DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Modelling under Uncertainty“

verfasst von / submitted by

Annemarie Martha Grass, BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin der Naturwissenschaften (Dr. rer. nat)

Wien, 2024 / Vienna 2024

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796 605 405
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Mathematik
Betreut von / Supervisor:	Univ. Prof. Dr. Mathias Beiglböck

Summary

Mathematical models are omnipresent and indispensable in our modern world, influencing a plethora of aspects of life one way or another. However, even a well-constructed model may eventually fall short of capturing reality with desired precision, giving rise to the issue of *model uncertainty* - the central subject under consideration of this thesis.

The first part of this thesis is concerned with model uncertainty in the pricing of financial derivatives. We investigate upper and lower *worst case bounds* on derivative prices considering only minimal assumptions on the possible underlying models. This, in turn, is connected to the identification of specific solutions to the *Skorokhod embedding problem* (SEP). In its simplest form, the SEP is to represent a given probability measure μ as a Brownian motion $(B_t)_{t\geq 0}$ stopped at a certain stopping time τ. We consider several solutions to the SEP which play and important role in the robust modelling framework, namely Root, Rost, and Perkins solutions.

For the first two, Root and Rost solutions, we provide a simple probabilistic explanation for a representation of these stopping times via solutions to an associated optimal stopping problem. We are then able to extend these results to a multi-marginal setting, that is, embedding a *sequence* of measure $\mu_1, \ldots, \mu_n$ into a Brownian motion. Most notably, this provides for the first time an optimal stopping representation of Rost's solution for multiple marginals. Utilising this representation we perform an empirical study of worst-case bounds on prices of variance options using market data.

The Perkins solution remained for a long time the only solution to the SEP where an extension to the case where the Brownian motion $(B_t)_{t\geq 0}$ is started according to an initial measure other than a Dirac was unknown. We are able to give such an extension and furthermore establish its uniqueness.

Motivated by the recent spread of COVID 19, the second part of this thesis is concerned with uncertainty in the context of infectious disease modelling.

Early on the importance of *superspreading*, i.e. the significant variability in the number of new infections caused by individuals, was acknowledged to play an important role in the progression of the pandemic. We are able to quantify the uncertainty in the estimation of the *reproduction number* $\mathcal{R}$, the average number of new infections that an infected individual produces, in the presence of superspreading.

Additionally, driven by the ongoing evolution of novel SARS-CoV-2 variants and progressing vaccination efforts we develop a simulation framework to investigate potential infection trajectories across diverse scenarios. We propose a dynamic compartment model suitable for simulation of the interplay among vaccination campaigns, non-pharmaceutical interventions (NPIs), and the emergence of new SARS-CoV-2 variants.

Zusammenfassung

Mathematische Modelle beeinflussen unser Leben auf vielfältige Weise, sie sind allgegenwärtig und unverzichtbar in unserer modernen Welt. Aber selbst ein exzellent konstruiertes Modell kann letzendlich maßgeblich von jener Realität abweichen, die es abbilden sollte. Dies führt zur Problematik der *Modellunsicherheit* - dem zentralen Thema dieser Dissertation.

Der erste Teil dieser Dissertation befasst sich mit Modellunsicherheit in der Bepreisung von Finanzderivaten. Und zwar suchen wir obere und untere *Worst-Case-Schranken* für solche Preise, wobei wir nur möglichst minimale Annahmen über unser zugrunde liegendes stochastisches Modell treffen möchten. Dies steht im Zusammenhang mit der Identifikation von spezifischen Lösungen des *Skorokhod-Einbettungsproblems* (SEP). In seiner einfachsten Form verlangt das SEP, eine gegebene Wahrscheinlichkeitsverteilung μ als Brown'sche Bewegung $(B_t)_{t\geq 0}$ gestoppt zu einer bestimmten Stoppzeit τ darzustellen. In dieser Arbeit betrachten wir mehrere Lösungen des SEP, die jeweils wichtige Rollen in der robusten Derivatbepreisung spielen, und zwar Root-, Rost- und Perkins-Lösungen.

Für die Lösungen von Root und Rost geben wir eine einfache, probabilistische Erklärung für die Darstellung dieser Stoppzeiten mittels Lösungen assoziierter Optimal-Stopping Problemen. Diese Darstellungen können wir auch auf den multi-marginalen Fall, also das Einbetten einer *Folge* von Maßen $\mu_1, \ldots, \mu_n$ in eine Brown'sche Bewegung, erweitern. Besonders ist, dass wir hiermit die erste Darstellung multi-marginaler Rost Lösungen mittels Optimal-Stopping Problemen finden. Mithilfe dieser Darstellungen ist es uns möglich, eine empirische Studie von Worst-Case-Schranken von Varianz-Options-Preisen basierend auf Marktdaten durchzuführen.

Die Perkins-Lösung war lange Zeit die einzige Lösung des SEP, für welche die Erweiterung für den Fall, dass die Brown'sche Bewegung $(B_t)_{t\geq 0}$ gemäß einer Anfangsverteilung anders als einem Dirac-Maß gestartet wird, unbekannt war. In dieser Arbeit finden wir eine solche Erweiterung und zeigen darüber hinaus noch deren Eindeutigkeit.

Motiviert durch die Verbreitung von COVID-19 im Jahr 2020, befasst sich der zweite Teil dieser Dissertation mit Unsicherheiten in der Modellierung von Infektionskrankheiten.

Bald nach Pandemieausbruch war bekannt, dass *Superspreading* - also signifikante Varianz in der Anzahl neuer Infektionen, die durch Einzelpersonen verursacht werden - eine wichtige Rolle für den Pandemieverlauf spielt. Uns ist es möglich, die Superspreading-bedingten Unsicherheiten in der Schätzung der *Reproduktionszahl* $\mathcal{R}$ - der durchschnittlichen Anzahl neuer Infektionen, die eine infizierte Person hervorruft - zu quantifizieren.

Des Weiteren, durch die fortlaufende Evolution neuer SARS-CoV-2-Varianten und voranschreitenden Impfbemühungen notwendig geworden, entwickeln wir ein Simulationsmodell, um potenzielle Epidemieverläufe in verschiedensten Szenarien zu untersuchen. Dieses Modell eignet sich für die Simulation der Wechselwirkungen zwischen Impfkampagnen, nicht-pharmazeutischen Interventionen (NPIs) und dem Auftreten neuer SARS-CoV-2-Varianten.

Acknowledgment

To Mathias and Alex, it took a while, but we got it done.

To Manu, Claus, Julio, Sigrid, Gudi, Martin, Astrid, Walter, Daniel, Bertram, Ben, Xin, Lorenz, Stefan, and whoever else I forgot who were here with me at one time or another.

To my family whose doubt that this thesis will be completed was a little bit smaller than mine.

To my extended family, Dani and Bartira, who generously endured all my moods, especially during this final phase.

To my many friends, Alge, Dani, Fritz, Irina, Lara, Laura, Manveer, Marc, Mathias, Paul, Sophie, Thomas and probably (hopefully) some others.

To my math teacher Michael Tscholl without whom I'd not have shown interest in mathematics, to Roland Zweimüller without whom I'd not have shown interest in probability theory, and my many teacher along the way who kept me here.

To my financiers whose financial support through the Austrian Science Fund (FWF) under the grants P35197, P28661, and Y0782, as well as the Vienna Science and Technology Fund (WWTF) under grant MA16-021 I gratefully acknowledge.

Annemarie M. Grass, Vienna, Jan 4, 2024

Contents

Introduction

Mathematical models are indispensable tools, helping to guide decisions in complex environments. The art of translating real-life problems into the language of mathematical equations is, to a large extend, a process of model selection. The use of such models is pervasive in modern-day society, and a plethora of aspects of life are influenced one way or another by their impact. Clearly, the goal is to rigorously test and validate these models to ensure they accurately represent the aspects of reality that they were meant to describe. However, there exists an unavoidable truth: No matter how well-constructed and stress-tested a mathematical model may be, there may come a moment when it fails to capture reality with the desired accuracy. We have arrived at the challenge of ***model uncertainty***.

The primary focus of this thesis lies on models used in stochastic finance and its applications. While such models are extensively used in various situations, they are not immune to failure. The infamous "Black Monday" of 1987 and the subsequent financial crises of 2001 and 2008 serve as reminders of the consequences that over-confidence in mathematical models and the lack of acknowledgment of model-uncertainty bear. In addition to questions of model-uncertainty in finance, this thesis is also concerned with the role of uncertainty in the simulation of COVID-19 outbreaks, which is often based on models which are related to the ones applied in stochastic finance.

In the following we briefly summarise the contributions to stochastic finance and then discuss the case of COVID-19 modelling.

Robust mathematical finance

A foremost challenge in mathematical finance is the *pricing problem*. Denote by $(P_t)_{t\in[0,T]}$ an underlying financial instrument which we want to investigate, typically the price process of a market index, a stock, or a foreign currency. Given a pay-off function F, the pricing problem is to assign a fair price F_T to the claim $F((P_t)_{t\in[0,T]})$, where $T > 0$ is the *maturity* of this claim.

In classical mathematical finance, one chooses a concrete stochastic model for the price process $(P_t)_{t\in[0,T]}$, commonly described by a system of stochastic differential equations. One then proceeds to compute a fair price F_T of the claim as the discounted, risk-neutral expectation of the payoff under said model, more precisely $F_T = \mathbb{E}_{\mathbb{Q}}\left[F((P_t)_{t\in[0,T]})\right]$ where $\mathbb{Q}$ denotes a risk-neutral pricing measure.

The financial community's most vigorously pursued approach to alleviate from model misspecification in this context is the ***worst case paradigm***. Following [9, 10], we briefly describe the idea.

In many situations, European call options are liquidly traded so that we do not need to consider them as derivatives of the underlying price process, but rather think of their prices as available data supplied by financial markets. In turn, mathematical models can use the prices of European options as inputs to assess the prices of more exotic derivatives. Specifically, given European call option prices on the underlying $(P_t)_{t\in[0,T]}$ for maturity T, the Breeden-Litzenberger theorem [2] provides the time-T marginal distribution μ_T of the underlying P_T. We can then proceed to identify an upper $\overline{F}$ and lower $\underline{F}$ *worst-case bound* for the fair price F_T by considering *all* such models, subject to only to the assumption that these models are continuous martingales agreeing with today's value P_0 and adhering to the given marginal distribution at time T. Let $\mathcal{M}_\mu^{P_0}$ denote the set of such martingales, then

$$\underline{F} := \inf_{M\in\mathcal{M}_\mu^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] \leq F_T \leq \sup_{M\in\mathcal{M}_\mu^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] =: \overline{F}. \tag{1}$$

We will refer to this chain of inequalities as the ***market-informed robust pricing problem***.

Hobson [9, 10] connected the robust pricing problem – specifically, the identification of models attaining the respective bounds while following the market-given marginals distribution – to solutions to the so-called *Skorokhod embedding problem*. In its most basic form, the problem can be stated as follows: Given a probability measure μ on $\mathbb{R}$ and a Brownian motion $(B_t)_{t\geq 0}$, find a stopping time τ such that

$$B_\tau \sim \mu \text{ and } B_{t\wedge\tau} \text{ is uniformly integrable.} \tag{SEP}$$

Given such a stopping time τ which embedds the time-T marginal μ_T into a Brownian motion started in P_0, it is easy to see that

$$P_t = B_{\frac{t}{T-t}\wedge\tau} \tag{2}$$

gives a plausible candidate price process as it is a continuous martingale which satisfies $P_T \sim \mu_T$.

In this thesis, two solutions to (SEP) will receive particular attention, namely the Root [12] and Rost [13] barrier-type solutions. These solutions are stopping times of the form $\tau_{\mathcal{R}} = \inf\{t \geq 0 : (t, B_t) \in \mathcal{R}\}$, that is, they can be represented as hitting times of sets $\mathcal{R} \subseteq \mathbb{R}_+ \times \mathbb{R}$ which are subject to some type of *barrier structure*, precisely:

$$\begin{aligned} &\bullet \text{ If } (t,x) \in \mathcal{R}^{Root} \text{ then also } (s,x) \in \mathcal{R}^{Root} \text{ for all } s \geq t. \\ &\bullet \text{ If } (t,x) \in \mathcal{R}^{Rost} \text{ then also } (s,x) \in \mathcal{R}^{Rost} \text{ for all } s \leq t. \end{aligned} \tag{3}$$

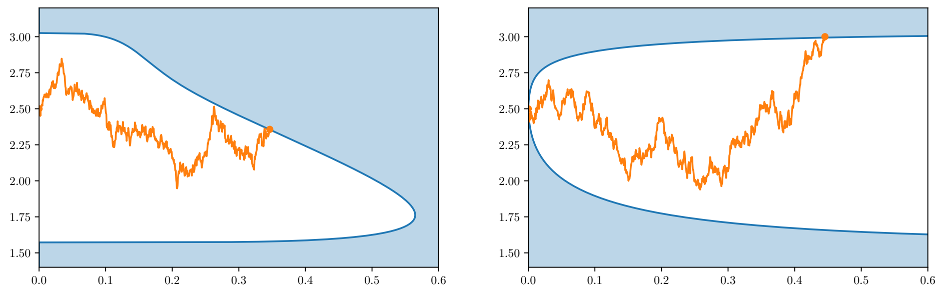


Figure 1: Illustration of a Root (left) resp. Rost (right) solution to (SEP)

.

Our interest in Root and Rost solutions is twofold. These solutions are classic and extensively studied in the context of (SEP), and generalisations thereof are interesting from a probabilistic point of view in their own right, as recognised by the extensive existing body of literature regarding Root and Rost solutions.

Yet, even more importantly for our modeling interest, Root and Rost solutions were identified as the solutions to (SEP) for which the corresponding models, defined as in (2), attain the lower and upper respective robust pricing bounds of options written on the realised variance of an underlying price process $(P_t)_{t\in[0,T]}$, known as *variance options*.

The robust pricing problem of variance options informed by the data given at maturity is a well researched topic, see for example [3] for a thorough numerical investigation of the extremal Root and Rost models using simulated data.

However, options are traded liquidly and frequently on thousands of stocks and indices. Exchanges like CBOE offer European call options expiring every month and, in some instances, much more often. Given this rich supply of data it becomes apparent that we should be able to use more input data for our robust pricing problem than the time T market-informed marginals, namely the marginal distributions given for all expiries quoted in between. Consequently, it is reasonable to assert that every plausible model should conform to all these intermediate marginal constraints, thereby tightening the robust bounds on the pricing problem.

We aim to extend the robust pricing framework to be able to incorporate all market-given maturities leading up to expiration. To formalise this, let $0 < T_1 < T_2 < \cdots < T_n = T$ denote the available maturities leading up to the expiration time T. Correspondingly, we denote the associated marginals recovered through the Breeden-Litzenberger theorem by $\mu_1, \ldots, \mu_n$.

This leads to a generalisation of (SEP), consisting of finding a sequence of stopping times $0 \leq \tau_1 \leq \cdots \leq \tau_n$ such that each of those stopping times solves the corresponding embedding problem, i.e. satisfy $B_{\tau_i} \sim \mu_i, i \leq n$. We call this a *multi-marginal* (SEP).

Owing to their inherent barrier structure, investigating multi-marginal Root and Rost models which attain the robust pricing problems on variance options informed by multiple maturities is especially interesting. These models are uniquely characterised by the fact that every possible sample path must pass through a sequence of barriers *consecutively*. Hence, analysing the structure of these respective barriers might offer further insights into these extremal models.

While the market-informed robust pricing theory is straightforward to extend to the incorporation of data given for multiple maturities as carried out in Article 3 of this thesis, the necessary existence results of multi-marginal Root and Rost solutions were for a long time only know in abstract, non-constructive terms.

Article 1 of this thesis gives representations of Root and Rost solutions that allow for numeric analysis while Article 2 extends this to multi-marginal solutions.

Prior research in [3, 4, 6, 7, 8] established that Root and Rost barriers which solve the (single-marginal) (SEP) can be recovered through solutions to a respective associated PDE in variational form. Such a PDE can be solved numerically, hence giving an algorithm for concrete numerical computation of such barriers.

In these papers it was also noted that due to connections between solutions to such PDEs in variational form and solutions to particular optimal stopping problems as know due to [1], Root and Rost barriers must admit a representation through optimal stopping problems. However, this was known only due to a relationship between PDEs in variational form and optimal stopping problems. An easy connection between Root and Rost solutions and their respective optimal stopping representation remained unknown.

The main objective of the first article of this thesis is to provide a simple probabilistic argument for this connection by establishing a specific switching identity. In a random walk regime an optimal stopping representation of Root and Rost solutions can be identified with concise martingale arguments. The Brownian case is then recovered by taking a Donsker type limit.

In the second article of this thesis we extend these arguments to the multi-marginal case. This recovers the optimal stopping representation of the multi-marginal Rost solution which is known due to [5]. Notably, we also establish a previously unknown optimal stopping characterisation of the multi-marginal Rost solution.

