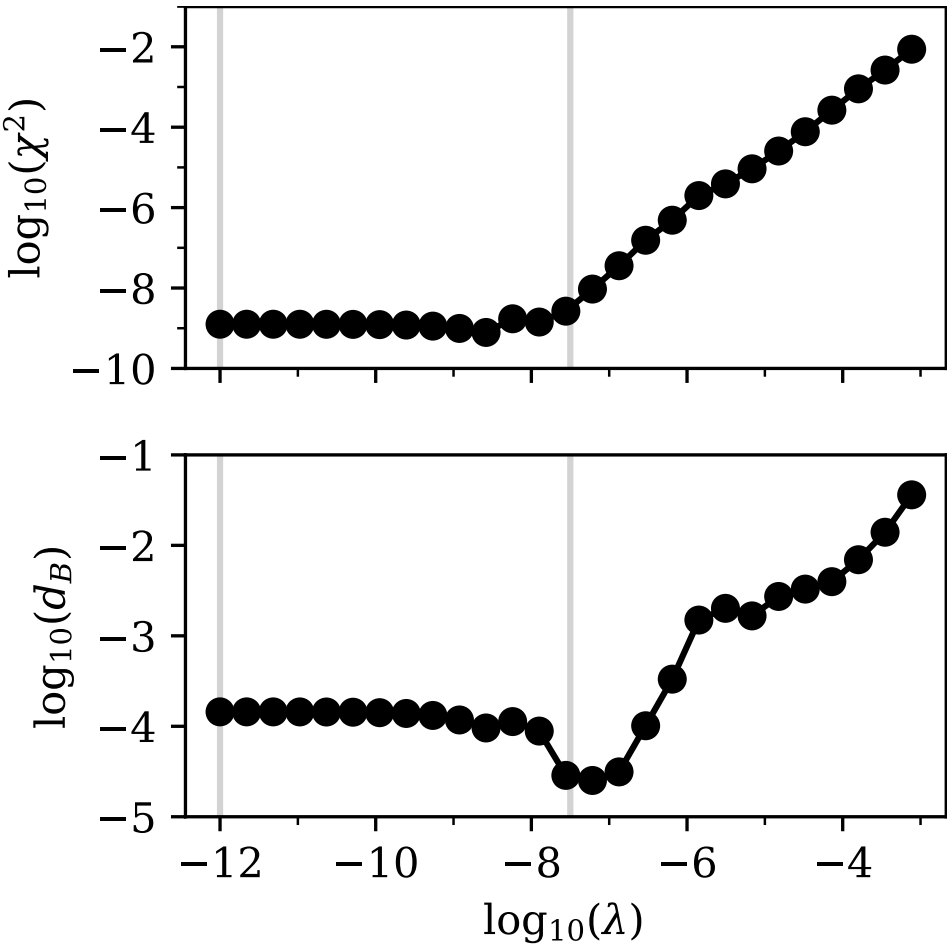


Figure 4. Log–log plot of the squared error in prices χ^2 (top panel) and the Bhattacharyya distance d_B (bottom panel). Both measures were calculated based on a comparison between the exact input data and our implied output data for prices and densities, respectively. The input option prices were based on a Bachelier model. The related input density was a normal distribution with $\sigma = 0.2$ and shifted to $\mu = S_0 \exp(r\tau) \approx 0.105$. The vertical lines mark the positions of $\lambda = 10^{-12}$ and $\lambda = 10^{-7.5}$, for which we show the implied densities in Figure 3.

For $\lambda \approx 10^{-7.5}$, the error in prices was very slightly higher, but the Bhattacharyya distance d_B, with respect to the known normal distribution, was actually minimal, since the oscillations were suppressed.

For larger values of λ, the error in prices and Bhattacharyya distance increased, since the implied density was a broadened version of the original normal distribution.

The effect of regularization on the optimized parameters ϕ' could also be visualized by counting the number of entries in ϕ', for which the absolute value was above a threshold a. The visualization for a few different threshold values is shown in Figure 5. Recall that the number of entries in ϕ' was $Q = 150$. Figure 5 shows that for $\lambda < 10^{-7.5}$ the problem was basically unregularized, and almost all parameters had a magnitude larger than $a = 10^{-2}$. As soon as the regularization became effective, the number of parameters with significant

magnitude drastically reduced, starting with those that had the least impact on the quality of the prices.

The point where the number of parameters sharply decreased was directly related to the minimum in the Bhattacharyya distance that we observed in Figure 4. For larger values of the regularization parameter, the fit simply became worse, since relevant components of ϕ' were strongly suppressed. The fit tried to compensate for this by increasing the number of other non-zero components of ϕ', but failed to achieve good accuracy.

The Bhattacharyya distance d_B can of course only be used to select an optimal solution when the original density is known. As we see in further examples, the minimum in the Bhattacharyya distance usually corresponds to the point where the error in prices starts to increase after showing a plateau for small values of the regularization parameter λ.

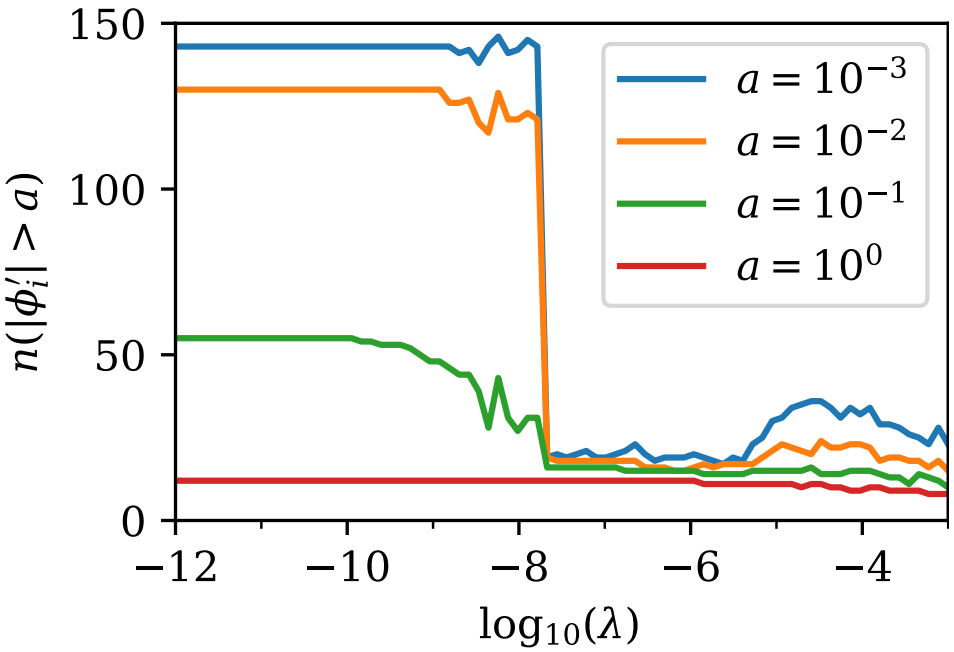


Figure 5. Visualization of the effect of regularization on the number of relevant parameters for the implied transformed density ϕ'. A larger regularization parameter λ led to fewer entries with significant magnitude in ϕ', i.e., regularization turned ϕ' into a sparse representation of the true density ϕ. The number of entries in ϕ' with a magnitude above the positive threshold a is denoted as $n(|\phi'_i| > a)$ and shown as a function of the regularization parameter λ. For λ we chose to show the axis in logarithmic scale. The figure is based on the same Bachelier model as that in Figures 3 and 4. The effect of regularization is clearly visible around $\lambda \approx 10^{-7.5}$, where there was a sharp decrease in the number of parameters with significant magnitude.

Hence, for practical use cases, where the density is not known, we suggest calculating the solutions for multiple values of λ, as we did here, and selecting the solution with the highest value of λ that still shows close to minimal error in prices χ^2. The effect of regularization on the parameters can also be verified by an analysis, similar to what we presented in Figure 5.

3.2. Log-Normal Density

A log-normal density corresponds to the classic Black–Scholes formula for option pricing. (Black and Scholes 1973) We fixed the initial underlying price to $S_0 = 0.5$, the volatility to $\sigma = 0.2$, the interest rate to $r = 0$ and the time to expiry to $\tau = 1$. We then calculated the prices of 200 Call and Put options for a uniformly discretized grid of strikes between $K_{\min} = 0.01$ and $K_{\max} = 1.0$.

We used the so-calculated option prices to imply the density $\phi(x)$ on a uniformly discretized grid with $x_{\min} = 0$, $x_{\max} = 1.5$ and $N = 1000$ using our method. We retained $Q = 150$ singular values.

In summary, we recovered the log-normal distribution of the Black–Scholes model to a high degree of accuracy (see Figure 6). Further evidence is shown in Figure 7, which depicts the error in prices χ^2 calculated from Equation (16) and the Bhattacharyya distance of $\phi(x)$ with respect to the log-normal distribution calculated via Equation (21).

It is clear that the results were optimal for $\lambda \approx 10^{-7.5}$, where the price error and the Bhattacharyya distance were simultaneously minimized. For smaller values of the regularization parameter λ we observed convergence issues in the `ECOS` solver, probably because we retained a large number of singular values Q, which allowed for many similarly good solutions. This could probably be resolved by either fine-tuning of numerical parameters in `ECOS` or by reducing the number of singular values.

Figure 6 compares the exact log-normal distribution and our implied discretization $\phi(x)$ for two values of the regularization parameter λ. The optimal solution ($\lambda = 10^{-7.5}$) closely followed the log-normal distribution. Even though the less optimal solution had a slightly broader shape, it still looked similar to a log-normal distribution. Finally, we verified that all densities we implied from option prices were non-negative.

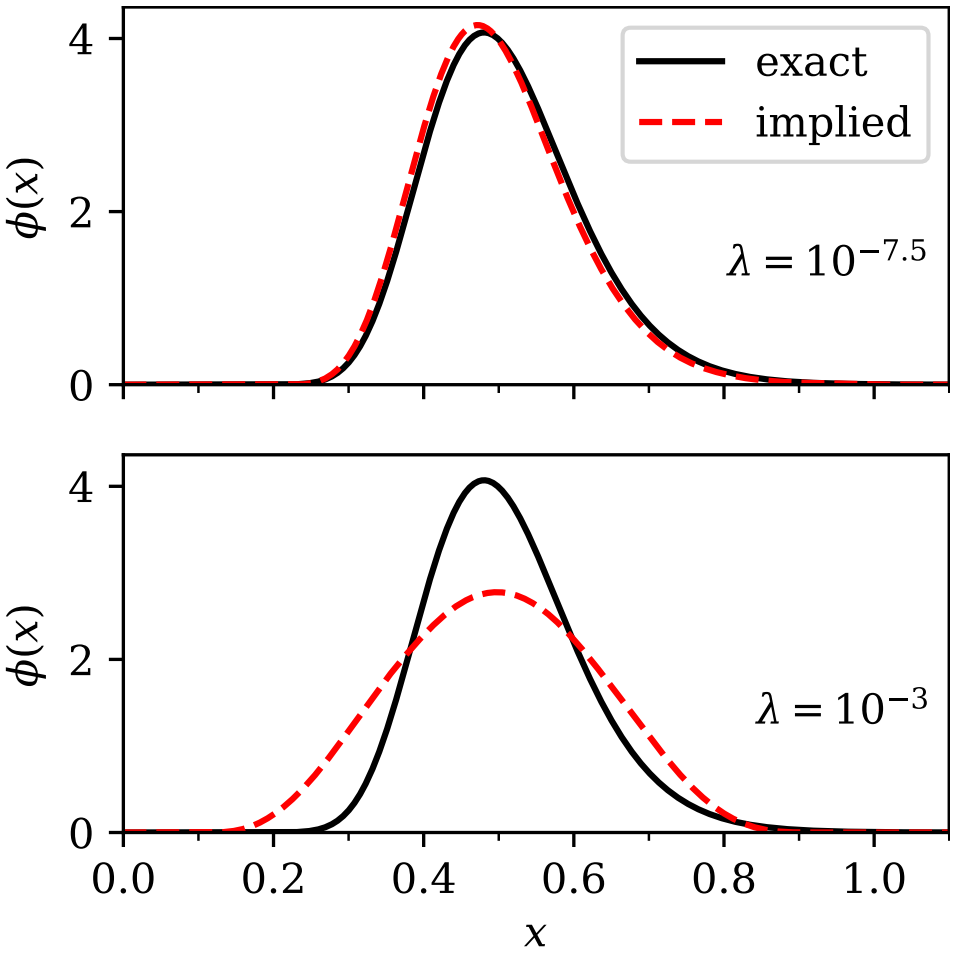


Figure 6. Comparison of the exact (bold line) and implied (dashed line) density $\phi(x)$ for two different values of the regularization parameter λ. The exact density was log-normal with $\sigma = 0.2$ and shifted to $\mu = S_0 = 0.5$. The top panel shows a close to optimal regularized implied density ($\lambda = 10^{-7.5}$), while the bottom panel shows an over-regularized implied density with $\lambda = 10^{-3}$.

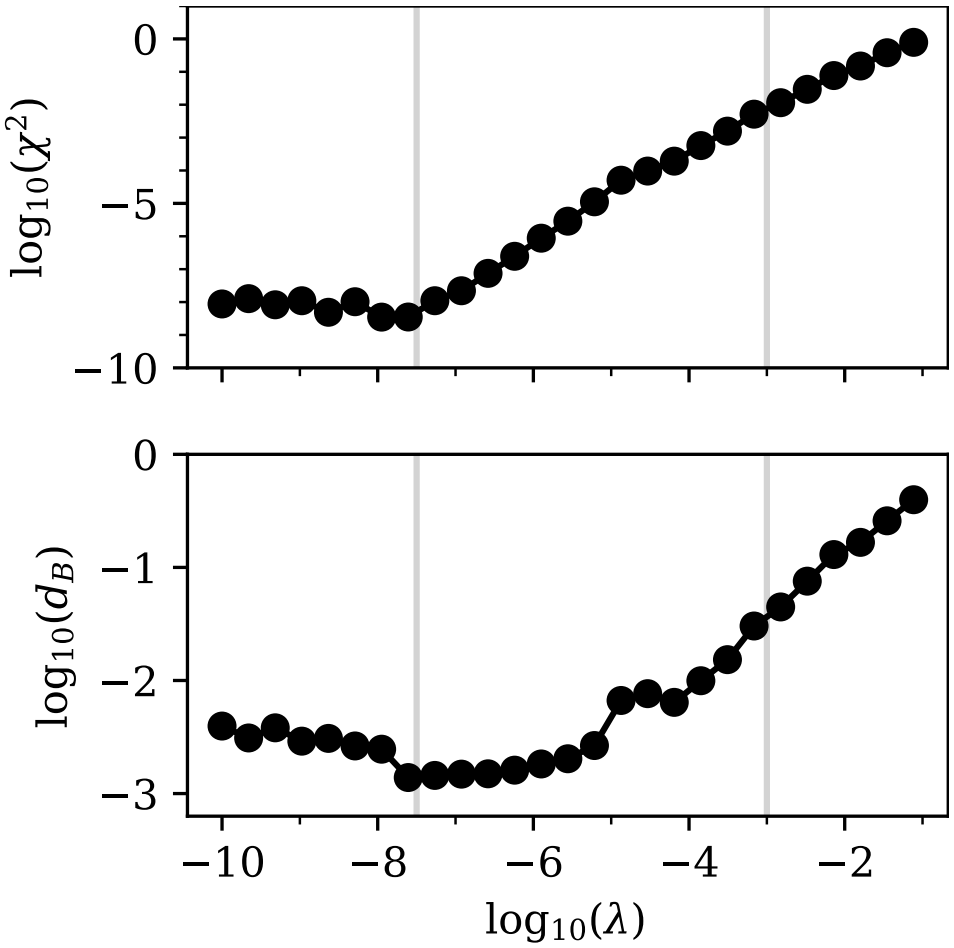


Figure 7. Log–log plot of the squared error in prices χ^2 (top panel) and the Bhattacharyya distance d_B (bottom panel). Both measures were calculated based on a comparison between exact input data and our implied output data for prices and densities, respectively. The input option prices were based on the Black–Scholes model. The related input density was a log-normal distribution with $\sigma = 0.2$ and shifted to $\mu = S_0 = 0.5$. The vertical lines mark the positions of $\lambda = 10^{-7.5}$ and $\lambda = 10^{-3}$, for which we show the implied densities in Figure 6.

3.3. Multimodal Density

Since it has been reported that several known methods struggle with multimodal densities (Le Floc'h and Osterlee 2019a), we wanted to verify that our method also performs well in such cases. For simplicity, we considered a linear combination of normal distributions $\phi_N(\mu, \sigma)$. Since the price could be calculated from an integral over the density (see Equation (1)), and since every integral was linear, we concluded that the price for an option on an underlying that was distributed according to a linear combination of normal distributions, could be calculated as the equivalent linear combination of Bachelier option prices (see Equations (23) and (24)).

We next considered a probability distribution ϕ_M, built from the linear combination of three normal distributions ϕ_N:

$$\phi_M = \sum_{i=1}^{3} c_i \phi_N(\mu_i, \sigma_i)\,. \tag{25}$$

For the parameters, we used the values given in Table 1. If we want ϕ_M to be a probability distribution, we must obviously require that $\sum_i c_i = 1$ and that all c_i are non-negative. Note that setting the mean value μ_i for each normal distribution implies the use of different values for the forward price $F = \mu_i$ in the Bachelier formulae for each term.

Table 1. Parameters for a multimodal density. These parameters are used in Equation (25) to generate a density which is a superposition of multiple normally distributed components.

i	c_i	μ_i	σ_i
1	0.50	−0.20	0.10
2	0.45	0.15	0.15
3	0.05	0.55	0.05

Again, we fixed the interest rate to $r = 0.05$ and the time to expiry to $\tau = 1$. We then calculated the prices of 200 Call and Put options for a uniformly discretized grid of strikes between $K_{\min} = -0.7$ and $K_{\max} = 0.7$.

We used the so-calculated option prices to imply the density $\phi(x)$ on a uniformly discretized grid with $x_{\min} = -0.9$, $x_{\max} = 0.9$ and $N = 1000$ using our method. We retained $Q = 150$ singular values.

In summary, multimodal distributions seem to pose no problem for our method. The original density is recovered with good accuracy (see Figure 8). Quantitative error estimates are shown in Figure 9, which depicts the error in prices χ^2 calculated from Equation (16) and the Bhattacharyya distance of $\phi(x)$ with respect to the linear combination of normal distributions calculated via Equation (21). Again, we observed that the minimum in the Bhattacharyya distance d_B corresponded to the minimum of the price error χ^2.

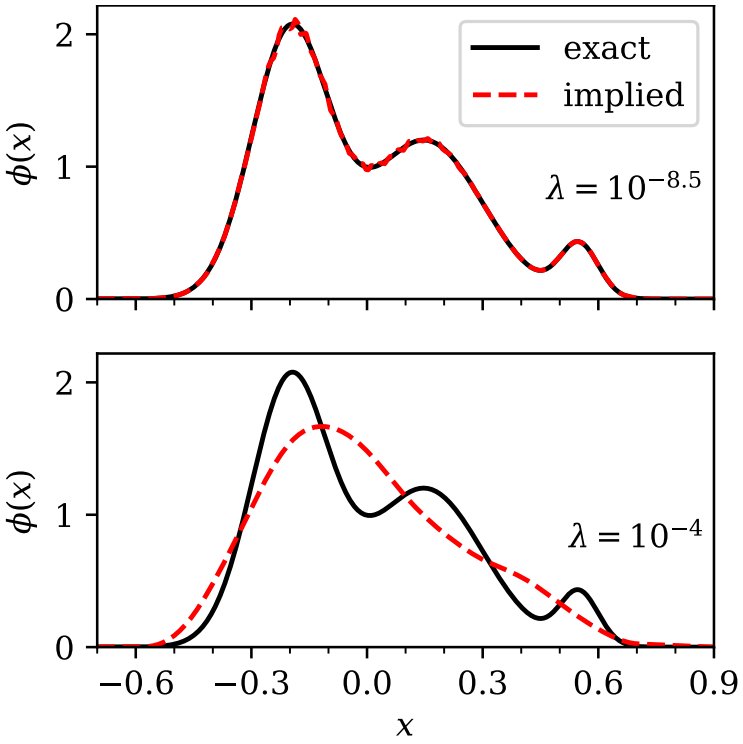


Figure 8. Comparison of the exact (bold line) and implied (dashed line) density $\phi(x)$ for two different values of the regularization parameter λ. The exact density is a linear combination of normal distributions, according to Equation (25), with parameters taken from Table 1. The top panel shows an optimal regularized implied density ($\lambda = 10^{-8.5}$), while the bottom panel shows an over-regularized implied density with $\lambda = 10^{-4}$.

In Figure 8 we show a comparison between the exact linear combination of normal distributions and our implied discretization $\phi(x)$ for two values of the regularization parameter λ. The optimal solution ($\lambda = 10^{-8.5}$) closely followed the original distribution. The solution for $\lambda = 10^{-4}$ was a version of our initial density, in which the features were broadened to become almost indistinguishable.

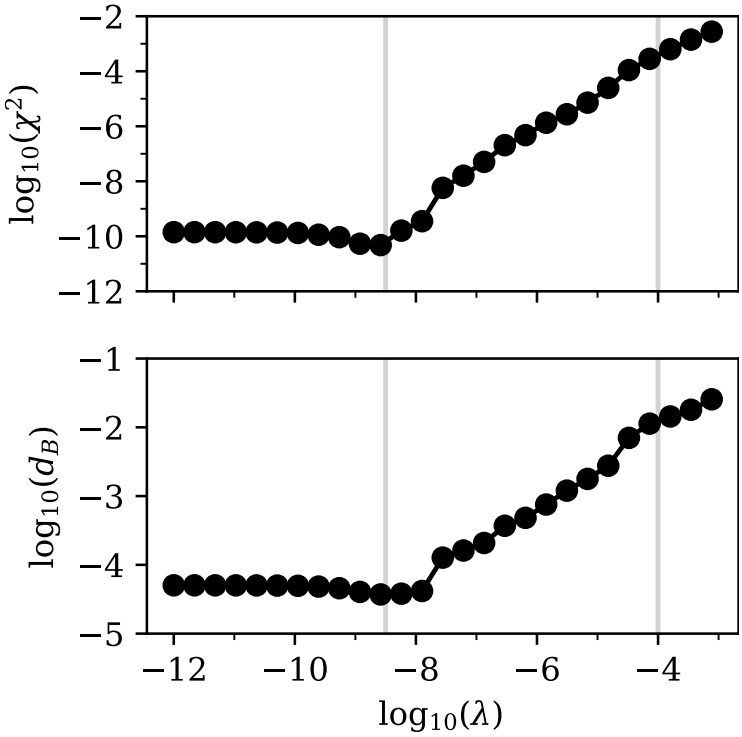


Figure 9. Log–log plot of the squared error in prices χ^2 (top panel) and the Bhattacharyya distance d_B (bottom panel). Both measures were calculated based on a comparison between exact input data and our implied output data for prices and density, respectively. The input density was given by Equation (25), with parameters taken from Table 1. The input option prices were calculated from an equivalent linear combination of Bachelier models with the same parameters, as in Table 1. The vertical lines mark the positions of $\lambda = 10^{-8.5}$ and $\lambda = 10^{-4}$, for which we show the implied densities in Figure 8.

3.4. Density Implied from Prices with Arbitrage

We now show that our method not only recovered known densities with high accuracy, but also carried out automatic de-arbitraging. Again, we set up a "density" as a linear combination of three normal distributions using Equation (25). However, we use quotation marks since we introduced one negative pre-factor, so that the resulting "density" ϕ_M contained negative "probabilities". The parameters used in this subsection can be found in Table 2.

Table 2. Parameters for a multimodal density. These parameters were used in Equation (25) to generate a density which was a superposition of multiple normally distributed components. This parameter set contained arbitrage, i.e., the resulting "density" contained negative "probabilities". This was due to the negative pre-factor.

i	c_i	μ_i	σ_i
1	0.55	0.80	0.10
2	−0.20	1.15	0.07
3	0.65	1.35	0.20

We used the Bachelier option pricing formulae (see Equations (23) and (24)) in the same way as in the multimodal case previously discussed. We set the interest rate to $r = 0.05$ and the time to expiry to $\tau = 1$. We then calculated the prices of 200 Call and Put options for a uniformly discretized grid of strikes between $K_{\min} = 0.3$ and $K_{\max} = 1.7$.

From the so-calculated option prices we implied the density $\phi(x)$ on a uniformly discretized grid with $x_{\min} = 0.1$, $x_{\max} = 2.2$ and $N = 1000$ using our method. We retained $Q = 150$ singular values.

Even prices that corresponded to partly negative "densities" could be processed using our method. In Figure 10 we show the error in prices χ^2 calculated from Equation (16) and the Bhattacharyya distance. Here, special care must be taken when calculating the Bhattacharyya distance, which is not defined for partly negative "densities". Therefore, we calculated the Bhattacharyya distance using Equation (21) with respect to the non-negative part of the input "density", that is $\phi_M^+(x) = \max(0, \phi_M(x))$. However, ϕ_M^+ is not truly a density. Recall that we fixed the sum of coefficients $\sum_i c_i = 1$ so that the integral over the density yielded unity. However, if there are negative regions, we know that the integral over the non-negative part ϕ_M^+ is larger than one. Therefore, the overlap in the Bhattacharyya formula may be larger than one, so that the Bhattacharyya distance d_B may become negative.

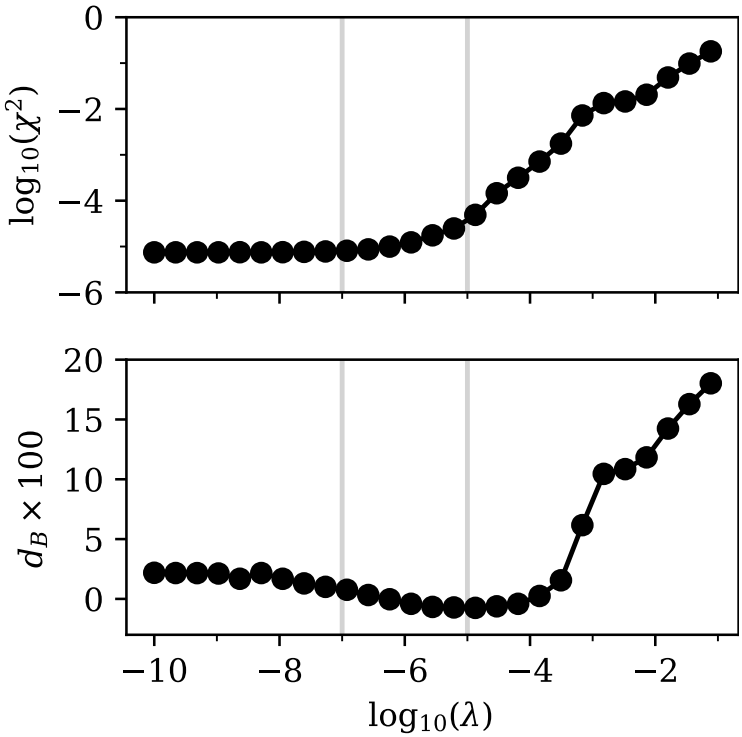


Figure 10. Log–log plot of the squared error in prices χ^2 (top panel) and log plot of the Bhattacharyya distance d_B (bottom panel). Both measures were calculated based on a comparison between exact input data and our implied output data for prices and densities, respectively. The input density was given by Equation (25) with parameters taken from Table 2. The input option prices were calculated from an equivalent linear combination of Bachelier models with the same parameters as those in Table 2. The vertical lines mark the positions of $\lambda = 10^{-7}$ and $\lambda = 10^{-5}$, for which we show the implied densities in Figure 11.

Of course, we could have normalized the non-negative part so that the integral over it was exactly one. However, we believe that such a situation may occur in practical applications and wanted to point out the consequences in detail. Therefore, we show, in Figure 10, the Bhattacharyya distance d_B directly, instead of its logarithm.

This time, any regularization led to an increase in the pricing error χ^2. This was quite logical, since the input prices were simply not reachable with a non-negative density, both because the inputs implied negative "probabilities" and because the integral over the positive part of the input "density" was not equal to one. Hence, there is no easy rule for selecting an appropriate regularization parameter. Any regularization increases

the pricing error, while it reduces the oscillations in the extracted density. Therefore, we suggest defining the necessary pricing accuracy and then selecting the largest possible regularization parameter λ that yields a lower pricing error.

In Figure 11 we show a comparison between the exact linear combination of normal distributions and our implied discretization $\phi(x)$ for two values of the regularization parameter λ. Which of these extracted densities was better suited for further processing depended on the specific use case.

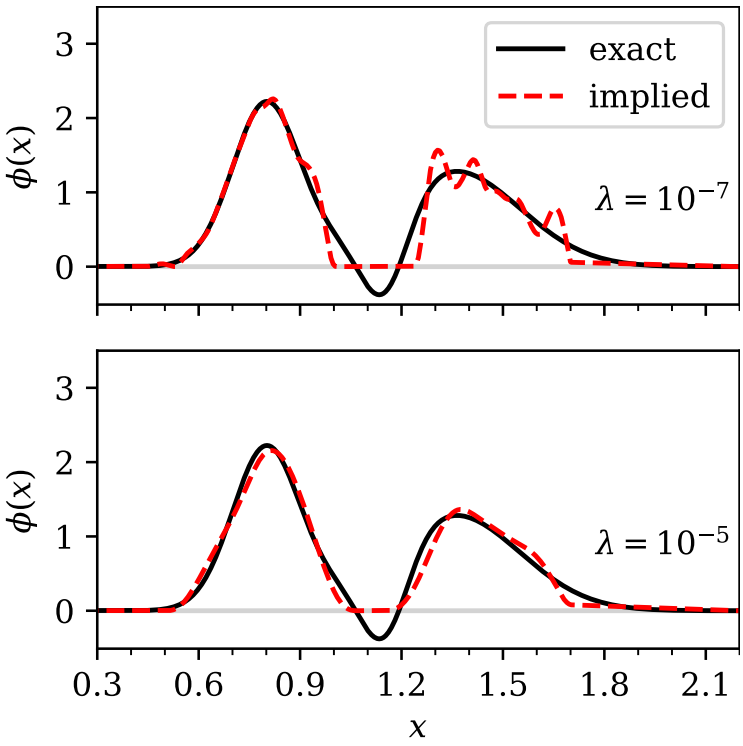


Figure 11. Comparison of the exact (bold line) and implied (dashed line) density $\phi(x)$ for two different values of the regularization parameter λ. The exact "density" is a linear combination of normal distributions according to Equation (25), using parameters from Table 2, which contains a region with negative probability. The top panel shows an under-regularized implied density ($\lambda = 10^{-7}$), while the bottom panel shows a well-regularized implied density with $\lambda = 10^{-5}$.

The automatic de-arbitraging feature of our method is also very useful when dealing with implied volatilities. Suppose we did not know the density from which input prices were generated. If we were to calculate implied volatilities we would usually use log-normal volatilities and simply calculate them by inverting the Black–Scholes model. This generates an implied volatility smile, which may, and in this case does, contain arbitrage. However, we could also generate an arbitrage-free volatility smile from the prices that we obtained from our optimization procedure. We calculated the implied volatilities using a simple bisection solver. We re-used the interest rate $r = 0.05$ and time to expiry $\tau = 1$. However, we now also needed an initial value of the underlying, which we arbitrarily set to $S_0 = 1.0$. Of course, in a realistic setting, this value would be known from the market.

We show the results of the log-normal implied volatility calculation in Figure 12. Clearly, the original volatility smile and our de-arbitraged version, calculated on the density with $\lambda = 10^{-5}$ (see Figure 11, bottom panel) were very similar. The arbitrage in the original volatility smile was not directly visible. This also shows why methods that work directly with the implied volatility may introduce arbitrage, which is not immediately apparent to the user.

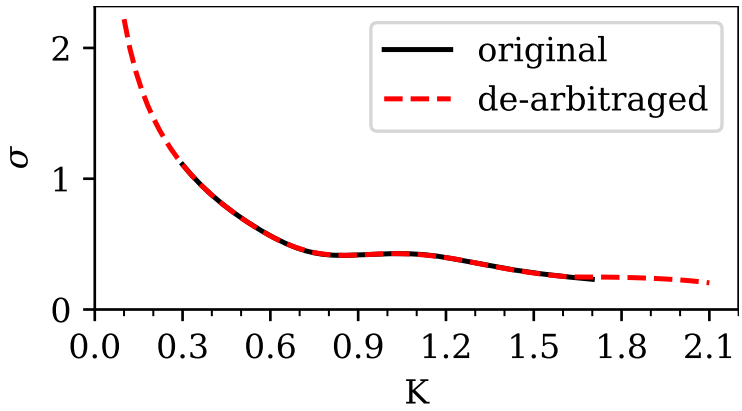


Figure 12. Comparison of the log-normal implied volatilities σ (as a function of the option strike K) calculated from the input prices containing arbitrage (bold line), which were based on Equation (25) and parameters from Table 2, and the de-arbitraged prices calculated from our method (dashed line) at $\lambda = 10^{-5}$. The calculation of de-arbitraged implied volatilities is based on the density shown in Figure 11 (bottom panel).

Note how our method also enables us to extrapolate beyond the range of known strikes, since it gives us access to smooth non-negative density in our range of choice. The results beyond the range of strikes with known prices are certainly speculative, but consistent with the known inputs. That we obtain sensible behavior in the wings of the volatility smile without any additional effort, is another advantageous property of the sparse modeling approach.

3.5. Density Implied from S&P 500 Option Prices

It has been mentioned in the literature (Le Floc'h and Osterlee 2019a) that short-term SPX500 options pose a challenge, particularly to stochastic volatility models and the similar SVI smile model (Gatheral and Jacquier 2014), since their volatility smiles are quite steep. We imported the market data from Table 11 in Le Floc'h and Osterlee (2019a), which corresponded to SPX500 1M (one month) options on 5 February 2018. We calculated Call and Put option prices from these market data for 75 strikes in the range between 1900 and 2900, using the Black (1976) model.

We noticed that the strikes in the thousands range led to numerical problems in the `ECOS` solver, so we divided all strikes in the inputs by 1000. The same transformation was also applied to the forward price. This simple rescaling of the problem solved the numerical issues we encountered with the original inputs.

We used the so-calculated option prices to imply the density $\phi(x)$ on a uniformly discretized grid with $x_{\min} = 1.4$, $x_{\max} = 3.4$ and $N = 1000$ using our method. We retained $Q = 70$ singular values, which was lower than in the previous cases, because we also had fewer quoted strikes available.

The original prices were reproduced to a high degree of accuracy. In Figure 13 we show the error in prices χ^2 calculated from Equation (16), which was negligible in the unregularized limit. Since the terminal density was truly unknown in this case, we could not measure the Bhattacharyya distance. As can be seen in Figure 13, any increase in the regularization parameter λ led to an increase in pricing error.

In Figure 14 we show the implied density $\phi(x)$ for two different values of the regularization parameter λ. For $\lambda = 10^{-7}$ the error in prices was still close to minimal and the density showed a pronounced spike around $x = 2800$, followed by a very sharp decrease. For stronger regularization, such as for $\lambda = 10^{-4}$, the features of the density were smeared out and the error in prices increased substantially. Again, the largest possible λ, which still gives an error in prices χ^2 close to the minimum, should be selected.

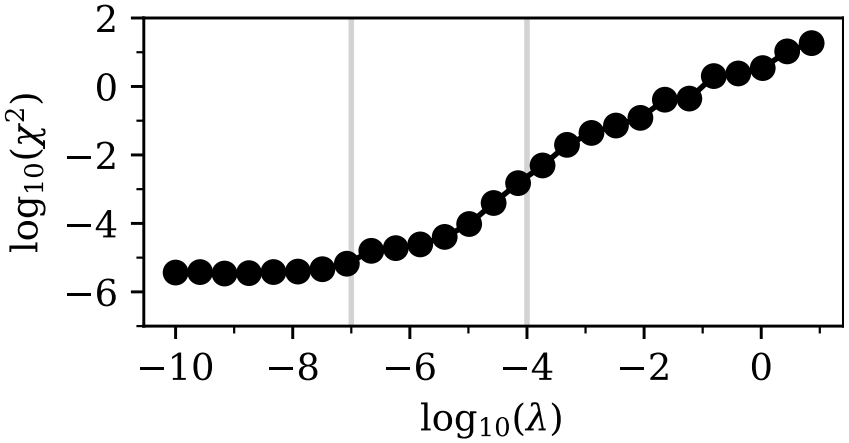


Figure 13. Log–log plot of the squared error in prices χ^2 as a function of the regularization parameter λ for SPX500 1M options as of 5 February 2018. The squared error was calculated from a comparison between exact input data and our implied output prices. The input option prices were calculated from the Black (1976) model with market data of Table 11 in Le Floc'h and Osterlee (2019a). The vertical lines mark the positions of $\lambda = 10^{-7}$ and $\lambda = 10^{-4}$, for which we show the implied densities in Figure 14.

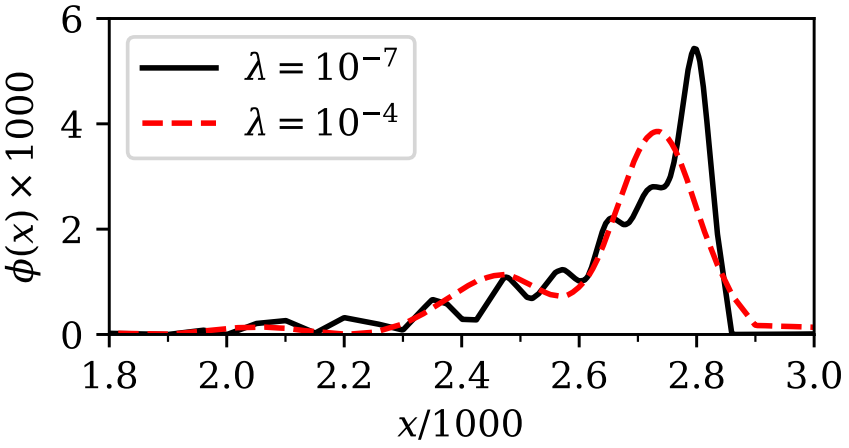


Figure 14. Comparison of implied densities $\phi(x)$ for SPX500 1M options as of 5 February 2018. A good compromise between accuracy and smoothness was achieved for $\lambda = 10^{-7}$ (bold line), while $\lambda = 10^{-4}$ yielded a density that contained fewer features (dashed line) and was potentially over-regularized. As explained in the main text, we rescaled both the price x of the underlying asset and the density $\phi(x)$, for numerical reasons, by a factor of $1/1000$ and 1000, respectively.

Since the original data in Le Floc'h and Osterlee (2019a) is given in terms of log-normal implied volatilities, we also calculated these implied volatilities from our density $\phi(x)$ at $\lambda = 10^{-7}$. The implied volatility was found by calculating option prices from the density using Equation (1) and then inverting the Black formula using a bisection solver for the volatility.

The comparison between input volatilities and the volatility smile extracted from our method is shown in Figure 15. The visible kink in the implied volatility around $K = 2800$ was well reproduced. Note again how our method enabled us to not only interpolate, but also extrapolate, implied volatilities even in challenging situations.

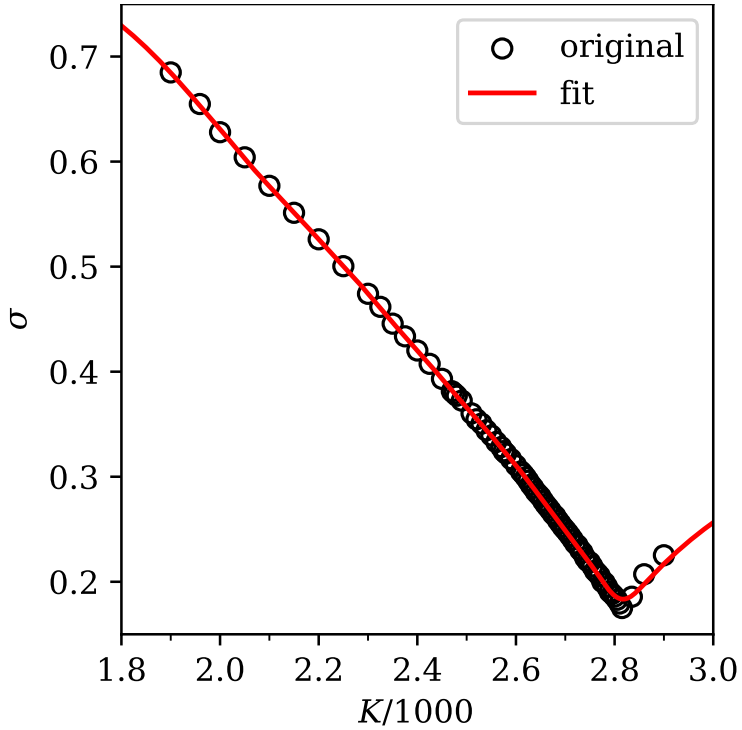


Figure 15. Comparison of input implied volatilities σ (open circles) and the volatility smile provided by our method at $\lambda = 10^{-7}$ (bold line) for SPX500 1M options as of 5 February 2018. The volatilities σ are shown as a function of the option strike K in units of thousands. Clearly, our method reproduced the inputs with a high degree of accuracy and, additionally, provided a sensible extrapolation of the available data.

4. Conclusions

We have presented a new method for implying terminal densities directly out of option prices. We showed that our method is able to produce arbitrage-free interpolations and extrapolations of both option prices and implied volatilities, while it does not require de-arbitraging of input prices or other pre-processing steps.

Our algorithm is based on singular value decomposition (SVD), which produces a transformation to a basis, in which the relation between prices and density is represented by a sparse model. In this sense, the number of parameters in the model Q is smaller than, or similar to, the number of input option prices M, while we may extract the density $\phi(x)$ with a much larger number of discretization points N.

This property is the hallmark of any sparse model. It enables us to formulate the optimization problem for finding the density based on optimizing a small number of parameters Q. We also showed how L_1-regularization helps in finding a density that is a good compromise between pricing error and smoothness.

Besides the trivial parameters $x_{\min}$ and $x_{\max}$ that define the boundaries of the interval in which the density is discretized, the only relevant parameters of our method that are visible to the user are the number of retained singular values Q, the number of input option prices M, the number of discretization points for the density N and the regularization parameter λ. The number of input prices M is fixed by the problem that is under investigation. We experienced good results when choosing $Q \lesssim M/2$. For N, we simply chose a large number so that we could re-calculate option prices with sufficient accuracy. For our purposes, $N = 1000$ always seemed sufficient.

In this sense, only the regularization parameter λ is truly up to the user's choice. We also presented a simple rule for selecting λ by scanning the error in prices for a number of different values for λ and choosing the highest possible one with close to minimal error χ^2. In the literature, this is often referred to as the "elbow method".

As far as we are aware, relying on optimization, and the subsequent need to choose a regularization parameter, seem to be the only drawbacks of the sparse modeling approach. Of course, our method cannot be used to directly price exotic options, while this is easily possible with stochastic volatility models once they have been calibrated. That, however, is not the topic of the present investigation.

Barring these restrictions, the advantages of our method are clear. The mathematics behind our method is simple and easy to understand. Using the numerical libraries mentioned, the algorithm is also easy to implement. Our own implementation in the Python programming language consists of fewer than 100 lines of code. The accuracy of our method proved to be excellent for all artificial and realistic examples we investigated.

Furthermore, our algorithm is robust against defects in the input data, such as arbitrage. Since it works directly with the terminal density, our method can also be used to extrapolate option prices and implied volatilities in an arbitrage-free manner far beyond the available range of market quotes. Having sensible behavior in the wings of the volatility smile, without any further effort, is quite rare an occurrence and is certainly a strong advantage of our method.

What makes our method stand out from other available approaches, is that choosing a polynomial or other basis for the regression is not necessary, because a suitable orthogonal basis is automatically constructed by the SVD. In this sense, our method is truly model-free.

At present, our algorithm works with options of multiple strikes, but with a single time to maturity. In future works, we would like to investigate extensions to multiple maturities, such that the whole volatility surface can be calibrated consistently.

Such an extension would enable us to use our method to calibrate a local volatility (Derman and Kani 1994; Dupire 1994) or local stochastic volatility model (Lipton 2002; Lipton et al. 2014), which we could then use to price options with American exercise (Andersen and Lake 2021; Andersen et al. 2016; Healy 2021), barrier options (Clark 2010; Guterding and Boenkost 2018) or other exotic products. Since our method is robust against defects in the market data, such an extension should be particularly helpful in situations where the market data are potentially stale or only a few strikes are quoted. For example, this could be the case in markets for cryptocurrency (Hou at al. 2020; Yang and Hamori 2021a; Zulfiqar and Gulzar 2021), energy (Benth et al. 2008; Fabbiani et al. 2020; Yang 2021b) or foreign exchange options (Clark 2010; Guterding and Boenkost 2018), and also for Equity options (Healy 2021) with less popular underlyings.

Funding: This research was funded by Technische Hochschule Brandenburg, University of Applied Sciences.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SVD	Singular Value Decomposition
LV	Local Volatility
LSV	Local Stochastic Volatility
SABR	Stochastic Alpha, Beta, Rho
SPX500	Standard & Poor's 500 Stock Index
ITM	In-The-Money
OTM	Out-Of-The-Money

Appendix A. Treatment of In-the-Money Options in the Error Function Calculation

When considering Call and Put options with arbitrary strike, the prices of the options may differ by several orders of magnitude if some options are in-the-money. In-the-money options are executed with large probability and, hence, have a high price.

However, if the prices in our problem differ by orders of magnitude, the calculated error function is dominated by the options with the largest price, which are, unfortunately, those that depend only on the tail of the probability distribution ϕ. This is not desirable, since we are usually equally interested in all regions of the density, or even more so in the density close to at-the-money. This problem can be solved by transforming prices of in-the-money options to prices of out-of-the-money options.

Let F define the forward price of the underlying asset at time T. Then a Call option with strike K is considered in-the-money if $K < F$. A Put option is considered in-the-money if $K > F$. Conversely, a Call option is out-of-the-money if $K > F$ and a Put option is out-of-the-money if $K < F$.

Let $\mathrm{Pr_C}$ denote the price of a European Call option with strike K and expiry at T and let $\mathrm{Pr_P}$ denote the price of a Put option with the same strike and expiry. If S_0 is the value of the underlying asset at $t = 0$, then, for European options with the same strike and the same time to expiry τ, Put–Call-parity holds:

$$\mathrm{Pr_C} + Ke^{-r\tau} = \mathrm{Pr_P} + S_0\,. \tag{A1}$$

So, for all in-the-money Call options, we may calculate the price of the respective out-of-the-money Put option from Equation (A1). Likewise, for all in-the-money Put options, we may calculate the price of the respective out-of-the-money Call option from Equation (A1).

In this way, we can easily restrict the prices in our optimization problem to out-of-the-money and at-the-money options, which all have prices with roughly the same order of magnitude. Hence, the error function is not dominated by the options that depend only on the tails of the probability distribution.

Although we did not have to apply this transformation from ITM to OTM options in the present manuscript, we believe it makes sense to document this idea, in case our readers encounter problems with ITM options when applying our method.

Appendix B. Ideas for Performance Optimization

The main tunable parameters that influence the performance of the algorithm we presented are the number of retained singular values Q and the number of discretization points for the density N. Since Q is also the number of parameters in the optimization problem, it is clear that reducing Q could lead to faster convergence of the optimization algorithm. Figure 5 clearly shows that after regularization only a few relevant parameters remained. Therefore, the remaining parameters with negligible weight could also be removed, before we even started the optimization process, by choosing a lower value of Q.

Based on our analysis of singular values in Figure 2 we could expect that a reduction of Q would first result in discarding of the less relevant parameters and would then progress to removing more relevant ones if Q further reduced. However, the manner by which this depends on input prices is not clear. Therefore, this issue needs more analysis before we can give any definitive conclusion, but this was not the main point of the present manuscript.

The second opportunity for speeding up the algorithm is to reduce the number of discretization points for the density N. The point of having more discretization points than input strikes is to have a high enough resolution to be able to recalculate the input option prices with sufficient accuracy, so as to enable use of the implied density for interpolation. Since N determines the effort for the partial SVD, the basis transformations and the checking of the constraints in the optimization problem, it is worth thinking about a reduction of N. We suggest first checking whether a reduction in N increases the error in prices χ^2. If the density is required for an interpolation of prices or implied volatilities, first interpolating

the density linearly and then calculating prices and implying volatilities, based on the interpolated densities, is always an option. Since a linear interpolation of a non-negative function is also non-negative, we can be sure that this procedure would not introduce negative densities, i.e., arbitrage. Linear interpolation of the density also does not violate the constraint that the trapezoidal integral over the density must be equal to unity.

If our method is used in a live environment, where the density is implied on every market data update, or on every couple of market data updates, it may be useful to warm-start the numerical solver from the previous solution ϕ' to accelerate convergence.

Note

[1] https://github.com/danielguterding/svdensity (accessed on 21 April 2023).

References

Agrawal, Akshay, Robin Verschueren, Steven Diamond, and Stephen Boyd. 2018. A rewriting system for convex optimization problems. *Journal of Control and Decision* 5: 42. [CrossRef]

Aït-Sahalia, Yacine, Peter J. Bickel, and Thomas M. Stoker. 2001. Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Economics* 105: 363. [CrossRef]

Andersen, Leif, and Mark Lake. 2021. Fast American Option Pricing: The Double-Boundary Case. *Wilmott* 2021: 30. [CrossRef]

Andersen, Leif, Mark Lake, and Dmitri Offengenden. 2016. High-performance American option pricing. *Journal of Computational Finance* 20: 39. [CrossRef]

Andreasen, Jesper, and Brian Norsk Huge. 2011. Volatility Interpolation. *Risk* 24: 76. [CrossRef]

Bachelier, Louis. 1901. Théorie mathématique du jeu. *Annales Scientifiques de l'Ecole Normale Supérieure* 18: 143. [CrossRef]

Baker, Glyn, Reimer Beneder, and Alex Zilber. 2004. FX Barriers with Smile Dynamics. Available online: https://ssrn.com/abstract=964627 (accessed on 21 April 2023). [CrossRef]

Benth, Fred Espen, Jūratė Šaltytė Benth, and Steen Koekebakker. 2008. *Stochastic Modeling of Electricity and Related Markets*. Singapore: World Scientific. ISBN 978–9-812-81230-8.

Bhattacharyya, Anil. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35: 99.

Black, Fischer. 1976. The pricing of commodity contracts. *Journal of Financial Economics* 3: 167. [CrossRef]

Black, Fischer, and Myron Scholes. 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81: 637. [CrossRef]

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2010. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning* 3: 1. [CrossRef]

Breeden, Douglas T., and Robert H. Litzenberger. 1978. Prices of State-Contingent Claims Implicit in Option Prices. *Journal of Business* 51: 621. [CrossRef]

Carr, Peter, and Roger Lee. 2009. Volatility Derivatives. *Annual Review of Financial Economics* 1: 319. [CrossRef]

Carr, Peter, Andrey Itkin, and Dmitry Muravey. 2022. Semi-analytical pricing of barrier options in the time-dependent Heston model. *arXiv*. [CrossRef]

Choi, Jaehyuk, Minsuk Kwak, Chyng Wen Tee, and Yumeng Wang. 2022. A Black-Scholes user's guide to the Bachelier model. *Journal of Futures Markets* 42: 959. [CrossRef]

Clark, Iain. 2010. *Foreign Exchange Option Pricing: A Practitioners Guide*. Chichester: John Wiley & Sons. ISBN 978-0-470-68368-2.

Derman, Emanuel, and Iraj Kani. 1994. Riding on a smile. *Risk* 7: 32.

Derman, Emanuel, and Michael. B. Miller. 2016. *The Volatility Smile*. Hoboken: John Wiley & Sons. ISBN 978-1-118-95916-9.

Diamond, Steven, and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17: 2909.

Domahidi, Alexander, Eric Chu, and Stephen Boyd. 2013. ECOS: An SOCP Solver for Embedded Systems. Paper presented at European Control Conference, Zurich, Switzerland, July 17–19. pp. 3071–3076.

Dupire, Bruno. 1994. Pricing with a smile. *Risk* 7: 18.

Egloff, Daniel, Markus Leippold, and Liuren Wu. 2010. The Term Structure of Variance Swap Rates and Optimal Variance Swap Investments. *Journal of Financial and Quantitative Analysis* 45: 1279. [CrossRef]

Fabbiani, Emanuele, Andrea Marziali, and Giuseppe De Nicolao. 2020. Vanilla-option-pricing: Pricing and market calibration for options on energy commodities. *Software Impacts* 6: 100043. [CrossRef]

Gatheral, Jim. 2006. *The Volatility Surface: A Practitioner's Guide*. Hoboken: Wiley Finance. ISBN 978-0-471-79251-2.

Gatheral, Jim, and Antoine Jacquier. 2014. Arbitrage-free SVI volatility surfaces. *Quantitative Finance* 14: 59. [CrossRef]

Guterding, Daniel. 2021. Inventory effects on the price dynamics of VSTOXX futures quantified via machine learning. *Journal of Finance and Data Science* 7: 126. [CrossRef]

Guterding, Daniel, and Wolfram Boenkost. 2018. The Heston stochastic volatility model with piecewise constant parameters—Efficient calibration and pricing of window barrier options. *Journal of Computational and Applied Mathematics* 343: 353. [CrossRef]

Hagan, Patrick, Andrew Lesniewski, and Diana Woodward. 2015. Probability Distribution in the SABR Model of Stochastic Volatility. In *Large Deviations and Asymptotic Methods in Finance*. Edited by Peter K. Friz, Jim Gatheral, Archil Gulisashvili, Antoine Jacquier and Josef Teichmann. Cham: Springer Proceedings in Mathematics & Statistics, vol. 110.

Hagan, Patrick, Deep Kumar, and Andrew Lesniewski. 2001. Managing Smile Risk. *Wilmott* 1: 84.

Healy, Jherek. 2021. *Applied Quantitative Finance for Equity Derivatives*. Lulu.com, ISBN 978-1-716-19039-1.

Heston, Steven L. 1993. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies* 6: 327. [CrossRef]

Hou, Ai Jun, Weining Wang, Cathy Y. H. Chen, and Wolfgang Karl Härdle. 2020. Pricing Cryptocurrency Options. *Journal of Financial Econometrics* 18: 250.