In order to apply the results of the second article to our multi-marginal market-informed robust pricing problem on variance options, Article 3 provides the PDEs in variational form corresponding to the respective multi-marginal optimal stopping characterisation. These PDEs can be solved numerically using standard methods. Additionally, Article 3 features an empirical study on variance options using real market data on the S&P500. Particularly intriguing about this empirical investigation is that it uncovers a potential inherent flaw of the robust pricing paradigm: In stark contrast to common intuition, incorporating additional market data into the pricing problem does *not*, in fact, tighten the bounds but leaves them unchanged. The barriers adhering to market-given data which emerge form the numerical analysis appear to be in perfect consecutive order, implying that the requirement for these barriers to be traversed consecutively imposes no additional constraints on the extremal models.

Remarkably, for the Rost model which attains the upper bound of the robust pricing problem on variance options this would imply the following highly implausible behaviour: A sample path that consecutively hits these ordered Rost barriers would attain a new running minimum or a new running maximum at every single intermediate maturity, a dynamic which appears extremely unrealistic.

This finding seems to call for a shift of paradigm in the robust pricing approach: To adequately assess model-risk the relative plausibility of the models under consideration also needs to be taken into account.

Article 4 of this thesis is concerned with another solution to (SEP) which is also of interest from a mathematical finance viewpoint: The Perkins solution (see [11]) minimises the law of the running maximum while simultaneously maximising the law of the running minimum $\underline{B}_t$, making it applicable to robust pricing problems for lookback options.

In the class of all known solutions to (SEP) which exhibit specific optimality properties, the Perkins solution stands out due to its geometric structure, which appears significantly different from the barrier-type behavior of other solutions. Moreover, it is the only optimal solution to (SEP) that, thus far, has been limited to the case of Brownian motion starting in a *Dirac distribution*.

In Article 4 of this thesis we provide, for the first time, an optimal solution to the Skorokhod embedding problem for the general (SEP) — where the Brownian motion is allowed to start according to a general initial distribution $B_0 \sim \lambda$ — which recovers the Perkins solution when applied to Brownian motion started in a Dirac. This solution to (SEP) also admits a novel geometric interpretation of the Perkins solution which better clarifies the relation to other optimal solutions of (SEP): We can represent the Perkins solution as a hitting time of a Rost inverse barrier if we identify it as the minimum of two stopping times given by two copies of the path simultaneously running on a specifically constructed space until either of them hits a Rost inverse barrier as depicted in Figure 2. This furthermore enables a Loynes-type uniqueness result.

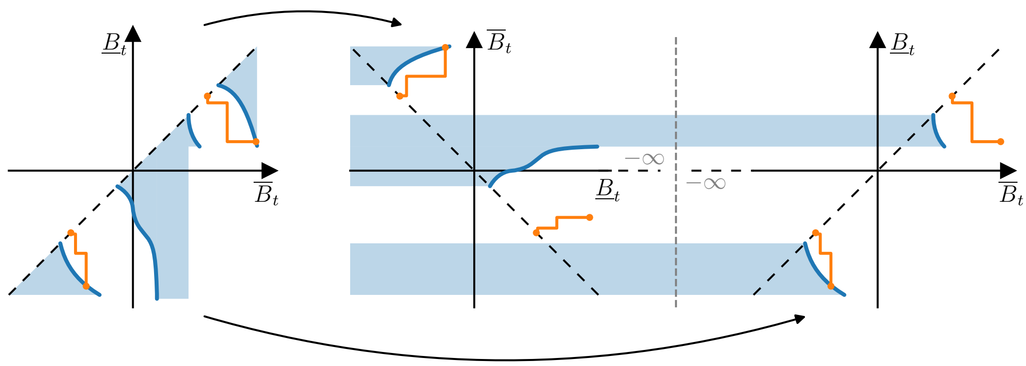


Figure 2: Geometric interpretation of Perkins solution and its transformation into a hitting time of a Rost-like barrier.

Uncertainty in the context of infectious disease modeling

Motivated primarily by the recent spread of COVID 19, Articles 5 and 6 of this thesis are concerned with the role of uncertainty in the spread of infectious diseases. A prime interest in disease modeling is to describe the number of infected individuals over time through (probabilistic) models. As it was the case in the previously presented articles of this thesis, we aim to better understand the dependence of the models' predictions on the underlying assumptions.

Article 5 of this thesis is concerned with uncertainty in the estimation of the *basic reproduction number* $\mathcal{R}$. Roughly put, at a given point in time, the reproduction number $\mathcal{R}$ denotes the average number of new infections that an infected individual produces. This value has significant implications for the disease progression, as $\mathcal{R} > 1$ corresponds to exponential growth of the number of infected individuals, while $\mathcal{R} < 1$ implies that this number decreases exponentially with time. There is significant literature suggesting that superspreading, i.e. the significant variability in number of new infections caused by individuals, plays an important role in the spread of COVID-19, see [14] for an recent overview of the literature. In Article 5, we consider the effect that such superspreading has on the estimation of the reproduction number and subsequently the estimation of the number of future infected individuals. Our results quantify the uncertainty in the estimation of the reproduction number depending on the dispersion parameter k which is used to measure the amount of superspreading: We demonstrate that the range of the prediction interval for the value of $\mathcal{R}$ has width proportional to $1/\sqrt{k}$. The presence of superspreading is associated to small values of k, with estimates for COVID-19 being as low as 0.04.

In Article 6 of this thesis we provide a framework to systematically explore the uncertainty of model-predictions with respect to a variety of input parameters and modeling assumptions. Motivated by the ongoing evolution of novel SARS-CoV-2 variants and the progressing landscape of vaccination efforts, we develop a simulation framework to investigate potential infection trajectories across diverse scenarios. We are concerned with the interplay among three key elements: vaccination campaigns, non-pharmaceutical interventions (NPIs), and the emergence of new SARS-CoV-2 variants. In this hierarchical framework, new infections are generated, considering random individual infectiousness, thereby encompassing the superspreading phenomenon witnessed in the COVID-19 pandemic. Our simulation operates as a dynamic compartment model, integrating an individual's infection history, vaccination status, and potential for reinfection to determine their resistance to subsequent infections. Introducing a measure of associated risk, denoted as $\rho^{\mathcal{V}}$, for each SARS-CoV-2 variant, which considers the level of immunity within a population we can explore how this risk evolves with increasing vaccination rates. Additionally, by examining diverse population compositions in terms of prior infections and vaccination types, we gain insights into variants that may pose varying risks to different countries.

References

[1] A. Bensoussan and J. Lions. Applications of variational inequalities in stochastic control. eng. Studies in mathematics and its applications ; 12. Amsterdam [u.a.]: North-Holland, 1982. ISBN: 0444863583.

[2] D. T. Breeden and R. H. Litzenberger. Prices of State-contingent Claims Implicit in Option Prices. *The Journal of Business* 51.4 (1978), pp. 621–51.

[3] A. M. G. Cox and J. Wang. Optimal robust bounds for variance options. *ArXiv e-prints* (Aug. 2013).

[4] A. M. G. Cox and J. Wang. Root's Barrier: Construction, Optimality and Applications to Variance Options. *Ann. Appl. Probab.* 23.3 (2013), pp. 859–894.

[5] A. M. G. Cox, J. Obłój, and N. Touzi. The Root solution to the multi-marginal embedding problem: an optimal stopping and time-reversal approach. *Probab. Theory Related Fields* 173.1-2 (2019), pp. 211–259.

[6] T. De Angelis. From optimal stopping boundaries to Rost's reversed barriers and the Skorokhod embedding. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.2 (2018), pp. 1098–1133.

[7] P. Gassiat, A. Mijatović, and H. Oberhauser. An integral equation for Root's barrier and the generation of Brownian increments. *Ann. Appl. Probab.* 25.4 (2015), pp. 2039–2065.

[8] P. Gassiat, H. Oberhauser, and G. dos Reis. Root's barrier, viscosity solutions of obstacle problems and reflected FBSDEs. *Stochastic Processes and their Applications* 125.12 (2015), pp. 4601–4631.

[9] D. Hobson. Robust hedging of the lookback option. *Finance and Stochastics* 2 (4 1998), pp. 329–347.

[10] D. Hobson. The Skorokhod embedding problem and model-independent bounds for option prices. *Paris-Princeton Lectures on Mathematical Finance 2010*. Vol. 2003. Lecture Notes in Math. Springer, Berlin, 2011, pp. 267–318.

[11] E. Perkins. The Cereteli-Davis solution to the H^1-embedding problem and an optimal embedding in Brownian motion. *Seminar on stochastic processes, 1985*. Springer. 1986, pp. 172–223.

[12] D. H. Root. The existence of certain stopping times on Brownian motion. *Ann. Math. Statist.* 40 (1969), pp. 715–718.

[13] H. Rost. Skorokhod stopping times of minimal variance. *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975)*. Berlin: Springer, 1976, 194–208. Lecture Notes in Math., Vol. 511.

[14] O. Wegehaupt, A. Endo, and A. Vassall. Superspreading, overdispersion and their implications in the SARS-CoV-2 (COVID-19) pandemic: a systematic review and meta-analysis of the literature. *BMC Public Health* 23.1003 (2023).

Publication and Authorship Information

Article 1

Switching Identities by Probabilistic Means

J. Backhoff-Veraguas, A. M. G. Cox, A. M. Grass and M. Huesmann

Séminaire de Probabilités, to appear (Accepted for publication on March 30, 2022)

Preprint available at: `https://arxiv.org/abs/2002.12840`

Contributions of the thesis-author:

The article was worked out in close collaboration.

Article 2

Delayed Switching Identities and Multi-Marginal Solutions to the Skorokhod Embedding Problem

A. M. G. Cox and A. M. Grass

ArXiv e-prints, under review at Probability Theory and Related Fields (Submission acknowledged on January 3, 2024)

Preprint available at: `https://arxiv.org/abs/2312.04218`

Contributions of the thesis-author:

The article was worked out in close collaboration.

Article 3

Robust Option Pricing with Volatility Term Structure — An Empirical Study for Variance Options

A. M. G. Cox, A. M. Grass

ArXiv e-prints, not yet submitted for publication

Preprint available at: `http://arxiv.org/abs/2312.09201`

Contributions of the thesis-author:

The article was worked out in close collaboration.

Article 4

Perkins Embedding for General Starting Laws

A. M. Grass

ArXiv e-prints, under review at Annals of Applied Probability (Submission acknowledged on August 3, 2023)

Preprint available at: `https://arxiv.org/abs/2307.03618`

Article 5

Disease Momentum: Estimating the Reproduction Number in the Presence of Superspreading

K. D. Johnson, M. Beiglböck, M. Eder, A. M. Grass, J. Hermisson, G. Pammer, J. Polechová, D. Toneian and B. Wölfl

Infectious Disease Modelling 6, April 2021, pp. 706–728

Preprint available at: `https://arxiv.org/abs/2012.08843`

Article available at: `https://doi.org/10.1016/j.idm.2021.03.006`

Contributions of the thesis-author:

The author contributed substantially to all aspects of the research, from conceptualization to manuscript finalization.

Article 6

Robust Models of SARS-CoV-2 Heterogeneity and Control

K. D. Johnson, A. M. Grass, D. Toneian, M. Beiglböck and J. Polechová

PLOS Global Public Health 2 (5), May 2022, pp. 1–35

Preprint available at: `https://arxiv.org/abs/2109.11156`

Article available at: `https://doi.org/10.1371/journal.pgph.0000412`

Contributions of the thesis-author:

The author contributed substantially to all aspects of the research, from conceptualization to manuscript finalization.

Switching identities by probabilistic means

J. Backhoff-Veraguas, A. M. G. Cox, A. M. Grass and M. Huesmann

Abstract

Switching identities have a long history in potential theory and stochastic analysis. In recent work of Cox and Wang, a switching identity was used to connect an optimal stopping problem with the Skorokhod embedding problem (SEP). Typically, switching identies of this form are derived using deep analytic connections. In this paper, we prove the switching identities using a simple probabilistic argument, which furthermore highlights a previously unexplored symmetry between the Root and Rost solutions to the SEP.

Keywords: Skorokhod embedding problem; Root; Rost; optimal stopping; switching identities.

MSC 2020 60G40, 60G42, 60H30.

1 Introduction

Let D be a rectangle of horizontal length T and let $(0,x),(0,y)$ be points on the left boundary of D. Let B be Brownian motion (started in x or y) and write σ for the first time at which (t,B_t) leaves the rectangle. As a particular case of Hunt's switching identities, we know that

$$\mathbb{E}^x\left[|B_\sigma - y|\right] = \mathbb{E}^y\left[|B_\sigma - x|\right]. \tag{1.1}$$

In this paper, we will be interested in generalisations of this identity, which are of particular relevance in the construction of solutions to the Skorokhod embedding problem. The Skorokhod embedding problem can be stated as follows: Given measures λ,μ on $\mathbb{R}$, and a Brownian motion B with $B_0 \sim \lambda$, find a stopping time τ such that

$$B_\tau \sim \mu \text{ and } (B_{t\wedge\tau})_{t\geq 0} \text{ is uniformly integrable.} \tag{SEP}$$

In this paper we will be interested in a sub-class of solutions to (SEP) which are given by the first hitting time of a right-barrier (resp. left-barrier) $\mathcal{R}$ which is a closed subset of $\mathbb{R}_+\times\mathbb{R}$ with the property that $(t,x)\in\mathcal{R} \implies (s,x)\in\mathcal{R}$, when $s>t$ (resp. $s<t$). The most prominent example of such a solution is the Root solution [16]. It establishes the existence of a right-barrier $\mathcal{R}^{Root}$ such that

$$\tau^{Root} := \inf\left\{t\geq 0 : (t,B_t)\in\mathcal{R}^{Root}\right\}$$

solves (SEP). There is an analogous result for left-barriers by Rost [17]. While Root's original contribution was non-constructive, it was shown in [4, 5] how to build the Root barrier. This was done by considering *optimal stopping problems* and establishing the crucial identity

$$-\mathbb{E}^\lambda\left[|B_{\tau^{Root}\wedge T} - y|\right] = \sup_{\sigma\leq T}\mathbb{E}^y\left[U_\mu(B_\sigma)\mathbb{1}_{\sigma<T} + U_\lambda(B_\sigma)\mathbb{1}_{\sigma=T}\right] \tag{1.2}$$

where the supremum is taken over all stopping times $\sigma\leq T$. Here U_m is the potential function associated with the measure m, i.e.

$$U_m(x) := -\int |y-x|\, m(\mathrm{d}y).$$

From a potential-theoretic perspective, the expression on the right-hand side of (1.2) is the *réduite* of a particular function on the time-space domain. The relevance of the réduite was clearly understood by Rost [17], although the connection to optimal stopping problems was not explicit until [8].

From formula (1.2) , the barrier region $\mathcal{R}^{Root}$ can be identified with the stopping region of the optimal stopping problem *after a reversal of the time component.* Together with the switching of the starting position between y and λ, this formula can be seen as a generalisation of (1.1). Furthermore, the articles [4, 5] also provide a construction for the Rost barrier similarly based on optimal stopping problems.

The equation (1.2) is well understood in potential theoretic terms, and can be seen as a deep consequence of Markov duality, applied to the time-space process (t,B_t). However this connection does not offer much insight into why such equations might arise in these applications. In [4, 5] this connection was made via viscosity theory and it was noted that a probabilistic explanation has yet to be given.

The aim of this short paper is to show that (1.2) *is a natural consequence of some simple probabilistic arguments.* As a bonus, our proof reveals a previously unexplored symmetry between the Root and Rost solutions to (SEP).

As a brief taste of the style of argument we will use, and to highlight its fundamental nature, let us first consider equation (1.1). Introduce a second Brownian motion W, independent of B, running from right to left and started on the right hand side of the rectangle D at the point (T, y). For any $s \in [0, T]$ we consider the stopping times

$$\sigma_s := \sigma \wedge (T - s)$$
$$\tau_s := \inf \{t \geq 0 : (T - t, W_t) \notin D\} \wedge s$$

and define

$$F(s) := \mathbb{E}^{B_0 = x, W_0 = y} [|B_{\sigma_s} - W_{\tau_s}|] .$$

Then $F(0) = \mathbb{E}^x [|B_\sigma - y|]$ and $F(T) = \mathbb{E}^y [|B_\sigma - x|]$. We will show that the function $s \mapsto F(s)$ is constant. In fact we will first show this for a discrete time version where we replace B by a random walk X and W by a random walk Y. For this discrete version, it is not difficult to see that $F(s) = F(s-1)$ (cf. Fig. 1 or Section 3.1 for details) so that F is constant. From there, (1.1) follows from an application of Donsker's theorem.

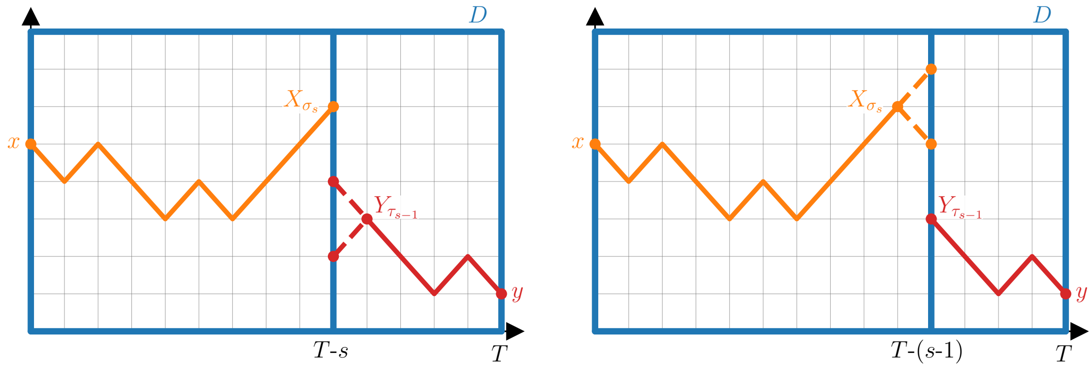


Figure 1: Illustration of $F(s) = F(s-1)$.

The Rost optimal stopping problem was subject of investigation in [14] by McConnell where it is derived via classical PDE methods and in [7] by De Angelis where a probabilistic proof is given relying on stochastic calculus. Furthermore, the Root optimal stopping problem was also derived by Gassiat, Oberhauser and Zou in [9] where a suitable extension for a much wider class of Markov processes is established using classical potential theoretic methods as well as by Cox, Obłój and Touzi in [6] where a multi-marginal extension of the problem is found.

1.1 Overview

In Section 3 we shall establish (1.2) in the context of simple symmetric random walks (SSRW) on the integer lattice, as well as a related identity pertaining Rost stopping times. Interestingly, the symmetry between Root and Rost cases will be obtained as a consequence of a simple time-reversal principle. Then in Section 4 we explore extensions to the multidimensional setting. In Section 5 we will give some remarks on the passage to continuous time and in Section 6 we will draw some future perspectives.

2 The Root and Rost optimal stopping problems

Recall the notion of Skorokhod embedding problem (SEP) from the introduction. This problem was first formulated and solved by Skorokhod [18, 19], and numerous new solutions have been found since. We refer to the surveys of Hobson [11] and Obłój [15] for an account of many of these solutions. To guarantee well-posedness, we assume thoroughout that λ, μ have finite first moment and are in convex order.

Let W denote a one-dimensional Brownian motion (in keeping with the introduction, we think of W as a Brownian motion running backwards; the reason for this will become clear in the following section). Suppose we are given initial and target distributions λ and μ, we want to study the Root [16] resp. Rost [17] solution to the corresponding (SEP). While Root and Rost solutions are most commonly given as hitting times of so called *barriers*, specific subsets of $\mathbb{R}^2$, keeping the previous notation, we will denote by D^{Root} (resp. D^{Rost}) the continuation set of the Root (resp. Rost) embeddings which can be seen as the complements of the barriers in $\mathbb{R}^2$.

Let us write μ^{Root} (resp. μ^{Rost}) for the law of the Brownian motion starting with distribution λ at the time it leaves D^{Root} (resp. D^{Rost}) and μ_T^{Root} (resp. μ_T^{Rost}) for the time it leaves $D^{Root} \cap ([0,T) \times \mathbb{R})$ (resp. $D^{Rost} \cap ([0,T) \times \mathbb{R})$). Recall that the potential of a measure m is denoted by $U_m(x) := -\int |x-y|\, m(\mathrm{d}y)$, and for a random variable Z we write U_Z for the potential of the law of Z. Throughout this note we consider *optimal stopping problems*, thus suprema taken over τ (resp. σ) will denote suprema over stopping times.

The relations of interest in our article, found in [4, 5], are

$$U_{\mu_T^{Root}}(x) = \mathbb{E}^x \left[U_{\mu^{Root}}(W_{\tau^*}) \mathbb{1}_{\tau^* < T} + U_\lambda(W_{\tau^*}) \mathbb{1}_{\tau^*=T} \right] \tag{2.1}$$

$$= \sup_{\tau \le T} \mathbb{E}^x \left[U_{\mu^{Root}}(W_\tau) \mathbb{1}_{\tau<T} + U_\lambda(W_\tau) \mathbb{1}_{\tau=T} \right], \tag{2.2}$$

where the optimizer is $\tau^* := \inf\left\{t \ge 0 : (T-t, W_t) \notin D^{Root}\right\} \wedge T$, and

$$U_{\mu^{Rost}}(x) - U_{\mu_T^{Rost}}(x) = \mathbb{E}^x \left[\left(U_{\mu^{Rost}} - U_\lambda \right)(W_{\tau_*}) \right] \tag{2.3}$$

$$= \sup_{\tau \le T} \mathbb{E}^x \left[\left(U_{\mu^{Rost}} - U_\lambda \right)(W_\tau) \right], \tag{2.4}$$

where the optimizer is $\tau_* := \inf\left\{t \ge 0 : (T-t, W_t) \notin D^{Rost}\right\} \wedge T$.

Remark that (2.1)-(2.2) are related to (1.2) in the introduction. On the other hand, (2.3)-(2.4) haven't been stated yet. One of the contributions of the article will be to obtain the latter as a consequence of the former, by means of a previously unexplored symmetry between Root and Rost solutions, as hinted at in the introduction.

In the coming section we establish the above results (2.1)-(2.4) in the context of simple symmetric random walks (SSRW) on the integer lattice.

Remark 2.1. In what follows, we will implicitly exploit the fact that Brownian motion and the SSRW are self-dual in the sense of duality of Markov processes. The results we present hold in greater generality (and in particular, for processes which are not in duality with themselves). In such instances, the potential kernels appearing e.g. in (2.1) should be replaced by the potential kernels of the dual processes, see e.g. [9].

3 A common one-dimensional random walk framework

Consider a set $D \subseteq \mathbb{Z} \times \mathbb{Z}$ satisfying

- If $(t,m) \in D$, then for all $s < t$ also $(s,m) \in D$.

This should be seen as a discretised version of the Root continuation set defined in the introduction. Likewise a Rost continuation set can be cast in the above form after reflection w.r.t. a vertical line.

Notation: Denote by X, Y two mutually independent SSRW on some probability space (Ω, P) which are started at possibly random initial positions. Given $x, y \in \mathbb{Z}$ we write P^x and P_y for the conditional distribution given $X_0 = x$ and $Y_0 = y$ resp. Similarly P_y^x means that we condition on both events simultaneously. We can consider a probability measure λ on $\mathbb{Z}$ a "starting" distribution by setting $P^\lambda := \sum_{x\in\mathbb{Z}} P^x \lambda(\{x\})$, etc. Let us then introduce the stopping time

$$\rho^{Root} = \inf\{t \in \mathbb{N} : (t, X_t) \notin D\},$$

where $\mathbb{N} = \{0, 1, \dots\}$. We define by μ^{Root} the law of $X_{\rho^{Root}}$ under P^λ, and assume henceforth that the martingale $\left(X_{\rho^{Root}\wedge t}\right)_{t\in\mathbb{N}}$ is uniformly integrable. We conveniently drop the dependence of μ^{Root} on λ. Given $T \in \mathbb{N}$ we write μ_T^{Root} for the P^λ-law of $X_{\rho^{Root}\wedge T}$. Note that this definition is equivalent to μ_T^{Root} being the law of X started with distribution λ at the time it leaves $D^{Root} \cap (\{0, \dots, T-1\} \times \mathbb{Z})$. We will first establish the following identity, which is a discrete-time version of (2.1)-(2.2)

$$U_{\mu_T^{Root}}(y) = \mathbb{E}_y\left[U_{\mu^{Root}}(Y_{\tau^*})\,\mathbb{1}_{\tau^*<T} + U_\lambda(Y_{\tau^*})\,\mathbb{1}_{\tau^*=T}\right] \tag{3.1}$$

$$= \sup_{\tau\le T}\mathbb{E}_y\left[U_{\mu^{Root}}(Y_\tau)\mathbb{1}_{\tau<T} + U_\lambda(Y_\tau)\mathbb{1}_{\tau=T}\right], \tag{3.2}$$

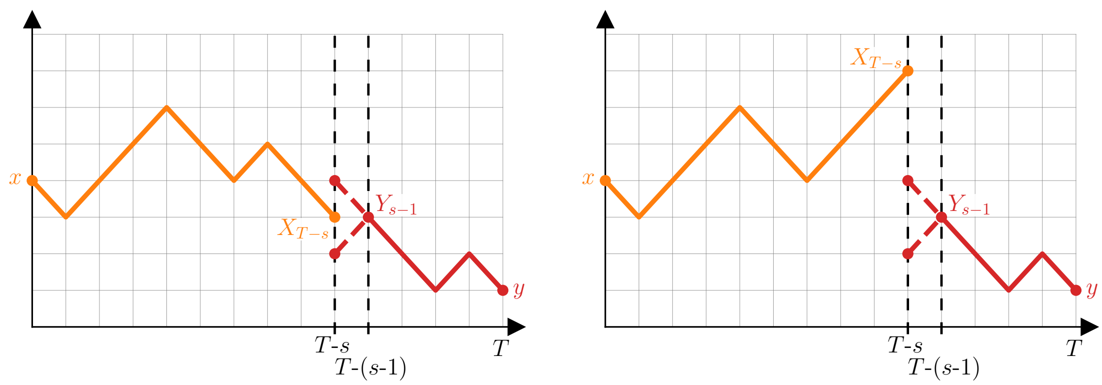
where the optimizer is $\tau^* := \inf\{t \in \mathbb{N} : (T-t, Y_t) \notin D\} \wedge T$.

From here we will derive the discrete-time analogue of (2.3)-(2.4).

3.1 Core argument

For convenience of the reader we present here the basis of the argument which we repeatedly use, namely that for $s \in \{1, \dots, T\}$

$$\mathbb{E}_y^x\left[|X_{T-s} - Y_s|\right] = \mathbb{E}_y^x\left[|X_{T-(s-1)} - Y_{s-1}|\right]. \tag{3.3}$$



(A) SSRW where $X_{T-s} = Y_{s-1}$ (B) SSRW where $X_{T-s} \neq Y_{s-1}$

Figure 2: Illustrating the appearance of the indicator function in the core argument.

This represents a formalisation of the discretised argument in the prelude. Indeed,

$$\begin{aligned}
\mathbb{E}_y^x\left[|X_{T-s} - Y_s|\right] &= \mathbb{E}_y^x\left[\mathbb{E}_y^x\left[|X_{T-s} - Y_s|\big|X_{T-s}, Y_{s-1}\right]\right] \\
&= \mathbb{E}_y^x\left[\mathbb{E}_y^x\left[|(X_{T-s} - Y_{s-1}) - (Y_s - Y_{s-1})|\big|X_{T-s}, Y_{s-1}\right]\right] \\
&= \mathbb{E}_y^x\left[|X_{T-s} - Y_{s-1}| + \mathbb{1}_{X_{T-s}=Y_{s-1}}\right] \\
&= \mathbb{E}_y^x\left[\mathbb{E}_y^x\left[|(X_{T-s} - Y_{s-1}) + (X_{T-s+1} - X_{T-s})|\big|X_{T-s}, Y_{s-1}\right]\right] \\
&= \mathbb{E}_y^x\left[\mathbb{E}_y^x\left[|X_{T-s+1} - Y_{s-1}|\big|X_{T-s}, Y_{s-1}\right]\right] \\
&= \mathbb{E}_y^x\left[|X_{T-(s-1)} - Y_{s-1}|\right],
\end{aligned}$$

which is best read from the top until the middle equality and then from the bottom until the same equality. More important than (3.3) is the reasoning above, especially the appearance of the indicator of the event $\{X_{T-s} = Y_{s-1}\}$, which stems from the fact that Y_{s-1} (resp. X_{T-s}) always splits into $Y_s \pm 1$ (resp. $X_{T-(s-1)} \pm 1$) with probability $1/2$. For an illustration of this phenomenon also see Figure 2.

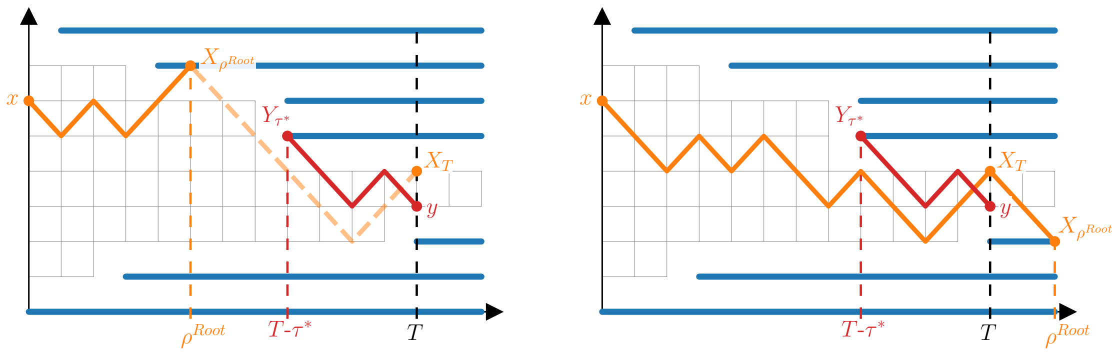
3.2 The Root case

Let $\tau^* := \min\{t \in \mathbb{N} : (T-t, Y_t) \notin D\} \wedge T$. We start with a useful observation:

Remark 3.1. The equality $U_{\mu^{Root}}(Y_{\tau^*})\,\mathbb{1}_{\tau^*<T} = U_{\mu_T^{Root}}(Y_{\tau^*})\,\mathbb{1}_{\tau^*<T}$ holds. Indeed, let $z = Y_{\tau^*}$ on $\{\tau^* < T\}$. Then $(T-\tau^*, z) \notin D$ and hence $(X_T - z)(X_{\rho^{Root}} - z) \geq 0$ on $\{\rho^{Root} > T\}$ as otherwise X would have left D before ρ^{Root}, see also Figure 3. This implies

$$
\begin{aligned}
-U_{\mu^{Root}}(z) &= \mathbb{E}^{\lambda}\left[|X_{\rho^{Root}} - z|\right] \\
&= \mathbb{E}^{\lambda}\Big[|X_{\rho^{Root}\wedge T} - z|\mathbb{1}_{\rho^{Root}\leq T} - (X_{\rho^{Root}} - z)\mathbb{1}_{\rho^{Root}>T,\, z>X_T} \\
&\qquad\qquad + (X_{\rho^{Root}} - z)\mathbb{1}_{\rho^{Root}>T,\, z\leq X_T}\Big] \\
&= \mathbb{E}^{\lambda}\Big[|X_{\rho^{Root}\wedge T} - z|\mathbb{1}_{\rho^{Root}\leq T} - (X_{\rho^{Root}\wedge T} - z)\mathbb{1}_{\rho^{Root}>T,\, z>X_T} \\
&\qquad\qquad + (X_{\rho^{Root}\wedge T} - z)\mathbb{1}_{\rho^{Root}>T,\, z\leq X_T}\Big] \\
&= \mathbb{E}^{\lambda}\left[|X_{\rho^{Root}\wedge T} - z|\right] = -U_{\mu_T^{Root}}(z). \qquad (3.4)
\end{aligned}
$$

Accordingly we may replace μ^{Root} by μ_T^{Root} in (3.1) (but we do not do so in (3.2)).



(A) SSRWs for which $\tau^* < T$ and $\rho^{Root} \leq T$

(B) SSRWs for which $\tau^* < T$ and $\rho^{Root} > T$

Figure 3: Illustrating properties of Root stopping times.

Given a Y-stopping time $\tau \leq T$, we define a stopping time $\sigma(\tau)$ of X as the first time before $T-\tau$ that X leaves D, i.e. $\sigma(\tau) := \rho^{Root} \wedge (T-\tau)$.* For any $y \in \mathbb{Z}$ we now introduce the crucial interpolating function

$$
F(s) \;:=\; F^{\tau^*}(s) \;:=\; \mathbb{E}_y^{\lambda}\left[|X_{\sigma(\tau^*\wedge s)} - Y_{\tau^*\wedge s}|\right] \quad \text{for } s \in \{0,\dots,T\}. \qquad (3.5)
$$

It may help to picture Y evolving "leftwards" from the lattice point (T, y) at time zero, so that its exit time τ^* from D before T is measured as $T - \tau^*$ for the "rightwards" process X.