Jäckel, Peter. 2014. Clamping down on arbitrage. *Wilmott* 2014: 54. [CrossRef]

Jiang, Yixiao. 2020. A Hausman Test for Partially Linear Models with an Application to Implied Volatility Surface. *Journal of Risk and Financial Management* 13: 287. [CrossRef]

Kahalé, Nabil. 2004. An arbitrage-free interpolation of volatilities. *Risk* 17: 102.

Le Floc'h, Fabien, and Cornelis W. Osterlee. 2019a. Model-free stochastic collocation for an arbitrage-free implied volatility: Part I. *Decisions in Economics and Finance* 42: 679. [CrossRef]

Le Floc'h, Fabien, and Cornelis W. Osterlee. 2019b. Model-free stochastic collocation for an arbitrage-free implied volatility: Part II. *Risks* 7: 1. [CrossRef]

Lipton, Alexander. 2002. The vol smile problem. *Risk* 15: 61.

Lipton, Alexander, Andrey Gal, and Andris Lasis. 2014. Pricing of vanilla and first-generation exotic options in the local stochastic volatility framework: Survey and new results. *Quantitative Finance* 14: 1899. [CrossRef]

Lorig, Matthew, Stefano Pagliarani, and Andrea Pascucci. 2017. Explicit Implied Volatilities for Multifactor Local-Stochastic Volatility Models. *Mathematical Finance* 27: 926. [CrossRef]

Mixon, Scott. 2002. Factors explaining movements in the implied volatility surface. *Journal of Futures Markets* 22: 915. [CrossRef]

O'Donoghue, Brendan, Eric Chu, Neal Parikh, and Stephen Boyd. 2016. Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding. *Journal of Optimization Theory and Applications* 169: 1042. [CrossRef]

Otsuki, Junya, Masayuki Ohzeki, Hiroshi Shinaoka, and Kazuyoshi Yoshimi. 2020. Sparse Modeling in Quantum Many-Body Problems. *Journal of the Physical Society of Japan* 89: 012001. [CrossRef]

Tian, Yu, Zili Zhu, Geoffrey Lee, Thomas Lo, Fima Klebaner, and Kais Hamza. 2014. Pricing Window Barrier Options with a Hybrid Stochastic-Local Volatility Model. Paper presented at 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), London, UK, March 27–28. [CrossRef]

Xing, Yuhang, Xiaoyan Zhang, and Rui Zhao. 2010. What Does the Individual Option Volatility Smirk Tell Us About Future Equity Returns? *Journal of Financial and Quantitative Analysis* 45: 641. [CrossRef]

Yang, Lu. 2021. Idiosyncratic information spillover and connectedness network between the electricity and carbon markets in Europe. *Journal of Commodity Markets* 25: 100185. [CrossRef]

Yang, Lu, and Shigeyuki Hamori. 2021. The role of the carbon market in relation to the cryptocurrency market: Only diversification or more? *International Review of Financial Analysis* 77: 101864. [CrossRef]

Zhu, Song-Ping, and Guang-Hua Lian. 2012. An analytical formula for VIX futures and its applications. *Journal of Futures Markets* 32: 166. [CrossRef]

Zulfiqar, Noshaba, and Saqib Gulzar. 2021. Implied volatility estimation of bitcoin options and the stylized facts of option pricing. *Financial Innovation* 7: 67. [CrossRef] [PubMed]

risks

MDPI

Article

BeVIXed: Trading Fear in the Volatility Complex

Chakravarthy Varadarajan [1] **and Klaus R. Schenk-Hoppé** [2,3,*]

[1] Department of Physics, School of Natural Sciences, University of Manchester, Manchester M13 9PL, UK; chakravarthy.varadarajan@student.manchester.ac.uk
[2] Department of Economics, School of Social Sciences, University of Manchester, Manchester M13 9PL, UK
[3] Department of Finance, NHH–Norwegian School of Economics, 5045 Bergen, Norway
* Correspondence: klaus.schenk-hoppe@manchester.ac.uk

Abstract: We explain the evolution of the volatility market and present the infamous day of 'Volmageddon' as an insightful case study. Our survey focuses on the pricing and trading of volatility-linked assets, highlighting the impact of mechanical hedging in markets for futures and higher-order derivatives. We supplement the vast statistical analysis of volatility derivatives with a financial economist's perspective.

Keywords: Volmageddon; volatility derivatives; hedging; VIX

Citation: Varadarajan, Chakravarthy and Klaus R. Schenk-Hoppé. 2023. BeVIXed: Trading Fear in the Volatility Complex. *Risks* 11: 86. https://doi.org/10.3390/risks11050086

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 12 April 2023
Revised: 25 April 2023
Accepted: 25 April 2023
Published: 4 May 2023

1. Introduction

Financial market volatility continues to reign supreme. In this paper, we outline volatility's progression from a theoretical risk measure to a tradable asset class, and describe the cataclysmic impact of mechanical hedging by major volatilty market participants.

The Chicago Board Options Exchange's (CBOE) Volatility Index (VIX) is a widely followed index that calculates the implied volatility of the U.S. stock market for the next 30 days. Colloquially referred to as the 'Fear Gauge', the VIX is updated every 15 s during a trading day. The VIX lies at the heart of the Volatility Complex, as illustrated in Figure 1, which influences how traders price the sprawl of volatility-linked assets.

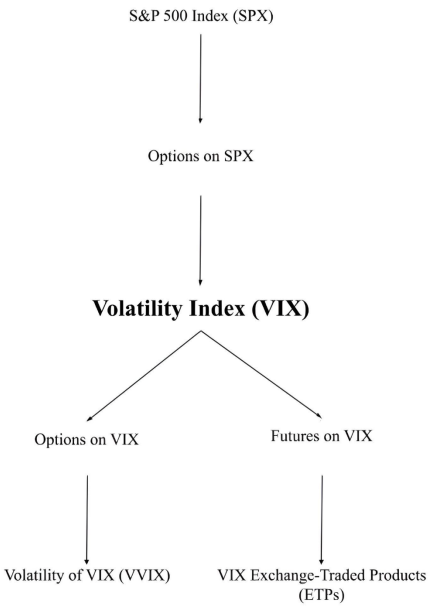


Figure 1. The Volatility Complex.

This paper is composed of two self-contained chapters. The first chapter builds the Volatility Complex from the ground up, beginning with the academic definition of volatility

and ending with the construction of a volatility-linked exchanged traded product. Along the way, we cover the appropriate mathematical and financial prerequisites to understand the VIX calculation methodology. One goal is to demystify the inner workings of the derivative and volatility markets. The second chapter delineates the growth in shorting, or betting against, volatility and the reason behind such a trading strategy's risk of backfiring. The carnage in financial products designed to short volatility on 5 February 2018 was later dubbed as 'Volmageddon'.

The prevalent explanation of Volmageddon, which we expound upon, posits that higher order volatility-linked derivatives sparked a negative feedback loop in tandem with underlying assets further down the volatility ladder because of the price-insensitive hedging carried out by major market participants. We postulate that a similar 'tail-wagging-the-dog' effect occurred in the SPX options market on 20 March 2020, again due to a mechanical hedging strategy.

The post-crisis regulation gave birth to the ongoing, protracted boom in volatility-sensitive investing. However, this came with an unintended consequence, the volatility is now more volatile. In particular, drawdowns in volatility-linked assets have increased in both frequency and intensity Peterseil and Kawa (2019). Many have valiantly tried (and failed) to hedge against such drawdowns using assets in the complex. Mistakes are inevitable if investors do not fully understand what they are actually trading. In light of this truism, we hope the paper will serve as a valuable guide to help the reader navigate through—and skilfully manage—fear in the Volatility Complex. As Mark Twain supposedly opined, "It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so". This paper may be of use to regulators, academics, and, of course, investors.

2. The Volatility Complex

Simply put, volatility is financial uncertainty. If volatility can be quantified, then so can an investment's risk profile. In 1952, Harry Markowitz explained volatility in statistical terms that allowed investors to better understand an investment portfolio's return as a function of risk. About 20 years later, in 1973, Black and Scholes' seminal paper 'The Pricing of Options and Corporate Liabilities' introduced a model to calculate implied, i.e., future, volatility from market prices of option contracts. Another 20 years later, in 1993, the CBOE launched the first version of the VIX, which computed the implied volatility of a basket of U.S. stocks with the help of the Black–Scholes–Merton model. In 2003, the CBOE changed the VIX's calculation methodology to make it 'model-free'. A couple of years later, derivatives on the VIX were rolled out and quickly picked up in popularity. Although the volatility market is relatively young (about 20 years), its rapid expansion provides an insight into how and why volatility derivatives became so popular with investors, ranging from retail to institutional.

2.1. Volatility as Risk

Any investment decision is based on two variables, risk and expected return. Risk is disliked (the less the better), and expected return is welcomed (the more the better). In practice, however, the investor is forced to make a trade-off between the two because an investment with higher expected return is usually accompanied with higher risk. So how is risk understood with respect to an expected return?

One way to measure risk is by the volatility of a financial asset's return. Indeed, when the return is normally distributed, return and volatility fully describe this return profile. In less simple scenarios, the frequency of extreme events (thick tails), asymmetry (skewness), time-dependence (e.g., caused by the business cycle or crises), the unthinkable (non-anticipated changes, such as the occurrence of a 'black swan'), and other dimensions, can also matter to the investor. Figure 2 illustrates how volatility corresponds to the degree of dispersion of an asset's return. In this example with a normal distribution, the expected

return is 8% for each asset. The distribution of (future) returns is symmetric about this expected return which means, for example, a future return of 12% is as likely as a return 4%.

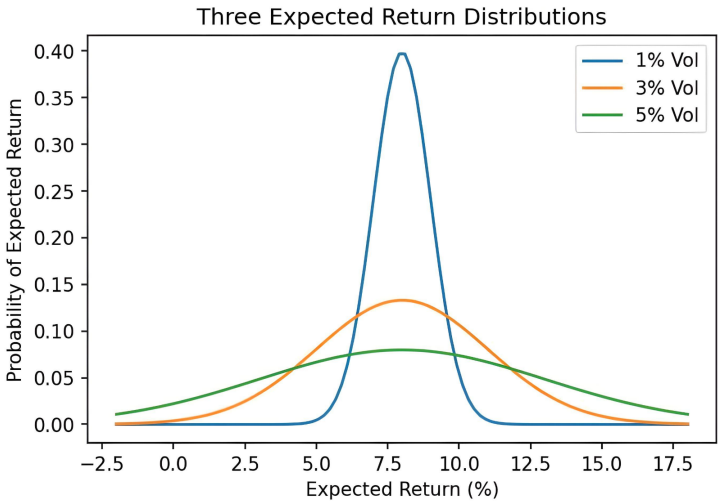


Figure 2. Three assets with expected return +8%. Volatilities are 1%, 3%, and 5%.

The bell-shaped curve posits that future returns closer to the expected return are more likely than those further toward the tail ends. As volatility σ increases, however, future returns away from the 8% mark increase in likelihood. Thus, the asset with 5% volatility is the riskiest of the three. In reality, however, one will find that the expected return increases with volatility.

There are subtleties in the semantics. An expected return is one you predict to see on average; a realized return is one you will observe in the future. There is a range of possible returns ex ante, but there will only be one return ex post. Risk, like expected return, is an ex ante concept because it is a feature of not-yet-realized returns. For example, consider an asset that promises you one of two returns, 20% or 10%, with equal likelihood. The expected return is 15%, but it will never be physically realized.[1]

One of the biggest challenges in quantitative finance is to come up with a fairly tenable distribution of an asset's future return. Are you able to provide a detailed account of the possible returns and their respective likelihoods of, say, Apple's stock price in one year's time? A more informed alternative to daft speculation is the projection of historical data into the future. Nevertheless, how likely is Apple's past performance a 'guarantee of its future results'? It is difficult to say, but the use of historical data to quantify a future return is a widely used rule-of-thumb by financial practitioners.

The S&P 500 index (SPX), which captures roughly 80% of the total U.S. stock market capitalization, is a measure of the near-term outlook of the United States' economy. It might be more intuitive to think about the SPX as a financial barometer that indicates the expected weather of the general U.S. economy in the coming months.

Therefore, an investment in the index is an investment in the broad U.S. economy. The SPX is merely a tracker and not directly investable; its index 'level' is a number calculated from all 500 stocks and their respective market capitalizations. So one can invest in other instruments that either proxy the SPX or that expose you to risks associated with it. That sounds more complicated than it actually is.

One standard investment vehicle that tracks, and delivers, the returns of an underlying index is called the exchange-traded fund (ETF). In the case of the SPX, an issuer of the ETF pools together millions or billions of dollars to buy shares of businesses comprising the index in the right proportions. The issuer will then passively manage this 'copy-of-SPX' portfolio on behalf of the investor in exchange for a management fee. Figure 3 depicts the past performance of the SPDR S&P 500 ETF Trust (SPY), a popular ETF on the S&P 500 index.

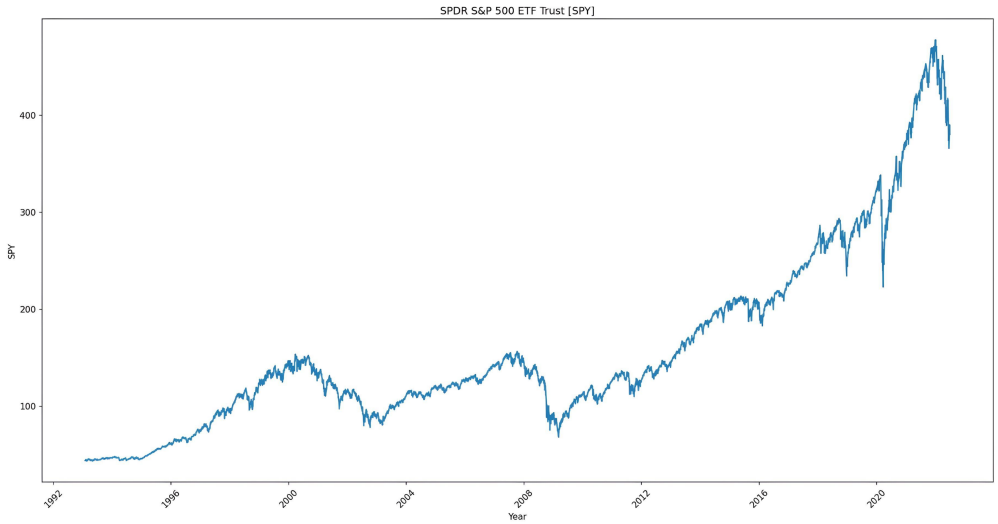


Figure 3. Price chart of the SPDR S&P 500 ETF Trust (SPY) since its inception.

Two popular alternative investments are options and futures on the underlying SPX. Each one is a derivative contract with specifications on when to trade the underlying and for how much. A futures contract specifies a future date when the buyer purchases the underlying asset from the seller at the predetermined 'future price'. Given the logistical challenges of delivering all 500 stocks associated with an SPX futures contract on the future date, the contract is instead settled on a cash basis. For example, if the SPX is higher than the predetermined future price on the future date, the buyer of the SPX futures receives the difference—in cash—from the seller.

A futures contract, unlike other derivatives, requires no payment upfront from the buyer to the seller.[2] Hence, the future price can be understood as the market's expectation of where the SPX is going to be on the predetermined future date. Figure 4 depicts a futures contract's payoff, from the point of view of the buyer and the seller. When a futures contract is agreed upon by both parties, the directional risk (i.e., whether the SPX will move up or down on the way to the future date) is taken up equally by both sides because the transaction is obligated to take place on the future date.

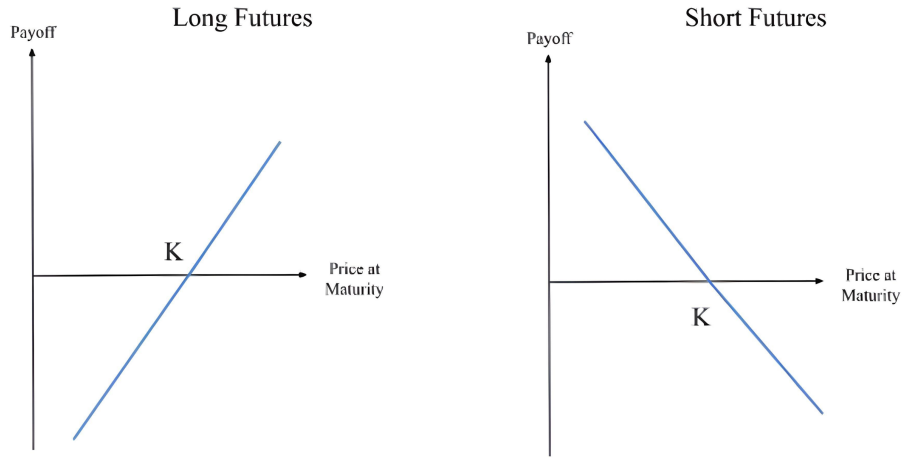


Figure 4. Long and Short Futures Payoff Structures.

An option contract differs in this regard by requiring an upfront payment in exchange for the option or choice to conduct the transaction on the future date.[3] An option contract's specifications are those of the futures', a future price and a future date. However, market players refer to an option's future price as its strike price and its future date as its expiration

date. The option to transact from the buyer's point of view, i.e., the right to buy on expiration date, is a call option. Likewise, the right to sell is a put option.

The buyer of a call option, therefore, gains if the SPX's level is higher than the strike price on the expiration date. The seller of the option will have to pay to the buyer the difference between the SPX and the strike price, i.e., cash settlement on expiration. However, if the SPX's level is below the strike price, no payment is made as the rational choice for the buyer would be to not exercise her right to purchase the SPX for a higher price. In options-speak, this call option expires out-of-the-money.

With put options, the payment on the expiration date is different. Since the buyer has the right to sell, the option is only exercised if the SPX is below the strike price. In that case, the put option seller pays the buyer the difference between the strike and the SPX. If the SPX is above the strike at the expiry date, the option is out-of-the-money and not exercised. To avoid confusion between the price of an option with the price of, say, the SPX, market players refer to the option price (aka the upfront payment) as option premium or simply premium.

A profit and loss (PL for short) is the price to hold the derivative starting now less the payoff at maturity. A long call's payoff, as shown in Figure 5, is only a neat approximation. In reality, its payoff is a convex function—i.e., a tilted smile of a curve above the approximation. For this reason, option contracts are a non-linear (i.e., convex) security. Likewise, futures contracts are a time-dependent linear derivative.

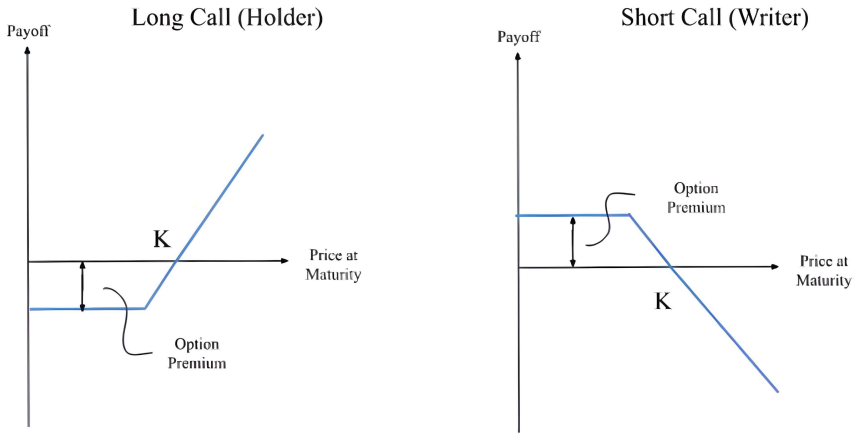


Figure 5. Call Option Payoff Structure.

Pricing an options contract is no mean feat. For example, as a seller of a call option, how do you determine the fair value of your premium?[4] Too high and there will be no buyers; too low and you will lose your shirt. Just right, and you would have received a Nobel prize. In 1997, Robert C. Merton and Myron S. Scholes were awarded the 'Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel' for the Black–Merton–Scholes (BSM[5]) option pricing formula.

The method is derived from an active trading strategy that is designed to perfectly replicate the payoff of an options contract using the underlying. The amount of cash required to mimic this payoff is simply the price of the option, i.e., the option premium, which is calculated for you by the formula. In technical terms, by dynamically trading with an amount equal to the option premium, one can fully hedge the option's payoff.

The trading strategy is simple as it only depends on the sensitivity of the premium with respect to the change in the price of the underlying. For example, if the option premium went up (down) by USD 0.75 and the price of the underlying went up (down) by USD 1, then the trading strategy advises the investor to hold exactly 0.75 units of the underlying. In technical terms, the Δ of an option provides the investor with an exact hedge. When a call option is hedged, the position in the underlying is positive (or, in finance terms,

long) and money needs to be borrowed. When a put option is hedged, the position in the underlying is negative (short) and money is deposited.

In practical terms, the hedge makes the seller of an options contract 'market-neutral', i.e., it removes the directional risk associated with selling an option. The bet from the seller's point of view is that the market goes down (up) and the call (put) option expires worthless.

Since hedging is of extreme importance to market participants, they have coined terms that express the underlying's current price relative to the option's strike price. An option is said to be 'out of the money' (OTM) if an expiration today will yield the contract worthless. Furthermore, an option is 'in the money' (ITM) if an expiration today will yield a positive payoff. Finally, an option is 'at the money' (ATM) if its strike price is the same as the current price of the underlying. Figure 6 presents how the delta value (Δ) of a call or a put varies with the moneyness of an option with respect to its strike. All most all options are issued with a strike close to the current price of the underlying.

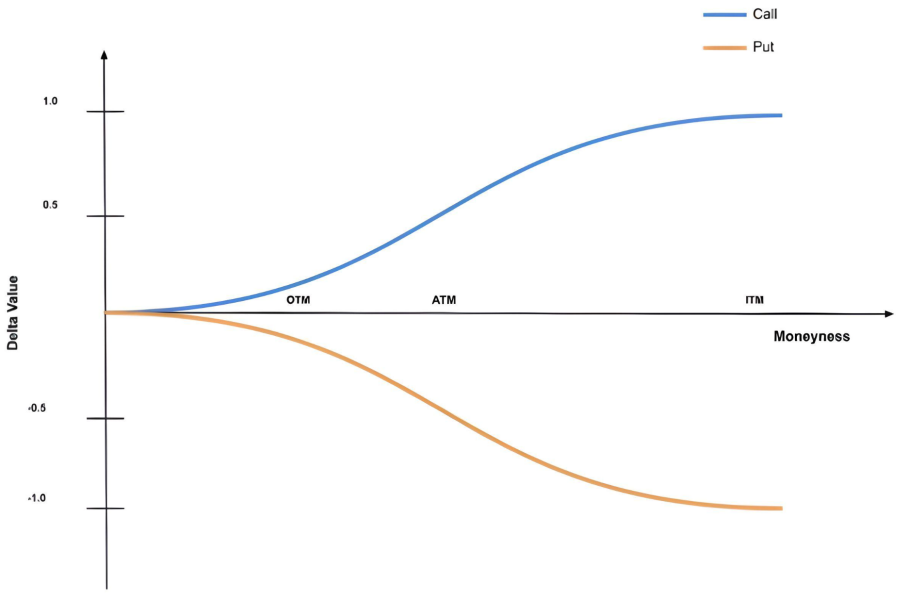


Figure 6. At the money, calls have a delta of 0.5 and puts have a delta of -0.5.

So who trades options and for what reasons? Well, there are a few market players whose intentions are clear. Retail investors would buy a call (put) if they are more (less) optimistic than the average investor about the future development of the underlying. They are classified as 'speculators', along with hedge funds, etc., because they make directional bets in the market. You will also find several investment banks who sell, or underwrite, options. Their aim is to collect the option premia and then hedge much less than what the BSM model prescribes. In fact, managing a large book of derivatives implies that many idiosyncratic risks offset each other, and so the investment bank only needs to manage the residual risk, i.e., what is leftover. Needless to say, computing this residual risk is not a walk in the park!

On the other hand, futures are often used by asset managers who want to hedge their investment. Suppose you manage a portfolio worth USD 100 billion in assets and whose risk-return profile closely resembles the S&P 500 index. If you are concerned about tomorrow's release of U.S. labour market data (i.e., the non-farm payroll), you can completely hedge your portfolio with an appropriate short futures position on the SPX. No matter what the markets tell you tomorrow, any change in the value of the S&P 500 will be matched (or closed out) by an opposite change in the value of the futures. You are, once again, market-neutral.

2.2. The CBOE Volatility Index

In 1993, the first volatility index was introduced by the CBOE to measure the 30-day implied volatility of the S&P 100 index. The old VIX, now called the VXO, took the average of the BSM implied volatilities[6] from eight near-the-money (≈ATM) S&P 100 index options with the two nearest expirations. In 2003, the calculation methodology was changed to provide a 'model-free' measure of S&P 500 index's 30-day expected volatility, Carr and Wu (2006). The methodology is in use to this day, and we aim to provide a sufficiently detailed discussion of it.

The VIX is quoted in volatility percentage points, when the VIX is at 15, it means that the (annualised) 30-day expected volatility of the S&P 500 is 15%. Rather counter-intuitively, it will make more sense to first understand how the VIX-squared works. The VIX^2 is modelled after a variance swap, which is a forward contract on annualised variance expressed in variance percentage points, see Diamond (2012). A buyer and a seller of a variance swap agree on the variance swap rate $\tilde{V}$ when exchanging contracts. On expiration day T, the buyer pays the variance swap rate $\tilde{V}$ and receives the realized variance V_T. The buyer makes a profit when the realized variance is higher than the swap rate, and a loss otherwise. In other words, the buyer locks in on the swap rate—which can be interpreted as the expected value of the variance.

Formally, the CBOE defines the VIX as

$$\text{VIX}^2 \equiv \tilde{V} \tag{1}$$

Here, the VIX^2 is the variance swap rate of a variance swap on the S&P 500. The variance swap rate $\tilde{V}$ in a variance swap is analogous to the predetermined future price in a futures contract; i.e., the variance swap rate $\tilde{V}$ is the expectation of total realized variance V_T on expiration. Mathematically, this is written as

$$\tilde{V} = \mathbb{E}[V_T] = \mathbb{E}\left[\frac{1}{T}\int_0^T \sigma_t^2 dt\right] \tag{2}$$

Rather remarkably, the fair value of $\tilde{V}$ can be calculated directly from out-of-the-money put and call options on the S&P 500 index, see Demeterfi et al. (1999). This is formally expressed as

$$\text{VIX}^2 \equiv \mathbb{E}\left[\frac{1}{T}\int_0^T \sigma_t^2 dt\right] = \frac{2e^{rT}}{T}\left[\int_0^{F_0} \frac{\text{put}_K}{K^2} dK + \int_{F_0}^{\infty} \frac{\text{call}_K}{K^2} dK\right]. \tag{3}$$

Here, F_0 is the forward price of the S&P 500 index at T days in the future, i.e., the market's best guess where the S&P 500 will be at that time. The VIX uses all put options at strikes lower than F_0 and all call options at strikes higher than F_0. All of these options are currently, at the present F_0, out-of-the-money. The price of each option, whether a put or a call, is divided by its strike price squared, K^2.

As the strike K becomes further out-of-the-money for both options, the scaling factor $1/K^2$ dampens the effect that these puts and calls have on the index's final calculation (see Figure 7). Likewise, as the prices of puts and calls clustered near the strike increase, the index weighs them approximately at their face value. It is important to note that it is extremely rare to have the VIX jump because of an increase in purchases of SPX calls near the strike Wang (2021). Instead, the VIX spikes when market players bid up the prices of SPX puts in anticipation that the underlying index will precipitously decline in value in the near future. It is for this reason that the Volatility Index is given the moniker 'fear gauge'.

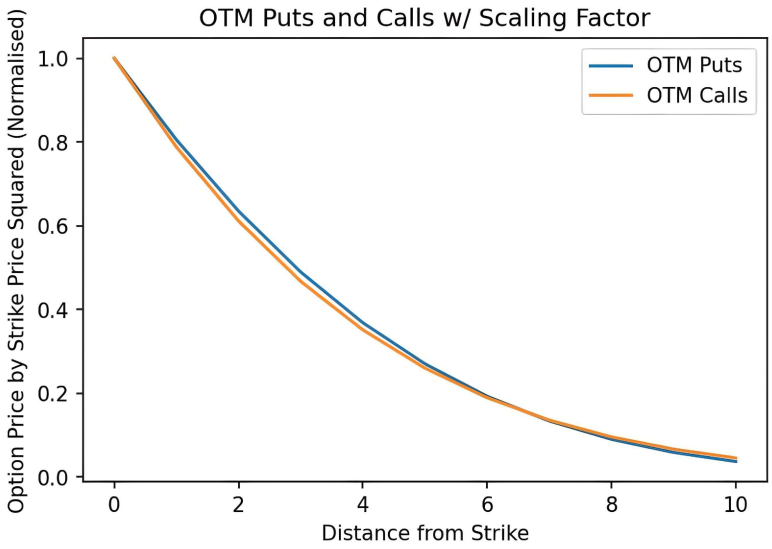


Figure 7. Option Price: USD 100. Volatility: 20%. Interest Rate: 2%. Expiration: 30 Days.

Since there are not infinitely many SPX options, the CBOE applies the following discrete approximation:[7]

$$\text{VIX}^2 = \left[\frac{2e^{rT}}{T}\sum_i \frac{Q(K_i)}{K_i^2}\Delta K_i - \frac{1}{T}\left(\frac{F_0}{K_0} - 1\right)^2\right]. \tag{4}$$

Let us develop a qualitative feel for CBOE's approximation. The integral over a continuum of options with strike prices K changes into a sum of currently traded options with their respective strike price K_i. The sum adds up prices of all available out-of-the-money puts and calls. The price is the midpoint of its bid-ask spread—expressed as the function $Q(K_i)$. The ΔK_i, a discrete modification to dK, is the sum of half the spread between its closest, neighbouring strike prices. There is, however, one catch: at K_0, the first strike below the forward price F_0 of the S&P 500 index, the call option is in-the-money. Indeed, the last term in (4) represents the correction needed to convert this in-the-money call into an out-of-the-money put.[8]

Remember that the new Volatility Index is defined as the square-root of a one-month variance swap rate on the SPX; however, the CBOE only makes use of SPX options with Friday expirations in their calculation. Here, SPX options with more than 23 days and less than 37 days to the Friday SPX expiration are weighted to yield a constant, 30-day measure of the expected volatility of the S&P 500 index.

Since the VIX cannot be traded or replicated (in contrast to stocks and stock market indices which can be either traded or replicated), the standard futures pricing relationship based on 'cash-and-carry' arbitrage[9] does not hold. Hence, any fair-value calculation of a futures contract on the VIX will always be model-dependent.

However, there are upper and lower price bounds on VIX futures. The VIX is defined as the square-root of a one-month variance swap rate, i.e., the variance swap rate expressed in volatility units, on the SPX; it is not the one-month volatility swap rate on the SPX. The difference between the two is given by Jensen's Inequality:

$$\text{VIX} = \sqrt{\mathbb{E}[V_T]} \geq \mathbb{E}\left[\sqrt{V_T}\right] = \text{Volatility Swap Rate} \tag{5}$$

A futures contract on the VIX, at time t with expiration T', is the expectation of square-root of expected total realized variance as shown below (Carr and Wu 2006).

$$\text{VIX futures} = \mathbb{E}_t\left[\text{VIX at time } T'\right] = \mathbb{E}_t\left[\sqrt{\mathbb{E}_{T'}\left[\frac{1}{T}\int_{T'}^{T'+T}\sigma_t^2 dt\right]}\right] \tag{6}$$

The mathematics here serves a symbolic purpose as it keeps the timeline in check. A typical volatility swap allows one to trade between realized and expected volatility. A forward volatility swap allows one to trade between future volatility percentages. While one fixes a volatility swap rate at the inception of a typical volatility swap, the fixing of a volatility swap rate on a forward volatility swap is set at some future time and the settlement day is on some time T' further in the future. A VIX futures contract is not your typical, over-the-counter volatility swap because it looks beyond the VIX, which itself is a forward-looking measure as it constantly gauges the expected square-root of variance of the SPX over the next 30 days. So then what is a VIX future?

By Jensen's inequality, the price of a VIX futures is bounded below by a forward volatility swap rate. Likewise, the price of a VIX futures is bounded above by a forward variance swap rate, expressed in volatility percentage points. In practice, however, the VIX future price is set approximately to a forward volatility swap rate on the VIX. In volatility parlance, a VIX futures is simply referred to as a forward volatility swap.[10] For cases where this is violated, see Van Tassel (2020).

2.3. Volatility as an Asset Class

The quantification of volatility via an index allowed investors to track expected volatility of the broad U.S. stock market. Soon, it led investors to crave for directional exposure on the VIX—but this is the equivalent of trading the market's expectation of future volatility. The initial VIX futures contracts were introduced in 2004, and VIX options followed after in 2006. Both volatility derivatives are settled in cash.

Historically, the correlation between daily returns of the S&P 500 index and daily changes in the VIX is around -0.8, see Macroption (2022). Figure 8 illustrates the negative relationship, leading up to and including Volmageddon.

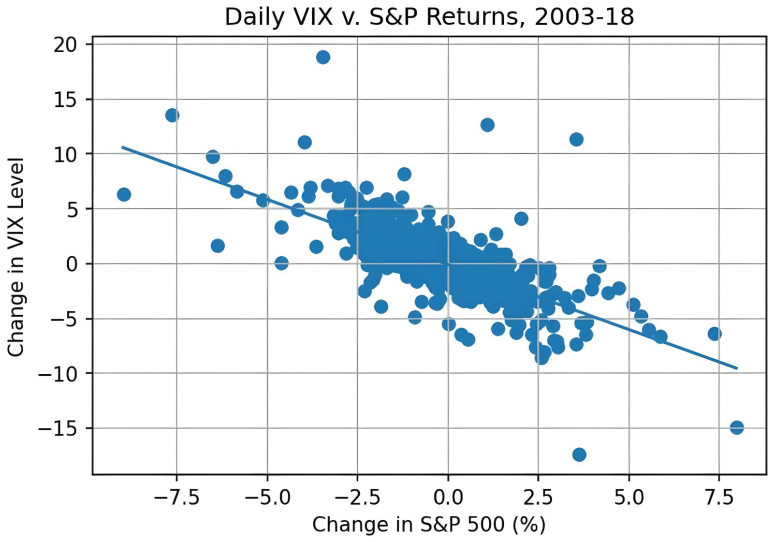


Figure 8. On most days, when the SPX rises (falls), the VIX falls (rises).

Such a correlation exists for a valid reason. A VIX future, in terms of contract specifications, is just like any other futures contract. Namely, the buyer of a VIX future profits if the VIX at some date at or before expiry exceeds the future price, i.e., if the market's

expectation of volatility has increased since the buyer entered the contract. Sudden market drawdowns tend to increase the market's expectation of volatility in the coming 30 days or so, therefore the price of near-term VIX futures shoots up. Likewise, the seller of a VIX futures profits if the market expects lower volatility in the short-term, which is usually associated with a rising market. Hence, the existence of the negative correlation between the VIX and SPX.

Of course correlation is not causality and there are several days where VIX and SPX move in opposite directions (all observations in the lower-left and upper-right quadrant in Figure 8). We also observe several outliers; the most extreme are a VIX movement of +19 resp. -17 while the SPX moved by less than $\pm 4\%$. The former will be discussed in detail in the next section on Volmageddon.

A VIX option contract's specifications adhere to those of any regular option on an underlying. Purchasing a call (put) means to speculate on an increase (decrease) in short-term volatility. Again, wanting limited downside risk as a buyer of an option would naturally entail the issuer demanding an upfront premium in return.

One drawback of entering a longer dated futures contract is that the exposure to changes in future volatility declines as the expiration date approaches. Implementing a futures trading strategy that maintains a defined exposure to the VIX can be challenging for the retail investor. For instance, it may be difficult to manage a portfolio of VIX futures that is supposed to provide a constant 30-day forward-looking exposure to the VIX. This sounds like a job for an index!

The index's aim is clear: track a time-varying basket of VIX futures that maintains an average of 30 days to expiration. This is obtained by continuously adjusting the relative proportion of VIX futures; each day, the index adds exposure to the second month futures contract and reduces exposure to the first (or front) month futures contract, thus maintaining 30 days to maturity. The so-called S&P 500 VIX Short-Term Futures Index (VIX Futures Index) does just that and is maintained by the S&P Dow Jones Indices LLC[11]. Like the S&P 500 index, this index cannot be directly invested in. Instead, VIX exchange-traded product (ETP) promises to track, and deliver, the returns of the VIX Futures Index.

Retail investors can buy ETPs from (and later sell them back to) their issuers, gaining access to an instrument that mimics changes in the 30-day VIX futures return. From the issuer's point of view, keeping the index as an underlying allows the issuer to hedge by taking the opposite side of the trade. It would ensure that the issuer avoided all directional exposure to volatility, but it is not that simple.

The recipe to construct a VIX ETP involves choosing the type of product, and the direction and magnitude of the volatility trade. The two options permitted in each case are, respectively, an exchange-traded note (ETN) or exchange-traded fund (ETF); long or 'inverse'; and leveraged or unleveraged.

For example, a 2X long leveraged VIX ETP with USD 100 million in assets would double the daily gains or losses for its investors by using a margin account to construct a USD 200 million notional position in VIX futures. Similarly, an unleveraged inverse VIX ETP would take short positions in VIX futures without any leverage.

On 29 January 2009, Barclays LLC (Barclays) launched the first VIX ETP—the iPath S&P 500 VIX Short-Term Futures ETN (VXX)—that started trading on 30 January 2009. By 2016, the VXX ETP would be the fifth-most actively traded security in the US stock markets (Wigglesworth 2017). An ETN is simply an unsecured debt obligation. The issuer promises to pay a 30-day VIX future price at a pre-specified maturity date. At the same time, Barclays also launched another ETN which gives exposure to a weighted average VIX futures maturity of 5 months (VXZ). Such ETNs are still actively traded at the time of this writing. On 19 January 2018, Barclays relaunched the iPath Series B S&P 500 VIX Short-Term Futures ETN (VXX) after the previous series hit its maturity the same year. VXX's new maturity is 23 January 2048 and has an annual management fee of 0.89%.

An ETN is different from an ETF. An ETF is secured by assets, e.g., an equity ETF, such as SPX, is backed by some proportion of the 500 stocks. When you own an ETF, you own

a portfolio; when you own an ETN, you own a promise because an issuer can (in theory) invest the investors' money in any asset. For the VXX, the investor is therefore exposed to credit risk, i.e., Barclays' ability to keep its promise. However, there are two main reasons to own an ETN, a favourable tax treatment and—at least in theory—no tracking error, i.e., the value of the ETN follows the underlying index when traded in the secondary market due to arbitrageurs correcting mispricing. ETN issuers tend to also hedge some of the risk they are exposed to (Damato 2009).

In 2011, inverse VIX ETPs, which provided investors with the opportunity to short volatility, followed suit. The most popular inverse VIX ETP at the time was issued by Credit Suisse Group AG (Credit Suisse) and was called the Velocity Shares Daily Inverse VIX Short-Term Futures ETN (XIV). A directional counterpart to the VXX, XIV promised to replicate the inverse one-day return of the VIX Futures Index. For example, if the VIX Futures Index dropped by 10%, XIV would post an end-of-the-day return of $+10\%$ to its investors (see Figure 9). The XIV offered to harvest the 'volatility risk premium' by betting that near-term future, or expected, volatility will be lower than realized volatility. In practice, the XIV carried out its mandate by mechanically selling the same basket of near-term VIX futures that the VIX Futures Index tracks.

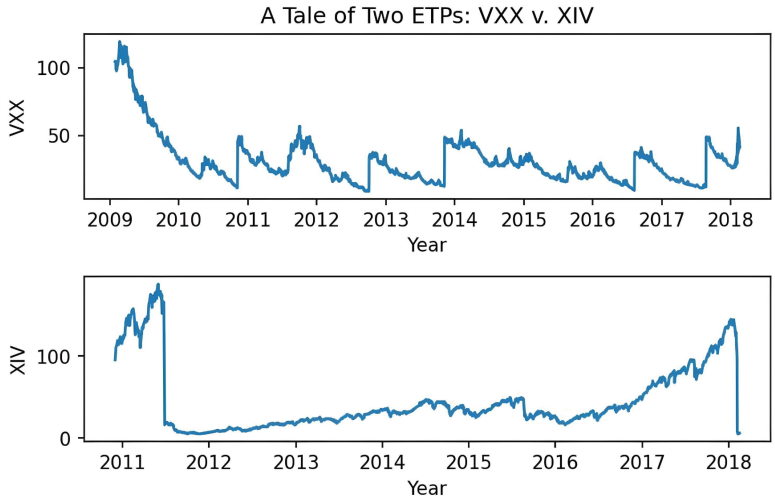


Figure 9. Their inceptions mark the birth of long and short volatility strategies, respectively.

3. Volmageddon

On Monday, 5 February 2018, a day with unremarkable weather in Chicago and New York, the S&P 500 index dropped 4%. Such an event is rare and happens about twice per year on average. Given the historical correlation of -0.8 between the VIX and the SPX, the VIX should have responded with a rise of about 3.2 points. Instead, the VIX spiked by about 20 points, from 17 at open to 37 at close, which marked the largest ever recorded increase in the index in a single day. The VIX's sudden skip, along with its tragic aftermath, would later aptly be dubbed 'Volmageddon'.

3.1. The Big Volatility Short

After the Global Financial Crisis in 2007/2008, investors became painfully aware of the need to hedge against 'tail risk'. This risk refers to the highly unlikely probability of extreme movements in asset prices.[12] If one is exposed to such risk (as most equity investors are), one can suffer drastic losses in a short time period.

Given the historically negative correlation between the VIX and the S&P 500 index, many market participants conjectured that VIX futures—if properly traded—could protect them against tail risk. A corollary is that longing the S&P 500 index is equivalent to (implicitly) shorting volatility. Therefore, a long VIX futures position can hedge away potential volatility shocks in a portfolio resembling the S&P 500 Index.

This hedging tactic gained traction after the the launch of the first long VIX exchange-traded product (VXX) in 2009. While VIX futures were introduced a full five years earlier, it required a financial crisis and a complex financial instrument to track them to make their presence known to risk managers. The previous strategy of manually 'rolling' VIX futures contracts, i.e., paying monthly insurance premiums to tame the tail risk, was seen as a messy way of trading them.

After a couple of years of low volatility in 2009–2011 (see Figure 10), insuring against an event that did not occur started to lose its appeal. Investors began to ask: why not switch sides? Shorting VIX futures contracts mean they would collect, as opposed to pay, periodic insurance premiums. The introduction of the first inverse VIX ETP (XIV), in 2011, made this trade available to the general public. Around this time, however, the European debt crisis led to considerable economic uncertainty, and the VIX reached a two-year high of 48. After this spike, the VIX would stay low for the next seven years until 2018.[13]

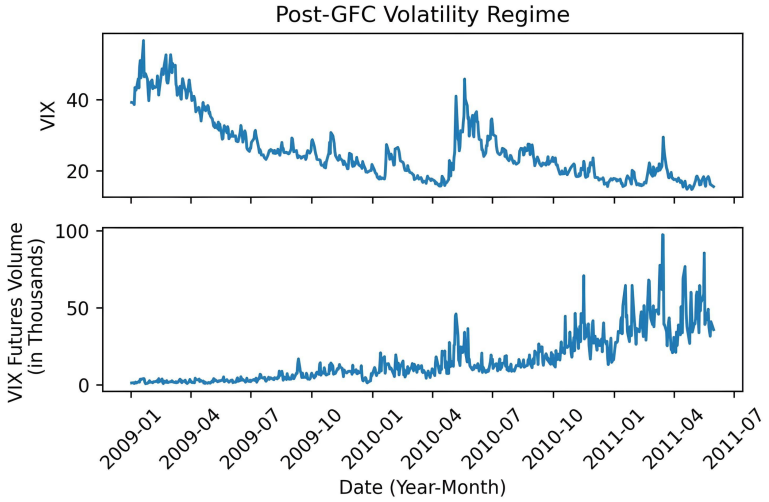


Figure 10. As the VIX level steadily declined, VIX futures volume picked up.

The start of the new volatility regime quickly conjured up a self-reinforcing prophecy: shorting the VIX—via options and futures—drives down the index and, thus, benefits those who are short, making betting against rising volatility even more attractive.

Conceptually, trading volatility with these products is similar to buying or selling a volatility swap; an investor who is looking to short volatility sells implied volatility and receives realized volatility at expiration. The volatility risk premium, i.e., expected volatility less realized volatility, is what the short seller seeks to capture or 'harvest'. Sellers of volatility collect periodic premiums, from buyers of volatility, in exchange for warehousing tail-risk. Unique to volatility markets, shorts make long-term bets while longs make short-term bets. However, the spread between implied and realized volatility narrows when selling insurance becomes a crowded trade.

Realistically, the dynamics of holding a short position against an index using futures contracts is different. Suppose you short 600 futures contracts, each with a notional value of USD 100, at a futures price of USD 10. If the futures index price goes up by 33%, your net asset value (NAV) drops from USD 60,000 to USD 40,000. However, the notional value of your position has become $-600 \times \text{USD } 100 \times 1.33 = -\text{USD } 80,000$. To restore the short position to your new NAV of USD 40,000, you have to buy futures at notional USD 40,000, i.e., USD 40,000/USD 100 = 400 contracts. If you are short, an adverse market move requires you to buy futures. If, instead, the market moves in your favour (falls), then you will have to sell more futures. In summary, an inverse VIX ETP (i.e., leverage -1) would sell (buy) if the price of futures goes down (up). A leveraged long VIX ETP (i.e., leverage > 1), rather surprisingly, works the exact same way! If the market moves up, the issuer has to go further long and buy futures contracts. If the market goes down, the issuer

has to sell futures to reduce the notional amount of the position. Therefore, whether the exchange-traded product is inverse or leveraged long, both trade in the same direction in the futures market, buying when broad market goes up and selling when it goes down. For this reason, both inverse and leveraged long are simply referred to as leveraged VIX ETPs (Augustin et al. 2021). It does not take a genius to realize that this strategy can create a feedback loop if the issuers of leveraged ETPs are large enough and all trade in the same direction.

At the time, it was slowly becoming clear to certain market players that issuers of leveraged VIX ETPs were, themselves, the largest traders of VIX futures. Since issuers also generate revenue from management fees, and not from directional bets on volatility, they are incentivised to issue big or go home. Market players also knew when and how often issuers of VIX ETPs would bid up or down the prices of VIX futures contracts: two minutes before the close at 4:00 pm ET, every trading day!

It was a smart trading strategy for those who knew what to do, given that leveraged VIX ETPs had to buy VIX futures in the event of a VIX spike, simply buy those VIX futures earlier for cheap and sell them to their issuers later for more.

The scene was set for the seven-year volatility regime to come to an abrupt end. All the VIX needed to do now was vault higher than the market expected to set off a catastrophic feedback loop. On Friday, 2 February 2018, U.S. stocks sold off as investors worried that the Federal Reserve might raise interest rates faster than needed (Kawa 2019). The VIX began to rise.

3.2. The Volatility Storm

On Monday, 5 February 2018, US stock prices fell sharply as the S&P 500 declined by more than 4%. Amidst the stock market volatility, the VIX experienced its largest ever daily jump, rising 115% from 17 to 37. Credit Suisse's XIV opened at USD 110 on that day, and closed around 10% lower at 4:00 pm ET. During after-hours, however, the price of XIV precipitously declined and had lost another 90% by around 6:30 pm ET (Franck 2018). So what happened? The XIV ETN became a victim of its own success.

We will use approximate numbers for a simpler illustration of what happened during market hours. For a more detailed account of that trading day, see Augustin et al. (2021). Recall that XIV was an exchange-traded note, with Credit Suisse as its issuer, that promised to track the inverse of the VIX short-term futures index. The prevalent theory is that XIV carried out the mechanical rebalancing strategy—as explained previously—at the end of the trading day (Augustin et al. 2021). As of Monday morning, the VIX stood at 15 and XIV had 15 million shares outstanding at a value of approximately USD 100 per share, or USD 1.5 billion in net asset value. Credit Suisse's XIV, tracking the inverse of the Index, maintained an appropriate short VIX futures position worth USD 1.5 billion. So, for instance, if VIX futures jumps from 15 to 20 (a 33% increase), XIV's net asset value drops by 33% to USD 1 billion. Post-jump, however, Credit Suisse's short position now has a 'notional' value of USD 2 billion. In short, there is a mismatch between XIV's NAV (which adjusts real-time as investors reprice the ETN's shares accordingly) and Credit Suisse's hedge (which needs to line up with NAV by the end of the trading day). In this specific VIX jump, Credit Suisse has to cut its volatility exposure in half by *buying* USD 1 billion VIX futures.

Towards the end of market's close, market players—aiming to take advantage of the end-of-day hedging by issuers—started to bid up the price of VIX futures at 3:30 pm ET. The snowball effect then began after the large mechanical nature of the rebalancing. By 4:15 pm, near the close of the futures market in Chicago, the prices of the VIX futures had spiked.

Drastic market movements are, by their very nature, extremely rare. A sharp downward spiral can sometimes have a computer protocol kick in and temporarily halt trading in order to prevent an all-out market panic. The VIX Futures Index, as per its provider S&P Dow Jones Indices LLC (SEC v. S&P Dow Jones Indices LLC 2021), used a software that

paused real-time updates in the event that the index moves by more than a pre-specified amount. The reasoning behind such a mechanism is obvious: if an index moves more than deemed 'normal', a human will need to intervene to ensure the price swing is not a computational mistake. It can be understood as a quality-control measure. On the day of 5 February 2018, prices of the underlying VIX futures fluctuated, and the index with it. Throughout the trading day there was no halting measure imposed by the index as there was no freakish market move. After market hours, however, between 4:00 pm and 5:08 pm, the VIX Futures Index told a different story (Levine 2018).

At 4:08 pm, in the span of one minute, 115,862 VIX futures contracts (or about 25% of the entire market) were transacted (Sushko and Turner 2018). A minute later, at 4:09 pm ET, the computer system stopped updating the S&P 500 VIX Short-Term Futures Index. For the next three minutes, the last reported value of 86 continued to be disseminated. At 4:12 pm ET, the index was updated in real time but froze again at 87 from 4:13–4:35 pm.

In this 22-minute interval, XIV's published intraday indicative value[14] was fixed at approximately USD 25 per share. This was not its true indicative value because the underlying index failed to update real-time changes in the VIX futures market. Investors were fooled into believing that the XIV had weathered the volatility storm on that day. From 4:35–5:08 pm ET, the XIV was sporadically updated with values between USD 24 and USD 27 per share. At 5:09 pm, XIV's closing indicative value was finally published. It was a meagre USD 4.22 per share, marking a 95% decline from its open on the day. Between 4:09 pm and 5:09 pm, USD 700 million in XIV shares were traded on the secondary market in the after-hours session. Buyers overpaid dramatically during this hour (Levine 2021). While VIX futures doubled and wiped out the value of XIV, investors of the note were completely oblivious.

The story of XIV's issuer during Volmageddon was not as tragic. After the VIX's unprecedented increase during the day, XIV's issuer scrambled to align their hedged positions with their ETP's net asset value by buying VIX futures at the day's close. Naturally, this inadvertently pushed up the price of futures, further decreasing their net asset values and forcing them to buy even more futures. In the most extreme case, the issuer would have mechanically rebalanced their hedge until their wealth hits zero. Fortunately, there is some legal leeway buried in their prospectus where it is stated that they can redeem shares (essentially returning capital to the investors and exiting the market) in the event that the underlying short-term VIX futures index moves by more than 80% in a single day (XIV Prospectus 2018). On 21 February 2018, Credit Suisse announced the termination event and redeemed notes at USD 5.99 each (Stempel 2021). Since they earn fees from managing the product on behalf of investors, issuers did not have much to complain about: they simply promised to track the underlying index, be it broken or not.

This misinformation about XIV's indicative value was an unintended consequence of the so-called 'Auto Hold' feature embedded in the S&P 500 Short-Term VIX Futures Index. The irregular pauses in the dissemination of real-time information caused substantial damage. Unfortunately, the index provider failed to disclose the existence of such a feature when it licensed the index to issuers. In May 2021, three years later, S&P Dow Jones Indices LLC agreed on a USD 9 million settlement with the U.S. Securities and Exchange Commission, see SEC v. S&P Dow Jones Indices LLC (2021).

What is clear is that trading VIX futures for hedging purposes can break the VIX Futures Index and ultimately affect the VIX note, which relies on the index to calculate its fair value. Additionally, hedging via VIX futures will always go against the performance of these notes. It is no surprise that investors who lost their shirts in February 2018 went on to accuse Credit Suisse of market manipulation. Although the initial class action lawsuit was dismissed, the appeal against that decision was successful (Stempel 2021). According to the case brought to the U.S. Court of Appeals for the Second Circuit (Chahal v. Credit Suisse Group AG 2021), the claimants say they lost USD 1.8 billion while the issuer made a profit of USD 475 million.

3.3. Vol 'Til You Bawl

Three years after the delisting of XIV, in October 2021, Dynamic Shares Trust was given permission by the SEC to list the Dynamic Short Short-Term Volatility Futures ETF (WEIX). Its promise is the exact same as that of the demised XIV, to provide investors with the inverse daily returns of the S&P 500 VIX Short-Term Futures Index. However, Dynamic Shares Trust assured investors that this product is actively managed to "provide better risk management than passively managed short VIX short-term futures ETFs" (WEIX Prospectus 2022). In particular, WEIX will calculate its closing indicative value by using a time-weighted average price of 15 min to 4:00 pm ET rather than the futures settlement price, which is determined 2 min to close. This broader rebalancing period is a safety measure applied to prevent a spike (once again) in the VIX futures prices near closing (Peterseil and Greifeld 2021).

It is difficult to say with certainty whether the new rebalancing method will prove to be an effective solution. Most forms of rebalancing, i.e., periodically rolling over contracts, are inherently price-insensitive, the buying and selling is carried out without any heed to the asset's price. If these mechanical trades are large enough, they will rapidly tear apart the very fabric of the underlying market, such as on 5 February 2018 and 19 October 1987.

We invite the reader to also look at the events during the week that ended on Friday, 20 March 2020, when the U.S. stock market found its floor. We postulate that a similar mechanical hedge rebalancing, this time performed by SPX options dealers, could have been the reason for the stock market drop. The so-called 'Gamma Hedging' is claimed to sway the S&P 500 index even to this day (Wang 2022). In options-speak, gamma measures how the delta of an option changes as the price of the underlying changes; in theory, a gamma hedge protects the investor from all of the underlying's price changes. An SPX options dealer, as a market maker, avoids directional exposure associated with the trade by hedging with respect to an option's gamma. This sounds awfully price-insensitive. Table 1 illustrates all of the similar hedging strategies deployed by VIX ETP issuers and SPX options dealers.