Remark 3.2. We see that $\sigma(0) = \rho^{Root} \wedge T$, so consequently

$$
F(0) = \mathbb{E}^{\lambda}\left[|X_{\rho^{Root}\wedge T} - y|\right] = -U_{\mu_T^{Root}}(y).
$$

*Formally, τ is a stopping time w.r.t. the filtration $\mathcal{G} = (\mathcal{G}_t)_{t\in\mathbb{N}}$, where $\mathcal{G}_t = \sigma(\{Y_u : u \leq t\})$. $\sigma(\tau)$ is a stopping time w.r.t. the filtration $\mathcal{F} = (\mathcal{F}_s)_{s\leq T}$, where $\mathcal{F}_s = \sigma(\{X_u, Y_t : u \leq s, t \leq T\})$

On the other hand, $\sigma(\tau^*) = \rho^{Root} \wedge (T - \tau^*)$, so

$$\begin{aligned}
F(T) &= \mathbb{E}_y^\lambda \left[|X_{\rho^{Root} \wedge (T-\tau^*)} - Y_{\tau^*}| \right] \\
&= \mathbb{E}_y^\lambda \left[|X_{\rho^{Root} \wedge (T-\tau^*)} - Y_{\tau^*}| \mathbb{1}_{\tau^* < T} + |X_{\rho^{Root} \wedge (T-\tau^*)} - Y_{\tau^*}| \mathbb{1}_{\tau^* = T} \right] \\
&= \mathbb{E}_y^\lambda \left[|X_{\rho^{Root} \wedge T} - Y_{\tau^*}| \mathbb{1}_{\tau^* < T} + |X_0 - Y_{\tau^*}| \mathbb{1}_{\tau^* = T} \right] \\
&= -\mathbb{E}_y \left[U_{\mu_T^{Root}} (Y_{\tau^*}) \mathbb{1}_{\tau^* < T} + U_\lambda(Y_{\tau^*}) \mathbb{1}_{\tau^* = T} \right],
\end{aligned}$$

by independence and by applying the appropriate analogue of the argument in Remark 3.1 for the third equality.

We can now prove (3.1) and (3.2); we treat the cases separately.

Lemma 3.3. *The function F is constant. Consequently*

$$U_{\mu_T^{Root}}(y) = \mathbb{E}_y \left[U_{\mu^{Root}} (Y_{\tau^*}) \mathbb{1}_{\tau^* < T} + U_\lambda (Y_{\tau^*}) \mathbb{1}_{\tau^* = T} \right]$$

Proof. Let $0 < s < T$. Define the stopping times $\tau_s^* := \tau^* \wedge s$ and $\sigma_s = \sigma(\tau^* \wedge s) = \rho^{Root} \wedge (T - \tau_s^*)$. Then

$$F(s) = \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_s^*}| \right].$$

Let us first prove that

$$\mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_s^*}| \right] = \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}^*}| + \mathbb{1}_{X_{\sigma_s} = Y_{\tau_{s-1}^*},\, \tau^* \geq s} \right]. \tag{3.6}$$

Since

$$\mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_s^*}| \right] = \mathbb{E}_y^\lambda \left[|X_{\rho^{Root} \wedge (T-s)} - Y_s| \mathbb{1}_{\tau^* \geq s} \right] + \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}^*}| \mathbb{1}_{\tau^* < s} \right],$$

and with the appropriate analogue of the core argument (3.3) (see also Figure 4)

$$\begin{aligned}
&\mathbb{E}_y^\lambda \left[|X_{\rho^{Root} \wedge (T-s)} - Y_s| \mathbb{1}_{\tau^* \geq s} \right] \\
&= \mathbb{E}_y^\lambda \left[\mathbb{E}_y^\lambda \left[|X_{\rho^{Root} \wedge (T-s)} - Y_s| \middle| X_{\rho^{Root} \wedge (T-s)}, Y_0, \dots, Y_{s-1} \right] \mathbb{1}_{\tau^* \geq s} \right] \\
&= \mathbb{E}_y^\lambda \left[\mathbb{E}_y^\lambda \left[|(X_{\rho^{Root} \wedge (T-s)} - Y_{s-1}) - (Y_s - Y_{s-1})| \middle| X_{\rho^{Root} \wedge (T-s)}, Y_0, \dots, Y_{s-1} \right] \mathbb{1}_{\tau^* \geq s} \right] \\
&= \mathbb{E}_y^\lambda \left[\left(|X_{\rho^{Root} \wedge (T-s)} - Y_{s-1}| + \mathbb{1}_{X_{\rho^{Root} \wedge (T-s)} = Y_{s-1}} \right) \mathbb{1}_{\tau^* \geq s} \right] \\
&= \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}^*}| \mathbb{1}_{\tau^* \geq s} + \mathbb{1}_{X_{\rho^{Root} \wedge (T-s)} = Y_{s-1}\, ,\, \tau^* \geq s} \right],
\end{aligned}$$

clearly (3.6) follows. Now let us similarly establish that

$$\begin{aligned}
\mathbb{E}_y^\lambda \left[|X_{\sigma_{s-1}} - Y_{\tau_{s-1}^*}| \right] &= \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}^*}| + \mathbb{1}_{X_{\sigma_s} = Y_{\tau_{s-1}^*},\, \tau^* \geq s,\, \rho^{Root} > T-s} \right] \qquad (3.7) \\
&= \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}^*}| + \mathbb{1}_{X_{\sigma_s} = Y_{\tau_{s-1}^*},\, \tau^* \geq s} \right]. \qquad (3.8)
\end{aligned}$$

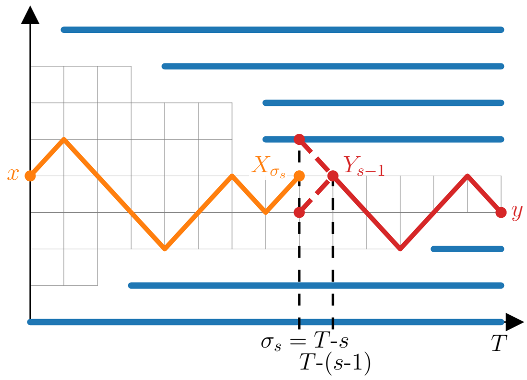
Indeed,

$$\begin{aligned}
\mathbb{E}_y^\lambda \left[|X_{\sigma_{s-1}} - Y_{\tau_{s-1}^*}| \right] &= \mathbb{E}_y^\lambda \left[|X_{T-s+1} - Y_{s-1}| \mathbb{1}_{\tau^* \geq s,\, \rho^{Root} > T-s} \right] \\
&\quad + \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}^*}| \mathbb{1}_{\{\tau^* < s\} \cup \{\rho^{Root} \leq T-s\}} \right],
\end{aligned}$$

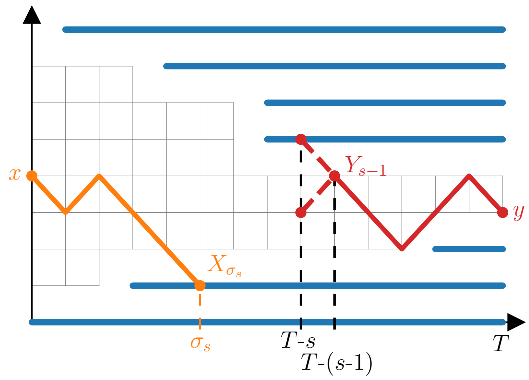
and again with the appropriate analogue of (3.3)

$$
\begin{aligned}
&\mathbb{E}_y^\lambda\left[|X_{T-s+1}-Y_{s-1}|\mathbb{1}_{\tau^*\geq s,\,\rho^{Root}>T-s}\right]\\
&\quad=\mathbb{E}_y^\lambda\left[\mathbb{E}_y^\lambda\left[|X_{T-s+1}-Y_{s-1}|\big|X_0,\dots,X_{T-s},Y_0,\dots,Y_{s-1}\right]\mathbb{1}_{\tau^*\geq s,\,\rho^{Root}>T-s}\right]\\
&\quad=\mathbb{E}_y^\lambda\left[\mathbb{E}_y^\lambda\left[|(X_{T-s}-Y_{s-1})\right.\right.\\
&\qquad\qquad\left.\left.+(X_{T-s+1}-X_{T-s})|\big|X_0,\dots,X_{T-s},Y_0,\dots,Y_{s-1}\right]\mathbb{1}_{\tau^*\geq s,\,\rho^{Root}>T-s}\right]\\
&\quad=\mathbb{E}_y^\lambda\left[\left(|X_{T-s}-Y_{s-1}|+\mathbb{1}_{X_{T-s}=Y_{s-1}}\right)\mathbb{1}_{\tau^*\geq s,\,\rho^{Root}>T-s}\right]\\
&\quad=\mathbb{E}_y^\lambda\left[|X_{\sigma_s}-Y_{\tau^*_{s-1}}|\mathbb{1}_{\tau^*\geq s,\,\rho^{Root}>T-s}+\mathbb{1}_{X_{\sigma_s}=Y_{\tau^*_{s-1}},\,\tau^*\geq s\,,\,\rho^{Root}>T-s}\right],
\end{aligned}
$$

also (3.7) follows. We then see that (3.8) holds true since on $\{X_{\sigma_s}=Y_{\tau^*_{s-1}},\tau^*\geq s\}$ we have $(T-(s-1),Y_{s-1})=(T-(s-1),Y_{\tau^*_{s-1}})=(T-s+1,X_{\sigma_s})\in D$, and so by definition of D necessarily $(\rho^{Root}\wedge(T-s),X_{\sigma_s})\in D$, thus $\rho^{Root}\geq T-s+1$ is fulfilled. The identities (3.6) and (3.8) now yield that F is constant. $\square$



(A) SSRWs for which $\tau^*\geq s$, $\sigma_s=T-s$ and $X_{\sigma_s}=Y_{s-1}$



(B) SSRWs for which $\tau^*\geq s$, $\sigma_s<T-s$ and $X_{\sigma_s}\neq Y_{s-1}$

Figure 4: Illustrating the core argument in the Root setting.

The proof of (3.2) follows similar lines. Given a Y-stopping time τ we consider the interpolating function

$$F^\tau(s)\quad:=\quad\mathbb{E}_y^\lambda\left[|X_{\sigma(\tau\wedge s)}-Y_{\tau\wedge s}|\right]\quad\text{ for } s\in\{0,\dots,T\}. \tag{3.9}$$

Lemma 3.4. *For every $\{0,\dots,T\}$-valued stopping time τ of Y, the function F^τ is increasing and*

$$U_{\mu_T^{Root}}(y)\geq\mathbb{E}_y\left[U_{\mu^{Root}}(Y_\tau)\,\mathbb{1}_{\tau<T}+U_\lambda(Y_\tau)\,\mathbb{1}_{\tau=T}\right]$$

Proof. Clearly $F^\tau(0)=-U_{\mu_T^{Root}}(y)$. On the other hand,

$$
\begin{aligned}
F^\tau(T)&=\mathbb{E}_y^\lambda\left[|X_{\rho^{Root}\wedge(T-\tau)}-Y_\tau|\mathbb{1}_{\tau<T}+|X_0-Y_\tau|\mathbb{1}_{\tau=T}\right]\\
&=-\mathbb{E}_y^\lambda\left[U_{X_{\rho^{Root}\wedge(T-\tau)}}(Y_\tau)\mathbb{1}_{\tau<T}+U_\lambda(Y_\tau)\mathbb{1}_{\tau=T}\right]\\
&\leq-\mathbb{E}_y^\lambda\left[U_{X_{\rho^{Root}}}(Y_\tau)\mathbb{1}_{\tau<T}+U_\lambda(Y_\tau)\mathbb{1}_{\tau=T}\right],
\end{aligned}
$$

where the inequality is a consequence of the potentials $s\mapsto U_{X_{\rho^{Root}\wedge s}}(z)$ being decreasing in s for each z (by Jensen's inequality and optional sampling) and the martingale $\left(X_{\rho^{Root}\wedge t}\right)_{t\in\mathbb{N}}$ being uniformly integrable. Thus if we show that $F^\tau(\cdot)$ is increasing, we can conclude.

Let $0 < s < T$. Define the stopping times $\tau_s := \tau \wedge s$ and $\sigma_s = \rho^{Root} \wedge (T - \tau_s)$. Then, analogous to the proof of Lemma 3.3, but replacing τ^* by τ, we get

$$\begin{aligned} F^\tau(s) &= \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}}| + \mathbb{1}_{X_{\sigma_s} = Y_{\tau_{s-1}}, \tau \geq s} \right] \\ &\geq \mathbb{E}_y^\lambda \left[|X_{\sigma_s} - Y_{\tau_{s-1}}| + \mathbb{1}_{X_{\sigma_s} = Y_{\tau_{s-1}}, \tau \geq s,\, \rho^{Root} \geq T-(s-1)} \right] = F^\tau(s-1), \end{aligned}$$

thus $F^\tau(\cdot)$ is increasing. □

Remark 3.5. The second part of the preceding proof, stating that the function F^τ is increasing, yields after trivial modifications that also

$$s \mapsto F_\sigma^\tau(s) := \mathbb{E}_y^\lambda \left[|X_{\sigma \wedge (T - \tau \wedge s)} - Y_{\tau \wedge s}| \right], \tag{3.10}$$

is increasing. We did not use the particular structure of ρ^{Root} there.

Remark 3.6. There may be many other interpolating functions (which must coincide when $\tau = \tau^*$ of course). For example, if we replace $\sigma(\tau \wedge s) = \rho^{Root} \wedge (T - \tau \wedge s)$ by

$$\sigma(\tau, s) := \begin{cases} \rho^{Root} & \text{if } \tau < s \\ \rho^{Root} \wedge (T - s) & \text{else} \end{cases},$$

and then define

$$\tilde{F}^\tau(s) \;:=\; \mathbb{E}_y^\lambda \left[|X_{\sigma(\tau,s)} - Y_{\tau \wedge s}| \right] \quad \text{for } s \in \{0, \ldots, T\}, \tag{3.11}$$

we have $\tilde{F}^\tau(0) = -U_{\mu_T^{Root}}(y)$ and $\tilde{F}^\tau(T) = -\mathbb{E}_y^\lambda \left[U_{X_{\rho^{Root}}}(Y_\tau) \mathbb{1}_{\tau < T} + U_\lambda(Y_\tau) \mathbb{1}_{\tau = T} \right]$, for each stopping time $\tau \in [0, T]$. This function can be seen to be increasing for each such τ and constant for τ^*.

3.3 The Rost case as a consequence of the Root case

Emboldened by the results in the Root case, we could proceed to establish (2.3)-(2.4) in a SSRW setting via interpolating functions as well. It is much more illuminating and elegant, however, to deduce the Rost case from the Root one. We thus keep the notation as in the previous part.

Proposition 3.7. *For each x, y, T, any stopping time τ for Y such that $\mathbb{E}_y[|Y_\tau|] < \infty$, and every $\{0, \ldots, T\}$-valued stopping time σ for X, we have*

$$\mathbb{E}_y \left[|x - Y_\tau| - |x - Y_{\tau \wedge T}| \right] \leq \mathbb{E}_y^x \left[|X_\sigma - Y_\tau| - |X_\sigma - y| \right]. \tag{3.12}$$

Suppose furthermore that

$$\tau = \inf\{t \in \mathbb{N} : (T - t, Y_t) \notin D\}, \tag{3.13}$$

and that $\sigma = \rho^{Root} \wedge T$. Then there is equality in (3.12).

Proof. We first prove the inequality

$$\mathbb{E}_y \left[|x - Y_\tau| - |x - Y_{\tau \wedge T}| \right] \leq \mathbb{E}_y^x \left[|X_\sigma - Y_\tau| - |X_{\sigma \wedge (T - \tau \wedge T)} - Y_{\tau \wedge T}| \right]. \tag{3.14}$$

This follows, on the one hand, by

$$\mathbb{E}_y \left[\left(|x - Y_\tau| - |x - Y_{\tau \wedge T}| \right) \mathbb{1}_{\tau < T} \right] = 0 \leq \mathbb{E}_y^x \left[\left(|X_\sigma - Y_\tau| - |X_{\sigma \wedge (T - \tau)} - Y_\tau| \right) \mathbb{1}_{\tau < T} \right],$$

where the inequality follows by Jensen's inequality and optional sampling. Similarly we also conclude by Jensen and optional sampling that

$$\mathbb{E}_y \left[\left(|x - Y_\tau| - |x - Y_{\tau \wedge T}| \right) \mathbb{1}_{\tau \geq T} \right] \leq \mathbb{E}_y^x \left[\left(|X_\sigma - Y_\tau| - |X_0 - Y_T| \right) \mathbb{1}_{\tau \geq T} \right].$$

We furthermore note that considering F_σ^τ as defined in (3.10) for the choice $\lambda = \delta_x$ we have that the r.h.s. of (3.12), resp. of (3.14), coincides with $\mathbb{E}_y^x[|X_\sigma - Y_\tau|] - F_\sigma^\tau(0)$, resp. $\mathbb{E}_y^x[|X_\sigma - Y_\tau|] - F_\sigma^\tau(T)$. We can now conclude from Remark 3.5, stating that F_σ^τ is increasing, the desired result (3.12).

In the case $\sigma = \rho^{Root} \wedge T$ and τ fulfilling (3.13), we obtain that $F^\tau_{\rho^{Root}\wedge T} = F^{\tau\wedge T}_{\rho^{Root}\wedge T} = F^{\tau^*} = F$ on $[0,T]$, by (3.13), which by Lemma 3.3 is constant. So to conclude we must show that

$$\mathbb{E}_y[|x - Y_\tau| - |x - Y_{\tau\wedge T}|] = \mathbb{E}_y^x\left[|X_{\rho^{Root}\wedge T} - Y_\tau|\right] - F(T).$$

We can use the arguments in Remark 3.1 resp. 3.2 to obtain

$$\mathbb{E}_y[(|x - Y_\tau| - |x - Y_{\tau\wedge T}|)\,\mathbb{1}_{\tau<T}] = 0 = \mathbb{E}_y^x\left[\left(|X_{\rho^{Root}\wedge T} - Y_\tau| - |X_{\rho^{Root}\wedge(T-\tau)} - Y_\tau|\right)\mathbb{1}_{\tau<T}\right].$$

Similarly also

$$\mathbb{E}_y[(|x - Y_\tau| - |x - Y_{\tau\wedge T}|)\,\mathbb{1}_{\tau\geq T}] = \mathbb{E}_y^x\left[\left(|X_{\rho^{Root}\wedge T} - Y_\tau| - |X_0 - Y_T|\right)\mathbb{1}_{\tau\geq T}\right],$$

which concludes the proof. □

A discrete time version of the Rost optimal stopping problem (2.3)-(2.4) can now be established as a consequence of Proposition 3.7. A Rost continuation set is a set $D^{Rost} \subset \mathbb{N}\times\mathbb{Z}$ satisfying

- If $(t,m) \in D^{Rost}$, then for all $s > t$ also $(s,m) \in D^{Rost}$.

Given such a set for each fixed $T \in \mathbb{N}$ we may define $D := \{(T-t,m) : (t,m) \in D^{Rost}\}$ which is a Root continuation set to which the previous result is applicable. Let us introduce

$$\rho^{Rost} := \inf\{t \in \mathbb{N} : (T-t, Y_t) \notin D\} = \inf\left\{t \in \mathbb{N} : (t, Y_t) \notin D^{Rost}\right\}, \tag{3.15}$$

and let μ^{Rost} (resp. μ_T^{Rost}) denote the law of a SSRW started with distribution λ and stopped at time ρ^{Rost} (resp. $\rho^{Rost} \wedge T$). See Figure 5 for an illustration of the connection between Root and Rost continuation sets and the respective hitting times. We assume uniform integrability of $(Y_{\rho^{Rost}\wedge t})_{t\in\mathbb{N}}$.

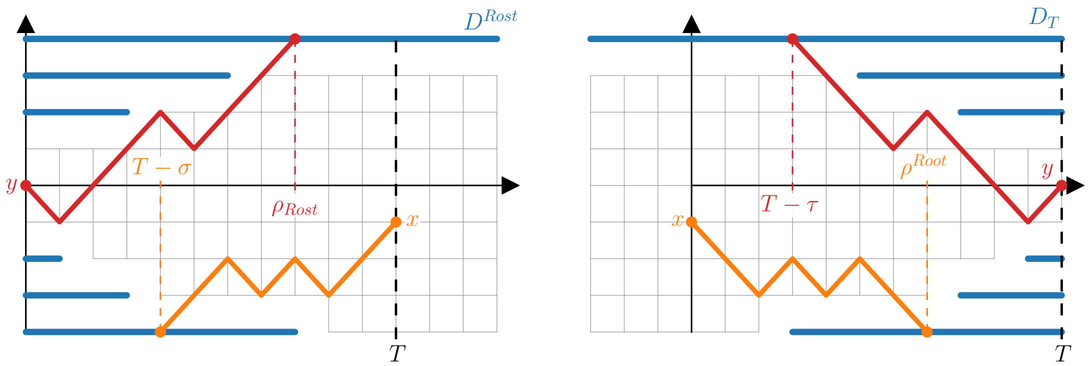


Figure 5: Illustration of the connection between D^{Rost} and D, resp. between ρ^{Rost} and ρ^{Root}.

Corollary 3.8. *We have*

$$U_{\mu^{Rost}}(x) - U_{\mu_T^{Rost}}(x) = \mathbb{E}^x\left[\left(U_{\mu^{Rost}} - U_\lambda\right)(X_{\sigma_*})\right] \tag{3.16}$$

$$= \sup_{\sigma\leq T}\mathbb{E}^x\left[\left(U_{\mu^{Rost}} - U_\lambda\right)(X_\sigma)\right], \tag{3.17}$$

where the optimizer is given by

$$\sigma_* := \rho^{Root} \wedge T = \inf\left\{t \in \mathbb{N} : (T-t, X_t) \notin D^{Rost}\right\} \wedge T.$$

Proof. For $y \in \mathbb{Z}$ let us first consider $Y_0 = y$, i.e. $\lambda = \delta_y$. Consider Proposition 3.7 for the stopping time $\tau = \rho^{Rost}$. As

$$\mathbb{E}_y\left[|x - Y_\tau| - |x - Y_{\tau\wedge T}|\right] = -\left(U_{\mu^{Rost}}(x) - U_{\mu_T^{Rost}}(x)\right),$$
$$\mathbb{E}_y^x\left[|X_\sigma - Y_\tau| - |X_\sigma - y|\right] = -\mathbb{E}^x\left[\left(U_{\mu^{Rost}} - U_\lambda\right)(X_\sigma)\right],$$

due to (3.12) we then have

$$U_{\mu^{Rost}}(x) - U_{\mu_T^{Rost}}(x) \le \sup_{\sigma\le T} \mathbb{E}^x\left[\left(U_{\mu^{Rost}} - U_\lambda\right)(X_\sigma)\right].$$

To prove (3.16) we note that $\tau = \rho^{Rost}$ satisfies (3.13). Thus, for $\sigma = \sigma_*$ we have equality in (3.12) which is precisely (3.16) and furthermore also gives (3.17). As this is true for arbitrary $y \in \mathbb{Z}$, the extension to general λ is clear due to identities of the form $\mathbb{E}_\lambda^x\left[|X_\sigma - Y_\tau|\right] = \sum_{y\in\mathbb{Z}} \mathbb{E}_y^x\left[|X_\sigma - Y_\tau|\right]\lambda(\{y\})$. □

4 The multidimensional case

We have established (2.1)-(2.4) for the integer lattice in one dimension. We shall extend this to the setting of the d-dimensional integer lattice $\mathbb{Z}^d$ for d arbitrary.

Let Z be a SSRW on $\mathbb{Z}^d$ and let

$$z \in \mathbb{Z}^d \mapsto G_n(z) := \mathbb{E}^{Z_0=0}[\#\{t \le n : Z_t = z\}],$$

denote the expected number of visits to site z of Z started in the origin, prior to n. We then consider the so-called *potential kernel* of the SSRW

$$z \in \mathbb{Z}^d \mapsto a(z) := \lim_{n\to\infty} G_n(0) - G_n(z),$$

which is finite in any dimensions and has the desirable property that

$$a(z) = -\mathbb{1}_{z=0} + \frac{1}{2d}\sum_{z'\sim z} a(z'). \tag{4.1}$$

Here $z' \sim z$ if z' is an immediate neighbour of z (corresponding to moving away from z along one coordinate only, so there are $2d$ of them). From this follows that $(a(Z_t))_{t\in\mathbb{N}}$ is a (Markovian) submartingale and by induction

$$\mathbb{E}\left[a(Z_{t+n})|Z_t\right] = a(Z_t) + \sum_{\ell=0}^{n-1} P(Z_{t+\ell} = 0|Z_t), \tag{4.2}$$

which is an identity we will repeatedly use. In the transient case ($d \ge 3$) we have that a is just the negative of the expected number of visits to a point up to an additive constant. In the one dimensional case we have $a(\cdot) = |\cdot|$. We refer to [12, Chapter 1] for a review of these concepts/facts.

We first observe that owing to (4.1) the core argument (3.3) in Section 3.1 is still valid, so for $s \in \{1,\dots,T\}$

$$\mathbb{E}_y^x\left[a(X_{T-s} - Y_s)\right] = \mathbb{E}_y^x\left[a(X_{T-s} - Y_{s-1}) + \mathbb{1}_{X_{T-s}=Y_{s-1}}\right] = \mathbb{E}_y^x\left[a(X_{T-(s-1)} - Y_{s-1})\right].$$

We shall see that all the computations we did using $z \mapsto |z|$ in the one-dimensional case are still valid for the potential kernel a. For a measure ν on $\mathbb{Z}^d$ let

$$A.\nu(y) := -\int a(y-x)\nu(dx).$$

As in the previous section, we denote by X, Y two independent SSRW in $\mathbb{Z}^d$.

Proposition 4.1. *Let λ be a starting distribution in $\mathbb{Z}^d$ and D^{Root} (resp. D^{Rost}) be Root-type (resp. Rost-type) continuation sets in $\mathbb{Z}^{d+1}$. Denote by μ^{Root} resp. μ^{Root}_T the law of a SSRW started with distribution λ and stopped upon leaving D^{Root} resp. $D^{Root} \cap \left(\{0,\dots,T-1\}\times\mathbb{Z}^d\right)$ (analogously for μ^{Rost} and μ^{Rost}_T), and assume that the SSRW stopped when leaving D^{Root} (resp. D^{Rost}) is uniformly integrable. Then*

$$A.\mu^{Root}_T(y) = \mathbb{E}_y\left[A.\mu^{Root}\left(Y_{\tau^*}\right)\mathbb{1}_{\tau^*<T} + A.\lambda\left(Y_{\tau^*}\right)\mathbb{1}_{\tau^*=T}\right] \tag{4.3}$$

$$= \sup_{\tau\le T}\mathbb{E}_y\left[A.\mu^{Root}(Y_\tau)\mathbb{1}_{\tau<T} + A.\lambda(Y_\tau)\mathbb{1}_{\tau=T}\right], \tag{4.4}$$

where the optimizer is $\tau^* := \inf\{t\in\mathbb{N} : (T-t, Y_t)\notin D^{Root}\}\wedge T$, and

$$A.\mu^{Rost}(x) - A.\mu^{Rost}_T(x) = \mathbb{E}^x\left[\left(A.\mu^{Rost} - A.\lambda\right)\left(X_{\tau_*}\right)\right] \tag{4.5}$$

$$= \sup_{\tau\le T}\mathbb{E}^x\left[\left(A.\mu^{Rost} - A.\lambda\right)\left(X_{\tau}\right)\right], \tag{4.6}$$

where the optimizer is $\tau_* := \inf\{t\in\mathbb{N} : (T-t, X_t)\notin D^{Rost}\}\wedge T$.

Proof. Let us first prove (4.3). In analogy to the previous section, we define an interpolating function

$$F(s) \;:=\; \mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}\wedge(T-\tau^*\wedge s)} - Y_{\tau^*\wedge s})\right] \quad \text{for } s\in\{0,\dots,T\}. \tag{4.7}$$

Then clearly $F(0) = -A.\mu^{Root}_T(y)$ and also

$$F(T) = \mathbb{E}^\lambda_y[a(X_{\rho^{Root}\wedge(T-\tau^*)} - Y_{\tau^*})\mathbb{1}_{\tau^*<T}] - \mathbb{E}_y[A.\lambda\left(Y_{\tau^*}\right)\mathbb{1}_{\tau^*=T}].$$

If we establish $-\mathbb{E}^\lambda_y[a(X_{\rho^{Root}\wedge(T-\tau^*)} - Y_{\tau^*})\mathbb{1}_{\tau^*<T}] = \mathbb{E}_y[A.\mu^{Root}\left(Y_{\tau^*}\right)\mathbb{1}_{\tau^*<T}]$ then (4.3) is implied by F being constant. Clearly it suffices to show that

$$\mathbb{E}^\lambda_y[a(X_{T-\tau^*} - Y_{\tau^*})\mathbb{1}_{\tau^*<T,\,\rho^{Root}>T-\tau^*}] = \mathbb{E}^\lambda_y[a(X_{\rho^{Root}} - Y_{\tau^*})\mathbb{1}_{\tau^*<T,\,\rho^{Root}>T-\tau^*}].$$

Indeed

$$\begin{aligned}
&\mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}} - Y_{\tau^*})\mathbb{1}_{\tau^*<T,\,\rho^{Root}>T-\tau^*}\right]\\
&\quad= \mathbb{E}^\lambda_y\left[\mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}} - Y_{\tau^*})\mathbb{1}_{\tau^*<T,\,\rho^{Root}>T-\tau^*}\middle| X_0,\dots,X_T,Y_0,\dots,Y_{T-1}\right]\right]\\
&\quad= \mathbb{E}^\lambda_y\Big[\big(a(X_{T-\tau^*} - Y_{\tau^*}) \qquad\qquad (4.8)\\
&\qquad\quad + \sum_{s=T-\tau^*}^{\rho^{Root}-1} P\left(X_s = Y_{\tau^*}\middle| X_0,\dots,X_T,Y_0,\dots,Y_{T-1}\right)\big)\mathbb{1}_{\tau^*<T,\,\rho^{Root}>T-\tau^*}\Big]\\
&\quad= \mathbb{E}^\lambda_y\left[a(X_{T-\tau^*} - Y_{\tau^*})\mathbb{1}_{\tau^*<T,\,\rho^{Root}>T-\tau^*}\right], \qquad\qquad (4.9)
\end{aligned}$$

where the last line holds since, given $\{X_0,\dots,X_T,Y_0,\dots,Y_{T-1}\}$ on $\{\tau^*<T,\,\rho^{Root}>T-\tau^*\}$,

We now prove that F is indeed constant. First we observe that

$$\begin{aligned}
F(s) &= \mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}\wedge(T-\tau^*\wedge s)} - Y_{\tau^*\wedge s})\right]\\
&= \mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}\wedge(T-\tau^*\wedge s)} - Y_{\tau^*\wedge(s-1)}) + \mathbb{1}_{X_{\rho^{Root}\wedge(T-s)}=Y_{s-1},\tau^*\ge s}\right].
\end{aligned}$$

To see this we consider the two cases $\{\tau^* < s\}$ and $\{\tau^*\ge s\}$ separately. While the former case is clear, on the latter we apply (4.2) where we condition on $\{X_0,\dots,X_{T-s},Y_0,\dots,Y_{s-1}\}$. Analogously but by splitting into $\{\tau^*<s\}\cup\{\rho^{Root}\le T-s\}$ and $\{\tau^*\ge s,\rho^{Root}>T-s\}$ we obtain

$$\begin{aligned}
F(s-1) &= \mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}\wedge(T-\tau^*\wedge(s-1))} - Y_{\tau^*\wedge(s-1)})\right]\\
&= \mathbb{E}^\lambda_y\left[a(X_{\rho^{Root}\wedge(T-\tau^*\wedge s)} - Y_{\tau^*\wedge(s-1)}) + \mathbb{1}_{X_{\rho^{Root}\wedge(T-s)}=Y_{s-1},\tau^*\ge s,\rho^{Root}>T-s}\right].
\end{aligned}$$

We conclude by observing that the two appearing indicator functions are equal, since on the set $\{X_{\rho^{Root}\wedge(T-s)} = Y_{s-1}, \tau^* \geq s\}$ we must necessarily have $\rho^{Root} > T - s$.

To show (4.4) define the multi dimensional equivalent of (3.9), that is for a $\{0, \ldots, T\}$-valued Y-stopping time τ define

$$F(s) \;:=\; \mathbb{E}_y^\lambda \left[a(X_{\rho^{Root}\wedge(T-\tau\wedge s)} - Y_{\tau\wedge s})\right] \quad \text{for } s \in \{0, \ldots, T\}. \tag{4.10}$$

Then clearly $F^\tau(0) = -A.\mu_T^{Root}(y)$. Again we can use (4.2) to show that F^τ is increasing and furthermore

$$F^\tau(T) \leq -\mathbb{E}_y \left[A.\mu^{Root}(Y_\tau)\mathbb{1}_{\tau<T} + A.\lambda(Y_\tau)\mathbb{1}_{\tau=T}\right].$$

The Rost case can be derived from the Root case by analogous arguments as in Section 3.3. A multidimensional version of Proposition 3.7 can be proved verbatim replacing the absolute value by the function a and the Jensen arguments by submartingale arguments. The equality case follows from (4.2) exploiting the barrier structure as it was done for (4.8). $\square$

5 From the random walk setting to the continuous case

While the passage to continuous time is in essence an application of Donsker-type results, we will give a more elaborate explanation using arguments established by Cox and Kinsley in [3] for the one-dimensional case. We note that all results and arguments in Section 3 are invariant under uniform scaling of the space-time grid. Thus for each $N \in \mathbb{N}$ we can consider a rescaled simple symmetric random walk Y^N with space step size $\frac{1}{\sqrt{N}}$ and time step size $\frac{1}{N}$ as it is done in [3]. The authors discretise an *optimal Skorokhod embedding problem*, an (SEP) featuring the following additional optimisation problem

$$\inf_{\tau \text{ solves (SEP)}} \mathbb{E}[F(B_\tau, \tau)]. \tag{OptSEP}$$

It is known that for *any* convex (resp. concave) function $f : \mathbb{R}_+ \to \mathbb{R}_+$, the (OptSEP) with $F(B_\tau, \tau) = f(\tau)$ is solved by a Root (resp. Rost) solution, see e.g. [1]. It is emphasised that the stopping time and the continuation set depend on the measures λ and μ alone and not the specific choice of f.