Table 1. Mechanical Hedge Rebalancing—VIX ETP Issuers and SPX Options Dealers.

	VIX ETP Issuers	**SPX Options Dealers**
Price-Insensitive Hedge	✓	✓
Buy into Rise, Sell into Decline	✓	✓(Short Gamma)
Rebalancing Period	End of Trading Day	Third Friday of the Month
Day of Major Event	5 February 2018	20 March 2020

Roughly USD 3.2 trillion options, of which USD 1.7 trillion were SPX options, expired on the third Friday of September 2022. The monthly options expiration (OpEx) forces holders to either mechanically roll over existing positions or open new ones. In recent years, OpEx has become an important market phenomenon for market players (of all sizes) to try and trade around (Wang 2022). It is unclear if the tail will wag the underlying dog, again; nevertheless, it is clear that the more things change in the financial markets, the more things seem to stay the same.

4. Conclusions

This survey paper threw light on the Volatility Complex and presented the ill-famed Volmageddon as an instructive case study for investors. We described the calculation methodology of the CBOE VIX and the pricing techniques applied to the index's higher order derivatives. We explained the reasoning behind the post-2008 strategy of harvesting the volatility risk premium via VIX futures and ETPs.

Our analysis sought to reveal the pernicious impact that price-insensitive hedge rebalancing by major market participants can have on the market microstructure. We

posited that systematic gamma hedging by options dealers may have created a similar negative feedback loop that resulted in the market bottom on the third Friday of March 2020. We hope our intuitive analysis of the SPX options market promotes future research in this area.

Author Contributions: Conceptualization, C.V. and K.R.S.-H.; methodology, C.V. and K.R.S.-H.; writing—original draft preparation, C.V. writing—review and editing, C.V. and K.R.S.-H.; supervision, K.R.S.-H.; funding acquisition, K.R.S.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by The University of Manchester's SEI programme.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. VIX Approximation

The VIX equation uses an infinite number of out-of-the-money call and put options. In reality there are far fewer as the strike ladder on the SPX is in 5 point increments (CBOE 2022), and not all out-of-the-money options are actively traded. Therefore, an approximation is needed. However, there is another snag. The separation between puts and calls is determined by the current forward price of the underlying index SPX, F_0, an observable market price. Denoting by K_0 the highest strike on the option strike ladder that does not exceed F_0, the CBOE approximation CBOE (2019) includes one in-the-money call if $F_0 > K_0$. In order not to throw away any data, one keeps this option price and adjusts the approximation accordingly.

An alternative way to approximate VIX is to change the integration limit to K_0 from F_0, work out the correction needed, and then translate into a discrete version that uses actual option prices.

Starting with Equation (3), one has

$$\mathrm{VIX}^2 = \frac{2e^{rT}}{T}\left[\int_0^{K_0} \frac{\mathrm{put}_K}{K^2}dK + \int_{K_0}^{\infty} \frac{\mathrm{call}_K}{K^2}dK\right] + \frac{2e^{rT}}{T}\left[\int_{K_0}^{F_0} \frac{\mathrm{put}_K - \mathrm{call}_K}{K^2}dK\right]. \tag{A1}$$

The first term in brackets is the one with changed integration bounds and the second term is the 'correction'. The put-call parity implies

$$\mathrm{put}_K - \mathrm{call}_K = e^{-rT}[K - F_0].$$

Therefore, the last bracketed term of (A1) can be written as

$$\frac{2}{T}\left[\int_{K_0}^{F_0} \frac{K - F_0}{K^2}dK\right] = \frac{2}{T}\left[1 - \frac{F_0}{K_0} + \ln(F_0/K_0)\right] \approx -\frac{1}{T}\left(\frac{F_0}{K_0} - 1\right)^2 \tag{A2}$$

where the last approximation is derived from the second-order Taylor expansion around 1. Define $f(x) = 1 - x + \ln(x)$. The second-order Taylor expansion at $\bar{x}$ is given by $f(\bar{x}) + f'(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2}f''(\bar{x}) \cdot (x - \bar{x})^2$. Since $f(1) = 0$, $f'(1) = 0$ and $f''(1) = -1$, one finds $f(F_0/K_0) \approx -(F_0/K_0 - 1)^2$ which means

$$1 - \frac{F_0}{K_0} + \ln(F_0/K_0) \approx -\frac{1}{2}(F_0/K_0 - 1)^2.$$

Thus, one obtains the CBOE definition (4). As the difference between F_0 and K_0 is usually small, the (second-order) approximation error is negligible.

Notes

[1] However, would you exchange an asset with a 15% guaranteed one-year return for this one? A 'risk-neutral' trader would say yes; the payoff in either case is 15% (realized or not).

[2] In reality, some money is required upfront. The 'initial margin' is a percentage of the purchase price that must be covered by the investor's own money. The 'maintenance margin requirement' is the amount of money the investor is required to maintain to keep her position open.

[3] This is a European option. If one wanted the greater privilege of conducting the transaction on or before the future date, one will have to purchase an American option via a larger upfront payment.

[4] Buying an option is like buying insurance, so the seller of the insurance collects an initial premium in exchange for paying up in case of the insured event happening in the future.

[5] The first name is in honour of Fischer Black, who passed away two years before in 1995 and, therefore, made him ineligible to receive the prize.

[6] The model works like this. There are five input variables, current underlying price, strike price, risk-free interest rate, time to maturity, and volatility. There is one output variable, option price. Given four out of five input variables and the current price of the option, one can reverse-engineer to obtain the implied (embedded) volatility in the option.

[7] The CBOE's VIX White Paper (CBOE 2019) provides a step-by-step calculation with a sample set of SPX options prices.

[8] An alternative derivation is provided in Appendix A.

[9] See Sinclair (2013) for more information.

[10] Alas, one has no choice but to become used to the financial 'Sprachspiel'!

[11] Similar to the CBOE's treatment of VIX, the level of the index is calculated by plugging the prices of VIX futures into a formula.

[12] Generally, this is defined as the (tiny) chance that the price of an asset swings by three standard deviations or more. For a normally distributed return, a tail-risk event has an equal 0.3% chance of occurring on the upside or downside.

[13] Mostly low would be more correct as another hop occurred in August 2015 after the devaluation of China's currency.

[14] The value was calculated by Janus Henderson Group Plc, which marketed XIV for Credit Suisse.

References

Augustin, Patrick, Ing-Haw Cheng, and Ludovic Van den Bergen. 2021. Volmageddon and the Failure of Short Volatility Products. *Financial Analysts Journal* 77: 35–51. [CrossRef]

Carr, Peter, and Liuren Wu. 2006. A Tale of Two Indices. *The Journal of Derivatives* 13: 13–29. [CrossRef]

CBOE. 2019. VIX White Paper. Available online: https://cdn.cboe.com/api/global/us_indices/governance/Volatility_Index_Methodology_Cboe_Volatility_Index.pdf (accessed on 1 September 2022).

CBOE. 2022. Delayed Quotes for SPX Options. Available online: https://www.cboe.com/delayed_quotes/spx/quote_table (accessed on 28 July 2022).

Chahal v. Credit Suisse Group AG. 2021. United States Court of Appeals for the Second Circuit. Available online: https://www.govinfo.gov/app/details/USCOURTS-ca2-19-03466 (accessed on 10 October 2022).

Damato, Karen. 2009. A Very Different Animal. *Wall Street Journal*, December 8. Available online: https://www.wsj.com/articles/SB10001424052748704402404574526042449793288 (accessed on 10 October 2022).

Demeterfi, Kresimir, Emanuel Derman, Michael Kamal, and Joseph Zou. 1999. More Than You Ever Wanted to Know about Volatility Swaps (But Less Than Can Be Said). Available online: https://emanuelderman.com/wp-content/uploads/1999/02/gs-volatility_swaps.pdf (accessed on 10 October 2022).

Diamond, Richard V. 2012. VIX as a Variance Swap. *SSRN Electronic Journal*. Available online: https://doi.org/10.2139/ssrn.2030292 (accessed on 1 September 2022).

Franck, Thomas. 2018. Obscure Security Linked to Stock Volatility Plummets 80% After Hours, Sparking Worries of Bigger Market Effect. *CNBC*, February 5. Available online: https://www.cnbc.com/2018/02/05/xiv-exchange-traded-security-linked-to-volatility-plummets-80-percent.html (accessed on 1 October 2022).

Kawa, Luke. 2019. The Day The VIX Doubled: Tales of 'Volmageddon'. *Bloomberg*, February 5. Available online: https://www.bloomberg.com/news/articles/2019-02-06/the-day-the-vix-doubled-tales-of-volmageddon (accessed on 1 October 2022).

Levine, Matt. 2018. Inverse Volatility Products Almost Worked. Bloomberg, February 9. Available online: https://www.bloomberg.com/opinion/articles/2018-02-09/inverse-volatility-products-almost-worked (accessed on 1 October 2022).

Levine, Matt. 2021. S&P Forgot to Update the XIV Index. Bloomberg, May 18. Available online: https://www.bloomberg.com/opinion/articles/2021-05-18/s-p-firm-forgot-to-update-the-xiv-index (accessed on 1 October 2022).

Macroption. 2022. VIX to S&P500 Correlation: Trends, Seasonality, Weekdays. Available online: https://www.macroption.com/vix-spx-correlation/ (accessed on 2 September 2022).

Peterseil, Yakob, and Katherine Greifeld. 2021. The Ghost of "Volmageddon" is Back to Haunt New Volatility Funds. *Bloomberg*, May 4. Available online: https://www.bloomberg.com/news/articles/2021-05-04/the-ghost-of-volmageddon-is-back-to-haunt-new-volatility-funds (accessed on 1 September 2022).

Peterseil, Yakob, and Luke Kawa. 2019. Stock Volatility Isn't Dead. It's Just Got Freakish and Extreme. *Bloomberg*, May 3. Available online: https://www.bloomberg.com/news/articles/2019-05-03/stock-volatility-isn-t-dead-it-s-just-got-freakish-and-extreme (accessed on 1 October 2022).

SEC v. S&P Dow Jones Indices LLC. 2021. Securities and Exchange Comission issues Cease and Desist against S&P Dow Jones Indices LLC. Available online: https://www.sec.gov/litigation/admin/2021/33-10943.pdf (accessed on 19 September 2022).

Sinclair, Euan. 2013. *Volatility Trading*, 2nd ed. Hoboken: John Wiley & Sons, Inc., Chapter 12, pp. 223–30.

Stempel, Jonathan. 2021. Credit Suisse Must Face Lawsuit Over U.S. 'Volatility' Crash. *Reuters*, April 27. Available online: https://www.reuters.com/business/finance/credit-suisse-must-face-lawsuit-over-us-volatility-crash-2021-04-27/ (accessed on 10 October 2022).

Sushko, Vladyslav, and Grant Turner. 2018. The Equity Market Turbulence of 5 February—The Role of Exchange-traded Volatility Products. *BIS Quarterly Review* 1: 4–6.

Van Tassel, Peter. 2020. The Law of One Price in Equity Volatility Markets. Available online: https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr953.pdf (accessed on 8 September 2022).

Wang, Lu. 2021. Options Craze Rewriting Rules of VIX, S&P 500 Relationship. *Bloomberg*, November 8. Available online: https://www.bloomberg.com/news/articles/2021-11-08/options-craze-is-rewriting-rules-of-vix-s-p-500-relationship (accessed on 1 October 2022).

Wang, Lu. 2022. A $2 Trillion Stock-Options Deadline Is Make-Or-Break Moment for Bulls. *Bloomberg*, August 18. Available online: https://www.bloomberg.com/news/articles/2022-08-18/a-2-trillion-options-deadline-is-make-or-break-moment-for-bulls (accessed on 2 October 2022).

WEIX Prospectus. 2022. Dynamic Shares Trust. Available online: https://sec.report/Document/0001771951-22-000021/ (accessed on 1 October 2022).

Wigglesworth, Robin. 2017. The Fearless Market Ignores Perils Ahead. *Financial Times*, April 18. Available online: https://www.ft.com/content/099ebfe2-2061-11e7-a454-ab04428977f9 (accessed on 1 October 2022).

XIV Prospectus. 2018. Credit Suisse AG Announces Event Acceleration of its XIV ETNs. Available online: https://www.credit-suisse.com/about-us-news/en/articles/media-releases/credit-suisse-announces-event-acceleration-xiv-etn-201802.html (accessed on 1 October 2022).

risks

MDPI

Article

A Diversification Framework for Multiple Pairs Trading Strategies

Kiseop Lee [1], Tim Leung [2] and Boming Ning [1,*]

[1] Department of Statistics, Purdue University, West Lafayette, IN 47906, USA; kiseop@purdue.edu
[2] Department of Applied Mathematics , University of Washington, Seattle, WA 98195, USA; timleung@uw.edu
* Correspondence: ningb@purdue.edu

Abstract: We propose a framework for constructing diversified portfolios with multiple pairs trading strategies. In our approach, several pairs of co-moving assets are traded simultaneously, and capital is dynamically allocated among different pairs based on the statistical characteristics of the historical spreads. This allows us to further consider various portfolio designs and rebalancing strategies. Working with empirical data, our experiments suggest the significant benefits of diversification within our proposed framework.

Keywords: pairs trading; Ornstein–Uhlenbeck process; diversification; portfolio allocation; mean reversion budgeting

Citation: Lee, Kiseop, Tim Leung, and Boming Ning. 2023. A Diversification Framework for Multiple Pairs Trading Strategies. *Risks* 11: 93. https://doi.org/10.3390/risks11050093

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 3 April 2023
Revised: 11 May 2023
Accepted: 12 May 2023
Published: 16 May 2023

1. Introduction

Pairs trading is a widely used strategy for traders and fund managers. It involves taking simultaneous positions in two correlated assets and speculating on the path behavior of the resulting spread. While it is difficult to fully model the dynamics of a single asset, a pair of assets or securities may exhibit mean reverting behavior that can be better captured by statistical models.

Each pairs trading strategy involves three main steps: (i) identification of assets, (ii) formation of spreads, and (iii) design of trading rules. While there are various ways to select assets for pairs trading, the two price processes for each pair should exhibit an adequate degree of comovement so that the resulting spread is mean reverting. In such a case, trading opportunities arise when the spread deviates from its mean.

We adopt a statistical approach to pairs trading. For any given pair of stocks, we construct the spread such that it best fits the Ornstein–Uhlenbeck (OU) model. Specifically, we determine the optimal ratio between two stocks, along with the model parameters, such that the resulting spread time series achieves the maximum likelihood. This method is an extension of the maximum likelihood estimation (MLE) approach which typically only determines the model parameters. With the spread formed, we apply a set of trading rules. The critical levels to enter and exit are set at a multiple of the standard deviation from the long-term mean of spread. For more details, we refer to Lee and Leung (2020) and references therein.

To our best knowledge, existing studies on pairs trading analyze the performance of a single pair, rather than the aggregate performance of multiple pairs. There are several practical benefits of trading multiple pairs and considering them together in a trading program. Firstly, more spreads may give rise to more trading opportunities at any point in time. Secondly, the spreads generated from different stocks are likely to be mostly uncorrelated, which makes it an ideal setting to take advantage of diversification. Lastly, trading multiple pairs also opens up a new direction for portfolio optimization, which does not exist when trading a single pair.

In this paper, we propose a novel framework for constructing diversified portfolios from multiple pairs trading strategies. Capital is allocated among different pairs based

on the statistical characteristics, such as the speed of mean reversion and volatility, of the historical spreads. Moreover, our approach is adaptive as portfolio weights are adjusted periodically. Among our allocation methods, we introduce the novel Mean Reversion Budgeting (MRB) and Mean Reversion Ranking (MRR) methods. The MRB method determines the portfolio weights based on the speeds of mean reversion, volatilities, and estimated average log-likelihood scores together. In contrast, the MRR method ranks spreads based on their estimated likelihood scores and speeds of mean reversion and assigns prespecified portfolio weights based on the rankings of the spreads.

Working with the empirical price data of six stock pairs, our experiment suggests that the proposed framework offers some desirable return profiles and portfolio features. We compare several allocation methods to the benchmark equal-weight portfolio and illustrate how dynamic rebalancing can improve portfolio performance.

The rest of the paper is organized as follows. We provide a review of related studies in Section 2. In Section 3, we outline the steps of portfolio construction and trading rules for mean reversion trading. As part of Section 4, we demonstrate how the volatility of spreads and speed of mean reversion affect the performance of simulated trades. In Section 5, we describe the proposed diversification framework. Section 6 describes the asset pairs included in our study, and our data collection process, and compares the performance of the proposed framework to the baseline. Concluding remarks are provided in Section 7.

2. Related Studies

Examples of mean-reverting spreads can be found in a number of empirical studies. They are generated from pairs of stocks and ETFs (Gatev et al. 2006; Avellaneda and Lee 2010; Montana and Triantafyllopoulos 2011; Leung and Li 2016), futures contracts (Brennan and Schwartz 1990; Dai et al. 2011), physical commodity and commodity stocks/ETFs (Kanamura et al. 2010), as well as cryptocurrencies (Leung and Nguyen 2019).

Gatev et al. (2006) provided one of the first in-depth studies on pairs trading. They proposed a commonly used Distance Method (DM) and test the CRSP stocks from 1962 to 2002. The DM method opens a position for a pair when the prices diverge from 0 by more than two historical standard deviations and closes the position at the next crossing of the prices. They reported an excess return of 1.3% for the top 5 pairs of the DM, and 1.4% for its top 20 pairs. Do and Faff (2012) examined the profitability of pairs trading accounting for transaction costs.

Other than the DM method, the cointegration test is commonly used in many alternative methods for mean reversion trading. Vidyamurthy (2004) detailed a cointegration framework for mean reversion trading based on the Engle and Granger's error correction model representation of cointegrated series presented in Engle and Granger (1987). Galenko et al. (2012) examined an active ETF trading strategy based on cointegrated time series. Leung and Nguyen (2019) constructed cointegrated portfolios of cryptocurrencies using the Engle–Granger two-step approach and Johansen cointegration test. Huck and Afawubo (2015) compared the performance of DM and cointegration method using the components of the S&P 500 index.

Another popular approach, the stochastic spread method, captures the path behavior of the spread through a stochastic process with mean-reverting property, such as the Ornstein–Uhlenbeck (OU) process. The construction of spreads and extraction of trading signals are typically derived from the analysis of parameters of the underlying model. Elliott et al. (2005) proposed a mean-reverting Gaussian Markov chain model to describe spread dynamics. The model's estimates are compared with observations of the spread to determine appropriate trading decisions. Do et al. (2006) further analyzed the method proposed by Elliott et al. (2005) and proposed a general stochastic residual spread method to model relative mispricing. Leung and Li (2015) studied the optimal timing strategies for mean-reverting trading under the OU process. They solved an optimal double-stopping problem to analyze the timing of entry and liquidation subject to trans-

action costs. Lee and Leung (2020) examined the performance of mean reversion trading using dynamically optimized entry and exit rules.

Additionally, there are methods that utilize copulas (Liew and Wu 2013; Xie et al. 2016), Principal Component Analysis (PCA) (Avellaneda and Lee 2010), and machine learning (Guijarro-Ordonez et al. 2021). In recent years, new optimization algorithms have been proposed to generate spreads with maximum in-sample mean reversion. Some of these methods can be automated to analyze a large number of stocks simultaneously (d'Aspremont 2011; Leung et al. 2020).

Finally, we note that all the studies mentioned above focus on the trading performance of a single pair. The proposed framework herein is designed for trading multiple pairs simultaneously and is optimized as a diversified portfolio.

3. Pairs Construction and Trading Rules

We construct a long-short position with two highly correlated assets S_t^1 and S_t^2. The portfolio value is given by

$$X_t = S_t^1 - BS_t^2,$$

where the coefficient B is called the hedge ratio and can be optimized. Following the procedure detailed in Leung and Li (2016, chp. 2), we determine the optimal pair ratio through maximum likelihood estimation (MLE), whereby the resulting historical spread time series best fits the Ornstein–Uhlenbeck (OU) model.

An OU process is a mean-reverting process, described by the following stochastic differential equation:

$$dX_t = \mu(\theta - X_t)dt + \sigma dW_t, \tag{1}$$

where $\mu \in \mathbb{R}$ represents the speed of mean reversion, $\theta \in \mathbb{R}$ is the long-term mean, and $\sigma > 0$ is the volatility parameter. Here, W_t is a standard Brownian motion under the historical measure $\mathbb{P}$.

3.1. Statistical Estimation for Optimized Mean Reversion

For any given hedge ratio b, consider the observed time series of the resulting spread $X_t = S_t^1 - bS_t^2$ up to time t_0. Then, we apply the method of maximum likelihood estimation to determine the optimal parameters for the OU model based on the historical data. Specifically, given $x_i = X_{t_i}$ with time increment $\Delta t = t_i - t_{i-1}$, define the average log-likelihood function of the past L observations as

$$\begin{aligned} l(\mu, \theta, \sigma, b) &= \frac{1}{L} \sum_{i=t_0-L+1}^{t_0-1} \log f(x_i | x_{i-1}; \mu, \theta, \sigma) \\ &= -\frac{1}{2}\ln(2\pi) - \ln(\tilde{\sigma}) - \frac{1}{2L\tilde{\sigma}^2} \sum_{i=t_0-L+1}^{t_0-1} [x_i - x_{i-1}e^{-\mu\Delta t} - \theta(1 - e^{-\mu\Delta t})]^2, \end{aligned} \tag{2}$$

where

$$\tilde{\sigma}^2 = \sigma^2 \frac{1 - e^{-2\mu\Delta t}}{2\mu}.$$

To express the OU parameter values that maximize the average log-likelihood in (2), we define the following:

$$X_x = \sum_{i=1}^{n} x_{i-1},$$

$$X_y = \sum_{i=1}^{n} x_i,$$

$$X_{xx} = \sum_{i=1}^{n} x_{i-1}^2,$$

$$X_{xy} = \sum_{i=1}^{n} x_{i-1} x_i,$$

$$X_{yy} = \sum_{i=1}^{n} x_{i-1}^2.$$

In turn, the optimal parameter estimates under the OU model are given explicitly by

$$\theta^* = \frac{X_y X_{xx} - X_x X_{xy}}{n(X_{xx} - X_{xy}) - (X_x^2 - X_x X_y)},$$

$$\mu^* = -\frac{1}{\Delta t} \ln \frac{X_{xy} - \theta^* X_x - \theta^* X_y + n(\theta^*)^2}{X_{xx} - 2\theta^* X_x + n(\theta^*)^2},$$

$$\begin{aligned}(\sigma^*)^2 = &\frac{2\mu^*}{n(1 - e^{-2\mu^* \Delta t})}(X_{yy} - 2e^{-2\mu^* \Delta t} X_{xy} + e^{-2\mu^* \Delta t} X_{xx} \\ &- 2\theta^*(1 - e^{-\mu^* \Delta t})(X_y - e^{-\mu^* \Delta t} X_x) + n(\theta^*)^2 (1 - e^{-\mu^* \Delta t})^2).\end{aligned}$$

At this stage, we maximize the average log-likelihood function over the three OU model parameters and denote the maximized average log-likelihood by

$$l^*(b) = l(\mu^*, \theta^*, \sigma^*, b).$$

Then, the maximized log-likelihood function $l^*(b)$ is further optimized over the ratio b to give the optimal hedge ratio

$$B = \arg\max_b \ l^*(b).$$

In practice, the implementation of this optimization can be done via a grid search. For instance, suppose we limit the absolute value of the hedge ratio to be 2. Then, we compute the maximized log-likelihood function $l^*(b)$ for

$$b = \{-2, -1.99, -1.98, \cdots, 1.99, 2\}.$$

The maximizer is denoted by B.

3.2. Trading Rules

We now describe the mechanism of the trading strategy in this study. Denote by S_t^1 and S_t^2 the prices of two assets at time t. Let C_t be the capital available at time t. The initial investment amount is C_0.

A number of indicators are required in order to set up the trading signals. We consider the M-day moving average of the spread X_t, denoted by $MA(X_t)$. Similarly, we define $SD(X_t)$ to be the standard deviation of the spread over the past M-days. We use the notations $POS(X_t)$ and $POS(S_t^2)$ for positions of S_t^1 and S_t^2. Finally, K is the trading threshold, r is the chosen risk level for the stop-loss rule, L is the length of the lookback window, and T is the length of the entire trading period.

In turn, the trading strategy is described as follows.

1. When entering trading period at time 0, use L-days historical prices of S_t^1 and S_t^2 to calculate the optimal pair ratio, given by

$$B = \arg\max_b l^*(b).$$

2. At each time point t, construct the spread by $X_t = S_t^1 - BS_t^2$.

3. Entry rule: If the current position is zero and that $X_t < MA(X_t) - K * SD(X_t)$, then enter a long position as follows:

$$POS(X_t) = C_t / S_t^1, \quad C_t = C_t - X_t \cdot POS(X_t).$$

If the current position is zero but $X_t > MA(X_t) + K * SD(X_t)$, then enter the short position:

$$POS(X_t) = -C_t / S_t^1, \quad C_t = C_t - X_t \cdot POS(X_t).$$

4. Exit rule: If $POS(X_t) > 0$ and $X_t > MA(X_t)$, then quit the long position:

$$C_t = C_t + X_t \cdot POS(X_t), \quad POS(X_t) = 0,$$

If $POS(X_t) < 0$ and $X_t < MA(X_t)$, then quit the short position:

$$C_t = C_t + X_t \cdot POS(X_t), \quad POS(X_t) = 0,$$

5. Stop-loss rule: Let X_0 denote the value of spread at entry time. If $POS(X_t) > 0$ and $X_t < X_0 * (1 - r)$, then quit the long position:

$$C_t = C_t + X_t \cdot POS(X_t), \quad POS(X_t) = 0.$$

If $POS(X_t) < 0$ and $X_t > X_0 * (1 + r)$, then quit the short position:

$$C_t = C_t + X_t \cdot POS(X_t), \quad POS(X_t) = 0.$$

6. If there is no signal indicating entry or liquidation, then stay put:

$$POS(X_t) = POS(X_t), \quad C_t = C_t.$$

In Step 3, a long position of X_t is established if no shares of X_t are held and the price of X_t drops below the past M-days average value minus K times standard deviation. The number of shares purchased is determined by the ratio of current cash C_t and price of S_t^1. That is, invest all the cash at hand to long S_t^1 and short the corresponding S_t^2 with the best pair ratio B.

As for the liquidation rule described in Step 4, the entire position is liquidated whenever the price of X_t rises beyond the past M-days average value. In short, we enter the long position if the price of spread drops sufficiently far from the long-term mean, and then wait for the price to revert and take profit. The trading rule for short positions follows a similar logic. For more details on this approach, we refer to Lee and Leung (2020) and the references therein.

The trading system includes four parameters: L is the length of look-back window length that determines how many data points are used to estimate the best ratio; T determines how long we will use the estimated best ratio; M is the number of days used to calculate the moving average and standard deviation of the spread, and K is the threshold that determines the trading boundary.

Intuitively, L and M control how much past data should be incorporated in the current trading decisions. The other parameter, K, will have direct influence on the timing and frequency of trades. For instance, a smaller K will result in more frequent trading.

4. Monte Carlo Simulation

In this section, we analyze how different degrees of mean reversion may impact trading performance via Monte Carlo simulation. This analysis will shed light on capital allocation across different spreads in our diversification framework.

We begin by considering trading a single spread, whose values are simulated according to the OU model. There are three parameters in the model. The speed of mean reversion is denoted by μ, the volatility of the process is denoted by σ, and the long-term mean is θ. We

generate 1000 sample paths with different combinations of parameters and simulate the trading transactions over one year ($T = 252$) for each path by following the rule described in the previous section.

We examine the trading performance through the Sharpe ratio, average daily return, volatility, total trading numbers, and cumulative return, which are calculated based on the average over 1000 paths. The other parameters of the trading system are $K = 1$ and $M = 30$ in this experiment. Since we generate the spread time series directly, there is no need to compute the optimal pair ratio.

The results are summarized in Table 1. Among the parameters, we observe that the speed of mean reversion μ has a significant effect on the performance. Raising μ with σ and θ constant, the daily returns, Sharpe ratio, and number of trades all increase while the standard deviation decreases slightly. This is intuitive since a fast mean-reverting spread offers more trading opportunities and reduces the risk associated with the trading strategy.

Table 1. The simulated one-year mean reversion trading performance under different configurations of OU model. The bolded numbers signify outperforming values. Parameters $\mu \in \{10, 20, 30, 40, 50\}$, $\sigma \in \{0.1, 0.5, 1.0, 1.5, 2.0\}$, and θ is chosen from $\{5, 10, 15, 20, 25\}$.

μ	σ	θ	Ret (%)	Std (%)	Sharpe	PnL (%)	No.Trade
10	1.0	15	0.0136	0.2951	0.0467	3.4067	15.410
20	1.0	15	0.0229	0.2903	0.0794	5.8520	18.092
30	1.0	15	0.0299	0.2881	0.1043	7.7184	20.992
40	1.0	15	0.0353	0.2841	0.1248	9.2014	23.563
50	1.0	15	**0.0400**	**0.2797**	**0.1431**	**10.4706**	**26.018**
30	0.1	15	0.0029	0.0286	0.1031	0.7385	20.826
30	0.5	15	0.0148	0.1433	0.1032	3.7566	20.879
30	1.0	15	0.0300	0.2872	0.1048	7.7462	20.899
30	1.5	15	0.0456	0.4295	**0.1064**	11.9390	**20.922**
30	2.0	15	**0.0600**	**0.5780**	0.1041	**15.7092**	20.698
30	1.0	5	0.0292	0.2779	0.1051	**7.9878**	20.960
30	1.0	10	0.0290	0.2821	0.1028	7.5805	**21.055**
30	1.0	15	0.0295	**0.2871**	0.1032	7.6169	20.699
30	1.0	20	**0.0304**	0.2888	**0.1053**	7.6606	20.706
30	1.0	25	0.0298	0.2834	0.1052	7.7975	20.787

Next, we examine the effect of the spread volatility σ. For different σ with $\mu = 30$ and $\theta = 15$, the Sharpe ratio is almost unchanged since both daily returns and volatility increase simultaneously as σ increases. Intuitively, higher volatility may offer more trading opportunities as the spread fluctuates more rapidly. However, higher volatility also leads to a wider trading band, thus increasing the risk exposure. Lastly, the long-term mean θ does not have a significant impact on trading performance as expected, given that trades are triggered by the deviation of the spread from the long-term mean.

Based on the simulation results, the speed of mean reversion has a significantly positive effect on performance as measured by average return and Sharpe ratio. Moreover, a faster spread of mean reversion also results in more frequent trades. On the other hand, the spread volatility increases the mean and standard deviation of returns simultaneously, so the Sharpe ratio is lower for very high volatility. In contrast, the long-term mean θ does not affect the trading performance materially. Hence, we consider the mean reversion rate as a critical factor when designing the diversification framework for trading multiple pairs.

5. Diversification Framework

We consider a diversified portfolio approach to trading multiple pairs simultaneously. This leads to the analysis of a number of different methods to allocate portfolio weights to the traded pairs.

The first step is to divide the entire trading period into stages with a defined schedule for adjusting the allocation. During the first trading stage, we apply equal weights to trade all the spreads in the portfolio. After that, empirical spreads and returns are recorded and analyzed to yield an optimal allocation for the next trading stage. Moving forward, we will periodically obtain updates on the portfolio weights and re-allocate the capital for the next trading period accordingly.

5.1. Portfolio Weights

In our framework, capital is allocated across different pairs in order to form a diversified portfolio. There are various methods to determine the portfolio weights dynamically over time. To that end, we consider several methods and examine their effects on portfolio performance.

5.1.1. Mean-Variance Analysis

Inspired by Mean-Variance Analysis (MVA), one method for determining the portfolio weights is by maximizing the Sharpe ratio. For any given lookback period, let $R \in \mathbb{R}^N$ be the vector of average returns of each pair and $\Sigma \in \mathbb{R}^{N\times N}$ be the corresponding co-variance matrix of daily returns of each pair. Then, the optimal weights are those that give the best historical in-sample Sharpe ratio. Precisely, we have

$$\omega^* = \arg\max_{\omega} \frac{w^T R}{w^T \Sigma w}, \quad s.t. \quad \|w\|_1 = 1. \tag{3}$$

This MVA-based method assumes that the weights that have produced the best Sharpe ratio in the recent past will lead to a good performance in the next trading period. This depends on the stability of the returns over time, which may not be a reliable assumption. To address this issue, we propose several alternative allocation methods based on the characteristics of the spreads, rather than returns.

5.1.2. Mean Reversion Budgeting

We propose an allocation method, called Mean Reversion Budgeting (MRB), which is based on the mean-reverting properties of the spreads. In essence, we seek to measure the degree of mean reversion for each spread and assign more weights to those that are considered highly mean-reverting.

The first index is the estimated maximum average log-likelihood function.

To measure the goodness of fit to an OU process for each spread, we consider the corresponding log-likelihood score. For each spread i, we denote by $\hat{l}_i$ the estimated maximum average log-likelihood score from the previous stage. Next, we do min-max normalization to make them all positive. Precisely, for each spread i, we define the index:

$$\tilde{l}_i = \frac{\hat{l}_i - \min\{\hat{l}_i\}}{\max\{\hat{l}_i\} - \min\{\hat{l}_i\}}.$$

In addition, we consider the speed of mean reversion μ. Intuitively, a more rapidly mean-reverting spread offers more trading opportunities, so it should be allocated with more portfolio weights. In order to account for the fluctuations due to volatility, we scale the speed of mean reversion by the volatility parameter. Precisely, we write

$$\hat{\mu}_i^r = \frac{\hat{\mu}_i}{\hat{\sigma}_i},$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are, respectively, the estimated speed of mean reversion and volatility for each pair. In turn, comparing across all spreads, we define the normalized relative speed of mean reversion for spread i as

$$\tilde{\mu}_i = \frac{\hat{\mu}_i^r - \min\{\hat{\mu}_i^r\}}{\max\{\hat{\mu}_i^r\} - \min\{\hat{\mu}_i^r\}}.$$

Lastly, we incorporate both $\tilde{l}_i$ and $\tilde{\mu}_i$ into the portfolio weight allocation. This results in the formula

$$w_i = \frac{\tilde{\mu}_i \tilde{l}_i}{\sum_{i=1}^{N} \tilde{\mu}_i \tilde{l}_i},\ i = 1, \cdots, N. \tag{4}$$

In summary, this means that more capital is allocated into pairs that are more OU-like and mean-revert more rapidly.

5.1.3. Mean Reversion Ranking

The MRB allocation method above may sometimes lead to extreme portfolio weights. For instance, if spread i has a likelihood score l_i that is significantly larger than the others, the method will allocate most of the capital to that pair based on (4), leading to portfolio concentration. Furthermore, the MRB method is susceptible to estimation errors since the portfolio weights are functions of the estimates.

These observations motivate us to propose an alternative allocation method, called Mean Reversion Ranking (MRR), which assigns fixed values based on the rankings of the product of the estimated likelihood score and speed of mean reversion. Assuming that n pairs are currently traded, we sort the n pairs in ascending order based on the product, $\tilde{l}_i$ and $\tilde{\mu}_i$, where $i = 1, \cdots, n$. Then, we calculate the weights of sorted pairs by

$$\left(\frac{n-1}{2n(n-1)}, \frac{n+1}{2n(n-1)}, \cdots, \frac{3(n-1)}{2n(n-1)}\right). \tag{5}$$

This method suggests that we allocate the fraction $\frac{n-1}{2n(n-1)}$ of the capital to the pair with the lowest likelihood value and speed of mean reversion, then another fraction $\frac{n+1}{2n(n-1)}$ to the pair with the second-ranked spread, and so on.

The intuition behind this approach is as follows. If we start with equal weights $\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$ and then reduce by half the weight of the pair with the lowest likelihood value and speed of mean reversion, then we invest the saved capital into the pair with the highest score. This yields the fractions $\frac{1}{2n}$ and $\frac{3(n-1)}{2n(n-1)}$ for the smallest and largest weights, respectively. The weights in the middle then increase uniformly based on their ranking indices. The ranking method offers a more moderate weight distribution than the MRB method. As our experiment shows, this method produces the highest Sharpe ratio for a given set of spreads.

5.2. Trading Rules under Diversification Framework

Next, we present all the stages within our trading framework. To start, we let $\{S_t^{11}, S_t^{12}, \ldots, S_t^{N1}, S_t^{N2}\}$ be the prices of N pairs of co-moved stocks at time t. With the initial investment C_0, we denote C_t to be the total cash at time t.

For each spread, we write $MA(X_t)$ and $SD(X_t)$ to represent the moving average and standard deviation of past M-days spread, respectively. At each time t, the position of X_t is given by $POS(X_t)$.

The length of the whole trading period is T and the length of lookback window is L. L_R represents the window length of each small stage, also referred to as rebalancing window size. Note that the trading period is divided into T/L_R small stages. Then, there are two types of transactions within the trading program: stage transition and intra-stage trading.

The complete trading procedure is described as followed:

1. **Pairs ratio calculation**. Entering trading period at time 0, use L-days historical prices of each pair $\{S_t^{i1}, S_t^{i2}\}$ to calculate the optimal pair ratio, given by

$$B^i = \arg\max_b l_i^*(b),$$

where $l_i^*(\cdot)$ is the maximized average likelihood function for the i-th pair and $1 \leq i \leq N$.

2. **Spreads construction**. At each time point t, construct N spreads by $X_t^i = S_t^{i1} - B^i S_t^{i2}$, $1 \leq i \leq N$.

3. **Stage transition**. Entering a new stage, liquidate all the positions and use past L_R-days trading information to update the weights ω by equation (3), (4) or (5). Allocate the total cash at hand based on weights:

$$C_t^i = \omega_i \cdot C_t.$$

4. **Entry of intra-stage trading**. Set a trading threshold K (a positive constant, e.g., 0.5, 1, or 1.25). Check the entry condition for each pair. For $i \in \{1, \cdots, N\}$, if $POS(X_t^i) = 0$ and $X_t^i < MA(X_t^i) - K * SD(X_t^i)$, then enter the long position:

$$POS(X_t^i) = C_t^i / S_t^{i1}, \quad C_t^i = C_t^i - X_t^i \cdot POS(X_t^i).$$

If $POS(X_t^i) = 0$ and $X_t^i > MA(X_t^i) + K * SD(X_t^i)$, then enter the short position:

$$POS(X_t^i) = -C_t^i / S_t^{i1}, \quad C_t^i = C_t^i - X_t^i \cdot POS(X_t^i).$$

5. **Liquidation of intra-stage trading.** Check the liquidation for each pair. For $i \in \{1, \cdots, N\}$, if $POS(X_t^i) > 0$ and $X_t^i > MA(X_t^i)$, then quit the long position:

$$C_t^i = C_t^i + X_t^i \cdot POS(X_t^i), \quad POS(X_t^i) = 0.$$

If $POS(X_t^i) < 0$ and $X_t^i < MA(X_t^i)$, then quit the short position:

$$C_t^i = C_t^i + X_t^i \cdot POS(X_t^i), \quad POS(X_t^i) = 0.$$

6. **Stop-loss rule of intra-stage trading.** Set a risk level r. For $i \in 1, \cdots, N$, let X_0^i denote the value of each spread at entry time. If $POS(X_t^i) > 0$ and $X_t^i < X_0^i * (1 - r)$, then quit the long position:

$$C_t^i = C_t^i + X_t^i \cdot POS(X_t^i), \quad POS(X_t^i) = 0.$$

If $POS(X_t^i) < 0$ and $X_t^i > X_0^i * (1 + r)$, then quit the short position:

$$C_t^i = C_t^i + X_t^i \cdot POS(X_t^i), \quad POS(X_t^i) = 0.$$

7. **No-trade scenario.** If there is no signal of entry or liquidation, then keep the positions unchanged:

$$POS(X_t^i) = POS(X_t^i), \quad C_t^i = C_t^i.$$

8. Conduct intra-stage mean reversion trading for L_R days and go back to step 3.

The entire framework can be briefly summarized as follows: (1) divide the trading period into several smaller stages; (2) when beginning a new trading stage, update the weights based on the speed of mean reversion and volatility of spreads from the previous stage; (3) liquidate all holding positions and re-allocate the total capital based on updated weights; (4) conduct mean reversion trading on each pair separately with the allocated

capital until the end of this stage; (5) record the historical spreads at the end and then move onto the next stage.

6. Backtesting

In this section, we examine the performance of the proposed diversified trading framework through a backtest. Different allocation methods are implemented and their results are compared. It is worth noting that we do not apply the stop-loss rule in our experiments.

6.1. Traded Assets and Price Data

To begin, we introduce the assets used in our experiment. Six pairs of stocks are selected from six different sectors in the US market. Each pair consists of two stocks from the same industry, ranging from airlines to banking. Table 2 provides a description of the companies, along with the ticker symbols of all the pairs: WM-RSG, UAL-DAL, V-MA, MS-GS, NVDA-AMD, and CVX-XOM.

For these six pairs, we collect the daily close prices for these stocks from 1 January 2021 to 31 December 2022 using the Yahoo! Finance API.[1] Figure 1 shows the price movements of each pair. Each price time series has been divided by its initial values for normalization. As we can see, the stock prices for each pair exhibit persistent comovement.

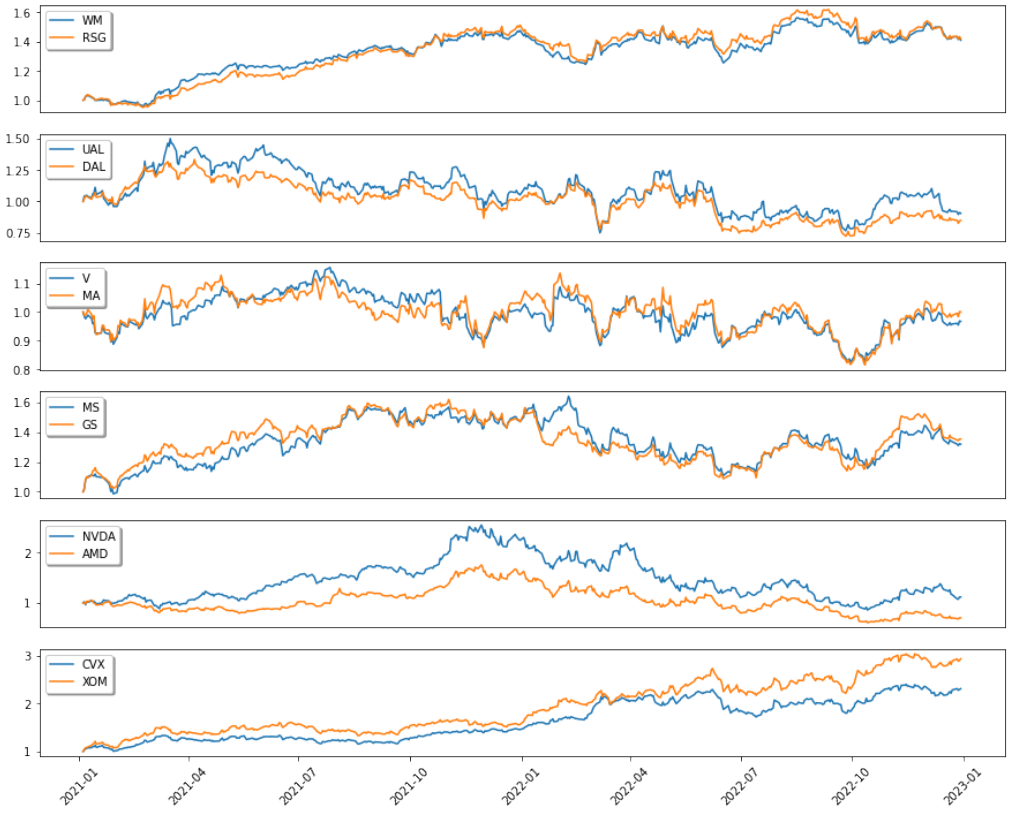


Figure 1. Normalized historical daily close prices of six pairs of stocks from 1 January 2021–31 December 2022. From top to bottom: WM-RSG, UAL-DAL, V-MA, MS-GS, NVDA-AMD, and CVX-XOM.

Pairs are constructed using the method detailed in Section 3 and the daily adjusted close prices from the first year. In the experiment, the subsequent year is the trading period. In each quarter, data are used to generate new trading signals. With quarterly rebalancing, we test several trading thresholds and analyze annual performance statistics.

Table 2. The stock pairs used for testing our mean-reversion trading strategies and the description of the company.

Pairs Symbols	Description
WM-RSG	Waste Management and Republic Services provide waste management and environmental services
UAL-DAL	United and Delta are two of the largest American airlines
V-MA	Visa and Mastercard are two dominant players in the payment industry
MS-GS	Morgan Stanley and Goldman Sachs are two large-cap stocks in the banking industry
NVDA-AMD	Nvidia and AMD are two American multinational semiconductor companies that develop computer processors
CVX-XOM	Chevron and Exxon Mobil are major companies in the energy sector

6.2. Trading on Each Pair

To demonstrate the efficacy of the proposed mean-reversion trading strategy, we first apply the strategy to six selected pairs separately. In the formation period, we use the data from 1 January 2021 to 31 December 2021 to compute the optimal ratio for each pair. Table 3 lists the estimated best ratio and likelihood score of each pair. The parameters are selected as $K = 1$, $L = 252$, $T = 252$, and $M = 63$.

Table 3. Optimal ratio and likelihood score for each pair estimated from 1 January 2021 to 31 December 2021.

S_t^1	S_t^2	Optimal Ratio b	Likelihood Score l
WM	RSG	0.98	5.7101
UAL	DAL	1.02	5.3507
V	MA	0.99	5.8283
MS	GS	1.00	6.1103
NVDA	AMD	1.31	2.8395
CVX	XOM	0.95	4.6977

Then, we construct spreads from each pair based on the best ratio. In turn, each spread is traded in the trading period. Figure 2 illustrates separately the spread, trading positions, and returns for the six pairs in 2022. From top to bottom in each panel, we record the movement of spread X_t, the position change $POS(X_t)$, and the cumulative return curve for each pair. As we can see, almost all of the pairs, except CVX-XOM, have positive annual profits, which demonstrates the effectiveness of the pairs trading strategy. As for the pair CVX-XOM, the main reason for failure is that the spreads constructed from the pairs do not display good mean reversion properties, as we can clearly observe a long-term downward trend in the spread. With this in mind, based on the spread behaviors and trading performance, it is intuitively optimal to allocate more weight to the first five pairs and less weight to CVX-XOM during the trading period. This echoes our motivation to propose a diversification framework to dynamically allocate weights on multiple pairs. It would be interesting to see whether any of the allocation methods can assign the weights accordingly.

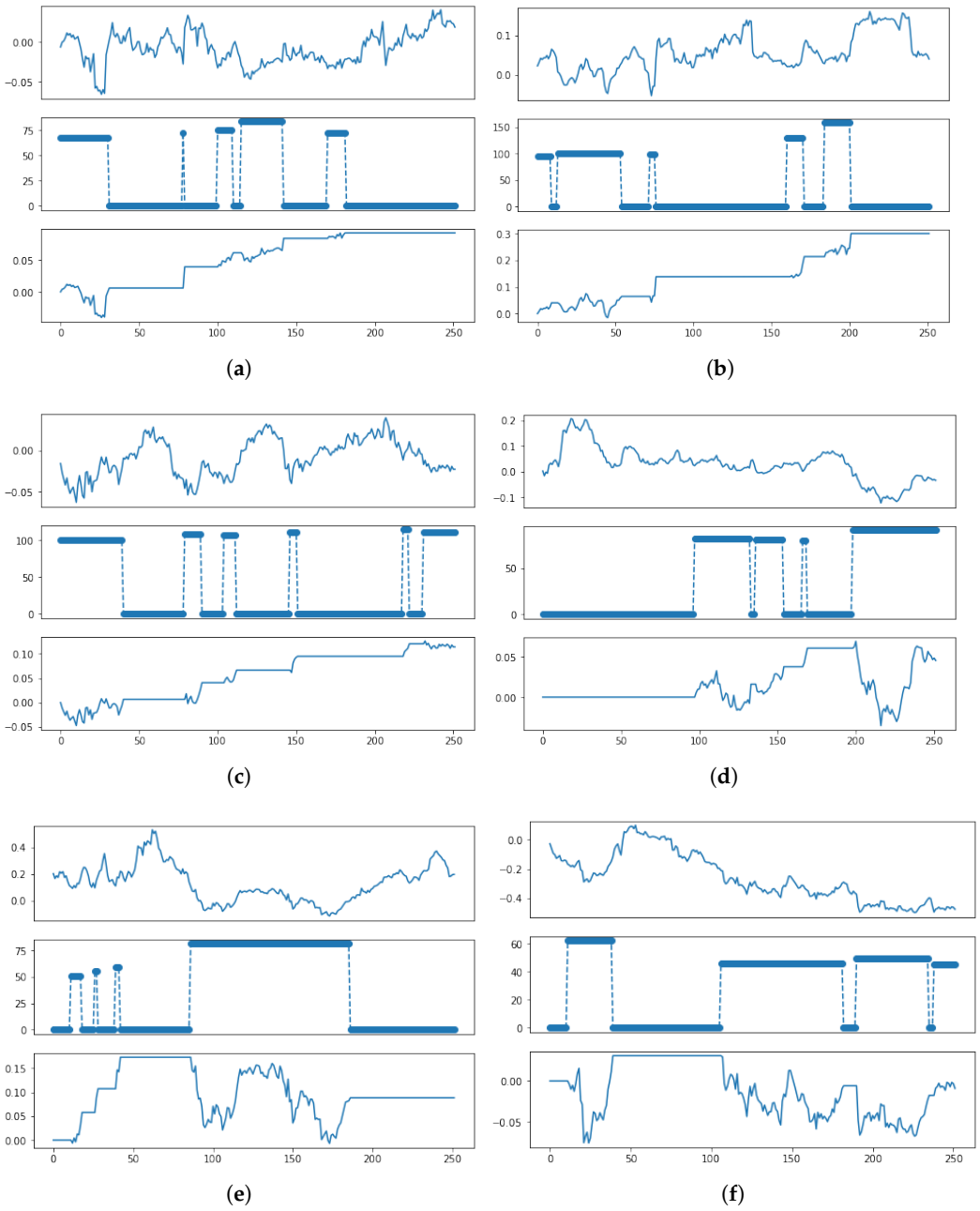


Figure 2. The separate trading performance on six pairs with $K = 1$, $L = 252$, $T = 252$, and $M = 63$ in 2022. For each pair, we show the spread, position, and cumulative return over time. (**a**) WM-RSG. (**b**) UAL-DAL. (**c**) V-MA. (**d**) MS-GS. (**e**) NVDA-AMD. (**f**) CVX-XOM.

As shown in Figure 3, the correlation coefficients between different spreads are generally small, with the most positive correlation coefficient being approximately 0.3. These low correlations suggest that there is a potential for diversification benefits in combining these spreads in a trading portfolio. By diversifying across multiple spreads, traders can potentially reduce portfolio volatility and increase the expected return. This motivates us to investigate the effectiveness of our proposed diversified trading framework.

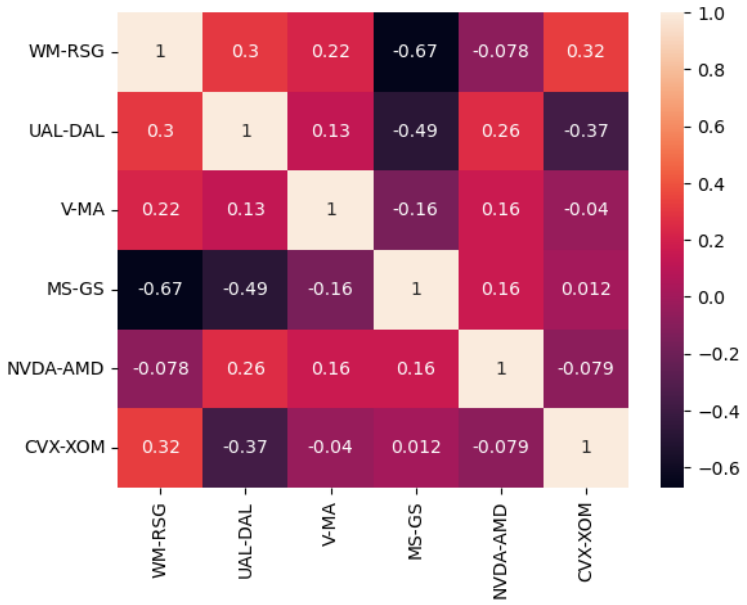


Figure 3. Correlation matrix of the constructed six spreads.

6.3. Equal-Weight Portfolio

The baseline diversification is to trade all pairs equally, which means we implement the trading strategies for all pairs with equally allocated initial cash during the entire trading period. This equal-weight portfolio is conceptually the most straightforward way to create a diversified portfolio. It is included here to compare against nondiversified trading of single pairs and against other diversification methods.

To start, we first compare the performance of the equal-weight portfolio to the performance of every single pair in this section. The trading period is the year 2022, and the average annual and daily performance statistics of the trading strategies are presented in Table 4. We consider different trading thresholds $K \in \{0.5, 0.75, 1.0, 1.25\}$. With a higher K, the spread needs to deviate further from its moving average in order to trigger a trade. The criteria presented include daily return (DailyRet), daily standard deviation (DailyStd), daily Sharpe ratio (DailySR), annual maximum drawdown (AnnMDD), and annual cumulative PnL (AnnPnL). In this experiment, we observe that the equal-weight portfolio generally outperforms single pairs in terms of daily Sharpe ratio, daily standard deviation, and annual maximum drawdown. This indicates the effectiveness of diversification for mean reversion trading.

Moreover, we observe that the standard deviations of individual pairs range from 0.4 to 1, while the standard deviation of our portfolio is lower than 0.3. This indicates that the equal-weight portfolio strategy has a lower level of risk compared to the individual pairs. This can be attributed to the diversification effect, as the portfolio strategy involves trading multiple pairs simultaneously, which helps reduce the overall risk. Therefore, our results demonstrate that an equal-weight portfolio strategy can generate a higher return and achieve a lower level of risk through diversification.

Table 4. Performance statistics of the equal-weight portfolio and trading of single pairs. The bolded numbers signify outperforming values. The parameters selected are: $K \in \{0.5, 0.75, 1.0, 1.25\}$, $L = T = 252$, and $M = 63$.