Let D be a Root (resp. Rost) continuation set and consider the corresponding measure $\mu = \mu^{Root}$ (resp. $\mu = \mu^{Rost}$). Following [3] we obtain for each $N \in \mathbb{N}$ a discretisation μ^N of μ such that $\mu^N \to \mu$ and moreover λ^N and μ^N are in convex order. Similarly a discretisation λ^N of λ can be found such that $\lambda^N \to \lambda$. The authors then propose and solve a discretised version of the (OptSEP) for λ^N and μ^N. The optimiser will again be of Root (resp. Rost) form, given as the first time a (scaled) random walk Y^N leaves a Root (resp. Rost) continuation set $\hat{D}^N$. Let D^N denote a time-continuous and rescaled completion of the discrete continuation set $\hat{D}^N$. In [3, Chapter 5] the authors then prove convergence of D^N to D. We note that in the more general setting considered in [3] a recovery of the initial continuation set D is not guaranteed. However, in the Root case this follows due to [13]. An analogous uniqueness result for Rost solutions is also true, see e.g. [10] for a generalisation.

By convergence of the continuation sets it is easy to see that for every $T \geq 0$ we have $\mu_T^N \to \mu_T$. As in this setting convergence of measures implies uniform convergence of potential functions, $U_{\mu_T^N} \to U_{\mu_T}$ (see [2] for details), we have established convergence of the l.h.s of (3.1) to the l.h.s of (2.1).

Let $\left(W_t^{(N)}\right)_{t\geq 0}$ denote the continuous version of the rescaled random walk Y^N. To avoid heavy usage of floor functions, we will assume $T \in I := \left\{\frac{m}{2^n} : m, n \in \mathbb{N}\right\}$. If limits are then taken along the subsequence $\left(Y^{2^n}\right)_{n\in\mathbb{N}}$ (resp. $\left(W^{(2^n)}\right)_{n\in\mathbb{N}}$) there exists an $N_0 \in \mathbb{N}$ such that T will always be a multiple of the step size $\frac{1}{2^n}$ for all $n \geq N_0$. For arbitrary $T > 0$ the results can be recovered via density arguments. We define the following stopping times

$$\begin{aligned} \hat{\tau}^{N*} &= \inf\left\{t \in \mathbb{N} : (NT - t, Y_t^N) \notin \hat{D}^N\right\} \wedge NT, \\ \bar{\tau}^{N*} &= \inf\left\{t > 0 : (T - t, W_t^{(N)}) \notin D^N\right\} \wedge T, \\ \tau^* &= \inf\left\{t > 0 : (T - t, W_t) \notin D\right\} \wedge T, \end{aligned} \tag{5.1}$$

and the functions

$$\begin{aligned} G^T(x,t) &:= U_\mu(x)\mathbb{1}_{t<T} + U_\lambda(x)\mathbb{1}_{t=T}, \\ G_N^T(x,t) &:= U_{\mu^N}(x)\mathbb{1}_{t<T} + U_{\lambda^N}(x)\mathbb{1}_{t=T}. \end{aligned}$$

The rescaled results of Section 3 then read

$$U_{\mu_T^N}(x) = \mathbb{E}^x\left[G_N^T\left(Y_{\hat{\tau}^{N*}}^N, \frac{\hat{\tau}^{N*}}{N}\right)\right] \tag{3.1*}$$

$$= \sup_{\frac{\tau}{N} \leq T} \mathbb{E}^x\left[G_N^T\left(Y_\tau^N, \frac{\tau}{N}\right)\right]. \tag{3.2*}$$

Or, as $\left(Y_{\hat{\tau}^{N*}}^N, \frac{\hat{\tau}^{N*}}{N}\right) = \left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right)$ we consider (3.1*) in $W^{(N)}$-terms

$$U_{\mu_T^N}(x) = \mathbb{E}^x\left[G_N^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right)\right]. \tag{3.1**}$$

By Lemma 5.5 and 5.6 of [3] we know $\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right) \xrightarrow{P} (W_{\tau^*}, \tau^*)$ as $N \to \infty$. To see convergence of (3.1**) to (2.1) we need to show

$$\begin{aligned} &\mathbb{E}^x\left[|G_N^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right) - G^T\left(W_{\tau^*}, \tau^*\right)|\right] \\ &\leq \mathbb{E}^x\left[|G_N^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right) - G^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right)|\right] \\ &+ \mathbb{E}^x\left[|G^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right) - G^T\left(W_{\tau^*}, \tau^*\right)|\right] \stackrel{N\to\infty}{\longrightarrow} 0. \end{aligned}$$

Convergence of the first term is clear due to the fact that uniform convergence of the potential functions implies uniform convergence of G_N^T to G^T. Thus it remains to show convergence of the second term. Note that G^T is usc, so it suffices to show that

$$\mathbb{E}^x[G^T(W_{\tau^*}, \tau^*)] \leq \liminf_{N\to\infty} \mathbb{E}^x\left[G^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right)\right]. \tag{5.2}$$

For this, given $\varepsilon > 0$ consider the auxiliary function

$$\tilde{G}^\varepsilon(x,t) := U_\mu(x)\mathbb{1}_{t\leq T-\varepsilon} + U_\lambda(x)\mathbb{1}_{T-\varepsilon<t\leq T}.$$

Then for any random variable X and stopping time τ we have

$$\mathbb{E}^x\left[|G^T(X,\tau) - \tilde{G}^\varepsilon(X,\tau)|\right] \leq c \cdot P\left[\tau \in (T-\varepsilon, T)\right].$$

Combining this with the fact that $\tilde{G}^\varepsilon$ is lsc and dominating G^T we get

$$\begin{aligned} \mathbb{E}^x\left[G^T\left(W_{\tau^*}, \tau^*\right)\right] &\leq \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty} \mathbb{E}^x\left[\tilde{G}^\varepsilon\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right)\right] \\ &\leq \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty} c\cdot P\left[\hat{\tau}^{N*} \in (T-\varepsilon, T)\right] + \liminf_{N\to\infty} \mathbb{E}^x\left[G^T\left(W_{\bar{\tau}^{N*}}^{(N)}, \bar{\tau}^{N*}\right)\right]. \end{aligned}$$

Thus we are left to show that $\lim_{\varepsilon\searrow 0}\liminf_{N\to\infty} P\left[\hat{\tau}^{N*}\in(T-\varepsilon,T)\right]=0$. To more easily see the arguments involving specific barrier structures, we consider the following stopping times

$$\begin{aligned}\bar{\rho}^N &= \inf\left\{t>0:(T-t,W_t^{(N)})\notin D^N\right\}=\inf\left\{t>0:(t,W_t^{(N)})\notin\tilde{D}^N\right\},\\ \rho &= \inf\left\{t>0:(T-t,W_t)\notin D\right\}=\inf\left\{t>0:(t,W_t)\notin\tilde{D}\right\},\end{aligned}$$

where $\tilde{D}^N$ resp. $\tilde{D}$ is the *Rost* continuation set we obtain by reflecting D^N resp. D along $\left\{\frac{T}{2}\right\}\times\mathbb{R}$. By [3, Chapter 5] we know that $\bar{\rho}^N\xrightarrow{P}\rho$. Note that we have $\bar{\rho}^N\mathbb{1}_{\bar{\rho}^N<T}=\bar{\tau}^{N*}\mathbb{1}_{\bar{\tau}^{N*}<T}$. For $0<\tilde{T}\le T$ consider

$$\begin{aligned}x_- &:= \sup\left\{y<x:(\tilde{T},y)\in\tilde{D}\right\},\\ x_+ &:= \inf\left\{y>x:(\tilde{T},y)\in\tilde{D}\right\}.\end{aligned}$$

Since $\tilde{D}$ is a Rost continuation set and ρ is its Brownian hitting time, we have

$$P[\rho=\tilde{T}]=P[W_{\tilde{T}}\in\{x_-,x_+\}]=0.$$

So, especially for any $\varepsilon>0$ we have $P[\rho=T-\varepsilon]=P[\rho=T]=0$. Altogether we have

$$\begin{aligned}\lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}P\left[\bar{\tau}^{N*}\in(T-\varepsilon,T)\right] &= \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}P\left[\bar{\rho}^N\in(T-\varepsilon,T)\right]\\ &= \lim_{\varepsilon\searrow 0}P\left[\rho\in(T-\varepsilon,T)\right]=0,\end{aligned}$$

which concludes the proof of (5.2), thus the proof of convergence of (3.1**) to (2.1). It only remains to show (2.2). So let $\bar{\tau}$ be an optimiser of (2.2). Lemma 5.2 in [3] then gives a discretisation $\tilde{\sigma}^N$ of $\bar{\tau}$ for which $Y^N_{\tilde{\sigma}^N}\xrightarrow{a.s.}W_{\bar{\tau}}$ and $\frac{\tilde{\sigma}^N}{N}\xrightarrow{P}\bar{\tau}$.

To obtain the other inequality, for $\varepsilon\in I$ define the function

$$\tilde{G}^\varepsilon_N(x,t):=U_{\mu^N}(x)\mathbb{1}_{t\le T-\varepsilon}+U_{\lambda^N}(x)\mathbb{1}_{T-\varepsilon<t\le T},$$

and by $\hat{\tau}^{N*}_\varepsilon$ resp. τ^*_ε consider the respective stopping times defined in (5.1), replacing T by $T-\varepsilon$. Then

$$\begin{aligned}\sup_{\tau\le T}\mathbb{E}^x\left[G^T(W_\tau,\tau)\right] &= \mathbb{E}^x\left[G^T\left(W_{\bar{\tau}},\bar{\tau}\right)\right]=\lim_{\varepsilon\searrow 0}\mathbb{E}^x\left[\tilde{G}^\varepsilon(W_{\bar{\tau}},\bar{\tau})\right] &(5.3)\\ &\le \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\mathbb{E}^x\left[\tilde{G}^\varepsilon_N\left(Y^N_{\tilde{\sigma}^N},\frac{\tilde{\sigma}^N}{N}\right)\right]\le\lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\sup_{\frac{\tau}{N}\le T}\mathbb{E}^x\left[\tilde{G}^\varepsilon_N\left(Y^N_\tau,\frac{\tau}{N}\right)\right] &(5.4)\\ &\le \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\sup_{\frac{\tau}{N}\le T-\varepsilon}\mathbb{E}^x\left[G^{T-\varepsilon}_N\left(Y^N_\tau,\frac{\tau}{N}\right)\right]=\lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\mathbb{E}^x\left[G^{T-\varepsilon}_N\left(Y^N_{\hat{\tau}^{N*}_\varepsilon},\frac{\hat{\tau}^{N*}_\varepsilon}{N}\right)\right] &(5.5)\\ &= \lim_{\varepsilon\searrow 0}\mathbb{E}^x\left[G^{T-\varepsilon}\left(W_{\tau^*_\varepsilon},\tau^*_\varepsilon\right)\right]=\lim_{\varepsilon\searrow 0}U_{\mu_{T-\varepsilon}}(x)=U_{\mu_T}(x)=\mathbb{E}^x\left[G^T\left(W_{\tau^*},\tau^*\right)\right]. &(5.6)\end{aligned}$$

The fact that $\lim_{\varepsilon\searrow 0}P\left[\bar{\tau}\in(T-\varepsilon,T)\right]=0$ gives (5.3) and that $\tilde{G}^\varepsilon$ is l.s.c gives (5.4). To see (5.5) consider the function

$$H^\varepsilon_N(x,t):=U_{\mu^N}(x)\mathbb{1}_{t<T-\varepsilon}+U_{\lambda^N}(x)\mathbb{1}_{T-\varepsilon\le t\le T}.$$

Then $H^\varepsilon_N(x,t)\ge G^\varepsilon_N(x,t)$ for all $(x,t)\in\mathbb{R}\times[0,T]$ and trivially

$$\sup_{\frac{\tau}{N}\le T}\mathbb{E}^x\left[G^\varepsilon_N\left(Y^N_\tau,\frac{\tau}{N}\right)\right]\le\sup_{\frac{\tau}{N}\le T}\mathbb{E}^x\left[H^\varepsilon_N\left(Y^N_\tau,\frac{\tau}{N}\right)\right].$$

Let $(Z)_{t\geq 0}$ be a martingale, then $\left(H_N^{\varepsilon}\left(Z_t,t\right)\right)_{t\in[T-\varepsilon,T]}=\left(U_{\lambda^N}\left(Z_t\right)\right)_{t\in[T-\varepsilon,T]}$ is a supermartingale as U_{λ^N} is a concave function. So for any stopping time τ we have

$$\mathbb{E}^x\left[H_N^{\varepsilon}\left(Z_{\tau\wedge(T-\varepsilon)},\tau\wedge(T-\varepsilon)\right)\right]\geq\mathbb{E}^x\left[H_N^{\varepsilon}\left(Z_{\tau\wedge T},\tau\wedge T\right)\right]. \tag{5.7}$$

We see that no optimiser of $\sup_{\frac{\tau}{N}\leq T}\mathbb{E}^x\left[H_N^{\varepsilon}\left(Y_\tau^N,\frac{\tau}{N}\right)\right]$ will stop after time $T-\varepsilon$, as this would decrease the value of the objective function. So we have

$$\sup_{\frac{\tau}{N}\leq T}\mathbb{E}^x\left[H_N^{\varepsilon}\left(Y_\tau^N,\frac{\tau}{N}\right)\right]=\sup_{\frac{\tau}{N}\leq T-\varepsilon}\mathbb{E}^x\left[H_N^{\varepsilon}\left(Y_\tau^N,\frac{\tau}{N}\right)\right]=\sup_{\frac{\tau}{N}\leq T-\varepsilon}\mathbb{E}^x\left[G_N^{T-\varepsilon}\left(Y_\tau^N,\frac{\tau}{N}\right)\right].$$

As we know that $\hat{\tau}_\varepsilon^{N*}$ is the optimiser of this optimal stopping problem, (5.5) follows. Lastly, (5.6) is due to the convergence result of (3.1*) to (2.1).

To prove convergence of the Rost optimal stopping problem replace the functions G^T and G_N^T above by the following functions

$$\begin{aligned}G^T(x,t)&=G(x):=U_\mu(x)-U_\lambda(x),\\ G_N^T(x,t)&=G_N(x):=U_{\mu^N}(x)-U_{\lambda^N}(x).\end{aligned}$$

We can now derive our convergence results analogous to the Root case.

6 Perspectives

We illustrated the elusive connection between Root and Rost's solutions to the (SEP) and optimal stopping problems. Specialising to the simplest possible setting, this note restricts itself to the case of SSRW and Brownian motion. In a recent article by Gassiat et. al. [9] the analytic connection between Root solutions to the (SEP) and solutions to optimal stopping problems was established for a much more general class of Markov processes. This suggests that our probabilistic arguments would also hold in this generalised setting. The extension to more general martingales should follow via analogous arguments to the extension made in Chapter 4 by using the appropriate potential kernel, however for non-martingales some arguments need to be replaced.

References

[1] M. Beiglböck, A. M. G. Cox, and M. Huesmann. Optimal transport and Skorokhod embedding. *Inventiones mathematicae* 208.2 (May 2017), pp. 327–400.

[2] R. V. Chacon. Potential processes. *Trans. Amer. Math. Soc.* 226 (1977), pp. 39–58.

[3] A. M. G. Cox and S. M. Kinsley. Discretisation and duality of optimal Skorokhod embedding problems. *Stochastic Processes and their Applications* 129.7 (2019), pp. 2376–2405.

[4] A. M. G. Cox and J. Wang. Optimal robust bounds for variance options. *ArXiv e-prints* (Aug. 2013).

[5] A. M. G. Cox and J. Wang. Root's Barrier: Construction, Optimality and Applications to Variance Options. *Ann. Appl. Probab.* 23.3 (2013), pp. 859–894.

[6] A. M. G. Cox, J. Obłój, and N. Touzi. The Root solution to the multi-marginal embedding problem: an optimal stopping and time-reversal approach. *Probab. Theory Related Fields* 173.1-2 (2019), pp. 211–259.

[7] T. De Angelis. From optimal stopping boundaries to Rost's reversed barriers and the Skorokhod embedding. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.2 (2018), pp. 1098–1133.

[8] N. El Karoui, J.-P. Lepeltier, and A. Millet. A probabilistic approach to the reduite in optimal stopping. *Probab. Math. Statist.* 13.1 (1992), pp. 97–121.

[9] P. Gassiat, H. Oberhauser, and C. Z. Zou. A free boundary characterization of the Root barrier for Markov processes. *Probability Theory and Related Fields* 180 (2021), pp. 33–69.

[10] A. M. Grass. Uniqueness properties of barrier type Skorokhod embeddings and Perkins embedding with general starting law. *Master Thesis* (2017). Available online at `https://www.mat.univie.ac.at/~grass/`.

[11] D. Hobson. The Skorokhod embedding problem and model-independent bounds for option prices. *Paris-Princeton Lectures on Mathematical Finance 2010.* Vol. 2003. Lecture Notes in Math. Springer, Berlin, 2011, pp. 267–318.

[12] G. F. Lawler. Intersections of random walks. Modern Birkhäuser Classics. Reprint of the 1996 edition. Birkhäuser/Springer, New York, 2013, pp. iv+223. ISBN: 978-1-4614-5971-2; 978-1-4614-5972-9.

[13] R. M. Loynes. Stopping times on Brownian motion: Some properties of Root's construction. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 16 (1970), pp. 211–218.

[14] T. R. McConnell. The two-sided Stefan problem with a spatially dependent latent heat. *Trans. Amer. Math. Soc.* 326.2 (1991), pp. 669–699.

[15] J. Obłój. The Skorokhod embedding problem and its offspring. *Probab. Surv.* 1 (2004), pp. 321–390.

[16] D. H. Root. The existence of certain stopping times on Brownian motion. *Ann. Math. Statist.* 40 (1969), pp. 715–718.

[17] H. Rost. Skorokhod stopping times of minimal variance. *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975).* Berlin: Springer, 1976, 194–208. Lecture Notes in Math., Vol. 511.

[18] A. V. Skorohod. Issledovaniya po teorii sluchainykh protsessov (Stokhasticheskie differentsialnye uravneniya i predelnye teoremy dlya protsessov Markova). Izdat. Kiev. Univ., Kiev, 1961, p. 216.

[19] A. V. Skorokhod. Studies in the theory of random processes. Translated from the Russian by Scripta Technica, Inc. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1965, pp. viii+199.

Delayed Switching Identities and Multi-Marginal Solutions to the Skorokhod Embedding Problem

A. M. G. Cox and A. M. Grass

Abstract

In this article, we consider a generalisation of the Skorokhod embedding problem (SEP) with a *delayed* starting time. In the delayed SEP, we look for stopping times which embed a given measure in a stochastic process, which occur after a given delay time. Our first contribution is to show that the switching identities introduced in a recent paper of Backhoff, Cox, Grass and Huesmann extend to the case with a delay.

We then show that the delayed switching identities can be used to establish an optimal stopping representation of Root and Rost solutions to the multi-marginal Skorokhod embedding problem. We achieve this by rephrasing the multi-period problem into a one-period framework with delay. This not only recovers the known multi-marginal representation of Root, but also establishes a previously unknown optimal stopping representation associated to the multi-marginal Rost solution. The Rost case is more complex than the Root case since it naturally requires randomisation for general initial measures, and we develop the necessary tools to develop these solutions. Our work also provides a comprehensive and complete treatment of *discrete* Root and Rost solutions, embedding discrete measures into simple symmetric random walks.

Keywords: Skorokhod embedding problem; Root barrier; Rost barrier; optimal stopping; switching identities.

MSC 2020 Primary 60G40, 60J65, 90C41; secondary 91G20

1 Introduction

We consider the *Skorokhod embedding problem*, that is given a Brownian motion $(W_t)_{t\geq 0}$ started according to an initial distribution $W_0 \sim \lambda$, and a probability distribution μ on $\mathbb{R}$, the question is to find a stopping time τ such that

$$W_\tau \sim \mu \text{ and } (W_{t\wedge\tau})_{t\geq 0} \text{ is uniformly integrable.} \tag{SEP}$$

While this problem was originally posed and solved by Skorokhod [37, 38] in the early 60s, it remains an active field of research to the present day. The 2004 survey paper [31] features more than 20 distinct solution to (SEP) and more recent contributions include [2, 3, 6, 7, 10, 11, 13, 14, 15, 18, 19, 20, 21, 22, 24, 29, 32] among others.

Of primary interest in this article will be an extension of the classical (SEP) to *multiple marginals*. Given a sequence of measures $\mu_1, \dots, \mu_n$ on the real line, the *multi-marginal Skorokhod embedding problem* is to find an increasing sequence of stopping times $\tau_1 \leq \cdots \leq \tau_n$ such that

$$W_{\tau_k} \sim \mu_k,\ k = 1, \dots, n \text{ and } (W_{t\wedge\tau_n})_{t\geq 0} \text{ is uniformly integrable.} \tag{MMSEP}$$

It is an application of Strassen's Theorem [39] and a well know fact that (MMSEP) will admit a solution if and only if the measures are in increasing convex order, that is $\lambda \leq_c \mu_1 \leq_c \cdots \leq_c \mu_n$. Throughout the paper we will assume this condition satisfied.

This extension to (MMSEP) is particularly motivated by applications in robust mathematical finance given the connections between the (SEP) and robust option pricing first proposed in Hobson's paper [25]; see also [26] for an extensive survey. These connections enable the computation of robust bounds on option prices while considering a market-given marginal distribution at the time of expiration. Extending this theory to multiple marginals would allow the incorporation of additional market data into the robust pricing problem. We develop this approach in a companion paper, [12].

The focal point of this paper are two fundamental and influential solutions to (SEP) known as Root [34] and Rost [35] barrier solutions. These stopping times can be represented as hitting times of subsets $R \subseteq \mathbb{R}_+ \times \mathbb{R}$, which are subject to some type of *barrier structure*, precisely:

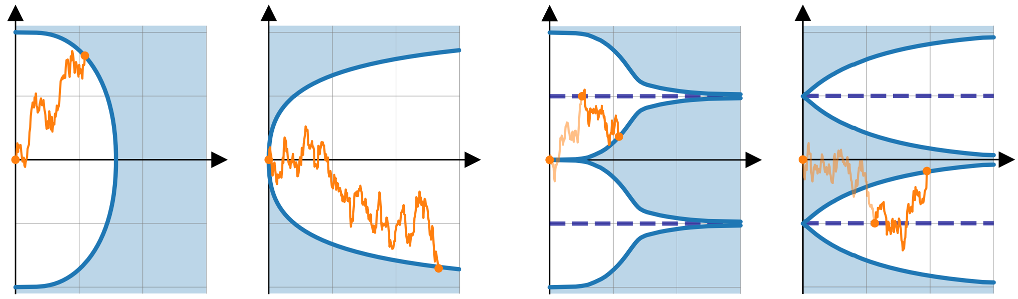
$$\begin{aligned} &\bullet \text{ If } (t,x) \in R^{Root} \text{ then } (s,x) \in R^{Root} \text{ for all } s > t. \\ &\bullet \text{ If } (t,x) \in R^{Rost} \text{ then } (s,x) \in R^{Rost} \text{ for all } s < t. \end{aligned} \tag{1.1}$$

See Figure 1 for an illustration.

Our interest in Root and Rost solutions is twofold. First, there is a substantial interest in the probability literature on these solutions, see for example [2, 4, 8, 11, 14, 15, 17, 18, 19, 20, 28, 33, 35]. Second, such solutions are related to robust bounds on variance options considering multiple market-given marginal distributions. In [12] we carry out an empirical exploration of such robust bounds based on the results presented here. See [9, 11, 15, 27, 30] for related work in the financial context.

While existence results of Root and Rost barriers were shown by their originators and also an extension to multiple marginals was given in [6], as was noted in Root's original work [34], concrete barriers are only known in the most trivial of situations. To actually construct a concrete barrier R given a probability measure μ or even more so a sequence barriers given a sequence of probability measures $\mu_1, \dots, \mu_n$ is a highly non trivial task.

Prior research in [11, 15, 17, 18, 19, 20] established that Root and Rost barriers which solve the (single-marginal) (SEP) can be recovered via solutions to a respective associated PDE in variational form. Solutions to such PDEs, in turn, are known to have a representation by means of an optimal stopping problem. While for a long time this optimal stopping representation of Root and Rost solutions was exclusively accessible through this detour over PDEs in variational form, a direct proof utilizing time-reversal arguments was given in [14], notably for the multi-marginal Root case.



(A) Root (left) and Rost (right) barriers solving (SEP) for $\lambda = \delta_0$ and $\mu = \mathsf{U}[-2, 2]$.

(B) Root (left) and Rost (right) barriers solving (MMSEP) for $\lambda = \delta_0$, $\mu_1 = \frac{1}{2}(\delta_{-1} + \delta_1)$ and $\mu_2 = \mathsf{U}[-2, 2]$.

Figure 1: Single-marginal (a) and multi-marginal (b) Root and Rost solutions embedding the measure $\mathsf{U}[-2, 2]$, one of the examples specifically mentioned [34] for its previously unknown barrier structure.

It was furthermore the aim of the article [1] to give a simple probabilistic argument connecting Root and Rost solutions to their respective optimal stopping representations in the single marginal case by establishing a specific switching identity. To the best of our knowledge, an optimal stopping representation associated to the multi-marginal Rost problem remained unknown to this day.

In order to establish such a multi-marginal representation, a crucial first step is to reduce the multi-step viewpoint to a single-marginal perspective where the initial marginal is now a space-time law instead of a spatial law at zero. This approach builds on an approach implicit in [14], and which we make explicit in our approach. Our solution is computed in terms of a *delay* stopping time inducing an initial space-time law after which we embed with a Root or Rost barrier respectively. Given such a representation it becomes apparent how a multi-marginal solution can be recovered inductively.

An initial step in our process is to extend the results established in [6] as well as [1] to allow for such a delay. There are two implications of this extension: First, it enables a more concise and intuitive recovery of the multi-marginal optimal stopping problem given in [14]. Second, we can provide a novel characterisation of the multi-marginal Rost embedding in terms of the solution to a given optimal stopping (or multiple stopping) problem. This lays the foundation for constructing Rost barriers such that their hitting times solve the (MMSEP) computable as demonstrated in Figure 1.

We briefly review the known results for the single marginal problem and then proceed to state its multi-marginal extension.

Summary of single-marginal results

Simple probabilistic arguments justifying the optimal stopping characterisation of the single marginal Root and Rost (SEP) were given in [1]. Let us denote by μ^{Root} (resp. μ^{Rost}) the law of the Brownian motion started with distribution λ at the time it hits R^{Root} (resp. R^{Rost}), and by μ_T^{Root} (resp. μ_T^{Rost}) the law at the time it hits the barrier $R^{Root} \cup ([T, \infty) \times \mathbb{R})$ (resp. $R^{Rost} \cup ([T, \infty) \times \mathbb{R})$).

Recall the potential of a measure m, denoted by

$$U_m(y) := -\int |y - x| \, m(\mathrm{d}x).$$

The single-marginal relations of interest, as provided in [1, 11, 15, 20] among others are

$$U_{\mu_T^{Root}}(y) = \mathbb{E}_y\left[U_{\mu^{Root}}\left(W_{\tau^*}\right)\mathbb{1}_{\tau^*<T} + U_\lambda\left(W_{\tau^*}\right)\mathbb{1}_{\tau^*=T}\right] \tag{1.2}$$

$$= \sup_{\tau\leq T}\mathbb{E}_y\left[U_{\mu^{Root}}(W_\tau)\mathbb{1}_{\tau<T} + U_\lambda(W_\tau)\mathbb{1}_{\tau=T}\right], \tag{1.3}$$

where the optimiser is $\tau^* := \inf\left\{t \geq 0 : (T-t, W_t) \in R^{Root}\right\} \wedge T$,
and

$$U_{\mu^{Rost}}(y) - U_{\mu_T^{Rost}}(y) = \mathbb{E}_y\left[\left(U_{\mu^{Rost}} - U_\lambda\right)\left(W_{\tau^*}\right)\right] \tag{1.4}$$

$$= \sup_{\tau\leq T}\mathbb{E}_y\left[\left(U_{\mu^{Rost}} - U_\lambda\right)\left(W_\tau\right)\right], \tag{1.5}$$

where the optimiser is $\tau^* := \inf\left\{t \geq 0 : (T-t, W_t) \in R^{Rost}\right\} \wedge T$.

Multi-marginal extensions

We are interested in Equations (1.2)–(1.5) in a multi marginal setting. While the multi marginal extension of the Root case (1.2)–(1.3) is known due to [14], the Rost equivalent appears to be novel. The results can be stated as follows.

Let $R_1, R_2, \ldots, R_n$ denote a sequence of Root (resp. Rost) barriers as defined in (1.1). Then consider the associated Root (resp. Rost) stopping times

$$\rho_1 := \inf\left\{t > 0 : (t, W_t) \in R_1\right\},$$
$$\rho_k := \inf\left\{t > \rho_{k-1} : (t, W_t) \in R_k\right\}, \text{ for } k \in \{2, \ldots, n\}.$$

Define the respective laws $\mu_i := \mathcal{L}^\lambda(W_{\rho_i})$ and $\mu_{i,T} := \mathcal{L}^\lambda(W_{\rho_i \wedge T})$. More precisely, μ_i will be the law of the Brownian motion starting with distribution λ at the time it consecutively hit all of the barrier sets $R_1, \ldots, R_i$ respectively, and $\mu_{i,T}$ will be the law at the time it consecutively hit the barrier sets $R_1 \cup ([T,\infty)\times\mathbb{R}), \ldots, R_i \cup ([T,\infty)\times\mathbb{R})$. Then these measures are in convex order, i.e. $\mu_1 \leq_c \mu_2 \leq_c \cdots \leq_c \mu_n$ as well as $\mu_{1,T} \leq_c \mu_{2,T} \leq_c \cdots \leq_c \mu_{n,T}$. Moreover, define the function,

$$U_{\mu_{0,T}}(x) := U_{\mu_0}(x) := U_\lambda(x) \quad \text{for all } T \geq 0. \tag{1.6}$$

The multi-marginal extension of our relations of interest are given as follows.

Theorem 1.1. *If (R_k) denotes a sequence of Root barriers, then*

$$U_{\mu_{k,T}}(y) = \mathbb{E}_y\left[U_{\mu_{k-1,T-\tau_k^*}}(W_{\tau_k^*}) + (U_{\mu_k} - U_{\mu_{k-1}})(W_{\tau_k^*})\mathbb{1}_{\{\tau_k^*<T\}}\right] \tag{1.7}$$

$$= \sup_{\tau\leq T}\mathbb{E}_y\left[U_{\mu_{k-1,T-\tau}}(W_\tau) + (U_{\mu_k} - U_{\mu_{k-1}})(W_\tau)\mathbb{1}_{\{\tau<T\}}\right], \tag{1.8}$$

where the optimiser is $\tau_k^* := \inf\left\{t \geq 0 : (T-t, W_t) \in R_k\right\} \wedge T$.

Moreover, given the interpolating potentials $U_{\mu_{k,T}}(y)$ for all $(T,y) \in [0,\infty)\times\mathbb{R}$ we can recover the Root barriers $R_1, \ldots, R_n$ embedding the measures $\mu_1, \ldots, \mu_n$ in the following way

$$R_k = \left\{(T,y) : \left(U_{\mu_{k,T}} - U_{\mu_{k-1,T}}\right)(y) = \left(U_{\mu_k} - U_{\mu_{k-1}}\right)(y)\right\}, \ k \in \{1, \ldots, n\}.$$

Theorem 1.2. *If $(\bar{R}_k)$ denotes a sequence of Rost barriers, then*

$$U_{\mu_k}(y) - U_{\mu_{k,T}}(y) = \mathbb{E}_y\left[U_{\mu_k}(W_{\tau_k^*}) - U_{\mu_{k-1,T-\tau_k^*}}(W_{\tau_k^*})\right] \tag{1.9}$$

$$= \sup_{\tau\leq T}\mathbb{E}_y\left[U_{\mu_k}(W_\tau) - U_{\mu_{k-1,T-\tau}}(W_\tau)\right] \tag{1.10}$$

where the optimiser is $\tau_k^* := \inf\left\{t \geq 0 : (T-t, W_t) \in \bar{R}_k\right\} \wedge T$.

Moreover, given the interpolating potentials $U_{\mu_{k,T}}(y)$ for all $(T,y) \in [0,\infty)\times\mathbb{R}$ we can recover the Rost barriers $\bar{R}_1, \ldots, \bar{R}_n$ embedding the measures $\mu_1, \ldots, \mu_n$ in the following way

$$\bar{R}_k = \left\{(T,y) : U_{\mu_{k,T}}(y) = U_{\mu_{k-1,T}}(y)\right\}, \ k \in \{1, \ldots, n\}.$$

Outline of the paper

In Section 2 we introduce the notion of *delayed* Skorokhod embeddings, which contain the multi-marginal Skorokhod embedding as a special case. We will establish a general existence of such solution following the approach in [6] and present an optimal stopping representation theorem.

Inspired by the framework proposed in [1], we first provide a analogous statements for simple symmetric random walks in Section 3. First we show general existence of delayed Root and Rost solution in this context, subsequently we state and prove the corresponding respective optimal stopping representation.

Section 4 is dedicated to providing a suitable discretisation of the continuous-time framework and of the respective optimal stopping representations such that the results of Section 3 apply. We then take a Donsker-type limit back to a Brownian motion, recovering the continuous times results stated in Section 2.

2 Delayed Root and Rost embeddings

For a more convenient and comprehensive treatment of (MMSEP), but also present results in greater generality we introduce a relatively recent extension of the classical (SEP) known as *delayed* embeddings. Although the concept of delayed embeddings has been implicitly explored in [14], to best of our knowledge the specific name was first mentioned in [8].