Index	Equal Weight	WM-RSG	UAL-DAL	V-MA	MS-GS	NVDA-AMD	CVX-XOM
				$K = 0.5$			
DailyRet (%)	0.0260	0.0461	**0.0741**	0.0266	0.0060	0.0251	−0.0181
DailyStd (%)	**0.3020**	0.4616	0.8667	0.5882	0.5935	1.0802	1.0132
DailySR	0.0861	**0.0999**	0.0855	0.0452	0.0100	0.0233	−0.0179
AnnMDD (%)	−1.7527	−4.0295	**−1.5036**	−4.7033	−6.5959	−3.6869	−10.1602
AnnPnL (%)	6.6218	11.9691	**19.3084**	6.4325	1.0577	4.9606	−5.6825
				$K = 0.75$			
DailyRet (%)	0.0405	0.0337	**0.0996**	0.0377	0.0140	0.0435	−0.0168
DailyStd (%)	**0.2702**	0.4425	0.8010	0.5480	0.5233	1.0543	0.9531
DailySR	**0.1499**	0.0762	0.1244	0.0689	0.0267	0.0412	−0.0177
AnnMDD (%)	−1.0363	−4.0295	−1.5036	−4.7033	−4.7938	**−1.0167**	−9.4947
AnnPnL (%)	10.5989	8.5626	**27.383**	9.5192	3.2131	9.9788	−5.2313
				$K = 1$			
DailyRet (%)	**0.0444**	0.0362	0.1075	0.0442	0.0191	0.0393	0.0005
DailyStd (%)	**0.2603**	0.4406	0.7962	0.5410	0.5197	1.0602	0.9026
DailySR	**0.1707**	0.0823	0.1350	0.0821	0.0368	0.0371	0.0005
AnnMDD (%)	−0.9259	−4.0295	−1.5036	−4.7033	−3.5482	**−0.6721**	−7.5687
AnnPnL (%)	**11.7003**	9.2573	29.9381	11.3813	4.5634	8.8227	0.9027
				$K = 1.25$			
DailyRet (%)	0.0339	0.0199	**0.0646**	0.0190	0.0240	−0.0156	0.0336
DailyStd (%)	**0.2377**	0.3816	0.6833	0.4912	0.4816	0.9437	0.8126
DailySR	**0.1428**	0.0523	0.0945	0.0386	0.0499	−0.0165	0.0413
AnnMDD (%)	**−2.1982**	−4.0295	−5.3303	−3.5172	−2.3069	−13.0758	−6.6102
AnnPnL (%)	8.8155	4.9415	**16.9198**	4.5624	5.9091	−4.9065	7.8850

6.4. Diversification Framework

We implement multiple mean reversion trading under the diversification framework using different allocation methods, and compare the performance with the baseline equal-weight portfolio. Since the trading period is one year (1/1/2022–12/31/2022) and the re-balancing window is set to be three months, the trading period is divided into four periods. As before, trading parameters are as follows: $K \in \{0.5, 0.75, 1.0, 1.25\}$, $L = T = 252$ days, $M = 90$ days.

Figure 4 displays the cumulative return curves with different allocation methods in the year 2022. Here, "MRB" stands for the Mean Reversion Budgeting method, "MRR" represents the Mean Reversion Ranking method, and "MVA" refers to the portfolio generated by the mean-variance analysis. The equal-weight portfolio serves as a baseline portfolio for comparison.

In terms of cumulative returns, the MRB method outperforms the baseline portfolio by a significant margin. On the other hand, the MRR method's returns are only slightly higher than the baseline. This is not surprising since the rank-based method tends to spread the weights more gradually from higher to lower-ranked pairs. In other words, the MRR method does not penalize the poorly fitted spreads severely and does not assign oversized weights to the spreads with the best mean-reverting properties.

However, taking volatility into account, subsequent experiments reveal that the ranking method achieves a higher Sharpe ratio than the baseline. In contrast, the MVA portfolio lags far behind the baseline. This is perhaps intuitive since the traditional mean-variance analysis does not take into account the mean-reverting properties of the spreads.

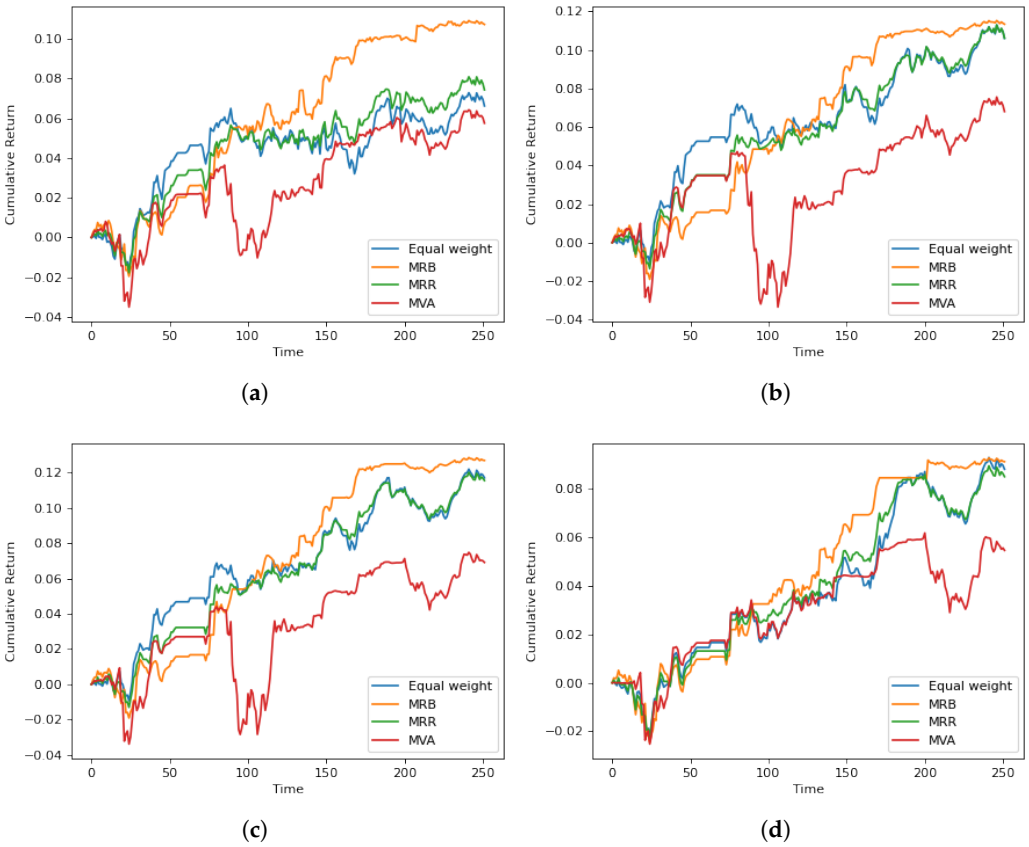


Figure 4. The accumulative returns for portfolios with different allocation methods. The trading period is the year 2022, and the parameters are: $L = T = 252$, $M = 90$, and $L_R = 63$ days. Each plot corresponds to a selected trading threshold $K \in \{0.5, 0.75, 1.0, 1.25\}$. (**a**–**d**) K = 0.5, 0.75, 1, 1.25.

To provide a comparative analysis of different portfolio allocation methods versus the equal-weight portfolio, we summarize the statistics of trading performance with different thresholds in Table 5. The statistics include daily returns (DailyRet), daily standard deviation (DailyStd), daily Sharpe ratio (DailySR), annual maximum drawdown (AnnMDD), and annual cumulative profit and loss (AnnPnL).

When compared to the baseline equal-weight portfolio, the MRB method significantly improves the annual return without increasing the standard deviation considerably. As a result, MRB yields a higher Sharpe ratio than the baseline portfolio. The MRR method significantly decreases volatility and also achieves a high Sharpe ratio. These observations support the notion that allocation based on mean reversion characteristics has the potential to improve portfolio performance.

Regarding the standard deviations of the differently weighted portfolios, it is notable that the MRR method exhibits the lowest standard deviation among all the portfolios, indicating that this particular combination of pairs is the least volatile. In addition, the MRB method also appears to decrease the standard deviation when compared to the equal weights and MVA methods. These findings provide further evidence for the advantages of employing a portfolio weights allocation approach to mitigate risk and enhance the overall performance of the portfolio.

Table 5. The comparison of different portfolio allocation methods vs. the baseline (equal-weight (EW)). The bolded numbers signify outperforming values. The trading period is the year 2022, and the parameters are: $L = T = 252$, $M = 90$, $L_R = 63$ days, and $K \in \{0.5, 0.75, 1.0, 1.25\}$.

Index	EW	MRB	MRR	MVA
		$K = 0.5$		
DailyRet (%)	0.0260	**0.0410**	0.0289	0.0230
DailyStd (%)	0.3020	0.2755	**0.2636**	0.3624
DailySR	0.0861	**0.1488**	0.1097	0.0633
AnnMDD (%)	**−1.7527**	−1.9555	−1.7722	−3.5034
AnnPnL (%)	6.6218	**10.7271**	7.4317	5.7548
		$K = 0.75$		
DailyRet (%)	0.0405	**0.0431**	0.0405	0.0272
DailyStd (%)	0.2702	0.2485	**0.2349**	0.4414
DailySR	0.1499	**0.1734**	0.1723	0.0616
AnnMDD (%)	**−1.0363**	−1.9116	−1.3602	−3.3486
AnnPnL (%)	10.5989	**11.3333**	10.6107	6.7990
		$K = 1$		
DailyRet (%)	0.0444	**0.0479**	0.0438	0.0277
DailyStd (%)	0.2603	0.2614	**0.2310**	0.4581
DailySR	0.1707	0.1834	**0.1897**	0.0604
AnnMDD (%)	**−0.9259**	−1.9116	−1.3052	−3.3865
AnnPnL (%)	11.7003	**12.6890**	11.5503	6.9090
		$K = 1.25$		
DailyRet (%)	0.0339	**0.0351**	0.0328	0.0216
DailyStd (%)	0.2377	0.2402	**0.2123**	0.2618
DailySR	0.1428	0.1461	**0.1543**	0.0823
AnnMDD (%)	**−2.1982**	−2.4710	−2.2249	−2.5295
AnnPnL (%)	8.8155	**9.1281**	8.5101	5.4669

6.5. MRB Portfolio Weights

In our experiment, the MRB method appears to be the best allocation method. In this section, we provide more details on how the weights vary under the MRB method during the trading period. Let us focus on the case where $K = 1$. Table 6 shows the estimated parameters using the data from 1 January 2022–31 December 2022. As we can see, the NVDA-AMD and CVX-XOM pairs have the lowest likelihood scores. Table 7 shows the changes in portfolio weights for each quarter during that year. Notice that the weights for the NVDA-AMD and CVX-XOM pairs are close to zero. This is consistent with the MRB method where weights are proportional to the likelihood scores.

Table 6. Parameters estimated from 1 January 2022 to 31 December 2022.

S_t^1	S_t^2	μ	σ	l
WM	RSG	0.1415	0.0193	5.7888
UAL	DAL	0.7284	0.0316	5.2948
V	MA	0.0108	0.0175	5.8850
MS	GS	0.0357	0.0244	5.5570
NVDA	AMD	2.7227	0.2594	3.1927
CVX	XOM	2.3461	0.3394	2.9276

Table 7. MRB portfolio weights for all six pairs over four quarters from 1 January 2022 to 31 December 2022.

S_t^1	S_t^2	Period 1	Period 2	Period 3	Period 4
WM	RSG	0.481	0.455	0.259	0.780
UAL	DAL	0.176	0.161	0.069	0.030
V	MA	0.096	0.309	0.249	0.142
MS	GS	0.220	0.065	0.422	0.048
NVDA	AMD	0.000	0.000	0.000	0.000
CVX	XOM	0.027	0.010	0.000	0.000

7. Conclusions

We have presented a trading program that dynamically allocates capital to multiple mean reversion trading strategies. The approach is designed for trading multiple pairs in order to achieve diversification effects. Moreover, the dynamic rebalancing is adaptive to the current model estimates and based on the relative performance or path behaviors of the pairs in the portfolio. Our empirical experiments have shown that, for a given set of pairs traded, the allocation method plays a significant role in the success of the diversification framework.

This paper provides portfolio managers and traders with a useful and flexible framework to test trading strategies and build portfolios involving multiple pairs. The approach discussed herein can also be applied to multiple futures spread trading that is common in other asset classes, such as interest rates, commodities, and currencies. There are plenty of examples of mean-reverting spreads between futures contracts in different markets, including Brent vs. WTI crude oil, soybean vs. soybean meal, gold vs. silver, and many more. Spread trading is very common in managed futures portfolios.

Future research directions include considering a risk-sensitive approach and incorporating additional risk controls into the trading problem. The effects on trading performance would be of practical interest. Another direction is to consider a variation of the mean-reverting model. In particular, regime-switching mean-reverting models have been used for futures trading (see Leung and Zhou 2019) and they can be suitable here for multiple spread trading.

Author Contributions: Methodology, K.L., T.L. and B.N.; software, B.N.; writing—original draft, B.N.; writing—review and editing, K.L. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OU	Ornstein–Uhlenbeck
MLE	Maximum likelihood estimation
DM	Distance method
SSD	Sum of squared deviations
PCA	Principal component analysis
ETF	Exchange traded funds
CNN	Convolutional neural network
MVA	Mean-variance analysis

Note

[1] https://pypi.org/project/yfinance/, accessed on 1 April 2023.

References

Avellaneda, Marco, and Jeong-Hyun Lee. 2010. Statistical arbitrage in the us equities market. *Quantitative Finance* 10: 761–82. [CrossRef]

Brennan, Michael J., and Eduardo S. Schwartz. 1990. Arbitrage in stock index futures. *Journal of Business* 63: S7–S31. [CrossRef]

Dai, Min, Yifei Zhong, and Yue Kuen Kwok. 2011. Optimal arbitrage strategies on stock index futures under position limits. *Journal of Futures Markets* 31: 394–406. [CrossRef]

d'Aspremont, Alexandre. 2011. Identifying small mean-reverting portfolios. *Quantitative Finance* 11: 351–64. [CrossRef]

Do, Binh, and Robert Faff. 2012. Are pairs trading profits robust to trading costs? *Journal of Financial Research* 35: 261–87. [CrossRef]

Do, Binh, Robert Faff, and Kais Hamza. 2006. A new approach to modeling and estimation for pairs trading. In *Proceedings of the 2006 Financial Management Association European Conference*. Stockholm: European Finance Association, vol. 1, pp. 87–99.

Elliott, Robert J., John Van Der Hoek, and William P. Malcolm. 2005. Pairs trading. *Quantitative Finance* 5: 271–76. [CrossRef]
Engle, Robert F., and Clive W. J. Granger. 1987. Co-integration and error correction: Representation, estimation, and testing. *Econometrica* 55: 251–76. [CrossRef]
Galenko, Alexander, Elmira Popova, and Ivilina Popova. 2012. Trading in the presence of cointegration. *The Journal of Alternative Investments* 15: 85–97. [CrossRef]
Gatev, Evan, William N. Goetzmann, and K. Geert Rouwenhorst. 2006. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies* 19: 797–827. [CrossRef]
Guijarro-Ordonez, Jorge, Markus Pelger, and Greg Zanotti. 2021. Deep learning statistical arbitrage. *arXiv* arXiv:2106.04028.
Huck, Nicolas, and Komivi Afawubo. 2015. Pairs trading and selection methods: Is cointegration superior? *Applied Economics* 47: 599–613. [CrossRef]
Kanamura, Takashi, Svetlozar T. Rachev, and Frank J. Fabozzi. 2010. A profit model for spread trading with an application to energy futures. *The Journal of Trading* 5: 48–62. [CrossRef]
Lee, Donovan, and Tim Leung. 2020. On the efficacy of optimized exit rule for mean reversion trading. *International Journal of Financial Engineering* 7: 2050024. [CrossRef]
Leung, Tim, and Hung Nguyen. 2019. Constructing cointegrated cryptocurrency portfolios for statistical arbitrage. *Studies in Economics and Finance* 36: 581–99. [CrossRef]
Leung, Tim, and Xin Li. 2015. Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical and Applied Finance* 18: 1550020. [CrossRef]
Leung, Tim, and Xin Li. 2016. *Optimal Mean Reversion Trading: Mathematical Analysis and Practical Applications*. Modern Trends in Financial Engineering. Hackensack: World Scientific Publishing Company.
Leung, Tim, and Yang Zhou. 2019. Dynamic optimal futures portfolio in a regime-switching market framework. *International Journal of Financial Engineering* 6: 1950034. [CrossRef]
Leung, Tim, Jize Zhang, and Aleksandr Aravkin. 2020. Sparse mean-reverting portfolios via penalized likelihood optimization. *Automatica* 111: 108651.
Liew, Rong Qi, and Yuan Wu. 2013. Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds* 19: 12–30.
Montana, Giovanni, and Kostas Triantafyllopoulos. 2011. Dynamic modeling of mean reverting spreads for statistical arbitrage. *Computational Management Science* 8: 23–49.
Vidyamurthy, Ganapathy. 2004. *Pairs Trading: Quantitative Methods and Analysis*. Hoboken: John Wiley & Sons, vol. 217.
Xie, Wenjun, Rong Qi Liew, Yuan Wu, and Xi Zou. 2016. Pairs trading with copulas. *The Journal of Trading* 11: 41–52. [CrossRef]

risks

MDPI

Article

Pricing of Pseudo-Swaps Based on Pseudo-Statistics †

Sebastian Franco *,‡ and Anatoliy Swishchuk *,‡

Department of Mathematics and Statistics, University of Calgary, Calgary, AB T2N 1N4, Canada
* Correspondence: sebastian.franco@ucalgary.ca (S.F.); aswish@ucalgary.ca (A.S.)
† This paper is dedicated to Peter Carr.
‡ These authors contributed equally to this work.

Abstract: The main problem in pricing variance, volatility, and correlation swaps is how to determine the evolution of the stochastic processes for the underlying assets and their volatilities. Thus, sometimes it is simpler to consider pricing of swaps by so-called pseudo-statistics, namely, the pseudo-variance, -covariance, -volatility, and -correlation. The main motivation of this paper is to consider the pricing of swaps based on pseudo-statistics, instead of stochastic models, and to compare this approach with the most popular stochastic volatility model in the Cox–Ingresoll–Ross (CIR) model. Within this paper, we will demonstrate how to value different types of swaps (variance, volatility, covariance, and correlation swaps) using pseudo-statistics (pseudo-variance, pseudo-volatility, pseudo-correlation, and pseudo-covariance). As a result, we will arrive at a method for pricing swaps that does not rely on any stochastic models for a stochastic stock price or stochastic volatility, and instead relies on data/statistics. A data/statistics-based approach to swap pricing is very different from stochastic volatility models such as the Cox–Ingresoll–Ross (CIR) model, which, in comparison, follows a stochastic differential equation. Although there are many other stochastic models that provide an approach to calculating the price of swaps, we will use the CIR model for comparison within this paper, due to the popularity of the CIR model. Therefore, in this paper, we will compare the CIR model approach to pricing swaps to the pseudo-statistic approach to pricing swaps, in order to compare a stochastic model to the data/statistics-based approach to swap pricing that is developed within this paper.

Keywords: volatility; variance; covariance; correlation swaps; pseudo-swaps; pseudo-statistics; Apple and Google data; CIR model

Citation: Franco, Sebastian, and Anatoliy Swishchuk. 2023. Pricing of Pseudo-Swaps Based on Pseudo-Statistics. *Risks* 11: 141. https://doi.org/10.3390/risks11080141

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 23 June 2023
Revised: 14 July 2023
Accepted: 28 July 2023
Published: 3 August 2023

Preface

The authors would like to dedicate this paper to Peter Carr, with whom they have had several discussions in the past about swaps, including variance, volatility, covariance, and correlation swaps.

The second author's first experience with swaps was in Vancouver in 2002 on a 5-day Industrial Problem Solving Workshop (IPSW) organized by PIMS. The problem was proposed by RBC Financial Group (see (RBC Financial Group Team 2002)), and it concerned the pricing of swaps involving the so-called pseudo-statistics, namely, the pseudo-variance, pseudo-covariance, pseudo-volatility, and pseudo-correlation. The team consisted of nine graduate students, with whom we solved the problem and prepared a report (see (Badescu et al. 2002)). I would like to thank them all for a very productive collaboration during this time. I would also like to mention that the idea to use the change in time method for solving swap pricing problems (see (Swishchuk 2004)) actually appeared to the second author during this workshop.

Peter was always humorous regarding any topic during our lighthearted conversations, and there is no exception for the latter one. I remember when I told him that we had some results on pricing pseudo-variance, pseudo-covariance, pseudo-volatility, and pseudo-correlation swaps (discretely defined as volatility derivatives), which we obtained during

this IPSW in Vancouver in 2002, together with nine graduate students and prepared a paper, he immediately sent me an email saying "Please, send me your pseudo-paper!".

1. Introduction

Volatility derivatives, such as variance, volatility, covariance, and correlation swaps, have been popular in the financial market for a long time (Carr and Madan 2001; Carr and Lee 2009; Demeterfi et al. 1999).

The simplest measure of a stock's risk or uncertainty is its volatility. The volatility, which we denote as σ_R, is the annualized standard deviation of the stock's returns during the period of interest, where the subscript R denotes the observed or "realized" volatility. Volatility swaps, or realized volatility forward contracts, are the easiest way to trade volatility because they provide pure exposure to volatility (and only to volatility). Therefore, volatility swaps are forward contracts on future realized stock volatility, and variance swaps are similar contracts on variance, the square of the future volatility. Both these instruments provide an easy way for investors to gain exposure to the future level of volatility.

Even though volatility is commonly talked about amongst options market participants, variance and volatility squared have more fundamental significance. A pricing covariance swap, from a theoretical point of view, is similar to a pricing variance swap. A covariance swap is a covariance forward contact of two underlying assets.

The following is in reference to the work of (Carr and Lee 2009). In the earlier part of the 1990s, one can find the first evidence of volatility derivatives being traded in the OTC market. Both variance and volatility derivatives were traded sporadically between 1993 and 1998. The first contracts to enjoy any liquidity were variance swaps. Due to historically high implied volatility in 1998, one can see the emergence of variance swaps in that year. The so-called covariance and correlation swaps have become available as well (Brockhaus and Long 2000). Some papers advocate for the creation of volatility indices and other financial products, whose payoff is tied to these indices, including the VIX index. We would also like to mention that (Gastineau 1977) and (Galai 1979) introduced option indices similar to stock indices. Futures and option contracts based on realized volatility indices were proposed by (Brenner and Galai 1989). (Whaley 1993) introduced derivative contracts written on the VIX index, while (Fleming et al. 1995) described the construction of the original VIX index. More information about the development of a volatility indices can be found in (Carr and Lee 2009).

Extending the Black Scholes model for stochastic volatility, some papers proposed a parametric process. Specifically, (Grünbichler and Longstaff 1996) valued options and futures considering a continuous time GARCH process for variance, while Brenner and Galai (1993, 1996) created an evaluation model for options on volatility via a binomial process. Some papers proposed an alternative non-parametric approach even before swaps on variance had been introduced. In this direction, the first paper was a working paper by Neuberger (Neuberger 1990), and it was published in 1994 (Neuberger 1994). Neuberger assumed that the price of stocks is continuous over time and that the limit of the sum of squared returns should exist, in place of assuming a specific stochastic process. Independently of Neuberger, (Dupire 1993) developed the same argument. Building on a prior working paper by (Dupire 1996), (Carr and Madan 2001) completed the task of developing a robust replicating strategy for continuously monitored variance swap. The next main breakthrough in the robust pricing of volatility derivatives occurred in path-breaking work by (Dupire 1996) and by (Derman et al. 1997). Moments swaps were studied in (Schoutens 2005). Modelling and pricing of variance swaps for stochastic volatilities with delay and for multi-factor stochastic volatilities with delay were considered in (Swishchuk 2005) and (Swishchuk 2006), respectively. Extensive overview on volatility derivatives may be found in (Carr and Lee 2009).

The Heston model is amongst the most popular stochastic models for the pricing of volatility swaps, but new approaches and modifications are continually being suggested. In (Salvi and Swishchuk 2014), the Heston model is used along with a new probabilistic

approach to study volatility swaps, while a variance drift-adjusted version of the Heston model is presented in (Swishchuk and Vadori 2014). Models and processes that incorporate jumps are also increasingly popular for modeling the fluctuations of financial markets. Amongst these jump processes is the Barndorff-Nielsen and Shephard (BN-S) model, which was used in (Benth et al. 2007) to analyze swaps written on powers of realized volatility. The arbitrage-free pricing of variance and volatility swaps under the BN-S model is examined in (Habtemicael and SenGupta 2016b), while covariance swaps under the BN-S model are looked at in (Habtemicael and Sengupta 2016a). (Issaka and SenGupta 2017) further examine the BN-S model by calculating the bounds of the arbitrage-free variance swap price. Derivative asset analysis and pricing of European style options under the BN-S model is found in (Sengupta 2016).

The main problem in pricing variance, volatility, and correlation swaps is how to determine the evolution of the stochastic processes for the underlying assets and their volatilities. Thus, sometimes it is simpler to consider the pricing of swaps by so-called pseudo-statistics, namely, the pseudo-variance, -covariance, -volatility, and -correlation (Badescu et al. 2002; RBC Financial Group Team 2002).

The main motivation of this paper is to consider the pricing of swaps based on pseudo-statistics, instead of stochastic models, and to compare this approach with the most popular stochastic volatility model in the Cox–Ingresoll–Ross (CIR) model. Within this paper, we will demonstrate how to value different types of swaps (variance, volatility, covariance, and correlation swaps) using pseudo-statistics (pseudo-variance, pseudo-volatility, pseudo-correlation, and pseudo-covariance). As a result, we will arrive at a method for pricing swaps that does not rely on any stochastic models for a stochastic stock price or stochastic volatility, and instead relies on data/statistics. A data/statistics-based approach to swap pricing is very different from stochastic volatility models such as the Cox–Ingresoll–Ross (CIR) model, which, in comparison, follows a stochastic differential equation (see (Swishchuk 2004)). Although there are many other stochastic models that provide an approach to calculating the price of swaps, we will use the CIR model for comparison within this paper, due to the popularity of the CIR model. Therefore, in this paper, we will compare the CIR model approach to pricing swaps to the pseudo-statistic approach to pricing swaps, in order to compare a stochastic model to the data/statistics-based approach to swap pricing that is developed within this paper.

In Section 2, we present the analytical closed-form formulas of pseudo-variance, pseudo-volatility, pseudo-correlation, and pseudo-covariance. Section 3 defines a pseudo-swap and presents the associated equations. In Section 4, four data sets that are to be used in the numerical examples that follow are defined. Section 5 defines the logarithmic return of a stock price, while in Section 6, we calculate the expected sample variance and expected sample volatility. The pseudo-variance, pseudo-volatility, pseudo-covariance, and pseudo-correlation are calculated in Sections 7, 8, 10, and 11, respectively. In Section 9, we calculate the expected sample covariance of the two stock data sets presented in Section 4. Section 12 outlines the data sets and purpose of the comparison between the pseudo-statistic approach and the approach based on the Cox–Ingresoll–Ross (CIR) model. The results of the numerical comparison between the variance and volatility swap payoffs when using the CIR model and the pseudo-statistic approach are shown in Sections 13 and 14, respectively. The paper is concluded in Section 15.

Within this paper, we will make use of statistical analysis, martingales, and stochastic calculus to provide analytical closed-form formulas of the pricing of swaps and calculate numerical values using these formulas. In addition, we have also used data analysis methods in the selection and preparation of the data sets that are presented in Section 4.

2. Pseudo-Statistics

The pseudo-statistics, as defined by (Badescu et al. 2002), are the estimators for the corresponding statistics. For example, if $R_i^{(k)} := \log(S_{t_i}^{(k)}/S_{t_{i-1}}^{(k)}), k = 1, 2, i = 1, 2, \ldots, n$, $t_0 = 0 < t_1 < t_2 < \ldots < t_n = T$, is the log return of the underlying asset $S_t^{(k)}, k = 1, 2$, then

the realized pseudo-variance is defined as

$$Var_n(S^{(k)}, T) := \frac{n}{(n-1)T} \sum_{i=1}^{n} (R_i^{(k)} - \bar{R}_n^{(k)})^2, \tag{1}$$

where $\bar{R}_n^{(k)} := \frac{1}{n} \sum_{i=1}^{n} R_i^{(k)}$, T is the maturity of the contract. It can be shown that the realized continuously sampled variance is given by $\sigma_R^2(S^{(k)}) = V^{(k)}(T) := \lim_{n \to +\infty} Var_n(S^{(k)}, T) = \frac{1}{T} \int_0^T \sigma_t^2(k) dt$, where $\sigma_t(k)$ is the volatility (stochastic, in general) of the underlying asset $S_t^{(k)}, k = 1, 2$. In the same way, as was shown in (Badescu et al. 2002), the other types of pseudo-statistics (such as pseudo-volatility, -covariance, and -correlation) can be defined as

$$Vol_n(S^{(k)}, T) := \sqrt{Var_n(S^{(k)}, T)} \tag{2}$$

for pseudo-volatility,

$$Cov_n(S^{(1)}, S^{(2)}, T) := \frac{n}{(n-1)T} \sum_{i=1}^{n} (R_i^{(1)} - \bar{R}_n^{(1)})(R_i^{(2)} - \bar{R}_n^{(2)}) \tag{3}$$

for pseudo-covariance, and

$$Corr_n(S^{(1)}, S^{(2)}, T) := \frac{Cov_n(S^{(1)}, S^{(2)}, T)}{\sqrt{Var_n(S^{(1)}, T)} \sqrt{Var_n(S^{(2)}, T)}} \tag{4}$$

for pseudo-correlation.

3. Pseudo-Swaps

3.1. Swaps

A stock volatility swap, as defined by (Demeterfi et al. 1999), is a forward contract on the annualized volatility. Its payoff at expiration is presented within (Demeterfi et al. 1999), and it is equal to

$$N(\sigma_R(S) - K_{vol}), \tag{5}$$

where σ_R is the realized stock volatility (quoted in annual terms) over the life of contract, K_{vol} is the annualized volatility delivery price, and N is the notional amount of the swap in dollars per annualized volatility point. The holder of a volatility swap at expiration receives N dollars for every point by which the stock's realized volatility σ_R has exceeded the volatility delivery price K_{vol}. The holder is swapping a fixed volatility K_{vol} for the actual (floating) future volatility σ_R. We note that usually $N = \alpha I$, where α is a converting parameter, such as \$1 per volatility, and I is a long/short index (+1 for long and -1 for short).

However, the variance or volatility squared are of higher significance, despite the talk of volatility by options market participants, as explained by (Demeterfi et al. 1999).

A variance swap, as defined by (Demeterfi et al. 1999), is a forward contract on annualized variance, the square of the realized volatility. Its payoff at expiration is presented within (Demeterfi et al. 1999), and it is equal to

$$N(\sigma_R^2(S) - K_{var}), \tag{6}$$

where $\sigma_R^2(S)$ is the realized stock variance (quoted in annual terms) over the life of the contract, K_{var} is the delivery price for variance, and N is the notional amount of the swap in dollars per annualized volatility point squared. The holder of variance swap at expiration receives N dollars for every point by which the stock's realized variance σ_R^2 has exceeded the variance delivery price K_{var}. Therefore, pricing the variance swap reduces to calculating the realized volatility square.

From a theoretical point of view, consistent with the information presented within (Salvi and Swishchuk 2014), pricing a covariance swap is similar to pricing variance swaps, using the following equation presented in (Brockhaus and Long 2000):

$$Cov_R(S^{(1)}, S^{(2)}) = 1/2[\sigma_R^2(S^{(1)}S^{(2)}) - \sigma_R^2(S^{(1)}) - \sigma_R^2(S^{(2)})], \tag{7}$$

where $S^{(1)}$ and $S^{(2)}$ are two given assets, $\sigma_R^2(S)$ is a variance swap for underlying assets, and $Cov_R(S^{(1)}, S^{(2)})$ is the realized covariance of the two underlying assets $S^{(1)}$ and $S^{(2)}$.

Within (Salvi and Swishchuk 2014), a covariance swap is defined as a covariance forward contract of the underlying rates $S^{(1)}$ and $S^{(2)}$, whose payoff at expiration is equal to

$$N(Cov_R(S^{(1)}, S^{(2)}) - K_{cov}), \tag{8}$$

where K_{cov} is a stock price, N is the notional amount, and $Cov_R(S^{(1)}, S^{(2)})$ is the covariance between two assets $S^{(1)}$ and $S^{(2)}$.

A correlation swap is defined by (Salvi and Swishchuk 2014) as a correlation forward contract of two underlying rates $S^{(1)}$ and $S^{(2)}$, whose payoff at expiration is equal to

$$N(Corr_R(S^{(1)}, S^{(2)}) - K_{corr}), \tag{9}$$

where $Corr(S^{(1)}, S^{(2)})$ is the realized correlation of two underlying assets $S^{(1)}$ and $S^{(2)}$, K_{corr} is a strike price, and N is the notional amount.

Finally, correlation $Corr_R(S^{(1)}, S^{(2)})$ is defined by (Salvi and Swishchuk 2014) as follows:

$$Corr_R(S^{(1)}, S^{(2)}) = \frac{Cov_R(S^{(1)}, S^{(2)})}{\sqrt{\sigma_R^2(S^{(1)})}\sqrt{\sigma_R^2(S^{(2)})}}, \tag{10}$$

where $Cov_R(S^{(1)}, S^{(2)})$ is defined in Equation (7) and $\sigma_R^2(S^{(1)})$ is the realized variance of asset $S^{(1)}$ (quoted in annual terms) over the life of the contract.

3.2. Pseudo-Swaps

A pseudo-variance swap is defined within (Badescu et al. 2002), as a forward contract with one unit of notional principal, a given maturity of $T > 0$, a strike K_{var}, and long/short index of I (+1 for long and -1 for short), such that the matured payoff of the contract, denoted by $C_{var}(S^{(k)}, T)$, is given by

$$C_{var}(S^{(k)}, T) := N_{var} I[Var_n(S^{(k)}, T) - K_{var}], \quad k = 1, 2, \tag{11}$$

where $Var_n(S^{(k)}, T)$ is defined in (1), and N_{var} is a converting parameter, such as \$1 per variance. It can be shown, using the framework presented in (Swishchuk 2004), that pricing the variance swaps reduces to calculating the expectation (with respect to the risk-neutral measure) of the pseudo-variance. In the same way, as was conducted in (Badescu et al. 2002), we can define the other types of pseudo-swaps, such as pseudo-volatility, pseudo-covariance, and pseudo-correlation swaps, based on pseudo-statistics defined in Equations (2)–(4). Therefore, we have the following formulas for pseudo-volatility, pseudo-covariance, and pseudo-correlation swaps, respectively:

$$C_{vol}(S^{(k)}, T) = N_{vol} I[Vol_n(S^{(k)}, T) - K_{var}], \quad k = 1, 2, \tag{12}$$

$$C_{cov}(S^{(k)}, T) = N_{cov} I[Cov_n(S^{(1)}, S^{(2)}, T) - K_{var}], \tag{13}$$

$$C_{corr}(S^{(k)}, T) = N_{corr} I[Corr_n(S^{(1)}, S^{(2)}, T) - K_{var}], \tag{14}$$

where $Vol_n(S^{(k)}, T)$, $Cov_n(S^{(1)}, S^{(2)}, T)$ and $Corr_n(S^{(1)}, S^{(2)}, T)$ are defined in Equations (2)–(4).

In this paper, we show how to price different types of swaps (variance, volatility, covariance, and correlation swaps) defined in Equations (5)–(9) using pseudo-statistics

(pseudo-variance, -volatility, -covariance, and -correlation) and pseudo-swaps defined in Equations (11)–(14).

4. Financial Data Used

For the following report, we have used real-life financial data from Yahoo Finance (https://ca.finance.yahoo.com/ (accessed on 7 June 2023)) of the publicly traded stocks belonging to

- Apple Inc. (AAPL)
- Alphabet Inc. Class C (GOOG).

In the sections that follow, we will make use of four data sets, which are described in the next two subsections.

4.1. 6-Month Data Sets

The first two data sets we will introduce contain 6 months of daily closing data for GOOG and AAPL, from 9 November 2022 to 8 May 2023. As a result, each data set has 123 entries, and we will define the following variables:

- Time in Years of First Data Entry: $T_s = 0$
- Time in Years of Last Data Entry: $T_e = 0.5$
- Duration of Data Collection in Years: $T_e - T_s = 0.5 - 0 = 0.5$.

We can then define each data entry in the data sets of AAPL and GOOG daily closing prices as follows:

AAPL Daily Closing Data at time t_i:

$$S^{(1)}_{t_i} \text{ for } t_i \in [T_e, T_s] \text{ and } T_s = t_0 < t_1 < \ldots < t_{122} = T_e.$$

Below, in Figure 1, a visualization of the daily closing price data of AAPL in the time period from 9 November 2022 to 8 May 2023 is provided.

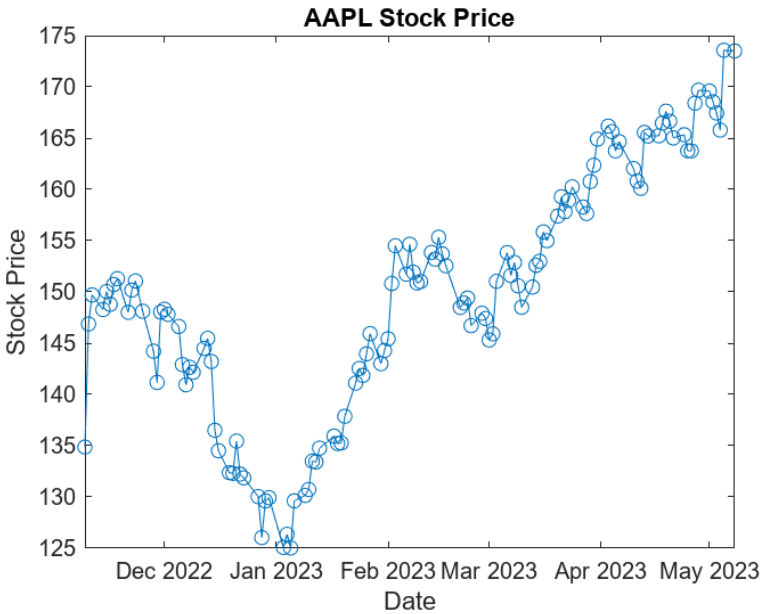


Figure 1. AAPL stock price over a 6-month period of November 2022 to May 2023.

GOOG Daily Closing Data at time t_i:

$$S^{(2)}_{t_i} \text{ for } t_i \in [T_e, T_s] \text{ and } T_s = t_0 < t_1 < \ldots < t_{122} = T_e.$$

In Figure 2, below, we present a visualization depicting the daily closing price data of GOOG in the 6-month time period from November 2022 to May 2023.

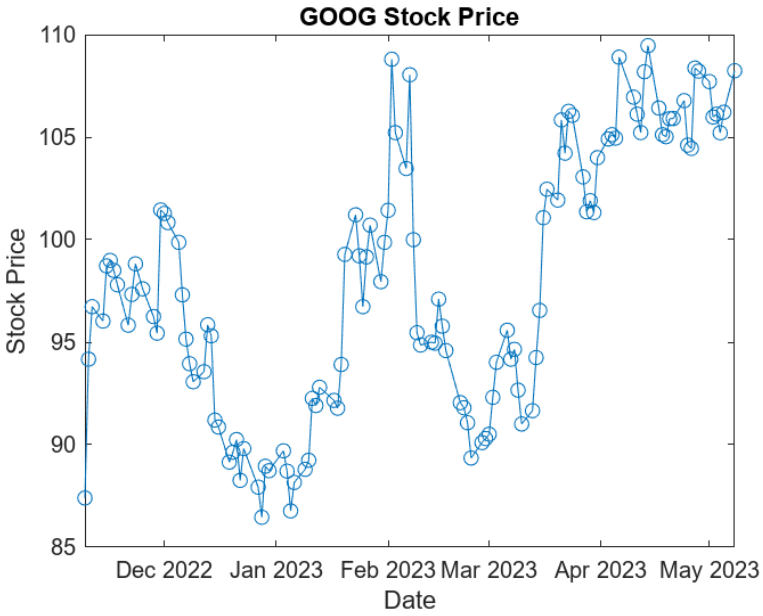


Figure 2. GOOG stock price over a 6-month period of November 2022 to May 2023.

4.2. One-Year Data Sets

In order to calculated the expected volatility and variance in 9 November, we will use two more data sets that contain 1 year of daily closing data for GOOG and AAPL, from 8 November 2021 to 7 November 2022. As a result, each data set has 252 entries, and we will define the following variables:

– Time in Years of First Data Entry: $T_s = 0$
– Time in Years of Last Data Entry: $T_e = 1$
– Duration of Data Collection in Years: $T_e - T_s = 1 - 0 = 1$.

We can then define each data entry in these data sets of AAPL and GOOG daily closing prices as follows:

AAPL Daily Closing Data at time t_i:

$$S^{(1)}_{t_i} \text{ for } t_i \in [T_e, T_s] \text{ and } T_s = t_0 < t_1 < \ldots < t_{251} = T_e.$$

Below, in Figure 3, a visualization of the daily closing price data of AAPL in the time period from November 2021 to November 2022.

GOOG Daily Closing Data at time t_i:

$$S^{(2)}_{t_i} \text{ for } t_i \in [T_e, T_s] \text{ and } T_s = t_0 < t_1 < \ldots < t_{251} = T_e.$$

In Figure 4, below, we present a visualization depicting the daily closing price data of GOOG in the 1-year time period from November 2021 to November 2022.

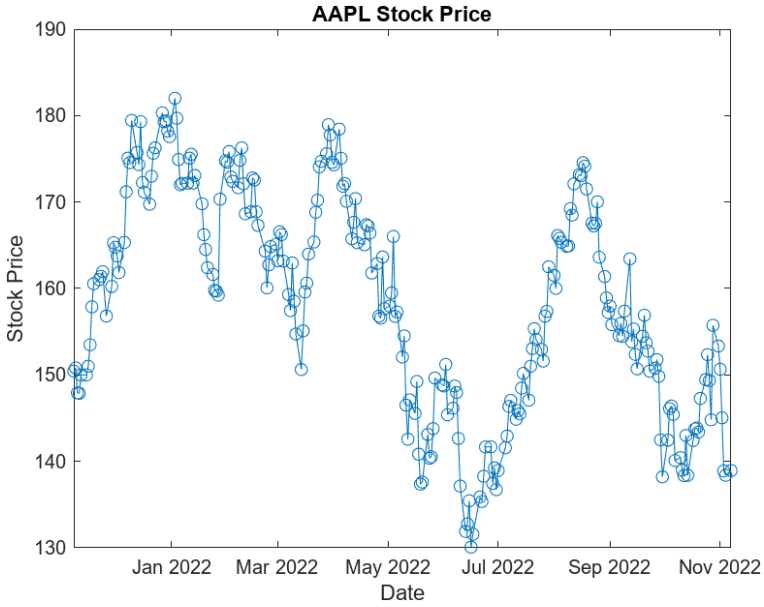


Figure 3. AAPL stock price over a 1-year period of November 2021 to November 2022.

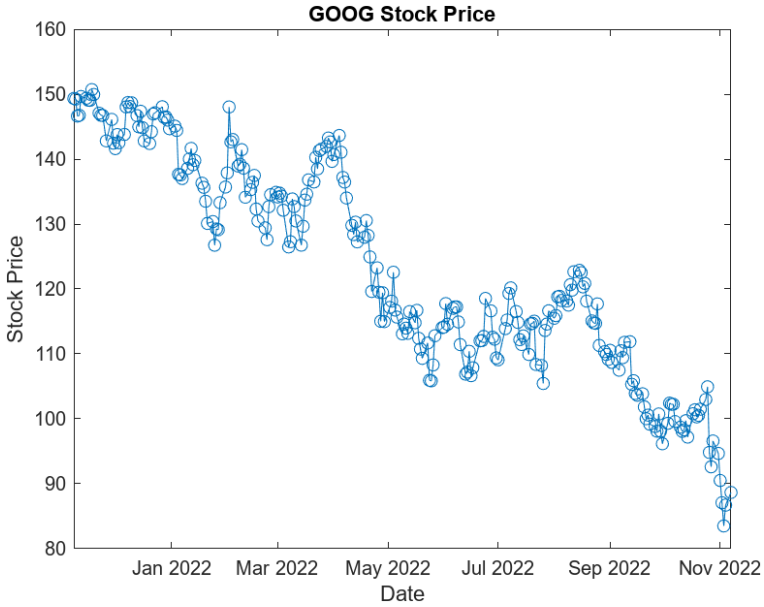


Figure 4. GOOG stock price over a 1-year period of November 2021 to November 2022.

5. Logarithmic Return of Stock Price

In order to calculate the realized pseudo-statistics and the expected pseudo-statistics for a future time period for the underlying stocks AAPL and GOOG, we must first calculate the logarithmic return of each stock price. Therefore, we need to define the logarithmic return and the arithmetic mean of the logarithmic returns. We will use the terminology and definitions found in (Badescu et al. 2002) Section 1.2, and the following equations, which can be found in (Badescu et al. 2002; Brockhaus and Long 2000; Carr and Lee 2009):

$$\text{Logarithmic Return: } R_i^{(k)} = ln\left(\frac{S_{t_i}^{(k)}}{S_{t_{i-1}}^{(k)}}\right), \tag{15}$$

where $t_i \in [T_s, T_e]$, $k = 1, 2$, and $i = 1, 2, \ldots, n$.

$$\text{Arithmetic Mean of Logarithmic Return: } \bar{R}_n^{(k)} = \frac{1}{n}\sum_{i=1}^{n} R_i^{(k)}, k = 1, 2, \tag{16}$$

where n is the number of logarithmic return data entries we need to calculate for each stock, and this number is 1 less than the total number of price data points. Using this information, along with Equations (15) and (16), we can now calculate the logarithmic returns and arithmetic mean of the logarithmic returns for the four data sets introduced in Section 4.

5.1. 6-Month Data Sets: Logarithmic Return and Arithmetic Mean

Using the data we described in Section 4.1, we can calculate the logarithmic return and the respective arithmetic means of the logarithmic returns of AAPL and GOOG over the 6-month period of November 2022 to May 2023, as follows:

Logarithmic Return of AAPL:

$$R_i^{(1)} = ln\left(\frac{S_{t_i}^{(1)}}{S_{t_{i-1}}^{(1)}}\right),$$

where $t_i \in [0, 0.5]$ and $i = 1, 2, \ldots, 122$.

Below, in Figure 5, we present an illustration depicting the daily logarithmic returns of AAPL that are calculated when Equation (15) is applied to the 6-month data set for AAPL described in Section 4.1.

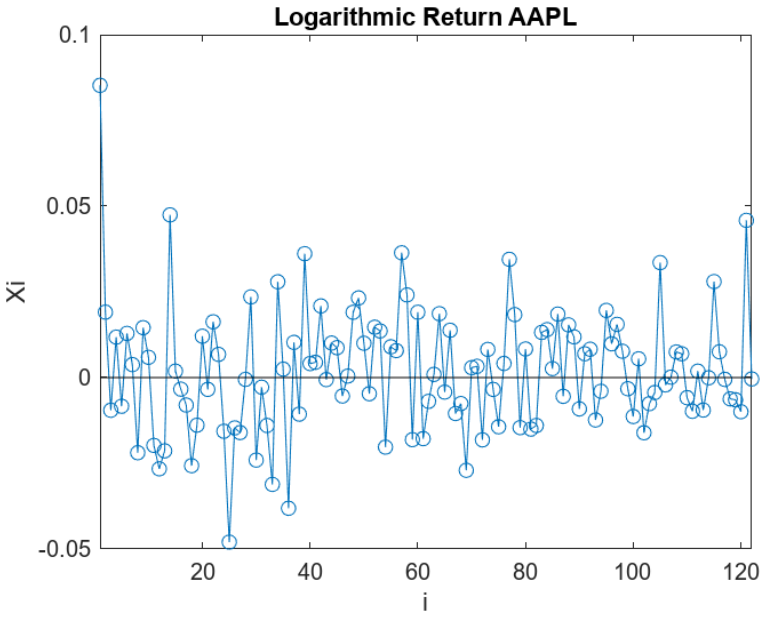


Figure 5. AAPL Logarithmic returns over a 6-month period of November 2022 to May 2023.

Arithmetic Mean of $R_i^{(1)}$:

$$\bar{R}_{122}^{(1)} = 0.0021.$$

Logarithmic Return of GOOG:

$$R_i^{(2)} = ln\left(\frac{S_{t_i}^{(2)}}{S_{t_{i-1}}^{(2)}}\right),$$

where $t_i \in [0, 0.5]$ and $i = 1, 2, \ldots, 122$.

Below, in Figure 6, we present the values of the daily logarithmic returns of GOOG calculated using Equation (15) and the 6-month data set for GOOG described in Section 4.1.

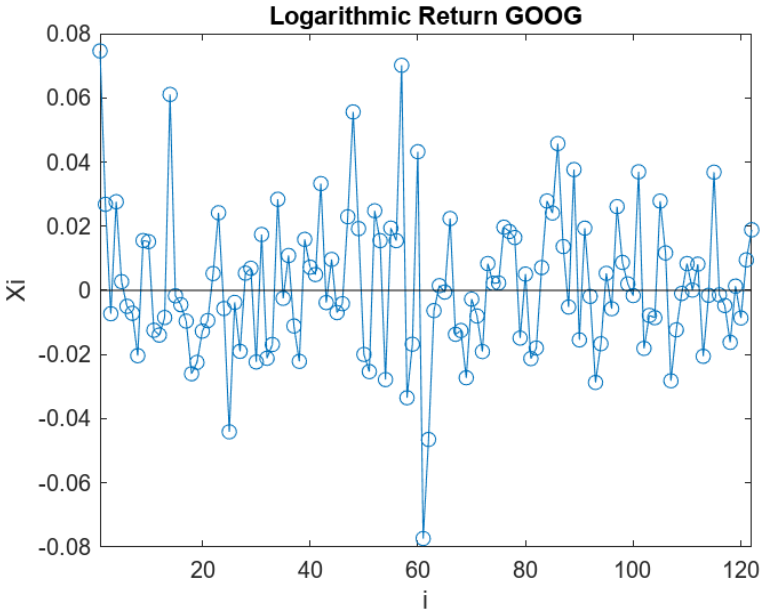


Figure 6. GOOG Logarithmic returns over a 6-month period of November 2022 to May 2023.

Arithmetic Mean of $R_i^{(2)}$:

$$\bar{R}_{122}^{(2)} = 0.0018.$$

5.2. One-Year Data Sets: Logarithmic Return and Arithmetic Mean

In a similar fashion, we will now use the data we described in Section 4.2 to calculate the logarithmic return and the respective arithmetic means of the logarithmic returns of AAPL and GOOG over the 1-year period of November 2021 to November 2022, as follows:

Logarithmic Return of AAPL:

$$R_i^{(1)} = ln\left(\frac{S_{t_i}^{(1)}}{S_{t_{i-1}}^{(1)}}\right),$$

where $t_i \in [0,1]$ and $i = 1, 2, \ldots, 251$.

Below, we present Figure 7, which depicts the daily logarithmic returns of AAPL over the 1-year time period from November 2021 to November 2022, which were calculated using the 1-year data set for AAPL and Equation (15).

Arithmetic Mean of $R_i^{(1)}$:

$$\bar{R}_{251}^{(1)} = -0.00031739.$$

Logarithmic Return of GOOG:

$$R_i^{(2)} = ln\left(\frac{S_{t_i}^{(2)}}{S_{t_{i-1}}^{(2)}}\right),$$

where $t_i \in [0,1]$ and $i = 1, 2, \ldots, 251$.

In Figure 8, below, we present the daily logarithmic returns of GOOG over the 1-year time period from November 2021 to November 2022, which were calculated using the 1-year data set for GOOG and Equation (15).

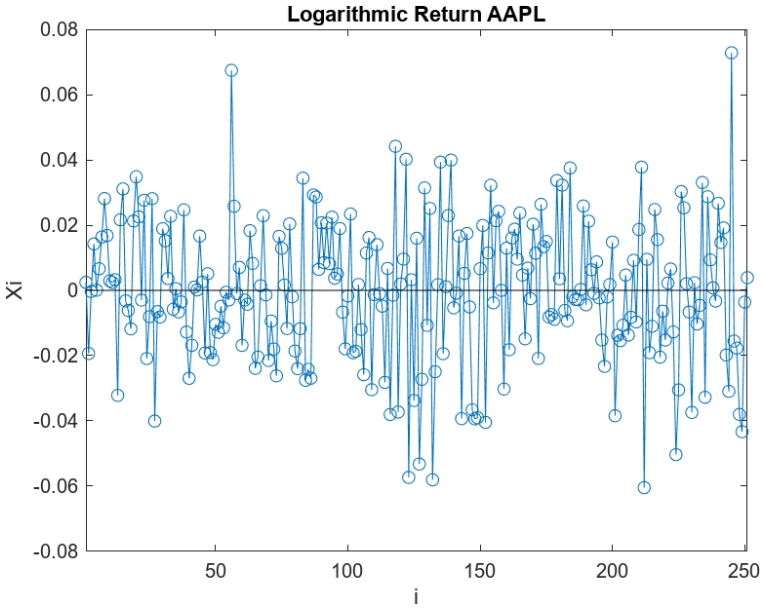


Figure 7. AAPL Logarithmic returns over a 1-year period of November 2021 to November 2022.

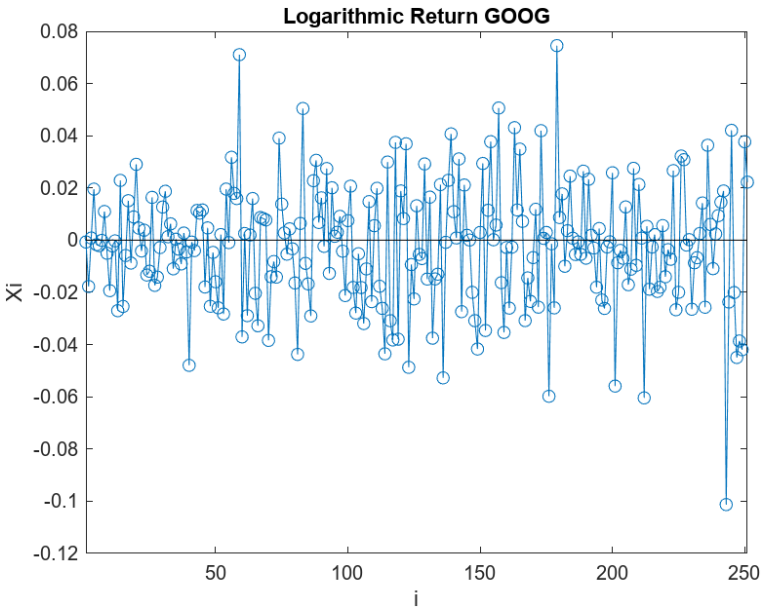


Figure 8. GOOG Logarithmic returns over a 1-year period of November 2021 to November 2022.

Arithmetic Mean of $R_i^{(2)}$:

$$\bar{R}_{251}^{(2)} = -0.0021.$$

6. Expected Sample Variance and Brockhaus–Long Approximation for Expected Sample Volatility

In order to calculate the payoffs of pseudo-variance and volatility swaps, we must first calculate the expected sample variance and the expected sample volatility at the start of the swap contract, on 9 November 2022. In order to calculate the expected sample variance and volatility, we will use the Brockhaus–Long approximation (Brockhaus and Long 2000), and run a GARCH(1,1) regression on the logarithmic return data generated from the 1-year data sets introduced in Section 4.2, as was conducted in (Javaheri et al. 2004), while defining the following parameters:

– Maturity Date in Years: T

- Number of Logarithmic Return Entries: n.
- GARCH(1,1) Constant: C
- Kurtosis of Logarithmic Returns: ξ.

Using the regression model and the parameters, we can then define the following formulas, which are dependent on the Heston model of securities markets, as described and derived in (Swishchuk 2004). We are using these formulas and the Heston model of securities markets, on which the formulas are based for comparison purposes, in order to obtain numerical values for the expected sample variance and volatility of a data set:

$$ARCH(1,1) = \alpha \tag{17}$$

$$GARCH(1,1) = \beta. \tag{18}$$

$$\text{Interval of Time in Years between Price Data Entries: } dt = \frac{T_e - T_s}{n}; \tag{19}$$

$$V = \frac{C}{1-\alpha-\beta}; \tag{20}$$

$$\text{Short Volatility: } \sigma_0 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2}; \tag{21}$$

$$\text{Long Volatility: } \theta = \frac{V}{dt}; \tag{22}$$

$$\text{Reversion Speed: } k = \frac{1-\alpha-\beta}{dt}; \tag{23}$$

$$\gamma = \alpha\sqrt{\frac{\xi-1}{dt}}; \tag{24}$$

$$\text{Expected Sample Variance: } E(V) = \frac{1-e^{-kT}}{kT}(\sigma_0^2 - \theta^2) + \theta^2; \tag{25}$$

$$Var(V) = \frac{\gamma^2 e^{-2kT}}{2k^3T^2}(2e^{2kT} - 4e^{kT}kT - 2)(\sigma_0^2 - \theta^2) + (2e^{2kT}kT - 3e^{2kT} + 4e^{kT} - 1)\theta^2; \tag{26}$$

$$\text{Convexity Adjustment: } \frac{Var(V)}{8E(V)^{3/2}}; \tag{27}$$

$$\text{Expected Volatility: } E(\sqrt{\sigma_R^2(S)}) \approx \sqrt{E(V)} - \text{Convexity Adjustment.} \tag{28}$$

Having defined the previous equations and the parameters that go with them, we can now apply these equations to the AAPL and GOOG data sets.

6.1. AAPL: Expected Variance and Volatility

We will use the 1-year data set of AAPL daily closing price data that we introduced in Section 4.2, while assuming a maturity date of 6 months, from 9 November 2022. As a result, we have the following parameters:

- Maturity Date in Years: $T = 0.5$
- Number of Logarithmic Return Entries: $n = 251$
- Kurtosis of AAPL Logarithmic Returns: $\xi = 3.3689$
- Short Volatility: $\sigma_0 = 0.0216$.

We then apply a GARCH(1,1) regression on the logarithmic AAPL stock return data, the results of which are presented in Table 1, as follows:

Table 1. GARCH(1,1) regression on AAPL logarithmic returns.

	Value	**Standard Error**	**T Statistic**	***p* Value**
Constant	3.9818×10^{-5}	6.0253×10^{-5}	0.66084	0.50871
GARCH{1}	0.87202	0.16901	5.1595	2.4762×10^{-7}
ARCH{1}	0.045118	0.045974	0.9814	0.3264

From the data presented in Table 1, along with the parameters defined above and Equations (17)–(28), we can calculate the following results:

Expected Sample Variance:

$$E(V) = 0.0132.$$

Expected Sample Volatility:

$$E(\sqrt{\sigma_R^2(S^{(1)})}) \approx 0.109791.$$

6.2. GOOG: Expected Variance and Volatility

Similarly, we will use the 1-year data set of GOOG daily closing price data that we introduced in Section 4.2, while assuming a maturity date of 6 months, from 9 November 2022. As a result, we have the following parameters:

– Maturity Date in Years: $T = 0.5$
– Number of Logarithmic Return Entries: $n = 251$
– Kurtosis of GOOG Logarithmic Returns: $\xi = 4.3086$
– Short Volatility: $\sigma_0 = 0.0233$.

We then apply a GARCH(1,1) regression on the logarithmic GOOG stock return data, which results in Table 2, as follows:

Table 2. GARCH(1,1) regression on GOOG logarithmic returns.

	Value	**Standard Error**	**T Statistic**	***p* Value**
Constant	1.8517×10^{-5}	2.5736×10^{-5}	0.71949	0.47184
GARCH{1}	0.94614	0.062937	15.033	4.4556×10^{-51}
ARCH{1}	0.024207	0.025447	0.95125	0.34148

From the data presented in Table 2, along with the parameters defined above and Equations (17)–(28), we can calculate the following results:

Expected Sample Variance:

$$E(V) = 0.0183.$$

Expected Sample Volatility:

$$E(\sqrt{\sigma_R^2(S^{(2)})}) \approx 0.126677.$$

7. Realized Pseudo-Volatility Square and Pseudo-Variance Swap Payoff

In order to calculate the realized pseudo-volatility square and the payoff of a pseudo-variance swap, we must first define the following parameters:

- Position Taken: $I = \begin{cases} 1 \text{ if Long Position Taken} \\ -1 \text{ if Short Position Taken} \end{cases}$
- Converting Parameter: α_{var}
- Strike Price: $\sum_K^2 = E(V)$.

Using these parameters, we can then utilize the following pair of equations found in (Badescu et al. 2002), in order to calculate the realized pseudo-volatility square and pseudo-variance swap payoff:

Realized Pseudo-Volatility Square:

$$\hat{\sum}_{(S)}^2(n; T_s, T_e) = \frac{n}{T_e - T_s}\left(\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R_n})^2\right). \tag{29}$$

Payoff of Pseudo-Variance Swap:

$$V_{var}(T) = \alpha_{var}\, I\left[\hat{\sum}_{(S)}^2(n; T_s, T_e) - \sum_K^2\right]. \tag{30}$$

We will use Equations (29) and (30), along with the defined parameters above, in the following sections, in order to calculate realized pseudo-volatility squares and pseudo-variance swap payoffs using the 6-month data sets introduced in Section 4.1.

7.1. AAPL: Realized Pseudo-Volatility Square and Pseudo-Variance Swap Payoff

Using the daily closing prices for AAPL in the 6-month period from November 2022 to May 2023, along with Equations (15), (16) and (29), we can calculate the realized pseudo-volatility square of AAPL to be the following:

Realized Pseudo-Volatility Square of AAPL:

$$\hat{\sum}_{(S^{(1)})}^2(122; 0, 0.5) = 0.0805.$$

As previously calculated in Section 5.1, the mean of $R_i^{(1)}$ is 0.0021, which is the value we used to calculate the realized pseudo-volatility square above. However, we can examine the relationship between the realized pseudo-volatility square of AAPL and the arithmetic mean of the logarithmic AAPL stock returns by letting $\bar{R}_{122}^{(1)}$ vary around its true value. As a result, we can produce Figure 9, which is below.

Now, using the realized pseudo-volatility square from November 2022 to May 2023 of the AAPL stock, which we calculated above, we can calculate the payoff of a pseudo-variance swap, whose underlying asset is the variance of the AAPL stock. We will assume the pseudo-variance swap has a maturity of 6 months, and that we are interested in the payoff associated with taking a long position on the swap. Additionally, we will assume that the payoff is calculated using a converting parameter of \$1 per unit of pseudo-statistic, and we will use the expected sample variance of AAPL calculated in Section 6.1 as the strike price. Therefore, we now have the following parameter values:

- Maturity Date: $T = 0.5$
- Position Taken: $I = 1$.
- Converting Parameter: $\alpha_{var} = 1$
- Strike Price: $\sum_K^2 = E(V) = 0.0132$.

Using these parameters, Equation (16), and the previously calculated realized pseudo-volatility square of AAPL, we can calculate the payoff of a pseudo-variance, swap whose underlying asset is the variance of the AAPL stock, to be the following:

Payoff of Pseudo-Variance Swap with Underlying of Variance of AAPL:

$$V_{var}(0.5) = 0.0673.$$

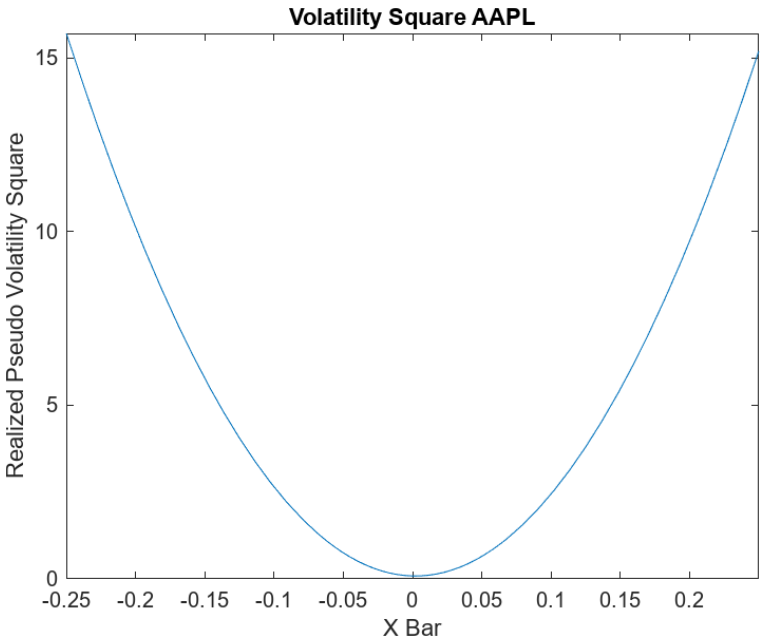


Figure 9. Realized pseudo−volatility square of AAPL from November 2022 to May 2023, as a function of $\bar{R}_{122}^{(1)}$.

7.2. GOOG: Realized Pseudo-Volatility Square and Pseudo-Variance Swap Payoff

Similarly, we use the 6 months of daily closing prices for GOOG, along with Equations (15), (16) and (29), to calculate the realized pseudo-volatility square of GOOG to be the following:

Realized Pseudo-Volatility Square of GOOG:

$$\hat{\sum}^2_{(S^{(2)})}(122; 0, 0.5) = 0.1269.$$

In Section 5.1, we found that $\bar{R}_{122}^{(2)}$ is 0.0018; however, we can allow this value to vary in order to examine the relationship between the realized pseudo-volatility square of GOOG and the arithmetic mean of the logarithmic GOOG stock returns. As a result, we can produce Figure 10, which is included below.

Once again, we will calculate the payoff of a long position on a 6-month pseudo-variance swap with a conversion factor of 1. However, the underlying asset of the variance swap will now be the variance of GOOG, and we will use the expected sample variance of GOOG calculated in Section 6.2 as the strike price. Therefore, we will now use the following parameter values:

- Maturity Date: $T = 0.5$
- Position Taken: $I = 1$
- Converting Parameter: $\alpha_{var} = 1$
- Strike Price: $\sum_K^2 = E(V) = 0.0183$.

Using these parameters, Equation (30) and the previously calculated realized pseudo-volatility square of GOOG, we can calculate the payoff of a pseudo-variance swap whose underlying asset is the variance of the GOOG stock to be

Payoff of Pseudo-Variance Swap with Underlying of Variance of GOOG:

$$V_{var}(0.5) = 0.1086.$$

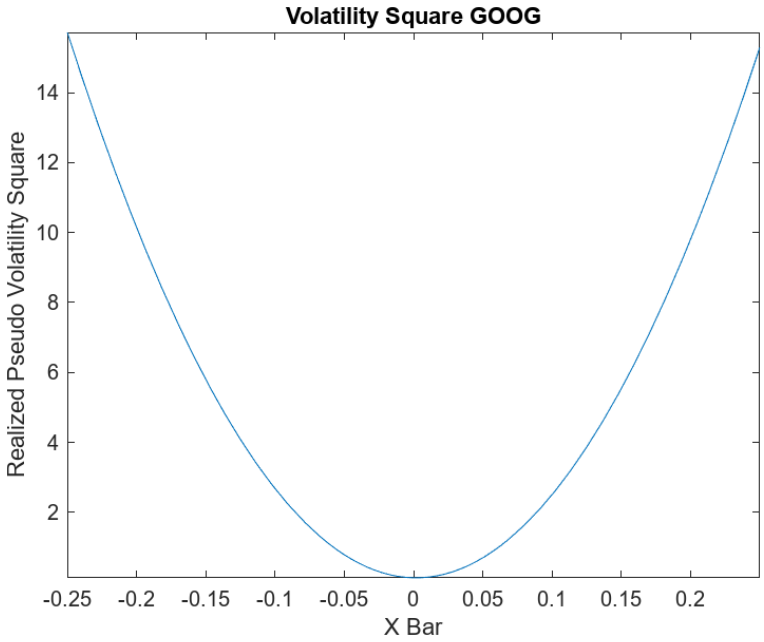


Figure 10. Realized pseudo−volatility square of GOOG from November 2022 to May 2023, as a function of $\bar{R}^{(2)}_{122}$.

8. Realized Pseudo-Volatility and Pseudo-Volatility Swap Payoff

To calculate the realized pseudo-volatility and the payoff of a pseudo-volatility swap, we must first define the following parameters:

– Position Taken: $I = \begin{cases} 1 \text{ if Long Position Taken} \\ -1 \text{ if Short Position Taken} \end{cases}$
– Converting Parameter: α_{vol}
– Strike Price: $\sigma_K = E(\sqrt{\sigma^2_{\bar{R}}(S)})$.

Using these parameters allows us to utilize the following two equations, found in (Badescu et al. 2002), to calculate the realized pseudo-volatility and pseudo-volatility swap payoff:

Realized Pseudo-Volatility:

$$\hat{\sigma}_{(S)}(n; T_s, T_e) = \sqrt{\frac{n}{T_e - T_s}\left(\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R_n})^2\right)}. \tag{31}$$

Payoff of Pseudo-Volatility Swap:

$$V_{vol}(T) = \alpha_{vol}\, I\left[\hat{\sigma}_{(S)}(n; T_s, T_e) - \sigma_K\right]. \tag{32}$$

We will use Equations (31) and (32), along with the defined parameters above, in the following two sections, to calculate realized pseudo-volatilities and pseudo-volatility swap payoffs using the data sets first introduced in Section 4.1.

8.1. AAPL: Realized Pseudo-Volatility and Pseudo-Volatility Swap Payoff

Using Equations (15), (16) and (31), along with the daily closing prices for AAPL in the 6-month period from November 2022 to May 2023, we can calculate the realized pseudo-volatility of AAPL to be the following:

Realized Pseudo-Volatility of AAPL:

$$\hat{\sigma}_{(S^{(1)})}(122; 0, 0.5) = 0.2838.$$

Previously, in Section 5.1, we found that $\bar{R}^{(1)}_{122}$ is equal to 0.0021, which we used above to calculate the realized pseudo-volatility. However, we can examine the relationship between the realized pseudo-volatility of AAPL and the arithmetic mean of the logarithmic AAPL stock returns by letting $\bar{R}^{(1)}_{122}$ vary around its true value. As a result, we can produce Figure 11 below.

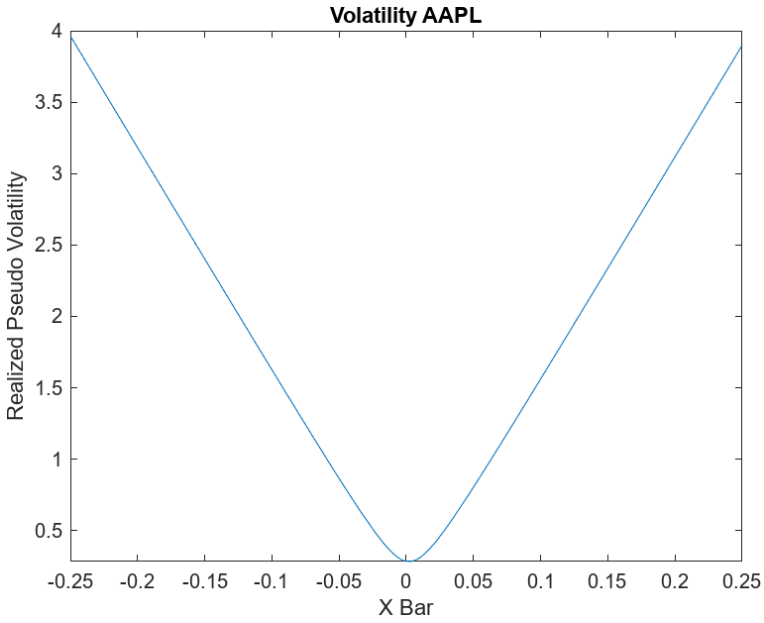


Figure 11. Realized pseudo−volatility of AAPL from November 2022 to May 2023, as a function of $\bar{R}^{(1)}_{122}$.

Using the realized pseudo-volatility calculated above, we can calculate the payoff of a pseudo-volatility swap, whose underlying asset is the volatility of the AAPL stock over the 6-month period from November 2022 to May 2023. We will assume the pseudo-variance swap has a maturity of 6 months, and that we are interested in the payoff associated with taking a long position on the swap. Additionally, we will assume that the payoff is calculated using a converting parameter of \$1 per unit of pseudo-statistic, and we will use the expected sample volatility of AAPL calculated in Section 6.1 as the strike price. Therefore, we can now write the following parameter values:

- Maturity Date: $T = 0.5$
- Position Taken: $I = 1$
- Converting Parameter: $\alpha_{vol} = 1$
- Strike Price: $\sigma_K = E(\sqrt{\sigma^2_R(S^{(1)})}) = 0.109791$.

Using these parameters along with Equation (32) and the previously calculated realized pseudo-volatility of AAPL, we can calculate the payoff of a pseudo-volatility swap, whose underlying asset is the volatility of the AAPL stock, to be the following:

Payoff of Pseudo-Volatility Swap with Underling of Volatility of AAPL:

$$V_{vol}(0.5) = 0.174009.$$

8.2. GOOG: Realized Pseudo-Volatility and Pseudo-Volatility Swap Payoff

Similarly, we can use the 6 months of daily closing prices for GOOG, along with Equation (15), (16) and (31), to calculate the realized pseudo-volatility of GOOG to be the following:

Realized Pseudo-Volatility of GOOG:

$$\hat{\sigma}_{(S^{(2)})}(122; 0, 0.5) = 0.3563.$$

In Section 5.1, we found that the mean of $R_i^{(2)}$ is equal to 0.0018, yet we can allow this value to vary in order to examine the relationship between the realized pseudo-volatility of GOOG and the arithmetic mean of the logarithmic GOOG stock returns. As a result, we can produce Figure 12 below.

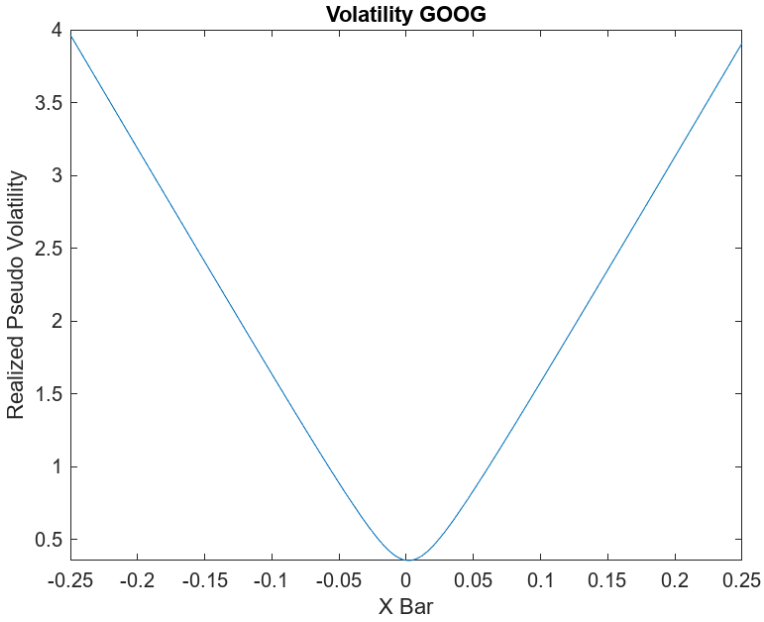


Figure 12. Realized pseudo−volatility of GOOG from November 2022 to May 2023, as a function of $\bar{R}_{122}^{(2)}$.

Once again, we will calculate the payoff of a long position on a 6-month pseudo-volatility swap with a conversion factor of 1. However, the underlying asset of the volatility swap will now be the volatility of GOOG, and we will use the expected sample volatility of GOOG calculated in Section 6.2 as the strike price. Therefore, we will now use the following parameter values:

- Maturity Date: $T = 0.5$
- Position Taken: $I = 1$
- Converting Parameter: $\alpha_{vol} = 1$
- Strike Price: $\sigma_K = E(\sqrt{\sigma_R^2(S^{(2)})}) = 0.126677$.

Using these parameters, Equation (32), and the previously calculated pseudo-volatility of GOOG, we can calculate the payoff of a pseudo-volatility swap, whose underlying asset is the volatility of the GOOG stock, to be:

Payoff of Pseudo-Volatility Swap with Underling of Volatility of GOOG:

$$V_{vol}(0.5) = 0.229623.$$

9. Expected Sample Covariance

In order to calculate the payoffs of pseudo-covariance swaps, we must first calculate the expected sample covariance at the start of the swap contract, on 9 November 2022. We will also use the equations and parameters we defined in Section 6, and we will again run a GARCH(1,1) regression on the logarithmic return data generated from the 1-year data sets introduced in Section 4.2. As a result, we can define the equation, found in (Swishchuk 2004), as follows:

Expected Sample Covariance:

$$E(Cov(S^{(1)}, S^{(2)})) = \frac{1}{4}(E(\sigma_R^2(S^{(1)}S^{(2)})) - E(\sigma_R^2(S^{(1)}/S^{(2)})), \tag{33}$$

where $E(\sigma_R^2(S)) = E(V)$ is calculated using Equation (25) defined in Section 6. Using these equations, we can now apply them to the AAPL ($S^{(1)}$) and GOOG ($S^{(2)}$), 1-year data sets defined in Section 4.2, in order to calculate the expected sample covariance.

9.1. $(S^{(1)}S^{(2)})$*: Expected Variance*

Using the 1-year data sets of daily closing prices of AAPL ($S^{(1)}$) and GOOG ($S^{(2)}$), and multiplying each entry within them, we can form a new data set, denoted $(S^{(1)}S^{(2)})$. This new data set, along with Equation (7) from Section 6, allows us to define the following parameters:

- Maturity Date in Years: $T = 0.5$
- Number of Logarithmic Return Entries: $n = 251$
- Kurtosis of GOOG Logarithmic Returns: $\xi = 3.0048$
- Short Volatility: $\sigma_0 = 0.0421$.

We then apply a GARCH(1,1) regression on the logarithmic $(S^{(1)}S^{(2)})$ return data, which results in Table 3, as follows:

Table 3. GARCH(1,1) regression on $(S^{(1)}S^{(2)})$ logarithmic returns.

	Value	**Standard Error**	**T Statistic**	***p* Value**
Constant	6.5445×10^{-5}	0.00010903	0.60024	0.54835
GARCH{1}	0.92637	0.083359	11.113	1.0847×10^{-28}
ARCH{1}	0.040856	0.035391	1.1544	0.24833

From the data presented in Table 3, along with the parameters defined above and Equations (15)–(28), we can calculate the following results:

Expected Sample Variance:

$$E(V) = 0.1916.$$

9.2. $(S^{(1)}/S^{(2)})$*: Expected Variance*

Similarly, we will use the 1-year data sets of daily closing prices of AAPL ($S^{(1)}$) and GOOG ($S^{(2)}$), and divide each entry within them, to form a new data set, denoted $(S^{(1)}/S^{(2)})$, which, along with Equation (7) from Section 6, allows us to define the following parameters:

- Maturity Date in Years: $T = 0.5$
- Number of Logarithmic Return Entries: $n = 251$
- Kurtosis of GOOG Logarithmic Returns: $\xi = 7.1870$

– Short Volatility: $\sigma_0 = 0.0157$.

We then apply a GARCH(1,1) regression on the logarithmic $(S^{(1)}/S^{(2)})$ return data, which results in Table 4, as follows:

Table 4. GARCH(1,1) regression on $(S^{(1)}/S^{(2)})$ logarithmic returns.

	Value	**Standard Error**	**T Statistic**	***p* Value**
Constant	0.00024031	0.012376	0.019416	0.98451
GARCH{1}	2×10^{-12}	51.535	3.8809×10^{-14}	1
ARCH{1}	2×10^{-12}	0.05022	3.9825×10^{-11}	1

From the data presented in Table 4, along with the parameters defined above and Equations (15)–(28), we can calculate the following results:

Expected Sample Variance:

$$E(V) = 0.0036.$$

9.3. Calculating the Expected Sample Covariance of AAPL and GOOG

Using the results calculated above in Sections 9.1 and 9.2, along with Equation (33), we can now calculate the expected covariance to be the following:

Expected Sample Covariance of AAPL and GOOG:

$$E(Cov(S^{(1)}, S^{(2)})) = 0.0470.$$

10. Realized Pseudo-Volatility Cross and Pseudo-Covariance Swap Payoff

In order to calculate the realized pseudo-volatility cross and the payoff of a pseudo-covariance swap, we must first define the following parameters:

– Position Taken: $I = \begin{cases} 1 \text{ if Long Position Taken} \\ -1 \text{ if Short Position Taken} \end{cases}$
– Converting Parameter: α_{cov}
– Strike Price: $\sum_K^2 = E(Cov(S^{(1)}, S^{(2)}))$.

Using these parameters allows us to utilize the following pair of equations, found in (Badescu et al. 2002), to calculate the realized pseudo-volatility cross and pseudo-covariance swap payoff:

Realized Pseudo-Volatility Cross:

$$\hat{\sum}^2_{(S^{(1)},S^{(2)})}(n; T_s, T_e) = \frac{n}{T_e - T_s}\left(\frac{1}{n-1}\sum_{i=1}^{n}\prod_{k=1}^{2}(R_i^{(k)} - \bar{R}_n^{(k)})\right). \tag{34}$$

Payoff of Pseudo-Covariance Swap:

$$V_{cov}(T) = \alpha_{cov}\, I\left[\hat{\sum}^2_{(S^{(1)},S^{(2)})}(n; T_s, T_e) - \sum_K^2\right]. \tag{35}$$

Using Equations (15), (16) and (34), along with the daily closing prices for AAPL and GOOG in the 6-month period from November 2022 to May 2023, we can calculate the realized pseudo-volatility cross of AAPL and GOOG to be the following:

Realized Pseudo-Volatility Cross of AAPL and GOOG:

$$\hat{\sum}^2_{(S^{(1)},S^{(2)})}(122; 0, 0.5) = 0.0712.$$

As previously calculated in Section 5.1, $\bar{R}_{122}^{(1)} = 0.0021$; thus, we can allow $\bar{R}_{122}^{(1)}$ to vary around its true value in order to allow us to examine the relationship between the realized pseudo-volatility cross of AAPL and GOOG over 6 months, and the arithmetic mean of the logarithmic AAPL stock returns. As a result, we can produce Figure 13, which is presented below.

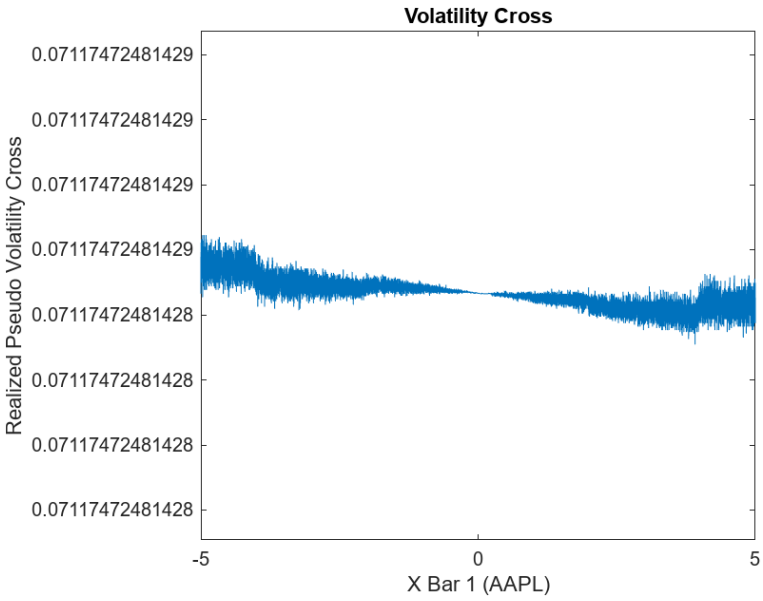


Figure 13. Realized pseudo−volatility cross of AAPL and GOOG from November 2022 to May 2023, as a function of $\bar{R}_{122}^{(1)}$.

Similarly, we can allow $\bar{R}_{122}^{(2)}$ to vary around its true value of 0.0018 to examine the relationship between the realized pseudo-volatility cross of AAPL and GOOG over 6 months, and the arithmetic mean of the logarithmic GOOG stock returns. As a result, we can produce Figure 14, which is presented below.

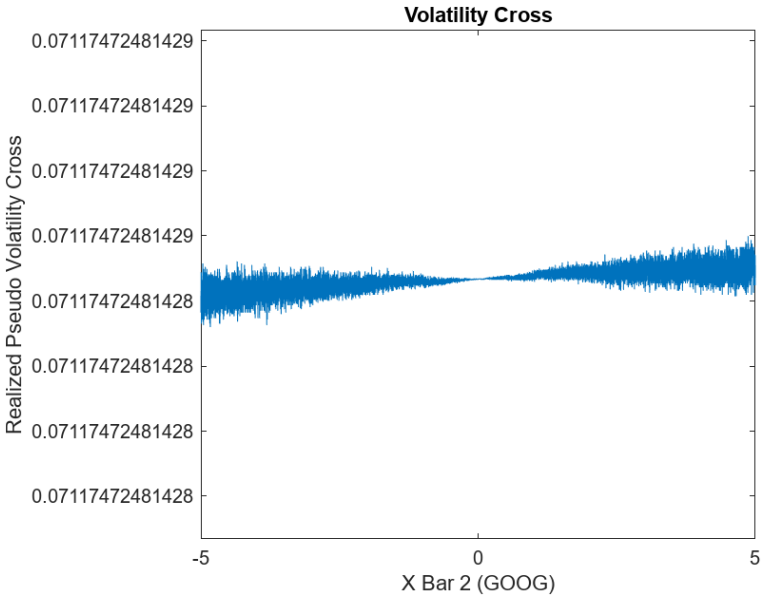


Figure 14. Realized pseudo−volatility cross of AAPL and GOOG from November 2022 to May 2023, as a function of $\bar{R}_{122}^{(2)}$.

Using the realized pseudo-volatility cross calculated above, we can calculate the payoff of a pseudo-covariance swap, whose underlying asset is the covariance of AAPL and

GOOG stocks over the 6-month period from November 2022 to May 2023. We will assume the pseudo-covariance swap has a maturity of 6 months, and that we are interested in the payoff associated with taking a long position on the swap. Additionally, we will assume that the payoff is calculated using a converting parameter of \$1 per unit of pseudo-statistic, and we will use the expected sample covariance calculated in Section 9.3 as the strike price. Therefore, we can now write the following parameter values:

- Maturity Date: $T = 0.5$
- Position Taken: $I = 1$
- Converting Parameter: $\alpha_{cov} = 1$
- Strike Price: $\sum_K^2 = E(Cov(S^{(1)}, S^{(2)})) = 0.0470$.

Using these parameters, Equation (21), and the previously calculated pseudo-volatility cross of AAPL and GOOG, we can calculate the payoff of a pseudo-covariance swap, whose underlying asset is the covariance of AAPL and GOOG, to be

Payoff of Pseudo-Covariance Swap with Underlying of Covariance of AAPL and GOOG:

$$V_{cov}(0.5) = 0.0242.$$

11. Realized Pseudo-Correlation and Pseudo-Correlation Swap Payoff

In order to calculate the realized pseudo-correlation and the payoff of a pseudo-correlation swap, we must first define the following parameters:

- Position Taken: $I = \begin{cases} 1 \text{ if Long Position Taken} \\ -1 \text{ if Short Position Taken} \end{cases}$
- Converting Parameter: α_{corr}
- Strike Price: ρ_K.

Using these parameters allows us to utilize the following two equations, found in (Badescu et al. 2002), to calculate the realized pseudo-correlation and pseudo-correlation swap payoff:

Realized Pseudo-Correlation:

$$\hat{\rho}_{(S^{(1)},S^{(2)})}(n; T_s, T_e) = \frac{\sum_{i=1}^{n} \prod_{k=1}^{2} (R_i^{(k)} - \bar{R}_n^{(k)})}{\prod_{k=1}^{2} \sqrt{\sum_{i=1}^{n} \left(R_i^{(k)} - \bar{R}_n^{(k)}\right)^2}}. \tag{36}$$

Payoff of Pseudo-Correlation Swap:

$$V_{corr}(T) = \alpha_{corr}\, I \left[\hat{\rho}_{(S^{(1)},S^{(2)})}(n; T_s, T_e) - \rho_K\right]. \tag{37}$$

Using Equations (15), (16) and (36), along with the data sets of daily closing prices for AAPL and GOOG introduced in Section 4.1, we can calculate the realized pseudo-correlation of AAPL and GOOG to be the following:

Realized Pseudo-Correlation of AAPL and GOOG:

$$\hat{\rho}_{(S^{(1)},S^{(2)})}(122; 0, 0.5) = 0.7039.$$

Back in Section 5.1, we found that $\bar{R}_{122}^{(1)} = 0.0021$; thus, in order to examine the relationship between the realized pseudo-correlation of AAPL and GOOG over 6 months, and the arithmetic mean of the logarithmic AAPL stock returns, we can allow $\bar{R}_{122}^{(1)}$ to vary around its true value. As a result, we can produce Figure 15 below.

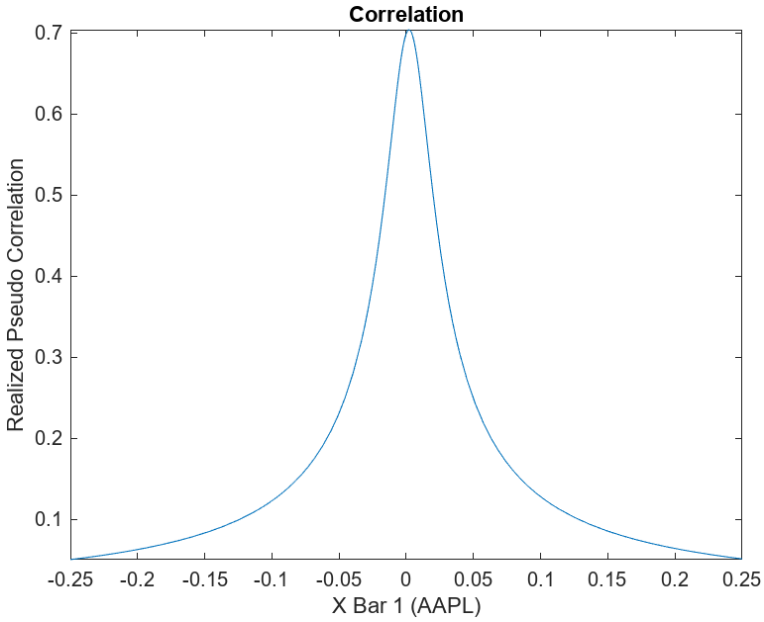


Figure 15. Realized pseudo−correlation of AAPL and GOOG from November 2022 to May 2023, as a function of $\bar{R}_{122}^{(1)}$.

Similarly, we can allow $\bar{R}_{122}^{(2)}$ to vary around its true value of 0.0018 to examine the relationship between the realized pseudo-correlation of AAPL and GOOG over 6 months, and the arithmetic mean of the logarithmic GOOG stock returns. In this case, we can produce Figure 16 below.

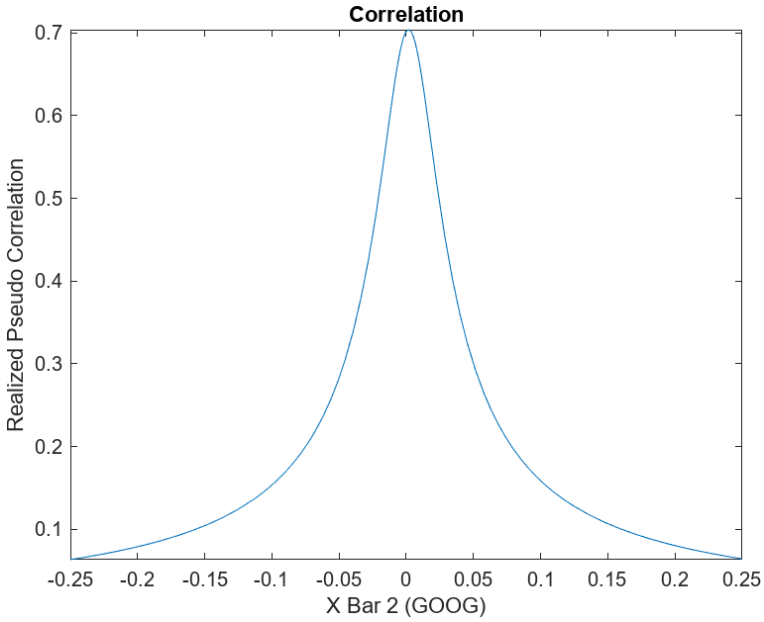


Figure 16. Realized pseudo−correlation of AAPL and GOOG from November 2022 to May 2023, as a function of $\bar{R}_{122}^{(2)}$.

Using the realized pseudo-correlation calculated above, we can calculate the payoff of a pseudo-correlation swap, whose underlying asset is the correlation of AAPL and GOOG stocks over the 6-month period from November 2022 to May 2023. We will assume the pseudo-covariance swap has a maturity of 6 months, and that we are interested in the payoff associated with taking a long position on the swap. Additionally, we will assume that the payoff is calculated using a converting parameter of \$1 per unit of pseudo-statistic.

For the strike price we will use the realized pseudo-correlation achieved over the last 15 market days before the start of the swap contract on 9 November. Therefore, we can now write the following parameter values:

- Maturity Date: $T = 0.5$
- Position Taken: $I = 1$
- Converting Parameter: $\alpha_{corr} = 1$
- Strike Price: $\rho_K = 0.7016$.

Using these parameters, Equation (37), and the previously calculated pseudo-correlation of AAPL and GOOG, we can calculate the payoff the payoff of a pseudo-correlation swap, whose underlying asset is the covariance of AAPL and GOOG, to be

Payoff of Pseudo-Correlation Swap with Underlying of Covariance of AAPL and GOOG:

$$V_{corr}(0.5) = 0.0023.$$

12. Comparing the Approach Based on the Cox–Ingresoll–Ross Model to the Realized Pseudo-Statistic Approach

The purpose of this section is to compare the payoffs of volatility and variance swaps, when the realized volatility and variance is calculated using the Cox–Ingresoll–Ross (CIR) model for variance rather than using the realized pseudo-statistic approach. The pseudo-statistic approach to calculating realized volatility and variance is a data/statistic approach that does not rely on a stochastic model for a stochastic stock price or stochastic volatility. Meanwhile, the Cox–Ingresoll–Ross (CIR) model is a stochastic volatility model that follows a stochastic differential equation (see (Swishchuk 2004)). Moreover, the Cox–Ingresoll–Ross (CIR) model is the most popularly used stochastic model for the calculation of variance, making it ideal for the comparison between a stochastic approach and a data/statistics approach to the pricing of swaps. Therefore, we will use the CIR model to create a comparison between a stochastic approach to pseudo-statistics and a data/statistical approach to pseudo-statistics in the following sections.

In the following two sections, we will calculate the realized variance and volatility of two stocks over a 6-month period, along with the payoffs of variance and volatility swaps, using the CIR model. Then, we will compare these calculated values to the payoffs of variance and volatility swaps associated with the same stocks over the same time period, which were calculated using the realized pseudo-statistic approach (see (Badescu et al. 2002)).

For the following report, we have used real-life financial data from Yahoo Finance (https://ca.finance.yahoo.com/ (accessed on 7 June 2023)) of the publicly traded stocks belonging to

- Apple Inc. (AAPL)
- Alphabet Inc. Class C (GOOG)

In the sections that follow, we will make use use of four data sets. The first two data sets contain 6 months of daily closing data for AAPL ($S^{(1)}$) and GOOG ($S^{(2)}$) from 9 November 2022 to 8 May 2023. The last two data sets contain 1 year of daily closing data for AAPL and GOOG from 8 November 2021 to 8 November 2022.

13. Realized Variance and Variance Swap Payoff

In order to calculate the realized variance using the Cox–Ingresoll–Ross (CIR) model, we will use the analytical closed form for discretely sampled variance, which is found in (Swishchuk 2004). The equation for the discretely sampled variance is as follows:

$$\text{Realized Discretely Sampled Variance: } Var_n(S) = \frac{n}{(n-1)T}\sum_{i=1}^{n} log^2 \frac{S_{t_i}}{S_{t_{i-1}}}, \qquad (38)$$

where T is the maturity date of the swap, n is one number less than the number of entries in the daily closing price data set, and S_{t_i} is the daily closing price at time i.

Using Equation (38), we can calculate the realized sample variance according to the CIR model for AAPL and GOOG in the 6-month time period from November 2022 to May 2023, while making note of the remark that follows.

Remark 1. *Using the same two 6-month data sets, one can find that the realized pseudo-volatility square of AAPL is* 0.0805 *and the realized pseudo-volatility square of GOOG is* 0.1269.

13.1. Calculating the Realized Discretely Sampled Variance Using the CIR Model

Using the 6-month daily closing price data set for AAPL described in Section 12, while using the following parameters:

– Maturity Date: $T = 0.5$
– Number of Logarithmic Return Data Points: $n = 122$.

It is possible to use Equation (38) to calculate the AAPL realized discretely sampled variance from November 2022 to May 2023 to be the following:

AAPL Realized Discretely Sampled Variance:

$$Var_{122}(S^{(1)}) = 0.0816.$$

Comparing this value to the realized variance of AAPL obtained through the realized pseudo-statistic approach (see (Badescu et al. 2002)), one can observe that the value calculated above represents a 1.37% increase over the realized variance of 0.0805 that is obtained with the realized pseudo-statistic approach.

Similarly, in order to calculate the realized variance over the 6 month period from November 2022 to May 2023 for GOOG, we will assume that

– Maturity Date: $T = 0.5$
– Number of Logarithmic Return Data Points: $n = 122$.

Then, we are able to use Equation (38) along with the 6-month daily closing price data set for GOOG described in Section 12, to calculate the GOOG realized discretely sampled variance to be

GOOG Realized Discretely Sampled Variance:

$$Var_{122}(S^{(2)}) = 0.1277.$$

Comparing this value to the realized variance of GOOG obtained through the realized pseudo-statistic approach (see (Badescu et al. 2002)), one can see that the value calculated above represents a 0.63% increase over the realized variance of 0.1269 that is obtained with the realized pseudo-statistic approach.

13.2. Variance Swap Payoffs

In order to calculate the payoff of a variance swap, we must first define the following equation, which is found in (Swishchuk 2004):

$$\text{Payoff of Variance Swap: } N(\sigma_R^2(S) - K_{var}), \tag{39}$$

where K_{var} is the strike price, $\sigma_R^2(S)$ is the realized variance calculated using Equation (38), and N is the notional amount, as previously defined in Section 3.1.

The strike price values for both GOOG and AAPL will be the expected 6-month variance calculated using a GARCH(1,1) regression on the 1-year data sets described in Section 12. This approach is explained in detail in (Javaheri et al. 2004) and (Swishchuk 2004). The strike prices have been calculated to be

– Strike Price AAPL $= E(V) = 0.0132$
– Strike Price GOOG $= E(V) = 0.0183$.

Then, using Equation (39), the strike prices above, and a notional amount of $N = 1$, along with the previously calculated realized variance for AAPL and GOOG from Section 13.1, we can calculate the payoffs of the volatility swaps using the CIR model to be

AAPL Variance Swap Payoff:

$$N(\sigma_R^2(S^{(1)}) - 0.0132) = 0.0684.$$

GOOG Variance Swap Payoff:

$$N(\sigma_R^2(S^{(2)}) - 0.0183) = 0.1094.$$

These values are different from the variance swap payoffs calculated using the pseudo-statistic approach for realized variance (see (Badescu et al. 2002)). This is true, as if the pseudo-statistic approach is used, then the payoff of a swap whose underlying asset is the variance of AAPL is calculated to be 0.0673, and if the underlying asset of the swap is the variance of GOOG, the swap payoff becomes 0.1086.

14. Realized Volatility and Volatility Swap Payoff

In order to calculate the realized volatility using the Cox–Ingresoll–Ross (CIR) model, we will use the analytical closed form for discretely sampled volatility. The equation for discretely sampled volatility is found in (Swishchuk 2004), and it can be defined as follows:

$$\text{Realized Discretely Sampled Volatility: } Vol_n(S) = \sqrt{Var_n(S)}, \tag{40}$$

where $Var_n(S)$ is the realized discretely sampled variance that was defined by Equation (38).

Using Equation (40), we can calculate the realized volatility according to the CIR model for AAPL and GOOG in the 6-month time period from November 2022 to May 2023, while making note of the remark that follows.

Remark 2. *The realized pseudo-volatility of AAPL over the same 6-month time period is* 0.2838*, while the realized pseudo-volatility of GOOG from November 2022 to May 2023 is* 0.3563*.*

14.1. Calculating the Realized Discretely Sampled Volatility Using the CIR Model

Using the 6-month daily closing price data set for AAPL described in Section 12, we previously calculated in Section 13.1 that the realized discretely sampled variance for AAPL from November 2022 to May 2023 is 0.0816. Therefore, we can use Equation (40) to calculate the realized discretely sampled volatility over the 6-month time period for AAPL to be:

AAPL Realized Discretely Sampled Volatility:

$$Vol_{122}(S^{(1)}) = 0.2856.$$

Comparing this value to the realized volatility of AAPL obtained through the realized pseudo-statistic approach (see (Badescu et al. 2002)), one can see that the value calculated above represents a 0.63% increase over the realized volatility of 0.2838 that is calculated if using the realized pseudo-statistic approach.

Similarly, in Section 13.1, we found that the realized discretely sampled variance of GOOG over the 6-month period from November 2022 to May 2023 is 0.1277. Therefore, using Equation (40), we can calculate the realized discretely sampled volatility over the 6-month time period of November 2022 to May 2023 to be:

GOOG Realized Discretely Sampled Volatility:

$$Vol_{122}(S^{(2)}) = 0.3574.$$

Comparing this value to the realized volatility of GOOG obtained through the realized pseudo-statistic approach (see (Badescu et al. 2002)), one can see that the value calculated above represents a 0.31% increase over the realized volatility of 0.3563 that is calculated if using the realized pseudo-statistic approach.

14.2. Volatility Swap Payoffs

In order to calculate the payoff of a volatility swap, we must first define the equation, found in (Swishchuk 2004), as follows:

$$\text{Payoff of Volatility Swap: } N(\sigma_R(S) - K_{vol}), \tag{41}$$

where K_{vol} is the strike price, $\sigma_R(S)$ is the realized volatility calculated using Equation (40), and N is the notional amount, as previously defined in Section 3.1.

The strike price values for both GOOG and AAPL will be the expected 6-month volatility calculated using a GARCH(1,1) regression (see (Swishchuk 2004) and (Javaheri et al. 2004)) and the Brockhaus–Long approximation (see (Brockhaus and Long 2000)) on the 1-year data sets described in Section 12. The strike prices have been calculated to be

– Strike Price AAPL $= E(\sqrt{\sigma_R^2(S^{(1)})}) = 0.1098$
– Strike Price GOOG $= E(\sqrt{\sigma_R^2(S^{(2)})}) = 0.1267$.

Using these strike prices, a notional amount of $N = 1$, Equation (41), and the previously calculated realized volatility for AAPL and GOOG from Section 14.1, we can calculate the payoffs of the volatility swaps, using the CIR model, to be:

AAPL Volatility Swap Payoff:

$$N(\sigma_R(S^{(1)}) - 0.1098) = 0.1758.$$

GOOG Volatility Swap Payoff:

$$N(\sigma_R(S^{(2)}) - 0.1267) = 0.2307.$$

These values differ from the volatility swap payoffs calculated using the pseudo-statistic approach for realized volatility (see (Badescu et al. 2002)). If the pseudo-statistic approach is used, then the payoff for a swap whose underlying asset is the volatility of AAPL is calculated to be 0.1740, and if the underlying asset of the swap is the volatility of GOOG, then the swap payoff becomes 0.2296.

15. Conclusions and Future Work

In accordance with the main motivation of this paper, which was described in Section 1, in the previous sections of this paper, we have considered the pricing of swaps based on pseudo-statistics, instead of stochastic models, and we have compared this approach with the most popular stochastic volatility model in the Cox–Ingresoll–Ross (CIR) model. Within this paper, we have demonstrated how to price various types of swaps (variance, volatility, covariance, and correlation swaps) using pseudo-statistics (pseudo-variance, pseudo-volatility, pseudo-correlation, and pseudo-covariance). We have also presented analytical closed-form formulas for both the pseudo-statistics and the pricing of swaps based on these pseudo-statistics. As a result, within this paper, we present a method for pricing swaps that does not rely on any stochastic models for a stochastic stock price or stochastic volatility, and instead relies on data/statistics. Consequently, using real-life data along with this data/statistical approach to swap pricing, we were able to calculate numerical values for the value of various swaps based on the volatility, variance, covariance, and correlation of AAPL and GOOG. Some of these results are summarized below in Table 5.

The data/statistics-based approach to swap pricing we have introduced is very different from stochastic volatility models such as the Cox–Ingresoll–Ross (CIR) model, which,

in comparison, follows a stochastic differential equation (see (Swishchuk 2004)). Therefore, in this paper, we have compared the CIR model approach to pricing swaps to the pseudo-statistic approach to pricing swaps, in order to compare a stochastic model to the data/statistics-based approach to swap pricing, which was developed within this paper. Although there are many other stochastic models that provide an approach to calculating the price of swaps, we have used the CIR model for comparison within this paper, due to the popularity of the CIR model.

After using Section 12 to define the four data sets that were to be used in the calculations of Sections 13 and 14, along with providing a short description of the approach that was to be used, we were able to use the Cox–Ingresoll–Ross (CIR) model to calculate the realized discretely sampled variance of AAPL and GOOG over a 6-month period of November 2022 to May 2023 in Section 13.1, the variance swap payoffs in Section 13.2, the realized discretely sampled volatility in Section 14.1, and the volatility swap payoffs in Section 14.2.

Consequently, we were able to compare these values to previously calculated values, which were obtained using the same time period and the same data sets, but with the realized pseudo-statistic approach (see (Badescu et al. 2002)). Although the values obtained through the realized pseudo-statistic approach were similar to the values obtained through the CIR model, differences between the values did exist, as can be seen in Table 5 below:

Table 5. Summary of results presented in sections above.

Value Obtained November 2022 to May 2023	Using CIR Model	Using Realized Pseudo-Statistic Approach
AAPL Realized Variance	0.0816	0.0805
GOOG Realized Variance	0.1277	0.1269
AAPL Variance Swap Payoff	0.0684	0.0673
GOOG Variance Swap Payoff	0.1094	0.1086
AAPL Realized Volatility	0.2856	0.2838
GOOG Realized Volatility	0.3574	0.3563
AAPL Volatility Swap Payoff	0.1758	0.1740
GOOG Volatility Swap Payoff	0.2307	0.2296

As a result, through the calculations contained within this report, we can conclude that realized volatility and variance, as well as the payoffs associated with volatility and variance swaps, can be calculated with either the CIR model or the realized pseudo-statistic approach, and similar results will be obtained. However, it is clear that even though the results will be similar depending on the approach used, the resulting values will not be identical.

In future work, we plan to further research the comparison between the data/statistical approach to swap pricing and the stochastic approach to swap pricing, beyond the Cox–Ingresoll–Ross (CIR) model. Therefore, future work could include comparing the pseudo-statistic approach to the non-Gaussian Ornstein–Uhlenbeck stochastic volatility model, using the information in (Benth et al. 2007) as a framework for calculating volatility under the Ornstein–Uhlenbeck model. Other stochastic models that should be compared to the data/statistical approach to calculating volatility include the BN-S stochastic volatility model (see (Sengupta 2016)), delayed volatility swaps (see (Swishchuk and Vadori 2014)), and Markov-modulated volatilities (see (Salvi and Swishchuk 2014)).

Author Contributions: Conceptualization, A.S.; methodology, A.S.; software, S.F.; validation, A.S.; formal analysis, S.F.; investigation, S.F.; data curation, S.F.; writing—original draft preparation, S.F.; writing—review and editing, A.S.; visualization, S.F.; supervision, A.S.; project administration, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The real-life financial data that were used within the calculations of this article can be found at (http://ca.finance.yahoo.com/ (accessed on 7 June 2023)). The 6-month data set for AAPL can be found at (https://ca.finance.yahoo.com/quote/AAPL/history?period1=1667952000&period2=1683590400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true (accessed on 7 June 2023)). The 6-month data set for GOOG can be found at (https://ca.finance.yahoo.com/quote/GOOG/history?period1=1667952000&period2=1683590400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true (accessed on 7 June 2023)). The 1-year data set for AAPL can be found at (https://ca.finance.yahoo.com/quote/AAPL/history?period1=1636329600&period2=1667865600&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true (accessed on 7 June 2023)). The 1-year data set for GOOG can be found at (https://ca.finance.yahoo.com/quote/GOOG/history?period1=1636329600&period2=1667865600&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true (accessed on 7 June 2023)).

Acknowledgments: The authors thank NSERC for continuing support.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
CIR	Cox–Ingresoll–Ross
AAPL	Apple Inc.
GOOG	Alphabet Inc. Class C

References

Badescu, Andrei, Anatoliy Swishchuk, Raymond Cheng, Stephan Lawi, Hammouda Mekki, Asrat Gashaw, Yuanyuan Hua, Marat Molyboga, Tereza Neocleous, and Yuri Petratchenko. 2002. Price Pseudo-Variance, Pseudo-Covariance, Pseudo-Volatility, and Pseudo-Correlation Swaps-In Analytical Closed-Forms. In *Sixth PIMS Industrial Problems Solving Workshop*. Vancouver: University of British Columbia, pp. 45–55.

Benth, Fred Espen, Martin Groth, and Rodwell Kufakunesu. 2007. Valuing Volatility and Variance Swaps for a Non-Gaussian Ornstein–Uhlenbeck Stochastic Volatility Model. *Applied Mathematical Finance* 14: 347–63. [CrossRef]

Brenner, Menachem, and Dan Galai. 1989. New Financial Instruments for Hedging Changes in Volatility. *Financial Analysts Journal* 45: 61–65. [CrossRef]

Brenner, Menachem, and Dan Galai. 1993. Hedging Volatility in Foreign Currencies. *The Journal of Derivatives* 1: 53–59. [CrossRef]

Brenner, Menachem, and Dan Galai. 1996. Options on Volatility. In *Option Embedded Bonds: Price Analysis, Credit Risk and Investment Strategies*. Edited by Israel Nelken. New York: Irwin Professional Pub, pp. 273–86.

Brockhaus, Oliver, and Douglas Long. 2000. Volatility Swaps Made Simple. *Risk-London Magazine Limited* 13: 92–96.

Carr, Peter, and Dilip Madan. 2001. Towards a theory of volatility trading. In *Option Pricing, Interest Rates and Risk Management, Handbooks in Mathematical Finance*. Cambridge: Cambridge University Press, vol. 22, pp. 458–76.

Carr, Peter, and Roger Lee. 2009. Volatility Derivatives. *Annual Review of Financial Economics* 1: 319–39. [CrossRef]

Demeterfi, Kresimir, Emanuel Derman, Michael Kamal, and Joseph Zou. 1999. A guide to volatility and variance swaps. *The Journal of Derivatives* 6: 9–32. [CrossRef]

Derman, Emanuel, Iraj Kani, and Michael Kamal. 1997. Trading and hedging local volatility. *Journal of Financial Engineering* 6: 233–68.

Dupire, Bruno. 1993. Model art. *Risk* 6: 118–24.

Dupire, Bruno. 1996. A unified theory of volatility. In *Derivatives Pricing: The Classic Collection*. London, United Kingdom: Risk Books London, pp. 185–96.

Fleming, Jeff, Barbara Ostdiek, and Robert E. Whaley. 1995. Predicting stock market volatility: A new measure. *The Journal of Futures Markets (1986–1998)* 15: 265.

Galai, Dan. 1979. A proposal for indexes for traded call options. *The Journal of Finance* 34: 1157–72. [CrossRef]

Gastineau, Gary L. 1977. An index of listed option premiums. *Financial Analysts Journal* 33: 70–75. [CrossRef]

Grünbichler, Andreas, and Francis A. Longstaff. 1996. Valuing futures and options on volatility. *Journal of Banking & Finance* 20: 985–1001.

Habtemicael, Semere, and Indranil Sengupta. 2016a. Pricing Covariance Swaps for Barndorff-Nielsen and Shephard Process Driven Financial Markets. *Annals of Financial Economics* 11: 1650012. [CrossRef]

Habtemicael, Semere, and Indranil Sengupta. 2016b. Pricing Variance and Volatility Swaps for Barndorff-Nielsen and Shephard Process Driven Financial Markets. *International Journal of Financial Engineering* 3: 1650027. [CrossRef]

Issaka, Aziz, and Indranil SenGupta. 2017. Analysis of Variance based Instruments for Ornstein–Uhlenbeck Type Models: Swap and Price Index. *Annals of Finance* 13: 401–34. [CrossRef]
Javaheri, Alireza, Paul Wilmott, and Espen Haug. 2004. GARCH and Volatility swaps. *Quantitative Finance* 4: 589–95. [CrossRef]
Neuberger, Anthony. 1994. The Log Contract. *The Journal of Portfolio Management* 20: 74–80. [CrossRef]
Neuberger, Anthony J. 1990. *Volatility Trading*. London: Institute of Finance and Accounting, London Business School.
Salvi, Giovanni, and Anatoliy V. Swishchuk. 2014. Covariance and Correlation Swaps for Financial Markets with Markov-Modulated Volatilities. *International Journal of Theoretical and Applied Finance* 17: 1450006. [CrossRef]
Schoutens, Wim. 2005. Moment swaps. *Quantitative Finance* 5: 525–30. [CrossRef]
Sengupta, Indranil. 2016. Generalized BN-S Stochastic Volatility Model for Option Pricing. *International Journal of Theoretical and Applied Finance* 19: 1650014. [CrossRef]
Swishchuk, Anatoliy. 2004. Modeling of Variance and Volatility Swaps for Financial Markets with Stochastic Volatilities. *Wilmott Magazine* 2: 64–72.
Swishchuk, Anatoliy. 2005. Modelling and Pricing for Stochastic Volatilities with Delay. *Wilmott Magazine* 19: 63–73.
Swishchuk, Anatoliy. 2006. Modeling and pricing of variance swaps for multi-factor stochastic volatilities with delay. *Canadian Applied Mathematics Quarterly* 14: 439–67.
Swishchuk, Anatoliy, and Nelson Vadori. 2014. Smiling for the Delayed Volatility Swaps: Smiling for the Delayed Volatility Swaps. *Wilmott* 2014: 62–73. [CrossRef]
RBC Financial Group Team. 2002. Price Pseudo-Variance, Pseudo-Covariance, Pseudo-Volatility, and Pseudo-Correlation Swaps-in Analytical Close Form. In *Proceedings of the 6th Annual PIMS Industrial Problem Solving Workshop*. Vancouver: University of British Columbia.
Whaley, Robert E. 1993. Derivatives on market volatility: Hedging tools long overdue. *The Journal of Derivatives* 1: 71–84. [CrossRef]

risks

MDPI

Article

A Semi-Static Replication Method for Bermudan Swaptions under an Affine Multi-Factor Model

Jori Hoencamp [1,*], Shashi Jain [2] and Drona Kandhai [1]

[1] Informatics Institute, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands; b.d.kandhai@uva.nl
[2] Indian Institute of Science, Department of Management Studies, Bangalore 560012, India; shashijain@iisc.ac.in
* Correspondence: j.h.hoencamp@uva.nl

Abstract: We present a semi-static replication algorithm for Bermudan swaptions under an affine, multi-factor term structure model. In contrast to dynamic replication, which needs to be continuously updated as the market moves, a semi-static replication needs to be rebalanced on just a finite number of instances. We show that the exotic derivative can be decomposed into a portfolio of vanilla discount bond options, which mirrors its value as the market moves and can be priced in closed form. This paves the way toward the efficient numerical simulation of xVA, market, and credit risk metrics for which forward valuation is the key ingredient. The static portfolio composition is obtained by regressing the target option's value using an interpretable, artificial neural network. Leveraging the universal approximation power of neural networks, we prove that the replication error can be arbitrarily small for a sufficiently large portfolio. A direct, a lower bound, and an upper bound estimator for the Bermudan swaption price are inferred from the replication algorithm. Additionally, closed-form error margins to the price statistics are determined. We practically study the accuracy and convergence of the method through several numerical experiments. The results indicate that the semi-static replication approaches the LSM benchmark with basis point accuracy and provides tight, efficient error bounds. For in-model simulations, the semi-static replication outperforms a traditional dynamic hedge.

Keywords: semi-static replication; Bermudan swaptions; affine term structure models

Citation: Hoencamp, Jori, Shashi Jain, and Drona Kandhai. 2023. A Semi-Static Replication Method for Bermudan Swaptions under an Affine Multi-Factor Model. *Risks* 11: 168. https://doi.org/10.3390/risks11100168

Academic Editors: Dan Pirjol and Lingjiong Zhu

Received: 25 August 2023
Revised: 13 September 2023
Accepted: 19 September 2023
Published: 26 September 2023

1. Introduction

The financial crisis of 2007–2008 firmly emphasized the importance of quantifying counterparty credit risk (CCR), which is the risk that the counterparty will default on the obligation and fail to fulfill its contractual agreements. Important indicators used to measure and price CCR include expected exposure (EE), potential future exposure (PFE), and various valuation adjustments (xVAs), which reflect credit, funding, and capital costs related to OTC derivative trading Gregory (2015). Most of these metrics depend on the distribution of the potential future losses resulting from a credit event. Due to the complex nature of these distributions, practitioners resort to numerical methods like Monte Carlo (MC) simulation to approximate the quantities. Typically, this involves scenario generation for the underlying risk factors and subsequent valuation of the contract for each time-step on each path Zhu and Pykhtin (2007). The latter is generally considered the most involved aspect because it needs to be carried out for full portfolios. This poses a major computational challenge to financial institutions. Efficient numerical methods for derivative valuation, both on spot and future simulation dates, are therefore highly relevant.

To address this problem, we extend the concept of (semi-)static replication, which has been extensively studied for, for example, equity derivatives, to interest rate derivatives. A traditional dynamic replication, such as a delta hedge, is achieved by constructing an asset portfolio that is rebalanced continuously through time as the market moves. A static replication on the other hand is an asset portfolio that mirrors the value of the

derivative without the need for rebalancing. The weights of the portfolio composition are so to speak static. In this work, we consider a semi-static hedge, which is a replicating portfolio that needs to be updated on only a finite number of instances. Considering a replication of vanilla products instead of the exotic derivative itself can greatly simplify its risk-assessment. Typically, ample machinery is available to analyze vanilla instruments, including closed-form prices and sensitivities.

In the equity world, the static replication problem has been addressed in the literature by, for example, Breeden and Litzenberger (1978), Carr and Bowie (1994), Carr et al. (1999), and Carr and Wu (2014). The main concept is to construct an infinite portfolio of short-dated European options with a continuum of different strike prices. A different but comparable approach is proposed in Derman et al. (1995). Here, a portfolio of European options with a continuum of different maturities is constructed to replicate the boundary and terminal conditions of exotic derivatives, such as knock-out options. The replication of an American-style option is challenging as it involves a time-dependent exercise boundary, giving rise to a free boundary problem. In Chung and Shih (2009), this is addressed by composing a portfolio of European options with multiple strikes and maturities, and, in Lokeshwar et al. (2022), a semi-static hedge is constructed using shallow neural network approximations. However, in the field of interest rate (IR) modeling, this topic has received little attention and the static replication of exotic IR derivatives remains largely an open problem. Where equity options depend on the realization of a stock, IR derivatives depend on the realization of a full term structure of interest rates, leveraging the complexity of the hedge. The articles of Pelsser (2003) and Hagan (2005) are among the few contributions to the literature, treating the static replication of guaranteed annuity options, and CMS swaps, caps, and floors, respectively, with a portfolio of European swaptions.

In this work, we study the replication problem of Bermudan swaptions under an affine term structure model, possibly multi-factor. Bermudan swaptions are a class of exotic interest rate derivatives that are heavily traded in the OTC market. We show that such a contract can be semi-statically replicated by a portfolio of short-maturity options, such as discount bond options. We propose a regress-later approach, which is introduced in Lokeshwar et al. (2022) for callable equity options. In Lokeshwar et al. (2022), the replication method combines the approximation power of artificial neural networks (ANNs) with the computational benefits of regress-later schemes. In traditional regress-now schemes, such as that of Longstaff and Schwartz (2001), sampled realizations of the continuation value are regressed against the realizations of the risk factors at the preceding monitor date. Advanced variations in this algorithm, where the polynomial regression functions are replaced by ANNs, include the work of Kohler et al. (2010), Lapeyre and Lelong (2019), and Becker et al. (2020). In contrast, in regress-later schemes, the sampled realizations of the continuation value are regressed against the realizations of the risk factors at *the same* date. The continuation value at the preceding monitor date is then obtained by evaluating the conditional expectation of this regression. An analysis and discussion of the benefits of this approach can be found in Glasserman and Yu (2004) and an example of such a scheme is presented in Jain and Oosterlee (2015).

Novel pricing algorithms that replace costly valuation functions with ANN-based approximations have been the subject of many recent papers. An early attempt to approximate option prices in the Black–Scholes model can be attributed to Hutchinson et al. (1994) and dates back to 1994. Since then, a great number of variations in this approach have been investigated. A comprehensive overview of articles devoted to this topic can be found in the literature review of Ruf and Wang (2020). An accessible introduction to neural networks and an application to derivative valuation is, for example, given in the work of Ferguson and Green (2018). A drawback of directly replacing value functions with ANNs is that the method continues to rely on external pricing methodologies to provide input to the training process. In that sense, it can accelerate, but not fully substitute, traditional valuation routines.

Other approaches in the literature consider an indirect use of ANNs and therefore do not depend on classical benchmarks for training. A noteworthy example is the development of deep backward SDE solvers, which, in a financial context, have been introduced by Henry-Labordere (2017). Where the dynamics of financial risk factors are typically captured by forward SDEs, option prices tend to be the solution to backward SDEs. An application to Bermudan swaption valuation is treated in Wang et al. (2018) and a generalization to a CCR management framework is proposed in Gnoatto et al. (2020). Another example is the development of the deep optimal stopping (DOS) algorithm by Becker et al. (2019). They propose an ANN-based method by directly learning the optimal stopping strategy of callable options, without depending on the approximation of continuation values. In the work of Andersson and Oosterlee (2021), the DOS algorithm is applied to compose exposure profiles for Bermudan contracts.

Our contribution to the existing literature is threefold. First, we propose a semi-static replication method for Bermudan swaptions under a multi-factor short-rate model. In the one-factor case, we argue that replication can be achieved with an options portfolio written on a single discount bond. In the multi-factor case, replication can be achieved with an options portfolio written on a basket of discount bonds. As such, we generalize the Black–Scholes-embedded method presented in Lokeshwar et al. (2022) to an interest rate modeling framework. Additionally we propose an alternative ANN design, such that a replication with vanilla options can also be achieved in the multi-factor case (as opposed to basket options). This facilitates highly efficient pricing, which is essential for credit risk applications, such as exposure, VaR, and xVAs, which rely on frequent re-evaluations of the portfolio.

Second, we propose a direct estimator and a lower and an upper bound estimator to the contract's value, which is implied by the semi-static replication. The lower bound results from applying a non-optimal exercise strategy on an independent set of Monte Carlo paths. The upper bound is based on the dual formulation of Haugh and Kogan (2004) and Rogers (2002), which, in contrast to other work, can be obtained without resorting to expensive nested simulations. We complement the study of Lokeshwar et al. (2022) by deriving analytical error margins to the lower and upper bound estimators. This provides a direct insight toward the approximation quality of the proposed estimators and proves their convergence as the regression errors of the ANNs diminish.

Thirdly, we prove that any desired level of accuracy can be achieved in the replication due to the universal approximating power of ANNs. We support this theoretical result with a range of representative numerical experiments. We demonstrate the pricing accuracy of the proposed algorithm by benchmarking to the established least-square method of Longstaff and Schwartz (2001). The regression error and convergence of the method is presented for different contract specifications. Lastly, we study the replication performance for different ANN designs.

The paper is organized as follows: Section 2 introduces the mathematical setting, describes the modeling framework, and provides the problem formulation. Section 3 provides a thorough introduction to the algorithm, motivates the use and interpretation of neural networks, and treats the fitting procedure. Section 4 introduces the lower bound and upper bound estimates to the true option price. In Section 5, we introduce the error bounds on the direct, lower bound, and upper bound estimates brought forth by the algorithm. We finalize the paper by illustrating the method through several numerical examples in Section 6 and providing a conclusion in Section 7.

2. Mathematical Background

In this section, we describe the general framework for our computations and give a detailed introduction to the Bermudan swaption pricing problem.

2.1. Model Formulation

We consider a continuous-time financial market, defined on finite time horizon $[0,\ \bar{T}]$. We additionally consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which represents all possible states of the economy, and let the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,\bar{T}]}$ represent all information generated by the economy up to time-t. The market is assumed to be frictionless and we ignore any transaction costs.

We let $B(t)$ denote the time$-t$ value of the *bank account*. Investments in the money market are assumed to compound a continuous, risk-free interest r_t, which we refer to as the *short rate*. $B(t)$ corresponds to the time-t value of a unit of currency invested in the money market at time-zero and we assume it is given by the following expression (see Andersen and Piterbarg 2010a or Brigo and Mercurio 2006):

$$B(t) := e^{\int_0^t r(u)du}, \qquad t \in [0, \bar{T}]$$

We denote by $\mathbb{Q}$ the risk-neutral measure equivalent to $\mathbb{P}$, which is associated to $B(t)$ as the numéraire. Attainable claims denominated by the numéraire are assumed to be martingales under $\mathbb{Q}$, which guarantees the absence of arbitrage Harrison and Pliska (1981).

We assume that the dynamics of the short-rate r are captured by an affine term structure model, in accordance with the set-up introduced in Duffie and Kan (1996) and Dai and Singleton (2000). The short rate itself is therefore considered to be an affine function of a—possibly multi-dimensional—latent factor $\mathbf{x}_t$, i.e.,

$$r(t) = \omega_1 + \boldsymbol{\omega}_2{}^\top \mathbf{x}_t \tag{1}$$

with ω_1, $\boldsymbol{\omega}_2$ denoting a scalar and a vector of time-dependent coefficients, respectively. We furthermore assume that the stochastic process $\{\mathbf{x}_t\}_{t\in[0,T]}$ is a bounded Markov process that takes values in $\mathbb{R}^d$, which represents all market influences affecting the state of the short rate. Let the dynamics of $\mathbf{x}_t$ be governed by an SDE of the form

$$d\mathbf{x}_t = \mu(t, \mathbf{x}_t)dt + \sigma(t, \mathbf{x}_t)d\mathbf{W}_t \tag{2}$$

where $\mathbf{W}_t$ denotes an $\mathbb{R}^d-$ valued Brownian motion under $\mathbb{Q}$ adapted to the filtration $\mathbb{F}$. The measurable functions $\mu : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are taken to satisfy the standard regularity conditions by which the SDE in Equation (2) admits a strong solution.

We let $P(t, T)$ denote the time$-t$ value of a zero-coupon bond contract that matures at T. A zero-coupon bond guarantees the holder one unit of currency at maturity, i.e., $P(T, T)$: = 1. Within the class of affine term structure models, zero-coupon bond prices are exponential affine in $\mathbf{x}_t$ Andersen and Piterbarg (2010b); Duffie and Kan (1996). Therefore, the value of $P(t, T)$ can be expressed as

$$P(t,T) := \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r_u du}\middle|\mathcal{F}_t\right] = \exp\Big\{A(t,T) - B(t,T)^\top x_t\Big\}$$

where the deterministic coefficients $A(t, T) \in \mathbb{R}$ and $B(T_1, T_2) \in \mathbb{R}^d$ can be found by solving a system of ODEs, which are of the form of the well-known Riccati equations; see Duffie and Kan (1996) or Filipovic (2009) for details. We consider this framework as it is still intensively used for risk management purposes. High-dimensional models, such as Libor market models, can be intractable for quantifying credit risk for large portfolios, particularly in a multi-currency setting. Multi-factor short-rate models are therefore popular amongst practitioners, providing a solid compromise between modeling flexibility and analytical tractability.

For simplicity, we will assume that the collateral rate used for discounting and the instantaneous rate used to derive term rates are both implied by the same short rate r_t.

Thus, we consider a classic single-curve model environment. As term rates, we consider simply compounded rates, which we refer to as LIBOR Brigo and Mercurio (2006)

$$L(t,T) := \frac{1 - P(t,T)}{\tau P(t,T)}$$

where τ denotes the year fraction between date t and T.

2.2. The Bermudan Swaption Pricing Problem

We consider the pricing problem of a Bermudan swaption. A Bermudan swaption is a contract that gives the holder the right to enter a swap with fixed maturity at a number of predefined monitor dates. Should the holder at any of the monitor dates decide to exercise the option, the holder immediately enters the underlying swap. The lifetime of this swap is assumed to be equal to the time between the exercise date and a fixed maturity date T_M.

As an underlying, we take a standard interest rate swap that exchanges fixed versus floating cashflows. For simplicity, we will assume that the contract is priced in a single-curve framework and that cashflow schemes of both legs coincide, yielding fixing dates $\mathcal{T}_f = \{T_0, \ldots, T_{M-1}\}$ and payment dates $\mathcal{T}_p = \{T_1, \ldots, T_M\}$. However, we stress that the algorithm is applicable to any industry standard contract specifications and is not limited to the simplifying assumptions that are made here. The time fraction between two consecutive dates is denoted as $\Delta T_m = T_m - T_{m-1}$. Let N be the notional and K the fixed rate of the swap. Assuming that the holder of the option exercises at T_m, the payments of the swap will occur at $T_{m+1}, \ldots, T_M$.

We consider the class of pricing problems, where the value of the contract is completely determined by the Markov process $\{\mathbf{x}_t\}_{t\in[0,T]}$ in $\mathbb{R}^d$ as defined in Section 2. Let $h_m : \mathbb{R}^d \to \mathbb{R}$ be the $\mathcal{F}_{T_m}$-measurable function denoting the immediate pay-off of the option if exercised at time T_m. Although the methodology holds for any generalization of the functions h_m, we will consider those in accordance with the contract specifications described above. This means that the functions h_m are assumed to be given by

$$h_m(\mathbf{x}_{T_m}) := \delta \cdot N \cdot A_{m,M}(T_m)(S_{m,M}(T_m) - K)$$

where the indicator $\delta = 1$ infers a payer and $\delta = -1$ infers a receiver swaption. The swap rate $S_{m,M}$ and the annuity $A_{m,M}$ are defined in the same fashion as Brigo and Mercurio (2006), given by the expressions

$$S_{m,M}(t) = \frac{\sum_{j=m+1}^{M} \Delta T_j P(t,T_j) F(t,T_{j-1},T_j)}{\sum_{j=m+1}^{M} \Delta T_j P(t,T_j)}, \quad A_{m,M}(t) = \sum_{j=m+1}^{M} \Delta T_j P(t,T_j)$$

where the function F denotes the simply compounded forward rate given by the expression

$$F\big(t,T_{j-1},T_j\big) = \frac{1}{\Delta T_j}\left(\frac{P\big(t,T_{j-1}\big)}{P\big(t,T_j\big)} - 1\right)$$

for any $j \in \{1, \ldots, M\}$. For details, we refer to Brigo and Mercurio (2006).

Now, let $\mathbb{T}$ denote the set of all discrete stopping times with respect to the filtration $\mathbb{F}$, taking values on the grid $\mathcal{T}_f \cup \{\infty\}$. Define the function h_τ as

$$h_\tau(\mathbf{x}_\tau) := h_{\tau(\omega)}(\mathbf{x}_\tau(\omega)) = \begin{cases} h_m(\mathbf{x}_{T_m}) & \text{if } \tau(\omega) = T_m \\ 0 & \text{if } \tau(\omega) = \infty \end{cases}, \qquad \omega \in \Omega \tag{3}$$

In this notation, $\tau(\omega) = \infty$ indicates that the option is not exercised at all. We aim to approximate the time-zero value of the Bermudan swaption, which satisfies the following equation:

$$V(0) = \sup_{\tau \in \mathbb{T}} \mathbb{E}^{\mathbb{Q}} \left[\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)} \middle| \mathcal{F}_0 \right] \tag{4}$$

Finding the optimal exercise strategy τ is typically a non-trivial exercise. Numerical approximations for $V(0)$ can, however, be computed by considering a dynamical programming formulation as given below, which is shown to be equivalent to (4) in, for example, Glasserman (2013). Let $t \in (T_m, T_{m+1}]$ for some $m \in \{0, \ldots, M-2\}$ and denote by $V(t)$ the value of the option, conditioned on the fact that it is not yet exercised prior to t. This value satisfies the equation (see Glasserman 2013)

$$V(t) = \begin{cases} \max\{h_{M-1}(\mathbf{x}_{T_{M-1}}),\ 0\} & \text{if } t = T_{M-1} \\ \max\left\{h_m(\mathbf{x}_t),\ B(t)\mathbb{E}^{\mathbb{Q}}\left[\frac{V(T_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_t\right]\right\} & \text{if } t = T_m,\ m \in \{0, \ldots, M-2\} \\ B(t)\mathbb{E}^{\mathbb{Q}}\left[\frac{V(T_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_t\right] & \text{if } t \in (T_m, T_{m+1}),\ m \in \{0, \ldots, M-2\} \end{cases} \tag{5}$$

We refer to the random variables $C_m(t) := B(t)\mathbb{E}^{\mathbb{Q}}\left[\frac{V(T_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_t\right]$ as the hold or continuation values. They represent the expected value of the contract if it is not being exercised up until t but continues to follow the optimal policy thereafter. Approximations of the dynamic formulation are typically obtained by a backward iteration based on simulations of the underlying risk factors. The objective is then to determine the continuation values as a function of the state of the risk factor $\mathbf{x}_t$. Popular numerical schemes based on regression have been introduced in, for example, Carriere et al. (1996) and Longstaff and Schwartz (2001).

Based on approximations of the continuation values, the optimal policy τ can be computed as follows. Assume that, for a given scenario $\omega \in \Omega$, the risk factor takes the values $\mathbf{x}_{T_0} = x_0, \ldots, \mathbf{x}_{T_{M-1}} = x_{M-1}$. Then, the holder should continue to hold the option if $C_m(T_m) > h_m(x_m)$ and exercise as soon as $C_m(T_m) \leq h_m(x_m)$. In other words, the exercise strategy can be determined as

$$\tau(\omega) = \min\left\{T_m \in \mathcal{T}_f \middle| C_m(T_m) \leq h_m(x_m)\right\}$$

Should, for some scenario, the continuation value be bigger than the immediate pay-off for each monitor date, then $\tau(\omega) = \infty$ and the option expires as worthless.

3. A Semi-Static Replication for Bermudan Swaptions

The main concept of our method is to construct static hedge portfolios that replicate the dynamical formulation in Equation (5) between two consecutive monitor dates. In this section, we introduce the algorithm for a Bermudan swaption that is priced under a multi-factor affine term structure model. The methodology is inspired by the algorithm presented in Lokeshwar et al. (2022) and utilizes a regress-later technique in which the intermediate option values are regressed against simple IR assets, such as discount bonds. The regression model is chosen deliberately to represent the pay-off of an options portfolio written on these assets. An important consequence is that the hedge can be valued in closed form. Throughout this work, we will use the terms semi-static hedge and semi-static replication interchangeably. A hedge in general refers to a trading strategy that reduces the exposure to market risk of an outstanding position. A replication refers to an asset portfolio that mirrors the value of a derivative, which is a common means to set up a hedge. As we see the efficient valuation properties in the context of credit risk quantification as the main application, rather than actual hedging, we will put emphasis on the term replication.

3.1. The Algorithm

The regress-later algorithm is executed in an iterative manner, backward in time. The outcome is a set of option portfolios $\{\Pi_{M-1}, \ldots, \Pi_0\}$ written on pre-selected IR assets. To be more precise, the algorithm determines the weights and strikes of each portfolio Π_m, such that it closely mirrors the Bermudan swaption after its composition at T_{m-1} until its expiry at T_m. The pay-off of Π_m exactly meets the cost of composing the next portfolio Π_{m+1} or the Bermudan's pay-off in case it is exercised. The methodology yields a semi-static hedging strategy as the portfolio compositions are constant between two consecutive monitor dates. Hence, there is no need for continuous rebalancing, as is the case for a dynamic hedging strategy. The algorithm can roughly be divided into three steps, presented below. Algorithm 1 summarizes the method.

Algorithm 1 The algorithm for a Bermudan swaption

Generate N risk factor scenarios for $\mathbf{x}_{T_m}$ for $m = 0, \ldots, M$
Compute N corresponding asset scenarios for z_m for $m = 0, \ldots, M$
$\tilde{V}\left(T_{M-1}; x^n_{T_{M-1}}\right) \leftarrow \max\left\{h_{M-1}\left(x^n_{T_{M-1}}\right), 0\right\}$ for $n = 1, \ldots, N$
Initialize G_{M-1} parameters ξ_{M-1} from independent uniform distributions
for $m = M-1, \ldots, 1$ **do**
  $\xi_m \leftarrow \underset{\xi \in \mathbb{R}^p}{\operatorname{argmin}}\ L(\xi | \hat{z}_m, \hat{x}_m)$ minimizing the MSE
  for $n = 1, \ldots, N$ **do**
    $\tilde{C}_{m-1}(T_{m-1}) \leftarrow B(T_{m-1})\mathbb{E}^{\mathbb{Q}}\left[\frac{G_m(z_m(T_m))}{B(T_m)}\middle|\mathcal{F}_{T_{m-1}}\right]$
    $\tilde{V}(T_{m-1}; x^n_{T_{m-1}}) \leftarrow \max\left\{\tilde{C}_{m-1}(T_{m-1}),\ h_{m-1}\left(x^n_{T_{m-1}}\right)\right\}$
  end for
  $\xi_{m-1} \leftarrow \xi_m$ initialize weights of G_{m-1}
end for
$\xi_0 \leftarrow \underset{\xi \in \mathbb{R}^p}{\operatorname{argmin}}\ L(\xi | \hat{z}_0, \hat{x}_0)$ minimizing the MSE
return $\mathbb{E}^{\mathbb{Q}}\left[\frac{G_0(z_0(T_0))}{B(T_0)}\middle|\mathcal{F}_0\right]$

3.1.1. Sample the Independent Variables

We start by sampling N realizations of the risk factor $\mathbf{x}_t$ on the time grid $\mathcal{T} = \{T_0, \ldots, T_{M-1}\}$. These realizations will serve as an input for the regression data. We will denote the data points as $\hat{x} := \left\{\left(x^n_{T_0}, \ldots, x^n_{T_{M-1}}\right)\right\}_{n=1}^{N}$. Different sample methodologies could be used, such as:

- Take a standard quadrature grid for each monitor date T_m, associated with the transition density of the risk factor. For example, if $\mathbf{x}_t$ has Gaussian dynamics, one could consider the Gauss–Hermite quadrature scaled and shifted in accordance with the mean and variance of $\mathbf{x}_t$. See, for example, Xiu (2010).
- Discretize the SDE of the risk factor and sample by the means of an Euler or Milstein scheme. Make sure that a sufficiently coarse time-stepping grid is used, which includes the M monitor dates. See, for example, Kloeden and Platen (2013) for details.

Secondly, we select an asset that will serve as the independent variable for the regression. We will denote this asset as $z_m(t)$. The choice for z_m can be arbitrary, as long as it meets the following conditions:

- The asset $z_m(T_m)$ should be a square integrable random variable that is $\mathcal{F}_{T_m}$ measurable, taking values in $\mathbb{R}^d$.
- The risk-neutral price of $z_m(t)$ should only be dependent on the current state of the risk factor and be almost surely unique; that is, the mapping $\mathbf{x}_{T_m} \mapsto z_m | \mathbf{x}_{T_m}$ should be

continuous and injective. This is required to guarantee a well-defined parametrization of the option value.

Examples for z_m would be a zero-coupon bond, a forward Libor rate, or a forward swap rate. For each sampled realization of the risk factor, the corresponding realization of the asset value will be computed and denoted as $\hat{z} := \left\{ \left(z_0^n, \ldots, z_{M-1}^n \right) \right\}_{n=1}^N$. This will serve as the regression data in the following step.

3.1.2. Regress the Option Value against an IR Asset

In this phase, we compose replication portfolios $\Pi_0, \ldots, \Pi_{M-1}$ by fitting M regression functions $G_0, \ldots, G_{M-1}$. We consider functions of the form $G_m : \mathbb{R}^d \to \mathbb{R}$, which assign values in $\mathbb{R}$ to each realization of the selected asset z_m. Fitting is performed recursively, starting at T_{M-1}, moving backwards in time, until the first exercise opportunity T_0. Approximations of the Bermudan swaption value at each monitor date serve as the dependent variable. At the final monitor date, the value of the contract (given it has not been exercised) is known to be

$$V\left(T_{M-1}; x_{T_{M-1}}^n\right) = \max\left\{ h_{M-1}\left(x_{T_{M-1}}^n\right),\ 0 \right\}, \qquad n = 1, \ldots, N$$

Now, assume that, for some monitor date T_m, we have an approximation of the contract value $\tilde{V}\left(T_m; x_{T_m}^n\right) \approx V\left(T_m; x_{T_m}^n\right)$. Let $\xi_m \in \mathbb{R}^p$ for some $p \in \mathbb{N}$ denote the vector of the unknown regression parameters. The objective is to determine ξ_m such that

$$G_m(z_m(T_m)) \approx V(T_m)$$

with the smallest possible error. This is carried out by formulating and solving a related optimization problem. In this case, we choose to minimize the expected square error, given by

$$\mathbb{E}^{\mathbb{Q}}\left[\left| G_m(z_m(T_m)) - V(T_m) \right|^2 \right] \tag{6}$$

There is no exact analytical expression available for the expectation of Equation (6). However, it can be approximated using the sampled regression data, giving rise to an empirical loss function L given by

$$L(\xi_m | \hat{z}_m, \hat{x}_m) = \frac{1}{N} \sum_{n=1}^{N} \left(G_m(z_m^n) - \tilde{V}\left(T_m; x_{T_m}^n\right) \right)^2 \tag{7}$$

The parameters ξ_m are then the result of the fitting procedure, such that

$$\xi_m \approx \underset{\xi \in \mathbb{R}^p}{\operatorname{argmin}}\ L(\xi | \hat{z}_m, \hat{x}_m)$$

If the regression model is chosen accordingly, $G_m(z_m)$ represents the pay-off at T_m of a derivative portfolio Π_m written on the selected asset z_m. Details on suggested functional forms of G_m, asset selection for z_m, and fitting procedures are subject of Section 3.2.

3.1.3. Compute the Continuation Value

Once the regression is completed, the last step is to compute the continuation value and subsequently the option value at the monitor date preceding T_m. For each scenario $n = 1, \ldots, N$, we approximate the continuation value as

$$\begin{aligned} \tilde{C}_{m-1}(T_{m-1}) &= B(T_{m-1}) \mathbb{E}^{\mathbb{Q}}\left[\frac{\tilde{V}(T_m)}{B(T_m)} \middle| \mathcal{F}_{T_{m-1}} \right] \\ &\approx B(T_{m-1}) \mathbb{E}^{\mathbb{Q}}\left[\frac{G_m(z_m(T_m))}{B(T_m)} \middle| \mathcal{F}_{T_{m-1}} \right] \end{aligned} \tag{8}$$

As G_m is chosen to represent the pay-off of a derivative portfolio Π_m written on z_m, we argue that computing C_{m-1} is in fact equivalent to the risk-neutral pricing of Π_m. In other words, we have

$$\tilde{C}_{m-1}(T_{m-1}) = B(T_{m-1})\mathbb{E}^{\mathbb{Q}}\left[\frac{\Pi_m(T_m)}{B(T_m)}\middle|\mathcal{F}_{T_{m-1}}\right] := \Pi_m(T_{m-1})$$

In Section 3.2, we treat examples for which Π_m can be computed in closed form.

Finally, the option value at the preceding monitor date T_{m-1} is given by

$$\tilde{V}\left(T_m; x^n_{T_m}\right) = \max\left\{\tilde{C}_{m-1}(T_{m-1}),\ h_{m-1}\left(x^n_{T_{m-1}}\right)\right\}, \qquad n = 1, \ldots, N$$

The steps are repeated recursively until we have a representation G_0 of the option value at the first monitor date. An estimator of the time-zero option value is given by

$$\tilde{V}(0) = \mathbb{E}^{\mathbb{Q}}\left[\frac{G_0(z_0(T_0))}{B(T_0)}\middle|\mathcal{F}_0\right]$$

We refer to this approximation as the *direct estimator*.

3.2. A Neural Network Approach to G_m

In this section, we propose to represent the regression functions G_m as shallow, artificial neural networks. The choices that are presented here are adapted to a framework of Gaussian risk factors, such as that presented in Section 2. The method, however, lends itself to be generalized to a broader class of models by considering an appropriate adjustment to the input or structure.

3.2.1. The 1-Factor Case

First, we discuss the case $d = 1$. Let $m \in \{0, \ldots, M-1\}$. As a regression function, we consider a fully connected, feed-forward neural network with one hidden layer, denoted as $G_m : \mathbb{R} \to \mathbb{R}$. The design with only a single hidden layer is graphically represented in Figure 1 and is chosen deliberately to facilitate the network's interpretation. As an input to the network (the asset z_m), we select a zero-coupon bond, which pays one unit of currency at T_M.

- The first layer consists of a single node and corresponds to the discount bond price, which serves as input. It is represented by the left node in Figure 1. The hidden layer has $q \in \mathbb{N}$ hidden nodes, represented by the center layer in Figure 1. The affine transformation acting between the first two layers is denoted $A_1 : \mathbb{R} \to \mathbb{R}^q$ and is of the form

  $$A_1 : x \mapsto \mathbf{w}_1 x + \mathbf{b}, \qquad \mathbf{w}_1 \in \mathbb{R}^{q\times 1}, \mathbf{b} \in \mathbb{R}^q$$

  As an activation function $\varphi : \mathbb{R}^q \to \mathbb{R}^q$ acting on the hidden layer, we take the ReLU-function, given by

  $$\varphi : (x_1, \ldots, x_q) \mapsto \big(\max\{x_1, 0\}, \ldots, \max\{x_q, 0\}\big)$$

  Note that the ReLU function corresponds to the pay-off function of a European option.
- The output of the network estimates contract value $\tilde{V}_m \in \mathbb{R}$ and therefore takes value in $\mathbb{R}$. It is represented by the right node in Figure 1. We consider a linear transformation acting between the second and last layer $A_2 : \mathbb{R}^q \to \mathbb{R}$, given by

  $$A_2 : x \mapsto \mathbf{w}_2 x, \qquad \mathbf{w}_2 \in \mathbb{R}^{1\times q}$$

  On top of that, we apply the linear activation, which comes down to an identity function, mapping x to itself.

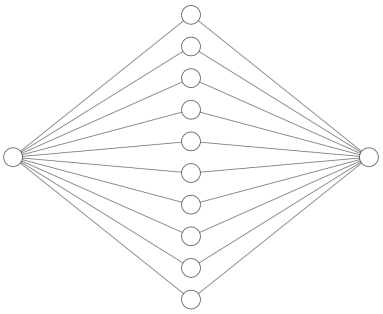


Figure 1. Suggested neural network design for $Dim(\mathbf{x}_t) = 1$.

Combined together, the network is specified to satisfy

$$G_m(\cdot) := A_2 \circ \varphi \circ A_1$$

and the trainable parameters can be presented by the list

$$\xi_m = \{w_{1,1}, b_{1,1}, \ldots, w_{1,q}, b_{1,q}\} \cup \{w_{2,1}, \ldots, w_{2,q}\}$$

3.2.2. Interpretation of the Neural Network

Now that we have specified the structure of the neural network, we will discuss how each function G_m can be interpreted as a portfolio Π_m. In the one-dimensional case, G_m can be expressed as follows:

$$G_m(z_m) := \sum_{j=1}^{q} w_{2,j} \max\{w_{1,j} z_m + b_j,\ 0\}$$

We can regard this as the pay-off of a derivative portfolio Π_m written on the asset z_m. The portfolio contains q derivatives that each have a terminal value equal to $w_{2,j} \max\{w_{1,j} z_m + b_j,\ 0\}$. In total, we can recognize four types of products, which depend on the signs of $w_{1,j}$ and b_j.

1. If $w_{1,j} > 0$ and $b_j > 0$, we have

   $$w_{2,j} \max\{w_{1,j} z_m + b_j,\ 0\} = w_{2,j} w_{1,j} z_m + w_{2,j} b_j$$

   which is the pay-off of a forward contract on $w_{2,j} w_{1,j}$ units in z_m and $w_{2,j} b_j$ units of currency.
2. If $w_{1,j} > 0$ and $b_j < 0$, we have

   $$w_{2,j} \max\{w_{1,j} z_m + b_j,\ 0\} = w_{2,j} w_{1,j} \max\left\{z_m - \frac{-b_j}{w_{1,j}},\ 0\right\}$$

   which is the pay-off corresponding to $w_{2,j} w_{1,j}$ units of a European call option written on z_m, with strike price $\frac{-b_j}{w_{1,j}}$.
3. If $w_{1,j} < 0$ and $b_j > 0$, we have

   $$w_{2,j} \max\{w_{1,j} z_m + b_j,\ 0\} = -w_{2,j} w_{1,j} \max\left\{\frac{b_j}{-w_{1,j}} - z_m,\ 0\right\}$$

   which is the pay-off corresponding to $-w_{2,j} w_{1,j}$ units of a European put option written on z_m, with strike price $\frac{b_j}{-w_{1,j}}$.

4. If $w_{1,j} < 0$ and $b_j < 0$, we have

$$w_{2,j} \max\{w_{1,j} z_m + b_j,\ 0\} = 0$$

which clearly represents a worthless contract.

The sign of the coefficient $w_{2,j}$ indicates if one has a short or long position of the product in the portfolio. Hence, under the assumption of a frictionless economy, the absence of arbitrage, and the Markov property for z_m, the portfolio Π_m replicates the original Bermudan contract over the period $(T_{m-1}, T_m]$. As the portfolio composition is constant between two consecutive monitor dates, the method described here can be interpreted as a semi-static hedging strategy.

3.2.3. The Multi-Factor Case

In the case $d \geq 2$, we propose that a basket of d zero-coupon bonds all maturing at different dates $T_m + \delta_1, \ldots, T_m + \delta_n$ is required as input to the regression. If the risk factor space is d-dimensional, it can only be parametrized by an at least d-dimensional asset vector.

To see why the above statement is true, simply consider n bonds $P(T_m, T_m + \delta_1), \ldots, P(T_m, T_m + \delta_n)$ and note that the following relation holds:

$$\begin{pmatrix} P(T_m, T_m + \delta_1) \\ \vdots \\ P(T_m, T_m + \delta_n) \end{pmatrix} = \begin{pmatrix} \exp\{A(T_m, T_m + \delta_1) - \sum_{j=1}^{d} B_j(T_m, T_m + \delta_1) x_j(T_m)\} \\ \vdots \\ \exp\{A(T_m, T_m + \delta_n) - \sum_{j=1}^{d} B_j(T_m, T_m + \delta_n) x_j(T_m)\} \end{pmatrix}$$

$$\implies \begin{pmatrix} B_1(T_m, T_m + \delta_1) & \ldots & B_d(T_m, T_m + \delta_1) \\ \vdots & \ddots & \vdots \\ B_1(T_m, T_m + \delta_n) & \ldots & B_d(T_m, T_m + \delta_n) \end{pmatrix} \begin{pmatrix} x_1(T_m) \\ \vdots \\ x_d(T_m) \end{pmatrix}$$

$$= \begin{pmatrix} A(T_m, T_m + \delta_1) - \log P(T_m, T_m + \delta_1) \\ \vdots \\ A(T_m, T_m + \delta_d) - \log P(T_m, T_m + \delta_n) \end{pmatrix}$$

$$\implies \mathbf{B}(T_m)\mathbf{x}_{T_m} = \boldsymbol{\alpha}$$

Since we have that $rank(\mathbf{B}(T_m)) = \min\{n, d\}$, it follows that if $n < d$, the image of $\mathbf{B}$ does not span the whole risk factor space, whereas if $n > d$, the image of $\mathbf{B}$ is still equal to the case $n = d$.

Concluding on the argument above, it would be an obvious choice to take a $d-$dimensional vector of bonds as the input and generalize the architecture of G_m by increasing the input dimension (i.e., the number of nodes in the first layer) from 1 to d. However, in that case, Π_m represents a derivatives portfolio written on a basket of bonds, by which the tractability of pricing Π_m would be lost. Therefore, we suggest two alternatives to the design of G_m, intended to preserve the analytical valuation potential of Π_m.

The basic specifications of the neural network will remain similar to the one-factor case. We consider a feed-forward neural network with one hidden layer of the form $G_m : \mathbb{R}^d \to \mathbb{R}$.

- The first layer consists of d nodes and the hidden layer has $q \in \mathbb{N}$ hidden nodes. The affine transformation and activation acting between the first two layers are denoted $A_1 : \mathbb{R}^d \to \mathbb{R}^q$ and $\varphi : \mathbb{R}^q \to \mathbb{R}^q$, respectively, given by

$$A_1 : x \mapsto \mathbf{w}_1 x + \mathbf{b}, \qquad \mathbf{w}_1 \in \mathbb{R}^{q \times d}, \mathbf{b} \in \mathbb{R}^q$$
$$\varphi : (x_1, \ldots, x_q) \mapsto (\max\{x_1, 0\}, \ldots, \max\{x_q, 0\})$$

- The output contains a single node. A linear transformation acts between the second and last layer $A_2 : \mathbb{R}^q \to \mathbb{R}$, together with the linear activation, given by

$$A_2 : x \mapsto \mathbf{w}_2 x, \qquad \mathbf{w}_2 \in \mathbb{R}^{1\times q}$$

- The network is given by $G_m(\cdot) := A_2 \circ \varphi \circ A_1$.

3.2.4. Suggestion 1: A Locally Connected Neural Network

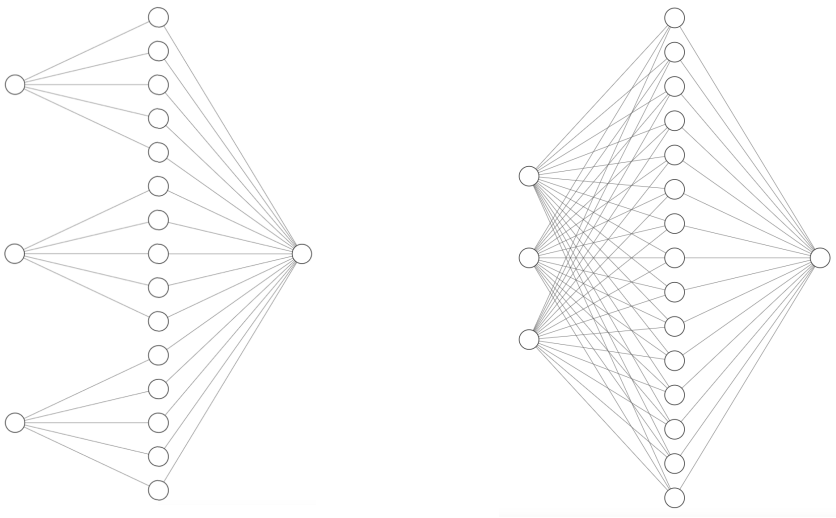
The outcome of each node in the hidden layer represents the terminal value of a derivative written on the asset $\mathbf{z}_m$, which, together, compose the portfolio Π_m. In the $d-$dimensional case, the outcome of the j^{th} node ν_j can be expressed as

$$\nu_j(\mathbf{z}) = \max\left\{ \sum_{k=1}^{d} w_{jk} z_k + b_j,\ 0 \right\}$$

which corresponds to the pay-off of an arithmetic basket option with weights $w_{j1}, \dots w_{jd}$ and strike price b_j. Such an exotic option is difficult to price. To overcome this issue, we constrain the matrix $\mathbf{w}_1$ to only admit a single non-zero value in each row. The architecture of this suggestion is graphically depicted in Figure 2a. Let the number of hidden nodes be a multiple of the input dimension, i.e., $q = n \cdot d$ for some $n \in \mathbb{N}$. The matrix $\mathbf{w}_1$ is set to be of the form

$$\mathbf{w}_1 = \begin{pmatrix} w_{1,1} & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ w_{1,n} & 0 & 0 & \cdots & 0 & 0 \\ 0 & w_{2,n+1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & w_{2,2n} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & w_{d,d\cdot n} \end{pmatrix}$$

As a result, none of the hidden nodes are connected to more than one input node (see Figure 2a). Therefore, the outcome of each node ν_j again represents a European option or forward written on a single bond, which can be priced in closed form (see Appendix A.1).



(**a**) (**b**)

Figure 2. Suggested neural network designs for $Dim(\mathbf{x}_t) \geq 2$. (**a**) Locally connected neural network. (**b**) Fully connected neural network.

We can recognize two drawbacks to this approach. First, the number of trainable parameters for a fixed number of hidden nodes is much lower compared to the fully connected case. This can simply be overcome by increasing q. Second, as the network is not fully connected, the universal approximation theorem no longer applies to G_m. Therefore, we have no guarantee that the approximation errors can be reduced to any desirable level. Our numerical experiments however indicate that the approximation accuracy of this design is not inferior to that of a fully connected counterpart of the same dimensions; see Section 6.

3.2.5. Suggestion 2: A Fully Connected Neural Network

Our second approach does not entail altering the structure or weights of the network, but suggests to take a different input. We hence consider a fully connected feed-forward neural network with one hidden layer of the form $G_m : \mathbb{R}^d \to \mathbb{R}$. The architecture is graphically depicted in Figure 2. As a consequence, each hidden node is connected to each input node. However, as an input, we use the log of n bonds, i.e.,

$$\mathbf{z}_m := (\log P(T_m, T_m + \delta_1), \dots, \log P(T_m, T_m + \delta_n))^\top$$

Therefore, each node ν_j can be compared to the pay-off of a geometric basket option written on n assets $\mathbf{z}_m$ equal to the log of $P(t, T_m + \delta_j)$. Under the assumption that the dynamics of the risk factor $\mathbf{x}_t$ are Gaussian, these options can be priced explicitly as we will show in Appendix A.2.

An advantage of this approach is that it employs a fully connected network that, by virtue of the universal approximation theorem Hornik et al. (1989), can yield any desired level of accuracy. A drawback is that the financial interpretation of the network as a replicating portfolio is not as strong as in suggestion 1 due to the required log in the payoff.

3.3. Training of the Neural Networks

In this section, we specify some of the main considerations related to the fitting procedure of the algorithm. The method requires the training of M shallow feed-forward networks as specified in Section 3.2, which we denote $G_0, \dots, G_{M-1}$. Our numerical experiments indicated that the normalization of the training set strongly improved the networks' fitting accuracy. Details for pre-processing the regression data are treated in Appendix B.

Optimization

The training of each network is performed in an iterative process, starting with G_{M-1} working backwards until G_0. The effectiveness of the process depends on several standard choices related to neural network optimization, of which some are listed below.

- As an optimizer, we apply AdaMax Kingma and Ba (2014), a variation of the commonly used Adam algorithm. This is a stochastic, first-order, gradient-based optimizer that updates weights inversely proportional to the L_∞-norm of their current and past gradient, whereas Adam is based on the L_2-norm. Our experiments indicate that AdaMax slightly outperforms comparable algorithms in the scope of our objectives.
- The batch size, i.e., the number of training points used per weight update, is set to a standard 32. The learning rate, which scales the step size of each update, is kept in the range 0.0001–0.0005.
- For the initial network, G_{M-1}, we use random initialization of the parameters. If the considered contract is a payer Bermudan swaption, we initialize the (non-zero) entries of $\mathbf{w}_1$ i.i.d. unif$(0, 1)$ and the biases $\mathbf{b}$ i.i.d. unif$(-1, 0)$. In the case of a receiver contract, it is the other way around. The weights $\mathbf{w}_2$ are initialized i.i.d. unif$(-1, 1)$.
- For the subsequent networks, $G_{M-2}, \dots, G_0$, each network G_m is initialized with the final set of weights of the previous network G_{m+1}.
- As a training set for the optimizer, we use a collection of 20,000 data-points.

Some specific choices for the hyperparameters are motivated by a convergence analysis presented in Appendix C.

4. Lower and Upper Bound Estimates

The algorithm described in Section 3.1 gives rise to a direct estimator of the true option price V. The accuracy of this estimator depends on the approximation performance of the neural networks at each monitor date. Should each regression yield a perfect fit, then the estimation error would automatically be zero. In practice, however, the loss function, defined in Equation (7), never fully converges to zero. As the networks are trained to closed-form exercise and continuation values, error measures such as MSE and MAE can be easily obtained. In particular, the mean absolute errors provide a strong indication of the error bounds on the direct estimator (see Section 5).

Although convergence errors put solid bounds on the accuracy of the estimator, they are typically quite loose. Therefore, they give rise to non-tight confidence bounds. To overcome this issue, we introduce a numerical approximation to a tight lower and upper bound to the true price, in the same spirit as Lokeshwar et al. (2022). These should provide a better indication of the quality of the estimate.

4.1. The Lower Bound

We compute a lower bound approximation by considering the non-optimal exercise strategy $\tilde{\tau}$ implied by the continuation values estimates introduced in Section 3.1. We define $\tilde{\tau}$ as

$$\tilde{\tau}(\omega) = \min\left\{ T_m \in \mathcal{T}_f \middle| \tilde{C}_m(T_m) \leq h_m(\mathbf{x}_{T_m}) \right\} \tag{9}$$

where $\tilde{C}_m$ refers to the approximated continuation value given in Equation (8). A strict lower bound is now given by

$$L(0) = \mathbb{E}^{\mathbb{Q}}\left[\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})} \middle| \mathcal{F}_0 \right] = P(0,T_M)\mathbb{E}^{T_M}\left[\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{P(\tilde{\tau},T_M)} \middle| \mathcal{F}_0 \right] \tag{10}$$

where $h_{\tilde{\tau}}$ corresponds to the definition given in Equation (3). The term on the right is obtained by changing the measure from $\mathbb{Q}$ to the T_M-forward measure $\mathbb{Q}^{T_M}$Geman et al. (1995). Under the T_M-forward measure, the lower bound can be estimated by simulating a fresh set of scenarios of the risk factor $\hat{x} := \left\{ \left(x^n_{t_1}, x^n_{t_2}, \ldots, x^n_{T_M} \right) \middle| n = 1, \ldots, N \right\}$. Denote by $P^n(t,T_M)$ the zero-coupon bond realization corresponding to x^n_t. Then, the lower bound cab be approximated as

$$\tilde{L}(0) = \frac{P(0,T_M)}{N} \sum_{n=1}^{N} \frac{h_{\tilde{\tau}^n}(x^n_{\tilde{\tau}^n})}{P^n(\tilde{\tau}^n,T_M)}$$

4.2. The Upper Bound

We compute an upper bound by considering a dual formulation of the price expression Equation (4) as proposed in Haugh and Kogan (2004) and Rogers (2002). Let $\mathcal{M}$ denote the set of all martingales M_t adapted to $\mathbb{F}$ such that $\sup_{t\in[0,T]}|M_t| < \infty$. An upper bound $U(0)$ to the true price $V(0)$ is obtained by observing that the following inequality holds (see Haugh and Kogan 2004):

$$V(0) \leq M_0 + \mathbb{E}^{\mathbb{Q}}\left[\max_{T_m \in \mathcal{T}_f} \left\{ \frac{h_m(\mathbf{x}_{T_m})}{B(T_m)} - M_{T_m} \right\} \middle| \mathcal{F}_0 \right] := U(0) \tag{11}$$

for any $M_t \in \mathcal{M}$. To find a suitable martingale that yields a tight bound, we consider the Doob–Meyer decomposition of the true discounted option price process $\frac{V(t)}{B(t)}$. As the price process is a supermartingale, we can write

$$\frac{V(t)}{B(t)} := Y_t + Z_t$$

where Y_t denotes a martingale and Z_t is a predictable, strictly decreasing process such that $Z_0 = 0$. Note that Equation (11) attains an equality if we set $M_t = Y_t$, i.e., the martingale part of the option price process. The bound will hence be tight if we consider a martingale M_t that is close to the unknown Y_t. Let $G_m(\cdot)$ denote the neural networks induced by the algorithm. In the spirit of Andersen and Broadie (2004) and Lokeshwar et al. (2022), we construct a martingale on the discrete time grid $\{0, T_0, \ldots, T_{M-1}\}$ as follows:

$$\begin{aligned} M_0 &= \mathbb{E}^{\mathbb{Q}}\left[\frac{G_0(z_0(T_0))}{B(T_0)}\middle|\mathcal{F}_0\right],\ M_{T_0} = \frac{G_0(z_0(T_0))}{B(T_0)} \\ M_{T_m} &= M_{T_{m-1}} + \frac{G_m(z_m(T_m))}{B(T_m)} - \mathbb{E}^{\mathbb{Q}}\left[\frac{G_m(z_m(T_m))}{B(T_m)}\middle|\mathcal{F}_{T_{m-1}}\right], \quad m = 1, \ldots, M-1 \end{aligned} \tag{12}$$

Clearly, the process $\{M_{T_m}\}_{m=0}^{M-1}$ yields a discrete martingale as

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}}\left[M_{T_m}\middle|\mathcal{F}_{T_{m-1}}\right] &= \mathbb{E}^{\mathbb{Q}}\left[M_{T_{m-1}} + \frac{G_m(z_m(T_m))}{B(T_m)} - \mathbb{E}^{\mathbb{Q}}\left[\frac{G_m(z_m(T_m))}{B(T_m)}\middle|\mathcal{F}_{T_{m-1}}\right]\middle|\mathcal{F}_{T_{m-1}}\right] \\ &= \mathbb{E}^{\mathbb{Q}}\left[M_{T_{m-1}}\middle|\mathcal{F}_{T_{m-1}}\right] + \mathbb{E}^{\mathbb{Q}}\left[\frac{G_m(z_m(T_m))}{B(T_m)} - \frac{G_m(z_m(T_m))}{B(T_m)}\middle|\mathcal{F}_{T_{m-1}}\right] \\ &= M_{T_{m-1}} \end{aligned}$$

Furthermore, the process M_t as defined above will coincide with Y_t if the approximation errors in $G_m(\cdot)$ equal zero, hence yielding an equality in Equation (11). Note that the recursive relation in Equation (12) can be rewritten as

$$M_{T_m} = \frac{G_0(z_0(T_0))}{B(T_0)} + \sum_{j=1}^{m}\left(\frac{G_j(z_j(T_j))}{B(T_j)} - \mathbb{E}^{\mathbb{Q}}\left[\frac{G_j(z_j(T_j))}{B(T_j)}\middle|\mathcal{F}_{T_{j-1}}\right]\right) \tag{13}$$

We can now estimate the upper bound by again simulating a set of scenarios of the risk factor $\left\{\left(x^n_{t_1}, x^n_{t_2}, \ldots, x^n_{T_M}\right)\middle| n = 1, \ldots, N\right\}$ and approximate $U(0)$ under the risk-neutral measure as

$$\tilde{U}(0) = M_0 + \frac{1}{N}\sum_{n=1}^{N}\max_{T_m \in \mathcal{T}_f}\left\{\frac{h_{T_m}\left(x^n_{T_m}\right)}{B^n(T_m)} - M^n_{T_m}\right\}$$

The upper bound can be approximated under the T_M-forward measure. In that case, the risk factor should be simulated under $\mathbb{Q}^{T_M}$ and the numéraire $B(t)$ should be replaced by $P(t, T_M)$. By carrying this out, we avoid the need to approximate the numéraire on a coarse simulation grid.

Note that by the deliberate choice of $G_m(\cdot)$, all the conditional expectations appearing in Equation (13) can be computed in closed form (see Appendix A). Hence, there is no need to resort to nested simulations, in contrast to, for example, Andersen and Broadie (2004) and Becker et al. (2020). Especially if simulations are performed under the T_M-forward measure, both lower and upper bound estimations can be obtained at minimal additional computational cost.

5. Error Analysis

In this section, we analyze the errors of the semi-static hedge, the direct estimator, the lower bound estimator, and the upper bound estimator, which are induced by the imprecision of the regression functions $G_0, \ldots, G_{M-1}$. We show that for a sufficiently large hedging portfolio, the replication error will be arbitrarily small. Furthermore, we will provide error margins for the price estimators in terms of the regression imprecision. We thereby show that the direct estimator, lower bound, and upper bound will converge to the true option price as the accuracy of the regressions increases. The cornerstone to the subsequent theorems is the universal approximation theorem, as presented in, for example, Hornik et al. (1989). Given that $\tilde{V}$ is a continuous function on the compact set $\mathcal{I}_d$, it guarantees that, for each $m \in \{0, \ldots, M-1\}$, there exists a neural network G_m such that

$$\sup_{x \in \mathcal{I}_d} B^{-1}(T_m) \left| \tilde{V}(T_m; x) - G_m(z_m(T_m)|x) \right| < \varepsilon$$

for arbitrary $\varepsilon > 0$. In other words, the regression error can be kept arbitrarily small on any compact domain of the risk factor.

5.1. Accuracy of the Semi-Static Hedge

Let $\mathcal{T}_f = \{T_0, \ldots, T_{M-1}\}$ denote the set of monitor dates. For the following theorem, we assume that $\mathbf{x}_t \in \mathcal{I}_d$ for some compact set $\mathcal{I}_d \subset \mathbb{R}^d$. As $\mathcal{I}_d$ can be arbitrarily large, this assumption is loose enough to account for a vast majority of the risk factor scenarios in a standard Monte Carlo sample. On top of that, $\mathcal{I}_d$ can be chosen as sufficiently large such that $\mathbb{E}^{\mathbb{Q}}\left[\left| \tilde{V}(T_m) - G_m(z_m) \right| \mathbb{1}_{\left\{ \mathbf{x}_{T_m} \notin \mathcal{I}_d \right\}} \middle| \mathcal{F}_0 \right]$ approaches zero. For the proof, we refer to Appendix D.

Theorem 1. *Let $\varepsilon > 0$ and $|\mathcal{T}_f| = M$. Denote by $\tilde{V}(t)$ the value of the replication portfolio for a Bermudan swaption, conditional on the fact that it is not exercised prior to time t. Assume that there exist M networks $G_m(\cdot)$ such that*

$$\sup_{x \in \mathcal{I}_d} B^{-1}(T_m) \left| \tilde{V}(T_m; x) - G_m(z_m(T_m)|x) \right| < \varepsilon, \qquad \forall_{m \in \{0,\ldots,M-1\}}$$

Then, for any $t \in [0, T_{M-1}]$, we have that

$$\sup_{x \in \mathcal{I}_d} B^{-1}(t) \left| V(t; x) - \tilde{V}(t; x) \right| < M\varepsilon$$

5.2. Error of the Direct Estimator

Theorem 1 bounds the hedging error of the semi-static hedge in terms of the maximum regression errors. This implicitly provides an error margin to the direct estimator under the aforementioned assumptions. Although the universal approximation theorem guarantees that the supremum errors can be kept at any desired level, in practice, they are substantially higher than, for example, the MSEs or MAEs of the regression function. This is due to inevitable fitting imprecision outside or near the boundaries of the finite training sets. In the following theorem, we propose that the error of the direct estimator can be bounded in terms of the discounted MAEs of the neural networks. These quantities are generally much tighter than the supremum errors and are typically easier to estimate.

The proof of the theorem follows a similar line of thought as the proof of Theorem 1. As the direct estimator at time-zero depends on the expectation of the continuation value at T_0, we can show by an iterative argument that the overall error is bounded by the sum of the mean absolute fitting errors at each monitor date. The error bound in the direct estimator therefore scales linearly with the number of exercise opportunities. For a complete proof, we refer to Appendix E.

Theorem 2. *Let $\varepsilon > 0$ and assume that $|\mathcal{T}_f| = M$. Denote by $\tilde{V}$ the time-zero direct estimator for the price of a Bermudan swaption V. Assume that, for each $T_m \in \{T_0, \ldots, T_{M-1}\}$, there is a neural network approximation $G_m(\cdot)$ such that*

$$\mathbb{E}^{\mathbb{Q}}\left[B^{-1}(T_m)\left|\tilde{V}(T_m) - G_m(z_m)\right|\middle|\mathcal{F}_0\right] < \varepsilon$$

where $\tilde{V}(T_m) := \max\left\{B(T_m)\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right], h_m(\mathbf{x}_{T_m})\right\}$ denotes the estimator at date T_m. Then, the error in $\tilde{V}$ is bounded as given below:

$$\left|V(0) - \tilde{V}(0)\right| < M\varepsilon$$

5.3. Tightness of the Lower Bound Estimate

A lower bound $L(t)$ to the true price can be computed by considering the non-optimal exercise strategy, implied by the direct estimator (see Section 4.1). This relies on the stopping time

$$\tilde{\tau}(\omega) = \min\left\{T_m \in \mathcal{T}_f \middle| \tilde{C}_m(T_m) \leq h_m(\mathbf{x}_{T_m})\right\} \tag{14}$$

In the following theorem, we propose that the tightness of $L(0)$ can be bounded by the discounted MAEs of neural network approximations.

The proof of the theorem relies on the fact that, conditioned on any realization of $\tilde{\tau}$ and τ, the expected difference between $L(0)$ and $V(0)$ is bounded by the sum of the mean absolute fitting errors at the monitor dates between $\tilde{\tau}$ and τ. In the proof, we therefore distinguish between the events $\tilde{\tau} < \tau$ and $\tilde{\tau} > \tau$. Then, by an inductive argument, we can show that the bound on the spread between $L(0)$ and the true price scales linearly with the number of exercise opportunities. For a complete proof, we refer to Appendix F.

Theorem 3. *Let $\varepsilon > 0$ and assume that $|\mathcal{T}_f| = M$. Denote by $L(0)$ the lower bound on the true Bermudan swaption price as defined in Equation (10). Assume that, for each $T_m \in \{T_0, \ldots, T_{M-1}\}$, there is a neural network approximation $G_m(\cdot)$, such that*

$$\mathbb{E}^{\mathbb{Q}}\left[B^{-1}(T_m)\left|\tilde{V}(T_m) - G_m(z_m)\right|\middle|\mathcal{F}_0\right] < \varepsilon$$

where $\tilde{V}(T_m) := \max\left\{B(T_m)\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right], h_m(\mathbf{x}_{T_m})\right\}$ denotes the estimator at date T_m. Then, the spread between $V(0)$ and $L(0)$ is bounded as given below:

$$|V(0) - L(0)| < 2(M-1)\varepsilon$$

5.4. Tightness of the Upper Bound Estimate

An upper bound $U(t)$ to the true price can be computed by considering a dual formulation of the dynamic pricing equation Haugh and Kogan (2004); see Section 4.2. From a practical point of view, the difference between the upper bound and the true price can be interpreted as the maximum loss that an investor would incur due to hedging imprecision resulting from the algorithm Lokeshwar et al. (2022). The overall hedging error at some monitor date T_m is the result of all incremental hedging errors occurring from rebalancing the portfolio at preceding monitor dates. As the incremental hedging errors can be bounded by the sum of the expected absolute fitting errors, we propose that the tightness of $U(t)$ can be bounded by the discounted MAEs of the neural networks and scales at most quadratically with the number of exercise opportunities.

The proof follows a similar line of thought as that presented in Andersen and Broadie (2004). There, it is noted that the difference between the dual formulation of the option and its true price is difficult to be bound. Here, we make a similar remark and propose a

theoretical maximum spread between $U(0)$ and $V(0)$ that is relatively loose. Our numerical experiments, however, indicate that the upper bound estimate is much tighter in practice. For a complete proof, we refer to Appendix G.

Theorem 4. *Let $\varepsilon > 0$ and assume that $|\mathcal{T}_f| = M$. Denote by $U(0)$ the upper bound on the true Bermudan swaption price as defined in Equation (11). Assume that, for each $T_m \in \{T_0, \ldots, T_{M-1}\}$, there is a neural network approximation $G_m(\cdot)$, such that*

$$\mathbb{E}^{\mathbb{Q}}\left[B^{-1}(T_m)\left|\tilde{V}(T_m) - G_m(z_m)\right|\middle|\mathcal{F}_0\right] < \varepsilon$$

where $\tilde{V}(T_m) := \max\left\{B(T_m)\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right], h_m(\mathbf{x}_{T_m})\right\}$ denotes the estimator at date T_m. Then, the spread between $V(0)$ and $U(0)$ is bounded as given below:

$$|U(0) - V(0)| < M(M-1)\varepsilon$$

6. Numerical Experiments

In this section, we treat several numerical examples to illustrate the convergence, pricing, and hedging performance of our proposed method. We will start by considering the price estimate of a vanilla swaption contract in a one-factor model. This is a toy example by which we can demonstrate the accuracy of the direct estimator in comparison to exact benchmarks. We continue with price estimates of Bermudan swaption contracts in a one-factor and a two-factor framework. The performance of the direct estimator will be compared to the established least-square regression method (LSM) introduced in Longstaff and Schwartz (2001), fine-tuned to an interest rate setting as described in Oosterlee et al. (2016). Additionally, we will approximate the lower and upper bound estimates as described in Section 4 and show that they are well inside the error margins introduced in Section 5. Finally, we will illustrate the performance of the static hedge for a swaption in a one-factor model and a Bermudan swaption in a two-factor model. For the one-factor case, we can benchmark the performance by the analytic delta hedge for a swaption, provided in Henrard (2003).

A $T_0 \times T_M$ contract (either European swaption or Bermudan swaption) refers to an option written on a swap with a notional amount of 100 and a lifetime between T_0 and T_M. This means that T_0 and T_{M-1} are the first and last monitor dates, respectively, in case of a Bermudan. The underlying swaps are set to exchange annual payments, yielding year fractions of 1 and annual exercise opportunities. All examples that are illustrated here have been implemented in Python using the Quant-Lib library Ametrano and Ballabio (2003) for standard pricing routines and Keras with Tensorflow backend Chollet et al. (2015) for constructing, fitting, and evaluating the neural networks.

6.1. 1-Factor Swaption

We start by considering a swaption contract under a one-dimensional risk factor setting. The direct estimator of the true $V(0)$ swaption price is computed similar to a Bermudan swaption, but with only a single exercise possibility at T_0. Therefore, only a single neural network per option needs to be trained to compute the option price. We have used 64 hidden nodes and 20,000 training points, generated through Monte Carlo sampling. We assume the risk factor to be captured by the Hull–White model with constant mean reversion parameter a and constant volatility σ. The dynamics of the shifted mean-zero process Brigo and Mercurio (2006) are hence given by

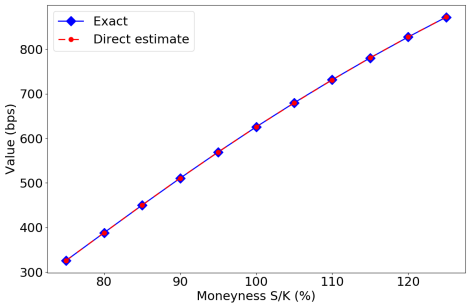
$$dx(t) = -ax(t)dt + \sigma dW(t), \qquad x(0) = 0 \tag{15}$$

For simplicity, we consider a flat time-zero instantaneous forward rate $f(0,t)$. The risk-neutral scenarios are generated using a discrete Euler scheme of the process above. Parameter values that were used in the numerical experiments are summarized in Table 1.

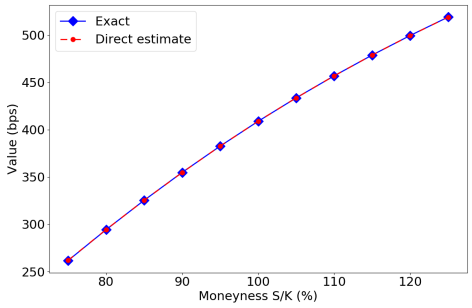
Table 1. Parameters 1F Hull–White model.

Parameter	a	σ	$f(0,t)$
Value	0.01	0.01	0.03

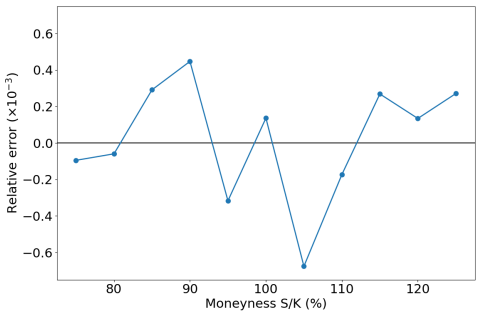
Figure 3a,b show the time-zero option values in basis points (0.01%) of the notional for a $5Y \times 10Y$ and a $10Y \times 5Y$ payer swaption as a function of the moneyness. The moneyness is defined as $\frac{S}{K}$, where K denotes the fixed strike and S the time-zero swap rate associated with the underlying swap. The exact benchmarks are computed by an application of Jamshidian's decomposition Jamshidian (1989). The relative estimate errors are shown in Figure 3c,d. We observe a close agreement between the estimates and the reference prices. The errors are in the order of several basis points of the true option price. In the current setting, the results presented serve mostly as a validation of the estimator. We however point out that this algorithm for swaptions is applicable in general frameworks, such as multi-factor, dual-curve, or non-overlapping payment schemes, for which exact routines are no longer available.



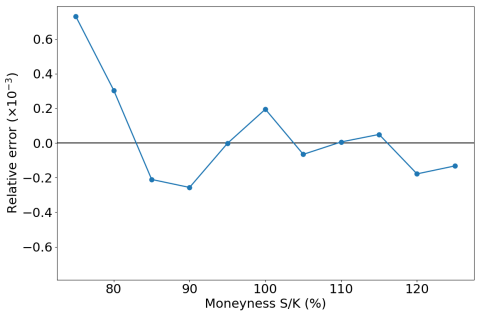
(**a**) $5Y \times 10Y$ swaption valuation



(**b**) $10Y \times 5Y$ swaption valuation



(**c**) Error $5Y \times 10Y$ swaption estimates



(**d**) Error $10Y \times 5Y$ swaption estimates

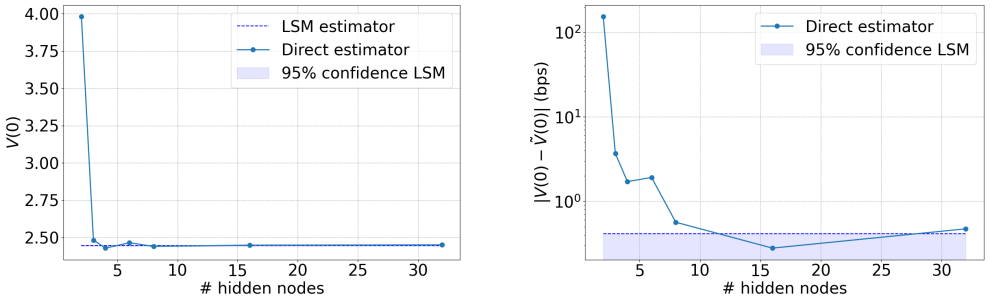
Figure 3. Accuracy of the direct estimator for vanilla swaptions. $S_{5Y\times10Y} \approx S_{10Y\times5Y} \approx 0.0305$.

6.2. 1-Factor Bermudan Swaption

As a second example, we consider a Bermudan swaption contract. The same dynamics for the underlying risk factor are assumed as discussed in the previous paragraph, using the parameter settings of Table 1. Monte Carlo scenarios are generated based on a discretized Euler scheme associated to the SDE in Equation (15), taking weekly time-steps.

We first demonstrate the convergence property of the direct estimator, which is implied by the replication portfolio. We consider a $1Y \times 5Y$ Bermudan swaption with strike $K = 0.03$. This strike is selected as it close to ATM, a moneyness level that is most likely to be liquid

in the market. For this analysis, the neural networks were trained to a set of 2000 Monte-Carlo-generated training points. Figure 4a shows the direct estimator as a function of the number of hidden nodes in each neural network, alongside an LSM-based benchmark. In Figure 4b, the error with respect to the LSM estimate is shown on a logscale. We observe that the direct estimator converges to the LSM confidence interval or slightly above, which is in accordance with the fact that LSM is biased low by definition. The analysis indicates that a portfolio of 16 discount bond options is sufficient to achieve a replication of a similar accuracy to the LSM benchmark.



(**a**) Convergence price (**b**) Convergence pricing error

Figure 4. Convergence of the direct estimator for the $1Y \times 5Y$ Bermudan swaption price as a function of hidden node count, with respect to the LSM benchmark under a 1-factor model.

Table 2 depicts numerical pricing results for a $1Y \times 5Y$, $3Y \times 7Y$ and $1Y \times 10Y$ receiver Bermudan swaption. For each contract, we consider different levels of moneyness, setting the fixed rate K of the underlying swap to, respectively, 80%, 100%, and 120% of the time-zero swap rate. The estimations of the direct, the upper bound, and the lower bound statistics are again reported alongside LSM-based benchmarks. Here, the neural networks have 64 hidden nodes and are fitted using a training set of 20,000 points. The lower and upper bound estimates, as well as the LSM estimates, are based on simulation runs of 200,000 paths each. The given lower and upper bounds are Monte Carlo estimates of the statistics defined in Equations (10) and (11) and are therefore subject to standard errors, which are reported in parentheses. The reference LSM results have been generated using $\{1, x, x^2\}$ as regression basis functions for approximating the continuation values. The standard errors and confidence intervals are obtained from ten independent Monte Carlo runs. The choice for hyperparameter settings is motivated by the analysis of Appendix C.

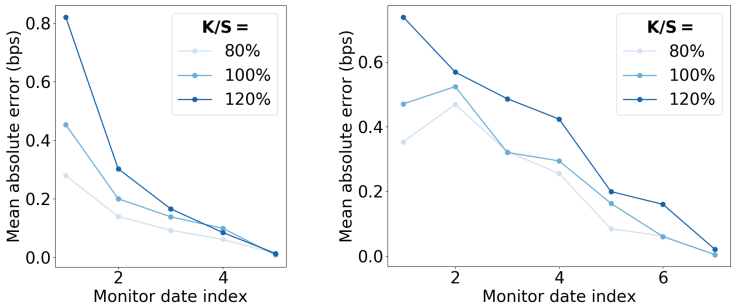
Table 2. Results of 1-factor model. $S_{1Y\times5Y} \approx S_{3Y\times7Y} \approx S_{1Y\times10Y} \approx 0.0305$. Standard errors are in parentheses, based on 10 independent MC runs of 2×10^5 paths each.

Type	K/S	Dir.est.	Lower bnd	Upper bnd	UB-LB	LSM est.	LSM 95% CI
$1Y \times 5Y$	80%	1.527	1.521 (0.001)	1.528 (0.000)	0.007	1.521 (0.001)	[1.518, 1.523]
	100%	2.543	2.534 (0.002)	2.542 (0.000)	0.008	2.534 (0.002)	[2.531, 2.538]
	120%	4.015	4.016 (0.002)	4.018 (0.000)	0.002	4.016 (0.002)	[4.012, 4.021]
$3Y \times 7Y$	80%	3.296	3.293 (0.002)	3.295 (0.000)	0.002	3.293 (0.002)	[3.290, 3.296]
	100%	4.767	4.755 (0.004)	4.761 (0.000)	0.006	4.755 (0.004)	[4.747, 4.762]
	120%	6.625	6.629 (0.004)	6.631 (0.000)	0.002	6.629 (0.004)	[6.621, 6.638]
$1Y \times 10Y$	80%	3.950	3.945 (0.005)	3.960 (0.000)	0.015	3.945 (0.005)	[3.935, 3.955]
	100%	5.818	5.811 (0.003)	5.818 (0.000)	0.007	5.811 (0.003)	[5.805, 5.816]
	120%	8.346	8.354 (0.005)	8.360 (0.000)	0.006	8.353 (0.005)	[8.344, 8.362]

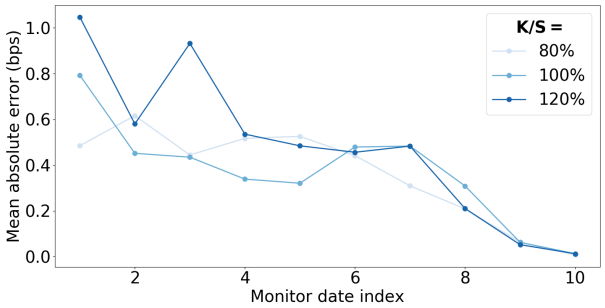
The spreads between the lower and upper bound estimates provide a good indication of the accuracy of the method. For the current setting, we obtain spreads in the order of several basis points up a few dozen of basis points. The lower bound estimate is typically very close to the LSM estimate, which itself is also biased low. Their standard errors are of the same order of magnitude. The upper bound estimates prove to be very stable and show a variance that is roughly two orders of magnitude smaller compared to that of the lower bound. The direct estimate is occasionally slightly less accurate. This can be explained by the fact that it depends on the accuracy of the regression over the full domain of the risk factor, whereas, for the lower bound, only a high accuracy near the exercise boundaries is required. In Figure 5, the mean absolute error of each neural network after fitting is presented as a function of the network's index. The errors are displayed in basis points of the notional. We observe that the errors are the smallest at maturity and tend to increase with each iteration backward in time. That the errors at the final monitor date are virtually zero can be explained by the fact that the pay-off at T_{M-1} is given by

$$\begin{aligned}\max\{h_{M-1}(\mathbf{x}_{T_{M-1}}), 0\} &= N \cdot \max\{A_{M-1,M}(T_{M-1}) \cdot (K - S_{M-1,M}(T_{M-1})), 0\} \\ &= N \cdot \max\{(\Delta T_M K + 1)P(T_{M-1}, T_M) - 1, 0\} \\ &\simeq w_2 \varphi(w_1 z - b)\end{aligned}$$

which can be exactly captured by a network with only a single hidden node. With each step backwards, the target function is harder to fit, yielding larger errors. We observe MAEs up to one basis point of the notional amount. The empirical lower–upper bound spreads remain well within the theoretical error margins provided in Sections 4.1 and 4.2. The spreads are mostly much lower than the sum of the MAEs, indicating that the bound estimates are in practice significantly tighter than their theoretical maximum spread.



(**a**) 1Y×5Y Bermudan (**b**) 3Y×7Y Bermudan



(**c**) 1Y×10Y Bermudan

Figure 5. Mean absolute errors of neural network fit per monitor date under a 1-factor model.

6.3. 2-Factor Bermudan Swaption

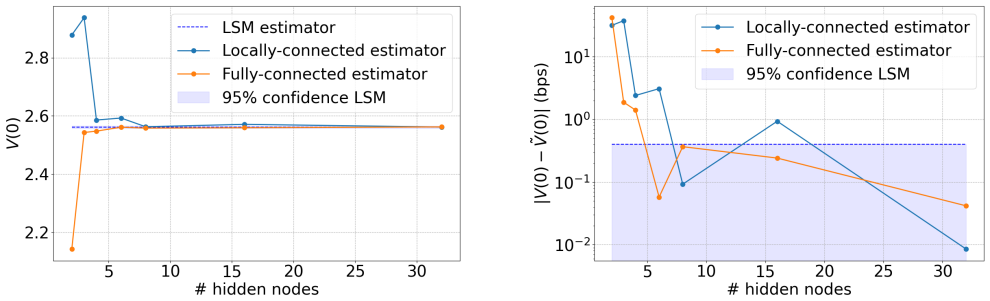
As a final pricing example, we consider a Bermudan swaption contract under a two-factor model. The dynamics of the underlying risk factors are assumed to follow a

G2++ model Brigo and Mercurio (2006). Monte Carlo scenarios are generated based on a discretized Euler scheme, taking weekly time-steps, based on the SDE below:

$$\begin{aligned} dx_1(t) &= -a_1 x_1(t)dt + \sigma_1 dW_1(t), \qquad x_1(0) = 0 \\ dx_2(t) &= -a_2 x_2(t)dt + \sigma_2 dW_2(t), \qquad x_2(0) = 0 \end{aligned}$$

where W_1 and W_2 are correlated Brownian motions with $d\langle W_1, W_2\rangle_t = \rho dt$. Parameter values that were used in the numerical experiments are summarized in Table 3.

We again start by demonstrating the convergence property of the direct estimator for both the locally connected and the fully connected neural network designs as specified in Section 3.2.3. The same $1Y \times 5Y$ Bermudan swaption with strike $K = 0.03$ is used and the networks are each fitted to a set of 6400 training points. Figure 6a shows the direct estimator as a function of the number of hidden nodes in each neural network, alongside an LSM-based benchmark. In Figure 6b, the error with respect to the LSM estimate is shown on a logscale. We observe a similar convergence behavior, where the direct estimators approach the LSM benchmark within the 95% confidence range. Here, it is noted that a portfolio of eight discount bond options is already sufficient to achieve a replication of a similar accuracy to the LSM estimator.



(**a**) Convergence price (**b**) Convergence pricing error

Figure 6. Convergence of the direct estimator for the $1Y \times 5Y$ Bermudan swaption price as a function of hidden node count, with respect to the LSM benchmark under a 2-factor model.

Table 3. Parameters 2F G2++ model.

Parameter	a_1	a_2	σ_1	σ_2	ρ	$f(0,t)$
Value	0.07	0.08	0.015	0.008	−0.6	0.03

In Table 4, numerical results for a $1Y \times 5Y$, $3Y \times 7Y$, and $1Y \times 10Y$ receiver Bermudan swaption are depicted for different levels of moneyness. We again report the direct, the upper bound, and the lower bound estimates for both neural network designs. In this case, all networks have 64 hidden nodes and are fitted to training sets of 20,000 points. As before, the lower bound, the upper bound, and the LSM estimates are the result of 10 independent Monte Carlo simulations of 200,000 scenarios.

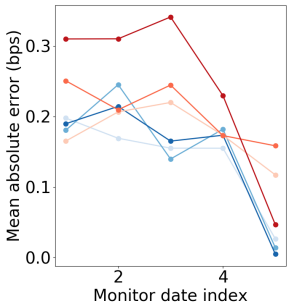
For the LSM algorithm, we used $\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}$ as basis functions. Note that the number of monomials grows quadratically with the dimension of the state space and, with that, the number of free parameters. For our method, this number grows at a linear rate. Choices for the hyperparameters are again based on the analysis of Appendix C. The results under the two-factor case share several features with the one-factor results. We observe spreads between the lower and upper bounds ranging from several basis points up to a few dozen basis points of the option price. The lower bound estimates turn out to be very close to the LSM estimates and the same holds for their standard errors. The upper bounds are again very stable with low standard errors and the direct estimator appears as slightly less accurate. If we compare the locally connected to the fully connected case, we observe that the results are overall in close agreement, especially the lower and upper

bound estimates. This is remarkable given that the fully connected case gives rise to more trainable parameters, by which we would expect a higher approximation accuracy. In the two-factor setting, the ratio of free parameters for the two designs is 3:4.

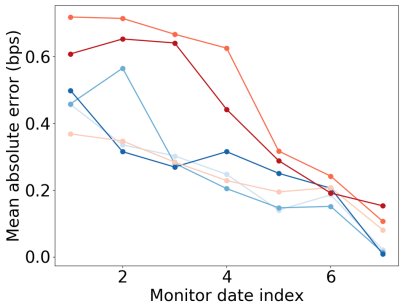
Table 4. Results of 2-factor model for the locally connected and fully connected neural network cases. $S_{1Y\times 5Y} \approx S_{3Y\times 7Y} \approx S_{1Y\times 10Y} \approx 0.0305$. Standard errors are in parentheses, based on 10 independent MC runs of 2×10^5 paths each.

Locally connected neural networks							
Type	K/S	Dir.est.	Lower bnd	Upper bnd	UB-LB	LSM est.	LSM 95% CI
$1Y \times 5Y$	80%	1.617	1.617(0.002)	1.619(0.000)	0.002	1.617(0.002)	[1.614, 1.621]
	100%	2.652	2.650(0.002)	2.654(0.000)	0.004	2.650(0.002)	[2.646, 2.654]
	120%	4.128	4.127(0.003)	4.131(0.000)	0.004	4.127(0.003)	[4.121, 4.132]
$3Y \times 7Y$	80%	3.073	3.076(0.004)	3.078(0.000)	0.002	3.077(0.004)	[3.069, 3.085]
	100%	4.554	4.553(0.004)	4.553(0.000)	0.000	4.552(0.004)	[4.545, 4.559]
	120%	6.444	6.448(0.004)	6.451(0.000)	0.003	6.446(0.005)	[6.435, 6.456]
$1Y \times 10Y$	80%	3.616	3.624(0.002)	3.626(0.000)	0.002	3.622(0.002)	[3.618, 3.627]
	100%	5.508	5.509(0.002)	5.514(0.000)	0.005	5.508(0.002)	[5.503, 5.512]
	120%	8.128	8.123(0.005)	8.130(0.000)	0.007	8.121(0.005)	[8.110, 8.132]
Fully connected neural networks							
Type	K/S	Dir.est.	Lower bnd	Upper bnd	UB-LB	LSM est.	LSM 95% CI
$1Y \times 5Y$	80%	1.617	1.617(0.002)	1.619(0.000)	0.002	1.617(0.002)	[1.614, 1.621]
	100%	2.651	2.650(0.002)	2.654(0.000)	0.004	2.650(0.002)	[2.646, 2.654]
	120%	4.129	4.127(0.003)	4.131(0.000)	0.004	4.127(0.003)	[4.121, 4.132]
$3Y \times 7Y$	80%	3.076	3.077(0.004)	3.078(0.000)	0.001	3.077(0.004)	[3.069, 3.085]
	100%	4.553	4.553(0.004)	4.554(0.000)	0.001	4.552(0.004)	[4.545, 4.559]
	120%	6.451	6.447(0.005)	6.451(0.000)	0.004	6.446(0.005)	[6.435, 6.456]
$1Y \times 10Y$	80%	3.616	3.624(0.002)	3.626(0.000)	0.002	3.622(0.002)	[3.618, 3.627]
	100%	5.506	5.509(0.002)	5.514(0.000)	0.005	5.508(0.002)	[5.503, 5.512]
	120%	8.124	8.123(0.005)	8.130(0.000)	0.007	8.121(0.005)	[8.110, 8.132]

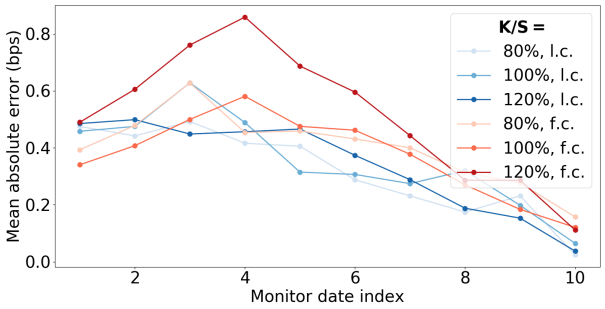
In Figure 7, the mean absolute errors of the neural networks after fitting are shown. The MAEs for the locally connected networks are in blue; the fully connected are in red. All are represented in basis points of the notional amount. We observe that the errors are mostly in the same order of magnitude as the one-dimensional case. The figures indicate that the locally connected networks slightly outperform the fully connected networks in terms of accuracy, although this does not appear to materialize in tighter estimates of the lower and upper bounds. For the locally connected case, we again observe that the errors are virtually zero at the last monitor date, for the same reasons as in the one-factor setting. In the fully connected representation, an exact replication might not exist, resulting in larger errors. We conjecture that this effect partially carries over to the networks at preceding monitor dates. The empirical lower–upper bound spreads remain well within the theoretical error margins, as the spreads are in all cases lower than the sum of the MAEs. Hence, also for the two-factor setting, we find that the bound estimates are tighter in practice than their theoretical maximum spreads.



(**a**) 1Y×5Y Bermudan



(**b**) 3Y×7Y Bermudan



(**c**) 1Y×10Y Bermudan

Figure 7. Accuracy of neural network fit per monitor date under a 2-factor model. Blue lines represent the locally connected (l.c.) case and the red lines represent the fully connected (f.c.) case. The legend in Figure (**c**) applies to all three graphs.

6.4. Performance Semi-Static Hedge

Finally, we consider the hedging problem of a vanilla swaption under the one-factor model and a Bermudan swaption under the two-factor model.

6.4.1. 1-Factor Swaption

Here, we compare the performance of a static hedge versus a dynamic hedge in the one-factor model. As an example, we take a $1Y \times 5Y$ European receiver swaption at different levels of moneyness. The model set-up is similar to that in Section 6.2, using the same set of parameters reported in Table 1. In the static hedge case, the option contract writer aims to hedge the risk using a static portfolio of zero-coupon bond options and discount bonds. The replicating portfolio is composed using a neural network with 64 hidden nodes, optimized using 20,000 training-points generated through Monte Carlo sampling. The portfolio is composed at time-zero and kept until the expiry of the option at $t = 1$ year. In the dynamic hedge case, the delta-hedging strategy is applied. The replicating portfolio is composed of units of the underlying forward-starting swap and investment in the money market. The dynamic hedge involves the periodic rebalancing of the portfolio. The delta for a receiver swaption under the Hull–White model (see Henrard 2003) is given by

$$\Delta(t) = \frac{\sum_{j=1}^{M} c_j P(t,T_j)\nu(t,T_j)\Phi(\kappa+\alpha_j) - P(t,T_0)\nu(t,T_0)\Phi(\kappa)}{\sum_{j=1}^{M} c_j P(t,T_j)\nu(t,T_j) - P(t,T_0)\nu(t,T_0)} \tag{16}$$

where κ is the solution of

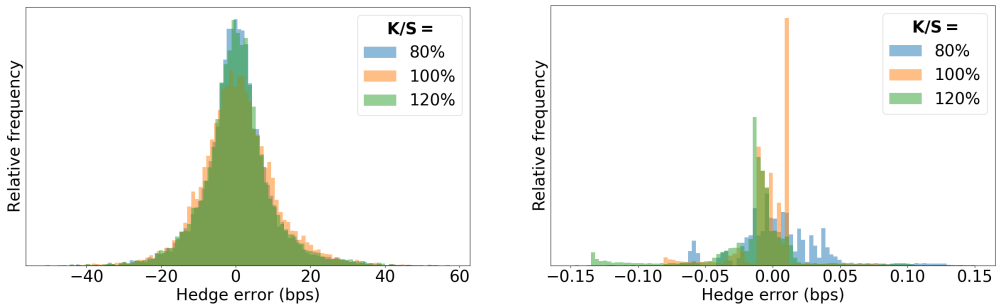
$$\sum_{j=1}^{M} c_j \frac{P(t,T_j)}{P(t,T_0)} \exp\left(-\frac{1}{2}\alpha_j^2 - \alpha_j \kappa\right) = 1$$

and

$$\alpha_j^2 := \int_0^{T_0} \big(\nu(u,T_j) - \nu(u,T_0)\big)^2 du$$

where Φ denotes the CDF of a standard normal distribution, $c_j = \Delta T_j K$ for $j = 1, \ldots, M-1$, and $c_M = 1 + \Delta T_M K$. The function $\nu(t,T)$ denotes the instantaneous volatility of a discount bond maturing at T, which, under Hull–White, is given by $\nu(t,T) := \frac{\sigma}{a}\Big(1 - e^{-a(T-t)}\Big)$. We validated the analytic expression above with numerical approximations of the Delta obtained by bumping the yield curve. Within the simulation, the dynamic hedge portfolio is rebalanced on a daily basis between time-zero and expiry of the option. In this experiment, that means it is updated on 255 instances at equidistant monitor dates.

The performance of both hedging strategies is reported in Table 5. The results are based on 10,000 risk-neutral Monte Carlo paths. The hedging error refers to the difference between the option's pay-off at expiry and the replicating portfolio's final value. The quantities are reported in basis points of the notional amount. The empirical distribution of the hedging error is shown in Figure 8. We observe that, overall, the static hedge outperforms the dynamic hedge in terms of accuracy, even though it involves only a quarter (64 versus 255) of the trades. Although it is not visible in Figure 8b, the static strategy does give rise to occasional outliers in terms of accuracy. These are associated with scenarios that reach or exceed the boundary of the training set. These errors are typically of a similar order of magnitude as the errors observed in the dynamic hedge. The impact of outliers can be reduced by increasing the training set and thereby broadening the regression domain.



(**a**) Hedge error dynamic strategy (**b**) Hedge error static strategy

Figure 8. Hedge error distribution for a $1Y \times 5Y$ receiver swaption, based on 10^4 MC paths. $S_{1Y\times 5Y} \approx 0.0305$.

Table 5. Hedging errors for static and dynamic hedging strategy for a $1Y \times 5Y$ receiver swaption, based on 10^4 MC paths. $S_{1Y\times 5Y} \approx 0.0305$.

Hedge Error (bps)	K/S	Static Hedge	Dyn. Hedge
Mean	80%	-1.9×10^{-2}	0.38
	100%	-2.2×10^{-3}	0.61
	120%	-1.5×10^{-2}	0.46
St. dev.	80%	2.5	9.1
	100%	3.1×10^{-2}	10.1
	120%	4.5×10^{-2}	9.4
95%-percentile	80%	6.6×10^{-2}	15.7
	100%	1.2×10^{-2}	17.9
	120%	2.0×10^{-2}	16.2

6.4.2. 2-Factor Bermudan Swaption

Here, we demonstrate the performance of the semi-static hedge for a $1Y \times 5Y$ receiver Bermudan swaption under a two-factor model. We compare the accuracy of the hedging strategy utilizing a locally connected network versus a fully connected neural network.

In the former, the replication portfolio consists of zero-coupon bonds and zero-coupon bond options. In the latter, the Bermudan is replicated with options written on hypothetical assets with a pay-off equal to the log of a zero-coupon bond (see Section 3.2.3). The model set-up is similar to that in Section 6.3, using the same set of parameters reported in Table 3. Both networks are composed with 64 hidden nodes and optimized using 20,000 training points generated through Monte Carlo sampling. The portfolio is set up at time-zero and updated at each monitor date of the Bermudan until it is either exercised or expired. We assume that the holder of the Bermudan swaption follows the exercise strategy implied by the algorithm, i.e., the option is exercised as soon as $\tilde{C}_m(T_m) \leq h_m(\mathbf{x}_{T_m})$. When a monitor date T_m is reached, the replication portfolio matures with a pay-off equal to $G_m(z_m(T_m))$. In case the Bermudan is continued, the price to set up a new replication portfolio is given by $\tilde{V}(T_m) = B(T_m)\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right]$, which contributes $G_m(z_m(T_m)) - \tilde{V}(T_m)$ to the hedging error. In case the Bermudan is exercised, the holder will claim $\tilde{V}(T_m) = h_m(\mathbf{x}_{T_m})$, which also contributes $G_m(z_m(T_m)) - \tilde{V}(T_m)$ to the hedging error. The total error of the semi-static hedge (HE) is therefore computed as

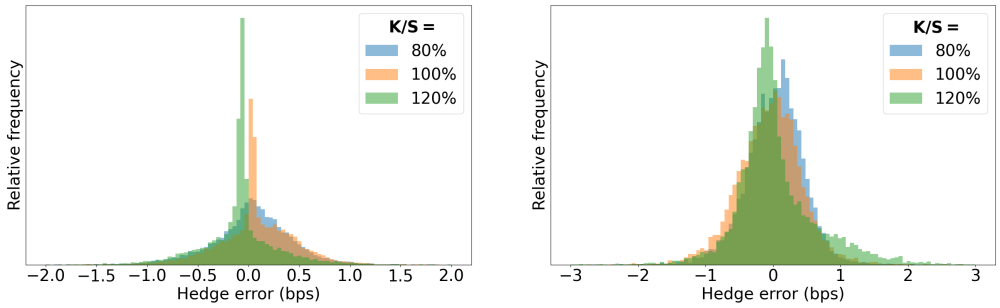
$$\text{HE} := \sum_{m=0}^{M-1} \left(G_m(z_m(T_m)) - \tilde{V}(T_m)\right) \mathbb{1}_{\{\tilde{\tau} \leq T_m\}}$$

where $\tilde{V}(T_m) := \max\left\{B(T_m)\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right], h_m(\mathbf{x}_{T_m})\right\}$ denotes the direct estimator at date T_m and $\tilde{\tau}$ denotes the stopping time, as defined in Equation (9).

The performance of the strategies related to locally and fully connected neural networks is reported in Table 6. The results are based on 10,000 risk-neutral Monte Carlo paths and reported in basis points of the notional amount. The empirical distribution of the hedging error is shown in Figure 9. We observe that both approaches yield an accuracy in the same order of magnitude, although the locally connected case slightly outperforms the fully connected case. This is in line with expectations, as the fitting performance of the locally connected networks is generally higher. For similar reasons to the one-factor case, the hedging experiments give rise to occasional outliers in terms of accuracy. These outliers can be in the order of several dozens of basis points. Again, the impact of outliers can be reduced by broadening the regression domain.

Table 6. Hedging errors of the semi-static hedging strategy for a $1Y \times 5Y$ receiver Bermudan swaption, based on 10^4 MC paths. $S_{1Y\times 5Y} \approx 0.0305$.

Hedge Error (bps)	K/S	Loc. conn. NN	Fully conn. NN
Mean	80%	3.2×10^{-2}	2.1×10^{-2}
	100%	7.9×10^{-2}	-5.5×10^{-2}
	120%	-9.4×10^{-2}	4.5×10^{-2}
St. dev.	80%	0.45	0.55
	100%	0.38	0.48
	120%	0.37	0.67
95%-percentile	80%	0.66	0.69
	100%	0.56	0.85
	120%	0.72	0.76



(**a**) Hedge error locally connected NN (**b**) Hedge error fully connected NN

Figure 9. Hedge error distribution for a $1Y \times 5Y$ receiver Bermudan swaption, based on 10^4 MC paths. $S_{1Y\times 5Y} \approx 0.0305$.

7. Conclusions

In this paper, we have proposed a semi-static replication algorithm for Bermudan swaptions under an affine term structure model. We have shown that Bermudan swaptions, an exotic interest rate derivative that is heavily traded in the OTC market, can be semi-statically replicated with an options portfolio written on a basket of discount bonds. The static portfolio composition is obtained by regressing the target option's value using a shallow, artificial neural network. The choice of the regression basis functions are motivated by their representation of an option's portfolio pay-off, implying an interpretable neural network structure. Leveraging the approximating power of ANNs, we proved that the replication can achieve any desired level of accuracy given that the portfolio is sufficiently large. We derived a direct estimator of the contract price, and an upper bound and lower bound estimate to this price can be computed at minimal additional computational cost.

The algorithm we presented is inspired by the work of Lokeshwar et al. (2022), which proposes a semi-static replication approach for callable equity options embedded in the Black–Scholes model. We contribute to the literature by extending the concept of (semi-)static replication to the field of interest rate modeling. Next, to a direct, lower bound, and upper bound estimator, we have derived analytical error margins for these statistics. This proves their convergence as the regression error diminishes and provides a direct insight toward the accuracy of the estimates. Additionally, we propose an alternative ANN design, which constrains the replication into a portfolio of vanilla bond options, even in the case of a multi-factor model. This guarantees efficiency in the portfolio valuation, which is key to many applications in credit risk management.

The performance of the method was demonstrated through several numerical experiments. We focused on Bermudan swaptions under a one- and two-factor model, which are popular amongst practitioners. The pricing accuracy of the method was determined through a benchmark to the established least-square method of Longstaff and Schwartz (2001). This reference is approached with basis point precision. A convergence analysis showed that a portfolio of 16 bond options suffices in achieving a replication with a similar accuracy to the LSM. Finally, the replication performance was studied through an in-model hedging experiment. This showed that the semi-static hedge outperforms a traditional dynamic replication in terms of hedging error.

As a look-out for further research, we consider applying the algorithm to the computation of credit risk measures and various value adjustments (xVAs). These metrics typically rely on generating forward value and sensitivity profiles of (exotic) derivative portfolios. We see the semi-static replication approach combined with the simple error analysis as an effective tool to address the computational challenges associated with these risk measures. The performance of the method in the context of quantifying CCR will therefore be studied in a forthcoming companion paper.

Author Contributions: Conceptualization, J.H., S.J. and D.K.; Formal analysis, J.H., S.J., D.K.; Investigation, J.H.; Writing—original draft, J.H.; Writing—review and editing, S.J. and D.K.; Visualization, J.H.; Supervision, S.J. and D.K.; Project administration, D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the NWO under the Industrial Doctorates grant. Grant Number: NWA.ID.17.029

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Disclosure: The opinions expressed in this work are solely those of the authors and do not represent in any way those of their current and past employers.

Appendix A. Evaluation of the Conditional Expectation

In this section, we will explicitly compute the conditional expectations related to the continuation values. We will distinguish two approaches associated with the two proposed network structures, i.e., the locally connected case (suggestion 1) and the fully connected case (suggestion 2).

For ease of computation, we will use a simplified, yet equivalent representation of the risk factor dynamics discussed in Section 2.1. This concerns a linear shift of the canonical representation of the latent factors as presented in Dai and Singleton (2000). We write $\mathbf{x}_t := (x_1(t), \ldots, x_n(t))^\top$, where each component x_i denotes a mean-reverting zero-mean process. The risk-neutral dynamics are assumed to satisfy

$$d\begin{pmatrix} x_1(t) \\ \vdots \\ x_d(t) \end{pmatrix} = -\begin{pmatrix} a_1(t)x_1(t) \\ \vdots \\ a_d(t)x_d(t) \end{pmatrix} dt + \begin{pmatrix} \sigma_{11}(t) & \ldots & \sigma_{1d}(t) \\ \vdots & \ddots & \vdots \\ \sigma_{d1}(t) & \ldots & \sigma_{dd}(t) \end{pmatrix} d\mathbf{W}(t), \quad \begin{pmatrix} x_1(0) \\ \vdots \\ x_d(0) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{A1}$$

where $\mathbf{W}$ denotes a standard d-dimensional Brownian motion with independent entries. By setting $\tilde{\sigma}_i(t) := \sqrt{\sum_{j=1}^d \sigma_{ij}^2(t)}$, the process above can be rewritten in terms of one-dimensional Itô processes Shreve (2004) of the form

$$dx_i(t) = -a_i(t)x_i(t)dt + \tilde{\sigma}_i(t)d\tilde{W}_i(t), \qquad i = 1, \ldots, d \tag{A2}$$

where $\tilde{W}_1, \ldots, \tilde{W}_d$ denote a set of one-dimensional, correlated Brownian motions under the measure $\mathbb{Q}$. The instantaneous correlation is denoted by ρ_{ij}, such that $d\langle \tilde{W}_i, \tilde{W}_j \rangle_t = \rho_{ij}(t)dt$.

Appendix A.1. The Continuation Value with Locally Connected NN

We consider the network $G_m(\cdot)$, which is trained to approximate $\tilde{V}(T_m)$. Let $t \in [T_{m-1}, T_m)$. In order to obtain $\tilde{V}(t)$, we need to evaluate $\mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du} G_m(\mathbf{x}_{T_m}) \middle| \mathcal{F}_t\right]$. As $G_m(\cdot)$ represents the linear combination of the outcome of q hidden nodes, we will focus on the conditional expectation of hidden node $i \in \{1, \ldots, q\}$. Our aim is then to compute the following:

$$H_i(t) := \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du} \varphi(\mathbf{w}_i^\top \mathbf{P}(T_m) + b_i) \middle| \mathcal{F}_t\right]$$

The map $\varphi : \mathbb{R} \to \mathbb{R}$ denotes the ReLU function defined as $\varphi(x) = \max\{x, 0\}$. The weight vector $\mathbf{w}_i$ (corresponding to hidden node i) and $\mathbf{P}(T_m)$ are defined as

$$\mathbf{w}_i = \begin{pmatrix} w_1^i \\ \vdots \\ w_d^i \end{pmatrix}, \qquad \mathbf{P}(T_m) = \begin{pmatrix} P(T_m, T_m + \delta_1) \\ \vdots \\ P(T_m, T_m + \delta_d) \end{pmatrix}$$

with $T_m < T_m + \delta_1 < \ldots < T_m + \delta_d \leq T_M$. Recall that, as a characteristic of the affine term structure model, the random variable $P(t,T)$ can be expressed as

$$P(t,T) = e^{A(t,T) - \sum_{i=1}^{d} B_i(t,T) x_i(t)}$$

for deterministic functions A and B_i, which are available in closed form (see Brigo and Mercurio 2006). By the structure of the network, the weight vector is constrained to have only a single non-zero entry, which we will denote to have index k. Therefore, we can rewrite

$$H_i(t) = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du} \max\left\{w_i^k P(T_m, T_m + \delta_k) + b_i,\ 0\right\} \middle| \mathcal{F}_t\right]$$

As we argued before, if w_i^k and b_i are both non-negative, $H_i(t)$ denotes the value of a forward contract. In that case, we have

$$\begin{aligned} H_i(t) &= \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du}\left(w_i^k P(T_m, T_m + \delta_k) + b_i\right) \middle| \mathcal{F}_t\right] \\ &= w_i^k \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du} \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_{T_m}^{T_m+\delta_k} r(u)du} \middle| \mathcal{F}_{T_m}\right] \middle| \mathcal{F}_t\right] + b_i \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du} \middle| \mathcal{F}_t\right] \\ &= w_i^k P(t, T_m + \delta_k) + b_i P(t, T_m) \end{aligned}$$

If, on the other hand, $b_i < 0 < w_i^k$ or $w_i^k < 0 < b_i$, we are dealing with a European call or put option, respectively. Closed-form expressions for European bond options are available based on Black's formula and have been treated extensively in the literature; see, for example, Musiela and Rutkowski (2005), Filipovic (2009), or Brigo and Mercurio (2006). In our case, we have

$$H_i(t) = \begin{cases} w_i^k P(t, T_m + \delta_k)\Phi(d_+) + b_i P(t, T_m)\Phi(d_-) & \text{if } b_i < 0 < w_i^k \\ -b_i P(t, T_m)\Phi(-d_-) - w_i^k P(t, T_m + \delta_k)\Phi(-d_+) & \text{if } w_i^k < 0 < b_i \end{cases}$$

where Φ denotes the CDF of a standard normal distribution, and we define

$$d_\pm := \frac{\log\left(-\frac{w_i^k P(t, T_m + \delta_k)}{b_i P(t, T_m)}\right) \pm \frac{1}{2}\Sigma(t, T_m)}{\sqrt{\Sigma(t, T_m)}}$$

and

$$\Sigma(t, T_m) := \int_t^{T_m} \|\nu(u, T_m + \delta_k) - \nu(u, T_m)\|^2 du$$

In the expression above, the function $\nu(t,T) \in \mathbb{R}^d$ refers to the instantaneous volatility at time t of a discount bond maturing at T. Under the dynamics of Equation (A1), ν is given by

$$\nu(t,T) = \begin{pmatrix} \sum_{i=1}^{d} B_i(t,T)\sigma_{i1}(t) \\ \vdots \\ \sum_{i=1}^{d} B_i(t,T)\sigma_{id}(t) \end{pmatrix} \tag{A3}$$

Appendix A.2. The Continuation Value with Fully Connected NN

Once again, we consider the network $G_m(\cdot)$, focus on the outcome of hidden node $i \in \{1, \ldots, q\}$, and let $t \in [T_{m-1}, T_m)$. Now, our aim is to evaluate the conditional expectation below, which, by a change in numéraire argument, can be rewritten as

$$\begin{aligned} &\mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du} \varphi(\mathbf{w}_i^\top \log \mathbf{P}(T_m) - b_i) \middle| \mathcal{F}_t\right] \\ &\qquad = P(t, T_m)\mathbb{E}^{T_m}\left[\max\left\{\mathbf{w}_i^\top \log \mathbf{P}(T_m) - b_i),\ 0\right\} \middle| \mathcal{F}_t\right] \end{aligned}$$

where the expectation on the right is taken under the T_m-forward measure, taking $P(t, T_m)$ as the numéraire. The weight vector $\mathbf{w}_i$ (corresponding to hidden node i) and $\log \mathbf{P}(T_m)$ are defined as

$$\mathbf{w}_i = \begin{pmatrix} w_1^i \\ \vdots \\ w_d^i \end{pmatrix}, \qquad \log \mathbf{P}(T_m) = \begin{pmatrix} \log P(T_m, T_m + \delta_1) \\ \vdots \\ \log P(T_m, T_m + \delta_d) \end{pmatrix}$$

with $T_m < T_m + \delta_1 < \ldots < T_m + \delta_d \leq T_M$. We set the input dimension equal to the number of risk factors (i.e., $d = n$). Therefore, we can write

$$\begin{aligned}
\mathbf{w}_i^\top \log \mathbf{P}(T_m) &= \sum_{j=1}^{d} w_j^i \log P(T_m, T_m + \delta_j) \\
&= \sum_{j=1}^{d} w_j^i A(T_m, T_m + \delta_j) - \sum_{j=1}^{d} w_j^i \sum_{k=1}^{d} B_k(T_m, T_m + \delta_j) x_k(T_m) \\
&= \begin{pmatrix} w_1^i & \ldots & w_d^i \end{pmatrix} \begin{pmatrix} A(T_m, T_m + \delta_1) \\ \vdots \\ A(T_m, T_m + \delta_d) \end{pmatrix} \\
&\quad - \begin{pmatrix} w_1^i & \ldots & w_d^i \end{pmatrix} \begin{pmatrix} B_1(T_m, T_m + \delta_1) & \ldots & B_d(T_m, T_m + \delta_1) \\ \vdots & \ddots & \vdots \\ B_1(T_m, T_m + \delta_d) & \ldots & B_d(T_m, T_m + \delta_d) \end{pmatrix} \begin{pmatrix} x_1(T_m) \\ \vdots \\ x_d(T_m) \end{pmatrix} \\
&= \mathbf{w}_i^\top \mathbf{A}(T_m) - \mathbf{w}_i^\top \mathbf{B}(T_m) \mathbf{x}_{T_m}
\end{aligned}$$

where we implicitly define

$$\mathbf{A}(T_m) := \begin{pmatrix} A(T_m, T_m + \delta_1) \\ \vdots \\ A(T_m, T_m + \delta_d) \end{pmatrix},$$

$$\mathbf{B}(T_m) := \begin{pmatrix} B_1(T_m, T_m + \delta_1) & \ldots & B_d(T_m, T_m + \delta_1) \\ \vdots & \ddots & \vdots \\ B_1(T_m, T_m + \delta_d) & \ldots & B_d(T_m, T_m + \delta_d) \end{pmatrix}$$

In order to compute the conditional expectation of Equation (A4), a change in measure is required to obtain the dynamics of $x_1, \ldots, x_n$ under the T_m-forward measure. Consider the Radon–Nikodym derivative process Beyna (2013), defined by

$$\frac{d\mathbb{Q}^{T_m}}{d\mathbb{Q}} \bigg| \mathcal{F}_t = \frac{B(t)}{B(T_m)} \frac{P(T_m, T_m)}{P(t, T_m)} = \exp\left\{ - \int_t^{T_m} \nu(u, T_m) \cdot d\mathbf{W}(u) - \frac{1}{2} \int_t^{T_m} \|\nu(u, T_m)\|^2 du \right\}$$

where ν refers to to the instantaneous volatility of the numéraire, given in Equation (A3). The dynamics of the risk factors under $\mathbb{Q}^{T_m}$ can be obtained by an application of Girsanov's theorem Musiela and Rutkowski (2005). Denote by $\sigma_i(t) := (\sigma_{i1}(t), \ldots, \sigma_{id}(t))$ the ith row of the volatility matrix of $\mathbf{x}_t$ and let $\tilde{W}_i^{T_m}$ be Brownian motions under $\mathbb{Q}^{T_m}$; then,

$$dx_i(t) = -a_i(t) x_i(t) dt - \sigma_i(t) \cdot \nu(t, T_m) dt + \tilde{\sigma}_i(t) d\tilde{W}_i^{T_m}(t), \qquad i = 1, \ldots, d \tag{A4}$$

Let $\Theta_i(t, T_m) = \int_t^{T_m} \sigma_i(s) \cdot \nu(s, T_m) e^{-\int_s^{T_m} a_i(u) du} ds$; then, the SDE above solves to

$$x_i(T_m) = x_i(t) e^{-\int_t^{T_m} a_i(u) du} - \Theta_i(t, T_m) + \int_t^{T_m} \tilde{\sigma}_i(s) e^{-\int_s^{T_m} a_i(u) du} d\tilde{W}_i^{T_m}(s), \quad i = 1, \ldots, d \tag{A5}$$

It follows that, as a property of the Itô integral, the risk factors $(x_1(T_m), \ldots, x_n(T_m))$ as presented in Equation (A5), conditional on $\mathcal{F}_t$, have a multivariate normal distribution under $\mathbb{Q}^{T_m}$. Their mean vector and co-variance matrix are, respectively, given by

$$\boldsymbol{\mu} := \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} := \begin{pmatrix} \mathbb{E}^{T_m}[x_1(T_m)|\mathcal{F}_t] \\ \vdots \\ \mathbb{E}^{T_m}[x_d(T_m)|\mathcal{F}_t] \end{pmatrix} = \begin{pmatrix} x_1(t)e^{-\int_t^{T_m} a_1(u)du} - \Theta_1(t, T_m) \\ \vdots \\ x_d(t)e^{-\int_t^{T_m} a_d(u)du} - \Theta_d(t, T_m) \end{pmatrix}$$

$$\mathbf{C} := \begin{pmatrix} c_{11} & \ldots & c_{1d} \\ \vdots & \ddots & \vdots \\ c_{d1} & \ldots & c_{dd} \end{pmatrix} := \begin{pmatrix} \mathrm{Cov}[x_1(T_m), x_1(T_m)|\mathcal{F}_t] & \ldots & \mathrm{Cov}[x_1(T_m), x_d(T_m)|\mathcal{F}_t] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_d(T_m), x_1(T_m)|\mathcal{F}_t] & \ldots & \mathrm{Cov}[x_d(T_m), x_d(T_m)|\mathcal{F}_t] \end{pmatrix}$$

$$c_{ii} = \int_t^{T_m} \tilde{\sigma}_i^2(s) e^{-2\int_s^{T_m} a_i(u)du} ds \qquad \forall_{i\in\{1,\ldots,d\}}$$

$$c_{ij} = \int_t^{T_m} \rho(s)\tilde{\sigma}_i(s)\tilde{\sigma}_j(s) e^{-\int_s^{T_m}\left(a_i(u)+a_j(u)\right)du} ds \qquad \forall_{i\neq j}$$

As a result, it should be clear that the random variable $Y := \mathbf{w}_i^\top \log \mathbf{P}(T_m)$ is normally distributed with mean and variance given, respectively, by

$$\mu_Y = \mathbf{w}_i^\top \mathbf{A}(T_m) - \mathbf{w}_i^\top \mathbf{B}(T_m)\boldsymbol{\mu}$$

and variance

$$\sigma_Y^2 = \mathbf{w}_i^\top \mathbf{B}(T_m)\mathbf{C}\mathbf{B}(T_m)^\top \mathbf{w}_i$$

As a result, we can compute

$$\mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^{T_m} r(u)du}\varphi(\mathbf{w}_i^\top \log \mathbf{P}(T_m) - b_i)\middle|\mathcal{F}_t\right] = P(t, T_m)\mathbb{E}^{T_m}\left[\max(Y - b_i, 0)\middle|\mathcal{F}_t\right]$$

where the conditional expectation on the right-hand side can be expressed in closed form following a similar analysis as presented in Musiela and Rutkowski (2005). Let $d_i := \frac{\mu_Y - b_i}{\sigma_Y}$ and denote by $\xi \sim N(0,1)$ a standard normal random variable. Then, it follows that

$$\begin{aligned}
\mathbb{E}^{T_m}\left[\max(Y - b_i, 0)\middle|\mathcal{F}_t\right] &= \mathbb{E}^{T_m}\left[(Y - b_i)\mathbb{1}_{\{Y>b_i\}}\middle|\mathcal{F}_t\right] \\
&= \mathbb{E}^{T_m}\left[(Y - \mu_Y)\mathbb{1}_{\{Y>b_i\}}\right] + (\mu_Y - b_i)\mathbb{Q}^{T_m}\left[Y > b_i\middle|\mathcal{F}_t\right] \\
&= \sigma_Y \mathbb{E}^{T_m}\left[\frac{Y-\mu_Y}{\sigma_Y}\mathbb{1}_{\left\{\frac{Y-\mu_Y}{\sigma_Y} > -d_i\right\}}\middle|\mathcal{F}_t\right] \\
&\quad + (\mu_Y - b_i)\mathbb{Q}^{T_m}\left[\frac{Y-\mu_Y}{\sigma_Y} > -d_i\middle|\mathcal{F}_t\right] \\
&= \sigma_Y \mathbb{E}\left[-\xi\mathbb{1}_{\{-\xi<d_i\}}\right] + (\mu_Y - b_i)\mathbb{P}[\xi < d_i] \\
&= \sigma_Y\phi(d_i) + (\mu_Y - b_i)\Phi(d_i)
\end{aligned}$$

where ϕ denotes the standard normal density function and Φ the standard normal cumulative density function.

Appendix B. Pre-Processing the Regression-Data

A procedure that significantly improves the fitting performance of the neural networks is the normalization of the training data. The linear rescaling of the input to the optimizer is a common form of data pre-processing Bishop et al. (1995). In the case of a multivariate input, the variables might have typical values in different orders of magnitude, even though that does not reflect their relative influence on determining the outcome Bishop et al. (1995). Normalizing the scale avoids the impact of a certain input being prioritized over another

input. Also, the transfer of the final weights in G_{m+1} to the initialization of G_m is more effective as the target variables are of roughly the same size at each time-step. In the default situation, the average continuation values would change in magnitude and the risk factor distribution would grow with each passing of a monitor date.

Another argument for pre-processing the input is that large data values typically induce large weights. Large weights can lead to exploding network outputs in the feed-forward process Goodfellow et al. (2016). Furthermore, it can cause an unstable optimization of the network, as extreme gradients can be very sensitive to small perturbations in the data Goodfellow et al. (2016).

In practice, we propose the following rescaling of the data. Denote by

$$\hat{z}(T_m) := \left\{ \begin{pmatrix} z_1(T_m) \\ \vdots \\ z_d(T_m) \end{pmatrix}_1, \ldots, \begin{pmatrix} z_1(T_m) \\ \vdots \\ z_d(T_m) \end{pmatrix}_N \right\}, \quad \hat{V}(T_m) := \left\{ \tilde{V}\left(T_m; x^1_{T_m}\right), \ldots, \tilde{V}\left(T_m; x^N_{T_m}\right) \right\}$$

the training points for the in- and output of network G_m. Define the standard sample mean and standard deviations as

$$\mu_{z_i}(T_m) := \frac{1}{N}\sum_{n=1}^{N} z_i^n(T_m), \qquad \mu_V(T_m) := \frac{1}{N}\sum_{n=1}^{N} \tilde{V}\left(T_m; x^n_{T_m}\right)$$

$$\sigma_{z_i}(T_m) := \frac{1}{N-1}\sum_{n=1}^{N} (z_i^n(T_m) - \mu_{z_i})^2, \quad \sigma_V(T_m) := \frac{1}{N-1}\sum_{n=1}^{N} \left(\tilde{V}\left(T_m; x^n_{T_m}\right) - \mu_V(T_m)\right)^2$$

We then perform a simple element-wise linear transformation to obtain the scaled data $\hat{z}^\dagger$ and $\hat{V}^\dagger$ given by

$$\hat{z}_i^\dagger(T_m) := \frac{\hat{z}_i(T_m) - \mu_{z_i}(T_m)}{\sigma_{z_i}(T_m)}, \qquad \hat{V}^\dagger(T_m) := \frac{\hat{V}(T_m)}{\sigma_V(T_m)}$$

With the transformations above in mind, it is important to adjust the associated composition of the replicating portfolio accordingly. For the two network designs, this has the following implications:

The locally connected NN case: Consider the outcome of the i^{th} hidden node ν_i and denote the input of the network as $\mathbf{z}$. Then, $\nu_i = \varphi\left(w_i^k z_k + b_i\right)$, where k is the index of the only non-zero entry of $\mathbf{w}_i$, the i^{th} row of weight matrix $\mathbf{w}_1$. The transformation $\mathbf{z} \mapsto \frac{\mathbf{z}-\mu_\mathbf{z}}{\sigma_\mathbf{z}}$ implies that

$$\nu_i \mapsto \varphi\left(w_i^k \frac{z_k - \mu_{z_k}}{\sigma_{z_k}} + b_i\right) = \varphi\left(\frac{w_i^k}{\sigma_{z_k}} z_k + \left(b_i - \frac{w_i^k \mu_{z_k}}{\sigma_{z_k}}\right)\right)$$

As a consequence, in the analysis of Appendix A.1, the transformations $w_i^k \mapsto \frac{w_i^k}{\sigma_{z_k}}$ and $b_i \mapsto b_i - \frac{w_i^k \mu_{z_k}}{\sigma_{z_k}}$ should be taken into account. Additionally, the transformation $\mathbf{w}_2 \mapsto \sigma_V \mathbf{w}_2$ is required to account for the scaling of $\hat{V}$.

The fully connected NN case: Again, consider the outcome of the i^{th} hidden node ν_i. This time, the transformation $\mathbf{z} \mapsto \frac{\mathbf{z}-\mu_\mathbf{z}}{\sigma_\mathbf{z}}$ implies that

$$\nu_i \mapsto \varphi\left(\mathbf{w}_i^\top \frac{\mathbf{z} - \mu_\mathbf{z}}{\sigma_\mathbf{z}} + b_i\right) = \varphi\left(\sum_{j=1}^{d} \frac{w_i^j}{\sigma_{z_j}} z_j + \left(b_i - \sum_{j=1}^{d} \frac{w_i^j \mu_{z_j}}{\sigma_{z_i}}\right)\right)$$

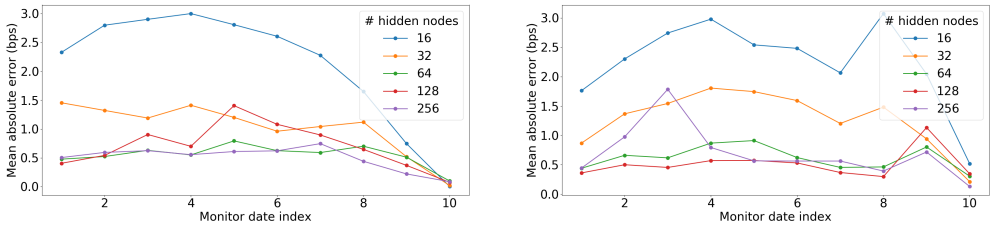
As a consequence, in the analysis of Appendix A.2, the transformations $\mathbf{w}_i \mapsto \left(\frac{w_i^1}{\sigma_{z_1}}, \ldots, \frac{w_i^d}{\sigma_{z_d}} \right)^\top$ and $b_i \mapsto b_i - \sum_{j=1}^d \frac{w_i^j \mu_{z_j}}{\sigma_{z_i}}$ should be taken into account. And, again, the transformation $\mathbf{w}_2 \mapsto \sigma_V \mathbf{w}_2$ is required to account for the scaling of $\hat{V}$.

Appendix C. Hyperparameter Selection

The accuracy of the neural network fitting procedure is dependent on the choice of several hyperparameters. For the numerical experiments reported in Section 6, the hyperparameters have been selected based on a convergence analysis. We focused on the following:

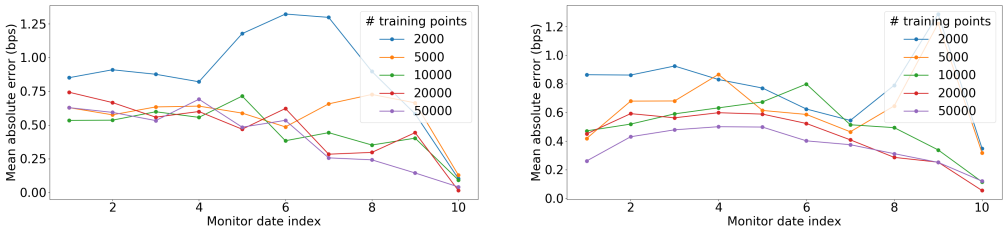
- Hidden node count: see Figure A1;
- Size training set: see Figure A2;
- Learning-rate: see Figure A3.

Several numerical experiments indicated that the batch size did not have a significant impact on the fitting accuracy and is therefore fixed at a default of 32. For the convergence analysis of the parameters listed above, we considered a $1Y \times 10Y$ receiver Bermudan swaption with a fixed rate of $K = 0.03$. Experiments were performed under the two-factor G2++ model using the model specifications depicted in Table 3. The figures show the mean absolute errors of the neural network fits per monitor date in basis points of the notional.



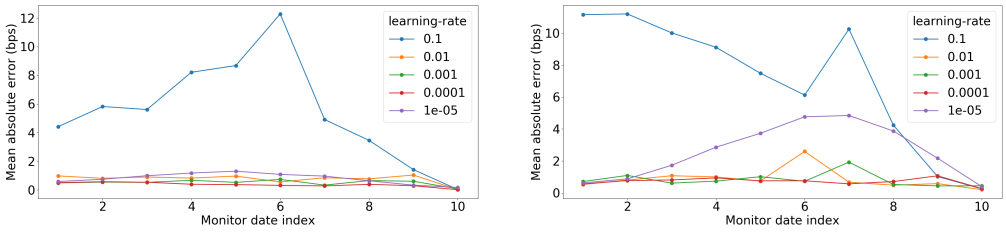
(**a**) Locally connected NN (**b**) Fully connected NN

Figure A1. Impact hidden node count: accuracy of the neural network fit per monitor date under a 2-factor model. # training points = 5000. Learning-rate = 0.0002.



(**a**) Locally connected NN (**b**) Fully connected NN

Figure A2. Impact size training set: accuracy of the neural network fit per monitor date under a 2-factor model. # hidden nodes = 64. Learning-rate = 0.0002.



(**a**) Locally connected NN (**b**) Fully connected NN

Figure A3. Impact learning-rate: accuracy of the neural network fit per monitor date under a 2-factor model. # hidden nodes = 64. # training-points = 10,000.

Appendix D. Proof of Theorem 1

Proof. We prove by induction on m. At the last exercise date of the Bermudan, i.e., $t = T_{M-1}$, we have $V(T_{M-1}; x) = \tilde{V}(T_{M-1}; x) := \max\{h_{M-1}(x), 0\}$, representing the final pay-off of the contract, which at T_{M-1} is exactly known. Hence, it should be obvious that

$$\sup_{x \in \mathcal{I}_d} B^{-1}(T_{M-1}) \left| V(T_{M-1}; x) - \tilde{V}(T_{M-1}; x) \right| = 0$$

For the inductive step, assume that, for some $T_{m+1} \in \mathcal{T}_f$, an approximation $\tilde{V}(T_{m+1})$ of the price is given, satisfying

$$\sup_{x \in \mathcal{I}_d} B^{-1}(T_{m+1}) \left| V(T_{m+1}; x) - \tilde{V}(T_{m+1}; x) \right| < k\varepsilon$$

We will show that it follows that, for all $t \in [T_m, T_{m+1})$,

$$\sup_{x \in \mathcal{I}_d} B^{-1}(t) \left| V(t; x) - \tilde{V}(t; x) \right| < (k+1)\varepsilon$$

First, consider the case $t \in (T_m, T_{m+1})$. It follows that

$$\begin{aligned}
\sup_{x \in \mathcal{I}_d} \left| \frac{V(t;x) - \tilde{V}(t;x)}{B(t)} \right| &= \sup_{x \in \mathcal{I}_d} \left| \frac{C_m(t;x) - \tilde{C}_m(t;x)}{B(t)} \right| \\
&= \sup_{x \in \mathcal{I}_d} \left| \mathbb{E}^{\mathbb{Q}} \left[\frac{V(T_{m+1})}{B(T_{m+1})} \middle| \mathbf{x}_t = x \right] - \mathbb{E}^{\mathbb{Q}} \left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})} \middle| \mathbf{x}_t = x \right] \right| \\
&\leq \sup_{x \in \mathcal{I}_d} \mathbb{E}^{\mathbb{Q}} \left[B^{-1}(T_{m+1}) \left| V(T_{m+1}) - G_{m+1}(z_{m+1}) \right| \middle| \mathbf{x}_t = x \right] \\
&= \sup_{x \in \mathcal{I}_d} \mathbb{E}^{\mathbb{Q}} \Big[B^{-1}(T_{m+1}) \big| V(T_{m+1}) - \tilde{V}(T_{m+1}) \\
&\qquad + \tilde{V}(T_{m+1}) - G_{m+1}(z_{m+1}) \big| \Big| \mathbf{x}_t = x \Big] \\
&\leq \sup_{x \in \mathcal{I}_d} \Big(\mathbb{E}^{\mathbb{Q}} \left[B^{-1}(T_{m+1}) \left| V(T_{m+1}) - \tilde{V}(T_{m+1}) \right| \middle| \mathbf{x}_t = x \right] \\
&\qquad + \mathbb{E}^{\mathbb{Q}} \left[B^{-1}(T_{m+1}) \left| \tilde{V}(T_{m+1}) - G_{m+1}(z_{m+1}) \right| \middle| \mathbf{x}_t = x \right] \Big)
\end{aligned}$$

In the last expression above, the first term is bounded due to the induction hypothesis, i.e., $B^{-1}(T_{m+1}) \left| V(T_{m+1}) - \tilde{V}(T_{m+1}) \right| < k\varepsilon$. The second term is bounded by assumption, i.e., there exists a network $G_{m+1}(\cdot)$ such that $B^{-1}(T_{m+1}) \left| \tilde{V}(T_{m+1}) - G_{m+1}(z_{m+1}) \right| < \varepsilon$. We hence conclude that

$$\sup_{x \in \mathcal{I}_d} B^{-1}(t) \left| V(t; x) - \tilde{V}(t; x) \right| < (k+1)\varepsilon, \qquad \forall_{t \in (T_m, T_{m+1})}$$

If, on the other hand, $t = T_m$, we have that

$$\sup_{x\in\mathcal{I}_d}\left|\frac{V(t;x)-\tilde{V}(t;x)}{B(t)}\right| = \sup_{x\in\mathcal{I}_d}\left|\frac{\max\{C_m(t;x),\, h_m(x)\}-\max\{\tilde{C}_m(t;x),\, h_m(x)\}}{B(t)}\right|$$

Denoting $H(x) := B^{-1}(t)\big|\max\{C_m(t;x),\, h_m(x)\}-\max\{\tilde{C}_m(t;x),\, h_m(x)\}\big|$ in the expression above, we can distinguish four cases for each $x \in \mathcal{I}_d$, which are

- $C_m(t;x), \tilde{C}_m(t;x) > h_m(x)$, then $H(x) = B^{-1}(t)\big|C_m(t;x)-\tilde{C}_m(t;x)\big| < (k+1)\varepsilon$;
- $C_m(t;x), \tilde{C}_m(t;x) < h_m(x)$, then $H(x) = B^{-1}(t)|h_m(x)-h_m(x)| = 0 < (k+1)\varepsilon$;
- $C_m(t;x) < h_m(x) < \tilde{C}_m(t;x)$, then $H(x) = B^{-1}(t)\big|h_m(x)-\tilde{C}_m(t;x)\big|$ $< B^{-1}(t)\big|C_m(t;x)-\tilde{C}_m(t;x)\big| < (k+1)\varepsilon$;
- $\tilde{C}_m(t;x) < h_m(x) < C_m(t;x)$, then $H(x) = B^{-1}(t)|C_m(t;x)-h_m(x)|$ $< B^{-1}(t)\big|C_m(t;x)-\tilde{C}_m(t;x)\big| < (k+1)\varepsilon$.

From all the cases, we can induce that

$$\sup_{x\in\mathcal{I}_d} B^{-1}(t)\big|V(t;x)-\tilde{V}(t;x)\big| \leq (k+1)\varepsilon$$

We conclude that, by induction on $m = M-1,\ldots,0$,

$$\sup_{x\in\mathcal{I}_d} B^{-1}(t)\big|V(t;x)-\tilde{V}(t;x)\big| < M\varepsilon$$

for all $t \in [0, T_{M-1}]$. □

Appendix E. Proof of Theorem 2

Proof. First, we fix some notation.

- Let $V_m := V(T_m)$ denote the true price of the Bermudan swaption at T_m conditioned on the fact that it is not yet exercised.
- Let $\tilde{C}_m := B(T_m)\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right]$ denote the estimator of the continuation value at T_m.
- Let $\tilde{V}_m := \max\{\tilde{C}_m, h_m(\mathbf{x}_{T_m})\}$ denote the estimator of V_m.
- Let $G_m := G_m(z_m)$ denote the neural network approximation of $\tilde{V}_m$.
- Let $B_m := B(T_m)$ denote the numéraire at T_m.
- Let $h_m := h_m(\mathbf{x}_{T_m})$.

Let $T_m \in \{T_0,\ldots,T_{M-1}\}$. We will prove the theorem by induction on m. For the base case, note that at time zero we have

$$\big|V(0)-\tilde{V}(0)\big| = \left|\mathbb{E}^{\mathbb{Q}}\left[\frac{V_0}{B_0}\middle|\mathcal{F}_0\right]-\mathbb{E}^{\mathbb{Q}}\left[\frac{G_0}{B_0}\middle|\mathcal{F}_0\right]\right| \leq \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_0-G_0}{B_0}\right|\middle|\mathcal{F}_0\right] \tag{A6}$$

which is induced by Jensen's inequality. For the inductive step, assume that, for some $m \in \{0,\ldots,M-1\}$, we have that

$$\big|V(0)-\tilde{V}(0)\big| < \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_m-G_m}{B_m}\right|\middle|\mathcal{F}_0\right] + m\cdot\varepsilon \tag{A7}$$

The expectation in (A7) can be rewritten using the triangular inequality

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_m-G_m}{B_m}\right|\middle|\mathcal{F}_0\right] &= \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_m-\tilde{V}_m+\tilde{V}_m-G_m}{B_m}\right|\middle|\mathcal{F}_0\right]\\ &\leq \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_m-\tilde{V}_m}{B_m}\right|\middle|\mathcal{F}_0\right]+\mathbb{E}^{\mathbb{Q}}\left[\left|\frac{\tilde{V}_m-G_m}{B_m}\right|\middle|\mathcal{F}_0\right]\end{aligned} \tag{A8}$$

The second term in (A8) is, by assumption, bounded by ε. Note that the first term in (A8) can be bounded as

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_m-\tilde{V}_m}{B_m}\right|\middle|\mathcal{F}_0\right] &= \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{\max\{C_m,h_m\}-\max\{\tilde{C}_m,h_m\}}{B_m}\right|\middle|\mathcal{F}_0\right] \\
&\leq \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{C_m-\tilde{C}_m}{B_m}\right|\middle|\mathcal{F}_0\right] \\
&= \mathbb{E}^{\mathbb{Q}}\left[\left|\mathbb{E}^{\mathbb{Q}}\left[\frac{V_{m+1}}{B_{m+1}}\middle|\mathcal{F}_{T_m}\right]-\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}}{B_{m+1}}\middle|\mathcal{F}_{T_m}\right]\right|\middle|\mathcal{F}_0\right] \\
&\leq \mathbb{E}^{\mathbb{Q}}\left[\mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_{m+1}-G_{m+1}}{B_{m+1}}\right|\middle|\mathcal{F}_{T_m}\right]\middle|\mathcal{F}_0\right] \\
&= \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_{m+1}-G_{m+1}}{B_{m+1}}\right|\middle|\mathcal{F}_0\right]
\end{aligned}$$

It follows that

$$\left|V(0)-\tilde{V}(0)\right| < \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_{m+1}-G_{m+1}}{B_{m+1}}\right|\middle|\mathcal{F}_0\right]+(m+1)\cdot\varepsilon$$

For the final step, note that if $m=M-1$, we have

$$\mathbb{E}^{\mathbb{Q}}\left[\left|\frac{V_m-G_m}{B_m}\right|\middle|\mathcal{F}_0\right] = \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{\max\{h_{M-1},0\}-G_{M-1}}{B_{M-1}}\right|\middle|\mathcal{F}_0\right] < \varepsilon$$

We conclude by induction on m that $\left|V(0)-\tilde{V}(0)\right| < M\varepsilon$ □

Appendix F. Proof of Theorem 3

Proof. We consider the following three events: $\{\tau=\tilde{\tau}\}$, $\{\tau<\tilde{\tau}\}$, and $\{\tau>\tilde{\tau}\}$. Note that

$$\begin{aligned}
V(0)-L(0) &= \mathbb{E}^{\mathbb{Q}}\left[\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}-\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\middle|\mathcal{F}_0\right] \\
&= \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}-\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{\{\tau=\tilde{\tau}\}}\middle|\mathcal{F}_0\right]+\mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}-\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{\{\tau<\tilde{\tau}\}}\middle|\mathcal{F}_0\right] \\
&\quad+\mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}-\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{\{\tau>\tilde{\tau}\}}\middle|\mathcal{F}_0\right] \\
&= E_1+E_2+E_3
\end{aligned}$$

We will bound the three terms above one by one.
Bounding $\mathbf{E_1}$:Starting with the event $\{\tau=\tilde{\tau}\}$, we observe that we can write

$$E_1 = \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}-\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}\right)\mathbb{1}_{\{\tau=\tilde{\tau}\}}\middle|\mathcal{F}_0\right]=0$$

Bounding $\mathbf{E_2}$:We continue with the event $\{\tau<\tilde{\tau}\}$. For this, we will introduce two types of sub-events: A_m: $=\{\tau=T_m\wedge\tilde{\tau}>T_m\}$ and $B_m:=\{\tau\leq T_m\wedge\tilde{\tau}>T_m\}$, where $\wedge$ denotes the logical AND operator. Also, we define the difference process e_m: $=\frac{\tilde{V}(T_m)}{B(T_m)}-\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}$. It should be clear that $\mathbb{1}_{\{\tau<\tilde{\tau}\}}=\sum_{m=0}^{M-1}\mathbb{1}_{A_m}$. Therefore, it holds that

$$E_2=\sum_{m=0}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)}-\frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{A_m}\middle|\mathcal{F}_0\right]\leq\sum_{m=0}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{A_m}\middle|\mathcal{F}_0\right]$$

where the inequality follows from the fact that the direct estimator has the property $\tilde{V}(T_m) = \max\{\tilde{C}_m, h_m\} \geq h_m$. Now, we will show by induction that $E_2 < (M-1)\varepsilon$. First, observe that $A_0 \equiv B_0$. Second, note that, for any $m \in \{0, \ldots, M-1\}$, we have that

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[e_m \mathbb{1}_{B_m} \middle| \mathcal{F}_0\right] &= \mathbb{E}^{\mathbb{Q}}\left[\left(\mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle|\mathcal{F}_{T_m}\right] - \frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{B_m}\middle|\mathcal{F}_0\right] \\
&= \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})} - \frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{B_m}\middle|\mathcal{F}_0\right] \\
&\leq \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})} - \frac{\tilde{V}(T_{m+1})}{B(T_{m+1})}\right|\mathbb{1}_{B_m}\middle|\mathcal{F}_0\right] + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_m}\middle|\mathcal{F}_0\right]
\end{aligned} \tag{A9}$$

The first equality follows from the fact that $\tilde{V}(T_m) = \tilde{C}_m$ in the event $\tilde{\tau} > T_m$. The second equality follows from the tower rule in combination with the fact that $\mathbb{1}_{B_m}$ is $\mathcal{F}_{T_m}-$measurable. The final inequality follows from an application of the triangle inequality. The first term in (A9) is, by assumption, bounded by ε. The second term in (A9) can be rewritten by observing that $\mathbb{1}_{B_m} := \mathbb{1}_{B_m^1} + \mathbb{1}_{B_m^2} := \mathbb{1}_{\{\tau \leq T_m \wedge \tilde{\tau} = T_{m+1}\}} + \mathbb{1}_{\{\tau \leq T_m \wedge \tilde{\tau} > T_{m+1}\}}$. We have that

$$\mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_m^1}\middle|\mathcal{F}_0\right] = \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_{m+1}(\mathbf{x}_{T_{m+1}})}{B(T_{m+1})} - \frac{h_{m+1}(\mathbf{x}_{T_{m+1}})}{B(T_{m+1})}\right)\mathbb{1}_{B_m^1}\middle|\mathcal{F}_0\right] = 0$$

Furthermore, we have that $\mathbb{1}_{B_m^2} + \mathbb{1}_{A_{m+1}} = \mathbb{1}_{B_{m+1}}$. Therefore we can infer that

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{B_m}\middle|\mathcal{F}_0\right] + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{A_{m+1}}\middle|\mathcal{F}_0\right] &< \varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_m^2}\middle|\mathcal{F}_0\right] + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{A_{m+1}}\middle|\mathcal{F}_0\right] \\
&= \varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_{m+1}}\middle|\mathcal{F}_0\right]
\end{aligned}$$

Together with the fact that $A_0 \equiv B_0$, we conclude by induction on m that

$$\begin{aligned}
E_2 &\leq \mathbb{E}^{\mathbb{Q}}\left[e_0\mathbb{1}_{B_0}\middle|\mathcal{F}_0\right] + \sum_{m=1}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{A_m}\middle|\mathcal{F}_0\right] \\
&< \varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_1\mathbb{1}_{B_1}\middle|\mathcal{F}_0\right] + \sum_{m=2}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{A_m}\middle|\mathcal{F}_0\right] \\
&\vdots \\
&< (M-1)\varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_{M-1}\mathbb{1}_{B_{M-1}}\middle|\mathcal{F}_0\right] = (M-1)\varepsilon
\end{aligned}$$

Bounding $\mathbf{E_3}$:We finalize the proof by considering the third event $\{\tau > \tilde{\tau}\}$. In a similar fashion as before, we introduce two types of sub-events: A_m: $= \{\tilde{\tau} = T_m \wedge \tau > T_m\}$ and B_m: $= \{\tilde{\tau} \leq T_m \wedge \tau > T_m\}$. Also, again define a difference process, this time given by e_m: $= \frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)} - \frac{\tilde{V}(T_m)}{B(T_m)}$. It should be clear that $\mathbb{1}_{\{\tau > \tilde{\tau}\}} = \sum_{m=0}^{M-1}\mathbb{1}_{A_m}$. Therefore, it holds that

$$E_3 = \sum_{m=0}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)} - \frac{h_{\tilde{\tau}}(\mathbf{x}_{\tilde{\tau}})}{B(\tilde{\tau})}\right)\mathbb{1}_{A_m}\middle|\mathcal{F}_0\right] = \sum_{m=0}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{A_m}\middle|\mathcal{F}_0\right]$$

where the second equality follows from the fact that the direct estimator has the property $\tilde{V}(\tilde{\tau}) = h_{\tilde{\tau}}$. Now, we will show by induction that $E_3 < (M-1)\varepsilon$. Note that, for any $m \in \{0, \ldots, M-1\}$, we have that

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[e_m \mathbb{1}_{B_m} \middle| \mathcal{F}_0\right] &\leq \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)} - \mathbb{E}^{\mathbb{Q}}\left[\frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\middle| \mathcal{F}_{T_m}\right]\right)\mathbb{1}_{B_m}\middle| \mathcal{F}_0\right] \\
&= \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_\tau(\mathbf{x}_\tau)}{B(\tau)} - \frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\right)\mathbb{1}_{B_m}\middle| \mathcal{F}_0\right] \\
&\leq \mathbb{E}^{\mathbb{Q}}\left[\left|\frac{\tilde{V}(T_{m+1})}{B(T_{m+1})} - \frac{G_{m+1}(z_{m+1})}{B(T_{m+1})}\right|\mathbb{1}_{B_m}\middle| \mathcal{F}_0\right] + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_m}\middle| \mathcal{F}_0\right]
\end{aligned} \tag{A10}$$

The first inequality follows from the fact that $\tilde{V}(T_m) = \max\{\tilde{C}_m, h_m\} \geq \tilde{C}_m$. The subsequent equality follows from the tower rule in combination with the fact that $\mathbb{1}_{B_m}$ is $\mathcal{F}_{T_m}-$measurable. The final inequality follows from an application of the triangle inequality. The first term in (A10) is, by assumption, bounded by ε. The second term in (A10) can be rewritten by observing that $\mathbb{1}_{B_m} := \mathbb{1}_{B_m^1} + \mathbb{1}_{B_m^2} := \mathbb{1}_{\{\tilde{\tau} \leq T_m \wedge \tau = T_{m+1}\}} + \mathbb{1}_{\{\tilde{\tau} \leq T_m \wedge \tau > T_{m+1}\}}$. We have that

$$\mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_m^1}\middle| \mathcal{F}_0\right] = \mathbb{E}^{\mathbb{Q}}\left[\left(\frac{h_{m+1}(\mathbf{x}_{T_{m+1}})}{B(T_{m+1})} - \frac{\tilde{V}(T_{m+1})}{B(T_{m+1})}\right)\mathbb{1}_{B_m^1}\middle| \mathcal{F}_0\right] \leq 0$$

where the inequality follows from the fact that $\tilde{V}(T_{m+1}) = \max\{\tilde{C}_{m+1}, h_{m+1}\} \geq h_{m+1}$. Furthermore, we have that $\mathbb{1}_{B_m^2} + \mathbb{1}_{A_{m+1}} = \mathbb{1}_{B_{m+1}}$. Therefore, we can once again infer that

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[e_m \mathbb{1}_{B_m}\middle| \mathcal{F}_0\right] + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{A_{m+1}}\middle| \mathcal{F}_0\right] &< \varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_m^2}\middle| \mathcal{F}_0\right] + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{A_{m+1}}\middle| \mathcal{F}_0\right] \\
&= \varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_{m+1}\mathbb{1}_{B_{m+1}}\middle| \mathcal{F}_0\right]
\end{aligned}$$

Together with the fact that $A_0 \equiv B_0$, we again conclude by induction on m that

$$\begin{aligned}
E_3 &\leq \mathbb{E}^{\mathbb{Q}}\left[e_0\mathbb{1}_{B_0}\middle| \mathcal{F}_0\right] + \sum_{m=1}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{A_m}\middle| \mathcal{F}_0\right] \\
&< \varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_1\mathbb{1}_{B_1}\middle| \mathcal{F}_0\right] + \sum_{m=2}^{M-1}\mathbb{E}^{\mathbb{Q}}\left[e_m\mathbb{1}_{A_m}\middle| \mathcal{F}_0\right] \\
&\vdots \\
&< (M-1)\varepsilon + \mathbb{E}^{\mathbb{Q}}\left[e_{M-1}\mathbb{1}_{B_{M-1}}\middle| \mathcal{F}_0\right] = (M-1)\varepsilon
\end{aligned}$$

Conclusion: We hence find that

$$V(0) - L(0) = E_1 + E_2 + E_3 < 0 + (M-1)\varepsilon + (M-1)\varepsilon = 2(M-1)\varepsilon$$

□

Appendix G. Proof of Theorem 4

Proof. The discounted true price process is a supermartingale under $\mathbb{Q}$. Therefore, we have that $\frac{V(t)}{B(t)} = Y_t + Z_t$ for a martingale Y_t and a predictable process Z_t, which starts at zero (i.e., $Z_0 = 0$) and is strictly decreasing. Define a difference process on $\mathcal{T}$, given

by $e_{T_m} = \frac{V(T_m) - G_m(z_m)}{B(T_m)}$. We can rewrite martingale M_t as defined in (13) in terms of e_t as follows:

$$\begin{aligned} M_{T_m} &= \frac{G_0(z_0)}{B(T_0)} + \sum_{j=1}^{m} \left(\frac{G_j(z_j)}{B(T_j)} - \mathbb{E}^{\mathbb{Q}} \left[\frac{G_j(z_j)}{B(T_j)} \middle| \mathcal{F}_{T_{j-1}} \right] \right) \\ &= Y_{T_m} - e_{T_0} - \sum_{j=1}^{m} \left(e_{T_j} - \mathbb{E}^{\mathbb{Q}} \left[e_{T_j} \middle| \mathcal{F}_{T_{j-1}} \right] \right) \end{aligned}$$

Substituting the expression for M_t into the definition of $U(0)$ yields

$$\begin{aligned} U(0) &= M_0 + \mathbb{E}^{\mathbb{Q}} \left[\max_{T_m \in \mathcal{T}_f} \left\{ \frac{h_m(\mathbf{x}_{T_m})}{B(T_m)} - M_{T_m} \right\} \middle| \mathcal{F}_0 \right] \\ &= \mathbb{E}^{\mathbb{Q}} \left[\frac{G_0(z_0)}{B(T_0)} \middle| \mathcal{F}_0 \right] + \mathbb{E}^{\mathbb{Q}} \left[\max_{m \in \{0,\ldots,M-1\}} \left\{ \frac{h_m(\mathbf{x}_{T_m})}{B(T_m)} - Y_{T_m} + e_{T_0} \right. \right. \\ &\qquad \left. \left. + \sum_{j=1}^{m} \left(e_{T_j} - \mathbb{E}^{\mathbb{Q}} \left[e_{T_j} \middle| \mathcal{F}_{T_{j-1}} \right] \right) \right\} \middle| \mathcal{F}_0 \right] \\ &\leq \mathbb{E}^{\mathbb{Q}} \left[\frac{V(T_0)}{B(T_0)} \middle| \mathcal{F}_0 \right] + \mathbb{E}^{\mathbb{Q}} \left[\max_{m \in \{0,\ldots,M-1\}} \left\{ \sum_{j=1}^{m} \left(e_{T_j} - \mathbb{E}^{\mathbb{Q}} \left[e_{T_j} \middle| \mathcal{F}_{T_{j-1}} \right] \right) \right\} \middle| \mathcal{F}_0 \right] \end{aligned}$$

The last step follows by merging $\mathbb{E}^{\mathbb{Q}}\left[e_{T_0} \middle| \mathcal{F}_0\right]$ with M_0 and by noting that $\frac{h_m(\mathbf{x}_{T_m})}{B(T_m)} - Y_{T_m} \leq \frac{V(T_m)}{B(T_m)} - Y_{T_m} = Z_{T_m} \leq 0$. The remaining inequality is not easy to bound Andersen and Broadie (2004). However, by taking the absolute values of the difference process, we can obtain a loose bound as follows:

$$\begin{aligned} U(0) &\leq V(0) + \mathbb{E}^{\mathbb{Q}} \left[\max_{m \in \{0,\ldots,M-1\}} \left\{ \sum_{j=1}^{m} \left| e_{T_j} \right| + \sum_{j=1}^{m} \left| \mathbb{E}^{\mathbb{Q}} \left[e_{T_j} \middle| \mathcal{F}_{T_{j-1}} \right] \right| \right\} \middle| \mathcal{F}_0 \right] \\ &\leq V(0) + \mathbb{E}^{\mathbb{Q}} \left[\sum_{j=1}^{M-1} \left| e_{T_j} \right| + \sum_{j=1}^{M-1} \left| \mathbb{E}^{\mathbb{Q}} \left[e_{T_j} \middle| \mathcal{F}_{T_{j-1}} \right] \right| \middle| \mathcal{F}_0 \right] \\ &\leq V(0) + 2 \sum_{j=1}^{M-1} \mathbb{E}^{\mathbb{Q}} \left[\left| e_{T_j} \right| \middle| \mathcal{F}_0 \right] \end{aligned}$$

Note that, as a consequence of Theorem 2, we have that $\mathbb{E}^{\mathbb{Q}}\left[|e_{T_m}| \middle| \mathcal{F}_0 \right] < (M-m)\varepsilon$. It follows that

$$|U(0) - V(0)| < 2 \sum_{m=1}^{M-1} (M-m)\varepsilon = M(M-1)\varepsilon$$

This concludes the proof. □

References

Ametrano, Ferdinando, and Luigi Ballabio. 2003. Quantlib—A Free/Open-Source Library for Quantitative Finance. Available online: https://github.com/lballabio/QuantLib (accessed on 1 March 2020).

Andersen, Leif, and Mark Broadie. 2004. Primal-dual simulation algorithm for pricing multidimensional american options. *Management Science* 50: 1222–34. [CrossRef]

Andersen, Leif B. G., and Vladimir V. Piterbarg. 2010a. *Interest Rate Modeling, Volume I: Foundations and Vanilla Models*. London: Atlantic Financial Press.

Andersen, Leif B. G., and Vladimir V. Piterbarg. 2010b. *Interest Rate Modeling, Volume II: Term Structure Models*. London: Atlantic Financial Press.

Andersson, Kristoffer, and Cornelis W. Oosterlee. 2021. A deep learning approach for computations of exposure profiles for high-dimensional bermudan options. *Applied Mathematics and Computation* 408: 126332. [CrossRef]

Becker, Sebastian, Patrick Cheridito, and Arnulf Jentzen. 2019. Deep optimal stopping. *Journal of Machine Learning Research* 20: 74.

Becker, Sebastian, Patrick Cheridito, and Arnulf Jentzen. 2020. Pricing and hedging american-style options with deep learning. *Journal of Risk and Financial Management* 13: 158. [CrossRef]
Beyna, Ingo. 2013. *Interest Rate Derivatives: Valuation, Calibration and Sensitivity Analysis*. Berlin/Heidelberg: Springer Science & Business Media.
Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.
Breeden, Douglas T., and Robert H. Litzenberger. 1978. Prices of state-contingent claims implicit in option prices. *Journal of Business* 51: 621–51. [CrossRef]
Brigo, Damiano, and Fabio Mercurio. 2006. *Interest Rate Models-Theory and Practice: With Smile, Inflation and Credit*. Berlin/Heidelberg: Springer, vol. 2.
Carr, Peter, and Jonathan Bowie. 1994. Static simplicity. *Risk* 7: 45–50.
Carr, Peter, Katrina Ellis, and Vishal Gupta. 1999. Static hedging of exotic options. In *Quantitative Analysis in Financial Markets: Collected Papers of the New York University Mathematical Finance Seminar*. Singapore: World Scientific, pp. 152–76.
Carr, Peter, and Liuren Wu. 2014. Static hedging of standard options. *Journal of Financial Econometrics* 12: 3–46. [CrossRef]
Carriere, Jacques F. 1996. Valuation of the early-exercise price for options using simulations and nonparametric regression. *Insurance: Mathematics and Economics* 19: 19–30. [CrossRef]
Chollet, François. 2015. Keras. Available online: https://keras.io (accessed on 1 May 2020).
Chung, San-Lin, and Pai-Ta Shih. 2009. Static hedging and pricing american options. *Journal of Banking & Finance* 33: 2140–49.
Dai, Qiang, and Kenneth J. Singleton. 2000. Specification analysis of affine term structure models. *The Journal of Finance* 55: 1943–78. [CrossRef]
Derman, Emanuel, Deniz Ergener, and Iraj Kani. 1995. Static options replication. *Journal of Derivatives* 2. [CrossRef]
Duffie, Darrell, and Rui Kan. 1996. A yield-factor model of interest rates. *Mathematical Finance* 6: 379–406. [CrossRef]
Ferguson, Ryan, and Andrew Green. 2018. Deeply learning derivatives. *arXiv* arXiv:1809.02233.
Filipovic, Damir. 2009. *Term-Structure Models. A Graduate Course.* Berlin/Heidelberg: Springer.
Geman, Helyette, Nicole El Karoui, and Jean-Charles Rochet. 1995. Changes of numeraire, changes of probability measure and option pricing. *Journal of Applied probability* 32: 443–58. [CrossRef]
Glasserman, Paul. 2013. *Monte Carlo Methods in Financial Engineering*. Berlin/Heidelberg: Springer Science & Business Media, vol. 53.
Glasserman, Paul, and Bin Yu. 2004. Simulation for american options: Regression now or regression later? In *Monte Carlo and Quasi-Monte Carlo Methods 2002*. Berlin/Heidelberg: Springer, pp. 213–26.
Gnoatto, Alessandro, Christoph Reisinger, and Athena Picarelli. 2023. Deep xva solver—A neural network based counterparty credit risk management framework. *SIAM Journal on Financial Mathematics* 14: 314–352. [CrossRef]
Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep Learning*. Cambridge: MIT Press Cambridge, vol. 1.
Gregory, Jon. 2015. *The xVA Challenge: Counterparty Credit Risk, Funding, Collateral and Capital*. Hoboken: John Wiley & Sons.
Hagan, Patrick S. 2005. Convexity conundrums: Pricing cms swaps, caps, and floors. *The Best of Wilmott* 305.. [CrossRef]
Harrison, J. Michael, and Stanley R. Pliska. 1981. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications* 11: 215–60. [CrossRef]
Haugh, Martin B., and Leonid Kogan. 2004. Pricing american options: A duality approach. *Operations Research* 52: 258–70. [CrossRef]
Henrard, Marc. 2003. Explicit bond option formula in heath–jarrow–morton one factor model. *International Journal of Theoretical and Applied Finance* 6: 57–72. [CrossRef]
Henry-Labordere, Pierre. 2017. Deep Primal-Dual Algorithm for Bsdes: Applications of Machine Learning to CVA and IM. Available online: https://ssrn.com/abstract=3071506 (accessed on 1 October 2020).
Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* 2: 359–66. [CrossRef]
Hutchinson, James M., Andrew W. Lo, and Tomaso Poggio. 1994. A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance* 49: 851–89. [CrossRef]
Jain, Shashi, and Cornelis W. Oosterlee. 2015. The stochastic grid bundling method: Efficient pricing of bermudan options and their greeks. *Applied Mathematics and Computation* 269: 412–31. [CrossRef]
Jamshidian, Farshid. 1989. An exact bond option formula. *The Journal of Finance* 44: 205–209. [CrossRef]
Kingma, Diederik P., and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv* arXiv:1412.6980.
Kloeden, Peter E., and Eckhard Platen. 2013. *Numerical Solution of Stochastic Differential Equations*. Berlin/Heidelberg: Springer Science & Business Media, vol. 23.
Kohler, Michael, Adam Krzyżak, and Nebojsa Todorovic. 2010. Pricing of high-dimensional american options by neural networks. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 20: 383–410. [CrossRef]
Lapeyre, Bernard, and Jérôme Lelong. 2019. Neural network regression for bermudan option pricing. *arXiv* arXiv:1907.06474.
Lokeshwar, Vikranth, Vikram Bharadwaj, and Shashi Jain. 2022. Explainable neural network for pricing and universal static hedging of contingent claims. *Applied Mathematics and Computation* 417: 126775. [CrossRef]
Longstaff, Francis A., and Eduardo S. Schwartz. 2001. Valuing american options by simulation: A simple least-squares approach. *The Review of Financial Studies* 14: 113–47. [CrossRef]
Musiela, Marek, and Marek Rutkowski. 2005. *Martingale Methods in Financial Modelling*. Berlin/Heidelberg: Springer Finance.

Oosterlee, Kees, Qian Feng, Shashi Jain, Patrik Karlsson, and Drona Kandhai. 2016. Efficient computation of exposure profiles on real-world and risk-neutral scenarios for bermudan swaptions. *Journal of Computational Finance* 20: 139–72. [CrossRef]

Pelsser, Antoon. 2003. Pricing and hedging guaranteed annuity options via static option replication. *Insurance: Mathematics and Economics* 33: 283–96. [CrossRef]

Rogers, Leonard C. G. 2002. Monte carlo valuation of american options. *Mathematical Finance* 12: 271–86. [CrossRef]

Ruf, Johannes, and Weiguan Wang. 2020. Neural networks for option pricing and hedging: A literature review. *Journal of Computational Finance*. *in press*. [CrossRef]

Shreve, Steven E. 2004. *Stochastic calculus for finance II: Continuous-time models*. Berlin/Heidelberg: Springer Science & Business Media, vol. 11.

Wang, Haojie, Han Chen, Agus Sudjianto, Richard Liu, and Qi Shen. 2018. Deep learning-based bsde solver for libor market model with application to bermudan swaption pricing and hedging. *arXiv* arXiv:1807.06622.

Xiu, Dongbin. 2010. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton: Princeton University Press.

Zhu, Steven H., and Michael Pykhtin. 2007. A guide to modeling counterparty credit risk. *GARP Risk Review, July/August*. Available online: https://ssrn.com/abstract=1032522 (accessed on 10 November 2020).

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Risks Editorial Office
E-mail: risks@mdpi.com
www.mdpi.com/journal/risks

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-2479-3