We consider a *delay stopping time* η and require our solution to (SEP) to wait for this delay before embedding the measure μ.

Formally, given a measure μ, a Brownian motion $(W_t)_{t\geq 0)}$ started according to an initial distribution $W_0 \sim \lambda$ and a (possibly randomised) delay stopping time η the *delayed Skorokhod embedding problem* is to find a stopping time τ such that

$$W_\tau \sim \mu \text{ where } \tau \geq \eta \text{ and } (W_{\tau\wedge t})_{t\geq 0} \text{ is uniformly integrable.} \tag{dSEP}$$

We will denote a (dSEP) by the triple (λ, η, μ) and assume throughout the paper that η is an integrable stopping time and that μ dominates $\mathcal{L}^\lambda(W_\eta)$ in the convex order. Trivially, the (dSEP) has no solution if either of these conditions are violated.

2.1 Root and Rost solutions to (dSEP)

We are interested in Root and Rost solutions to (dSEP). In [5] the necessary theory and tools were given to identify specific solutions to (SEP) as stopping times that are optimisers over appropriate additional optimisation criteria. Such (SEP) with additional optimisation criteria are denoted by (OptSEP). In [6] this concept was extended to multiple marginals and denoted by (OptMSEP). Especially multi-marginal Root and Rost solutions were recovered as solutions to appropriate (OptMSEP).

We can further ask whether we can recover the delayed Root and Rost embeddings as optimisers over an appropriate optimisation criteria. The answer to this is positive, and is a relatively simple application of the results of [6]. We work in the class of *randomised multi-stopping times* introduced in [6], RMST_2^1, which we understand to be probability measures on the space $\mathbb{R} \times C_0(\mathbb{R}_+) \times \Xi^2$, where $\Xi^2 = \{(s_1, s_2) \in \mathbb{R}_+^2, s_1 \leq s_2\}$, and where the measure projected on its first two coordinates is equal to the measure $\lambda \times \mathbb{W}$, with $\mathbb{W}$ the Brownian path measure on $C_0(\mathbb{R}_+)$. The Ξ^2 variables represent the first and second stopping times respectively. The measure has further constraints on its structure (see [6], Definition 3.9), which ensure that the stopping time properties are respected.

We are interested in optimising a functional of the form $\mathbb{E}[F(W_{\tau_2}, \tau_2)]$ over the class of randomised stopping times (τ_1, τ_2) which satisfy the embedding constraint, $W_{\tau_2} \sim \mu$, and for which $\tau_1 = \eta$ is the given (randomised) delay stopping time embedding $\alpha_X = \mathcal{L}^\lambda(W_\eta)$.

Note that η can be understood in the sense of randomised multi-stopping times, see Lemma 3.11 of [6].

Moreover, we argue that for the specific form of $F(x,t) = h(t)$ for a convex/concave function h, then the stopping time recovered as the optimiser is exactly the delayed Root/Rost stopping time. We can do this using a slight modification of the arguments in [6].

We proceed as follows:

Let $\mathrm{RMST}_2^1(\eta,\mu)$ denote the subset of RMST_2^1 such that the projection onto $\mathbb{R}\times C_0(\mathbb{R}_+)\times\Xi$ is exactly η. Then:

1. $\mathrm{RMST}_2^1(\eta,\mu) \subseteq \mathrm{RMST}(\lambda,\alpha_X,\mu)$ as defined as in [6, Definition 3.18] which is a compact set (Proposition 3.19);

2. $\mathrm{RMST}_2^1(\eta,\mu)$ is closed, as the projection map is continuous, hence it is also compact;

3. The proof of a suitably modified version of Theorem 5.2 now goes through essentially verbatim to give the following theorem.

Theorem 2.1. *Assume that $\gamma : S^{\otimes 2} \to \mathbb{R}$ is Borel measurable. Assume that* (OptMSEP) *is well posed and that $\xi \in \mathrm{RMST}_2^1(\eta,\mu)$ is an optimiser. Then there exists a (γ,ξ)-monotone Borel set Γ such that $r_1(\xi)(\Gamma) = 1$.*

Here, Γ being (γ,ξ)-monotone is as defined in [6, Definition 5.1] , without the constraint on the projections of the respective sets. The final conclusion of the Theorem can be applied as in [6, Theorem 2.7], [6, Theorem 2.10] (or more transparently, the corresponding results in [6]) to recover the Root/Rost forms of the optimisers.

While Root barrier solutions will always be adapted to the filtration generated by the underlying Brownian motion, it is a well known fact that Rost barrier solutions will require some form of external randomisation in the case where the initial and the terminal distribution share some mass. The story is similar for delayed Root and Rost embeddings and will be summed up in the following two theorems.

Theorem 2.2. *Consider a* (dSEP) *given by (λ,η,μ). Then there exists a Root barrier R such that the stopping time $\rho \geq \eta$ defined via*

$$\rho = \inf\{t \geq \eta : (t, W_t) \in R\}$$

embeds the measure μ.

The proof of this theorem follow analogously to the proof of [6, Theorem 2.7], or similarly of [5, Theorem 2.1] heeding the modifications given above. Moreover, it is an easy application of the standard Loynes argument [28] to see that such a Root stopping time ρ must be unique in the sense that for any other Root stopping time $\tilde{\rho}$ solving the (dSEP) we would have $\mathbb{P}^\lambda[\rho = \tilde{\rho}] = 1$. More details on how to generalise the Loynes uniqueness argument can be found in [23, Section 4].

We proceed to Rost solutions. In order to give better insights on what will happen in the case of delayed Rost embeddings, we quickly review the case of an undelayed (SEP) determined by the initial measure λ and terminal measure μ. It is well known that all mass shared between the two measures λ and μ must be stopped immediately in the following way. Let $\bar{\rho}$ be a Rost stopping time solving (SEP), then $\mathbb{P}^\lambda[\bar{\rho} = 0, W_0 \in A] = (\lambda\wedge\mu)(A)$ where the infimum of measures is defined as

$$(\lambda\wedge\mu)(A) := \inf_{B\subseteq A,\, B \text{ measurable}} \big(\lambda(B) + \mu(A\setminus B)\big).$$

On the set $\{\bar{\rho} > 0\}$ the stopping time $\bar{\rho}$ will be the first hitting time of a Rost barrier $\bar{R}$.

On a set where $(\lambda\wedge\mu)(A) < \lambda(A)$ we will have to randomise our initial stopping rule in A in order stop only sufficiently much mass. If $(\lambda\wedge\mu)(A) < \mu(A)$ then, at least in some points in A, we need to stop all mass and in fact we will need to extend the Rost barrier beyond 0 at those points to stop more mass later on. But in this case, all mass will stop immediately in those points *independent* of the choice of randomisation rule, as a consequence of the recurrence of Brownian motion. Thus any randomisation rule will give the same stopping rule although an obvious "maximal" rule is to stop all paths starting in A due to the randomisation. If $(\lambda \wedge \mu)(A) = \mu(A) = \lambda(A)$, the randomisation rule that stops all mass is the only viable stopping rule since there will be no further barrier in A beyond time 0.

In the delayed framework something similar needs to happen while it is apparent that things are complicated by the fact that not all mass will be released at time 0 at once, but rather over time, according to the measure $\mathcal{L}^\lambda(\eta)$. It is, however, clear that such external randomisation in only feasible in the *atoms* of the measure $\mathcal{L}^\lambda(\eta)$, of which there can only be countably many $\{t_0, t_1, \dots\}$. A *delayed, randomised* Rost solution will therefore consist of both a Rost barrier $\bar{R}$ and some description of this randomisation in the atoms of the delay stopping time. Hence, we choose to represent such a solution as follows.

Definition 2.3. *A delay stopping time η which distribution has atoms in $\{t_0, t_1, \dots\}$, a Rost barrier $\bar{R}$ and a family of probability densities $\{\varphi_0, \varphi_1, \dots\}$ determine a* delayed, randomised *Rost stopping time $\bar{\rho}^\varphi$ in the following way.*

$$\bar{\rho}^\varphi := \inf \left\{ t > \eta : (t, W_t) \in \bar{R} \right\} \wedge Z$$

where

$$Z = \min_{k \in \mathbb{N}} Z_k \quad \textit{for} \quad Z_k := \begin{cases} t_k & \textit{with probability } \mathbb{P}^\lambda \left[Z_k = t_k \middle| W_\eta, \eta = t_k \right] = \varphi_k \left(W_\eta \right) \\ \infty & \textit{otherwise.} \end{cases}$$

Let

$$\alpha := \mathcal{L}^\lambda((\eta, W_\eta)) \in \mathcal{P}([0, \infty) \times \mathbb{R})$$

denote the space-time distribution induced by the delay stopping time η. Then we abbreviate by

$$\alpha_t := \alpha\left(\{t\} \times \cdot\right)$$

its spatial marginal at time t.

A crucial fact to note about Rost solutions is that at any time point $t \geq 0$ we are able to tell exactly how much mass was already embedded in a given measurable set $A \subseteq \mathbb{R}$ but more importantly we are furthermore able to decide that no more mass will be embedded in this set going forward in time. These properties allow us to define Rost solutions in an iterative fashion and for this purpose, given a Rost solution $\bar{\rho}$ we define the following quantity

$$\nu_t := \mu - \mathbb{P}^\lambda \left[W_{\bar{\rho}} \in \cdot, \bar{\rho} < t \right],$$

the "mass not embedded before time t".

Let t_k denote an atom of the measure $\mathcal{L}^\lambda(\eta)$, then note that the quantity α_{t_k} is given while the quantity ν_{t_k} is fully determined by the Rost barrier *before* time t_k, i.e. $\bar{R} \cap ([0, t_k) \times \mathbb{R})$ and the randomisation densities $\{\varphi_l\}_{t_l < t_k}$.

We consider

$$\psi_k(x) := \frac{\mathrm{d}\left(\alpha_{t_k} \wedge \nu_{t_k}\right)}{\mathrm{d}\left(\alpha_{t_k}\right)}(x), \tag{2.1}$$

the Radon-Nikodym density of $\alpha_{t_k} \wedge \nu_{t_k}$ with respect to α_{t_k} and state the following theorem.

Theorem 2.4. *Consider* (dSEP) *given by (λ, η, μ). Let $\{t_0, t_1, \dots\}$ denote the set of (at most countably many) atoms of $\mathcal{L}^\lambda(\eta)$. Then there exists a Rost barrier $\bar{R}$ such that the stopping time $\bar{\rho} \geq \eta$ defined via*

$$\bar{\rho} = \inf \left\{ t > \eta : (t, W_t) \in \bar{R} \right\} \wedge Z$$

where

$$Z = \min_{k\in\mathbb{N}} Z_k \quad \textit{for} \quad Z_k := \begin{cases} t_k & \textit{with probability } \mathbb{P}^\lambda\left[Z_k = t_k \middle| W_\eta, \eta = t_k\right] = \psi_k\left(W_\eta\right) \\ \infty & \textit{otherwise} \end{cases}$$

is a solution to (dSEP)*.*

As in the Root case, the existence of a Rost barrier will follow analogously to the proof of [6, Theorem 2.10] or similarly of [5, Theorem 2.4] heeding the modifications given above.

Outside of the atoms of the delay distribution, more precisely conditional on the event $\{\eta \notin \{t_0, t_1, \dots\}\}$, two Rost stopping times $\bar\rho$ and $\tilde\rho$ solving the same (dSEP) are given as hitting times of two Rost barriers, hence the same Loynes argument as mentioned in the Root case will provide a.s. equality.

The density of the random variables Z_k does not necessarily need to be unique. There is, however, a sort of maximality about the chosen densities $\psi_0, \psi_1, \dots$ in Theorem 2.4 which will be clarified with the following lemma.

Lemma 2.5. *Consider a* (dSEP) *given by* (λ, η, μ)*. Let* $\bar R$ *be a Rost barrier and* $\{\varphi_0, \varphi_1, \dots\}$ *be a family of densities such that the delayed, randomised Rost stopping time* $\bar\rho^\varphi$ *given as in Definition 2.3 solves the* (dSEP)*. Then for any* $k \in \mathbb{N}$ *we can replace* φ_k *by* ψ_k *given by* (2.1) *to obtain the stopping time* $\bar\rho^\psi$ *for which we will have* $\mathbb{P}^\lambda\left[\bar\rho^\varphi = \bar\rho^\psi\right] = 1$*.*

Proof. Let us consider an arbitrary atom t_k of the delay distribution $\mathcal{L}^\lambda(\eta)$. First of all we clarify that in order for φ_k to be feasible we must have $\varphi_k \le \psi_k$, as for all measurable sets $A \subseteq \mathbb{R}$ we have

$$\mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in A, \bar\rho^\varphi = t_k, \eta = t_k\right] \le \begin{cases} \mathbb{P}^\lambda\left[W_{t_k} \in A, \eta = t_k\right] = \alpha_{t_k}(A), \\ \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in A, \bar\rho^\varphi = t_k\right] \le \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in A, \bar\rho^\varphi \ge t_k\right] = \nu_{t_k}(A). \end{cases}$$

Denote by $A_k := \bigcap_{\varepsilon>0}\{x : (t_k + \varepsilon, x) \in \bar R\}$ all spatial points in the barrier $\bar R$ at time t_k. Then clearly due to the recurrence of Brownian motion we have

$$\mathbb{P}^\lambda\left[\bar\rho^\varphi = t_k \middle| \eta = t_k, W_{t_k} \in A_{t_k}\right] = 1 = \mathbb{P}^\lambda\left[\bar\rho^\psi = t_k \middle| \eta = t_k, W_{t_k} \in A_{t_k}\right].$$

Hence, the specific values of φ and ψ are irrelevant and we might replace $\varphi(x)$ with $\psi(x)$ for all $x \in A_k$.

Due to the Rost barrier structure, on the set A_k^c, no mass will be embedded at a later time point,

$$\mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in A_{t_k}^c, \bar\rho^\varphi > t_k\right] = 0 = \mathbb{P}^\lambda\left[W_{\bar\rho^\psi} \in A_{t_k}^c, \bar\rho^\psi > t_k\right]. \tag{2.2}$$

Consider now the set $B_k := \{x : \varphi_k(x) < \psi_k(x)\} \cap A_k^c$ and let us assume that $\alpha_{t_k}(B_k) > 0$ as otherwise both $\varphi_k(x)$ and $\psi_k(x)$ must be 0 for all $x \in B_k$ anyway. Our claim is then proven if we are able to establish $\int_{B_k} \varphi(x)\mathrm{d}x = \int_{B_k} \psi(x)\mathrm{d}x$.

It is a consequence of (2.2) that

$$\mu(B_k) = \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in B_k, \bar\rho^\varphi < t_k\right] + \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in B_k, \bar\rho^\varphi = t_k\right],$$

or equivalently that

$$\begin{aligned} \nu_{t_k}(B_k) &= \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in B_k, \bar\rho^\varphi = t_k\right] \\ &= \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in B_k, \bar\rho^\varphi = t_k, \eta = t_k\right]. \end{aligned}$$

Here the second equality follows from the specific structure of Rost solutions, as all mass that is already released before time t_k and stopped in B_k will have done so *before* time t_k almost surely.

Then

$$\begin{aligned}\nu_{t_k}(B_k) = \mathbb{P}^\lambda\left[W_{\bar\rho^\varphi} \in B_k, \bar\rho^\varphi = t_k, \eta = t_k\right] &= \mathbb{P}^\lambda\left[W_{t_k} \in B_k, \bar\rho^\varphi = t_k, \eta = t_k\right] \\ &= \alpha_{t_k}(B_k)\int_{B_k}\varphi(x)\mathrm{d}x \\ &\le \alpha_{t_k}(B_k)\int_{B_k}\psi(x)\mathrm{d}x = (\alpha_{t_k}\wedge\nu_{t_k})(B_k)\end{aligned}$$

In particular, this implies that

$$(\alpha_{t_k}\wedge\nu_{t_k})(B_k) = \nu_{t_k}(B_k),$$

hence we must have $\int_{B_k}\varphi(x)\mathrm{d}x = \int_{B_k}\psi(x)\mathrm{d}x$. □

Here we observe that if the set of atoms $t_0, t_1, \dots$ is either finite or infinite but can be ordered we can assume $t_0 \le t_1 \le \cdots$ and construct the densities ψ_k consecutively. When this is not the case, meaning $t_0, t_1, \dots$ is infinite and cannot be ordered, while our result will still hold, we cannot expect to derive an explicit formula for the functions ψ_k in this scenario.

An explicit Root solution to (MMSEP) is clear given Theorem 2.2 while an explicit Rost solution can be recovered from Theorem 2.4 in the following way.

Corollary 2.6. *Consider a starting measure λ and a sequence of measures $\mu_1, \dots, \mu_n$ such that $\lambda \le_c \mu_1 \le_c \cdots \le_c \mu_n$. Then there exists a sequence or Rost barriers $\bar R_1, \dots, \bar R_n$ such that the stopping times $\bar\rho_1 \le \bar\rho_2 \le \cdots \le \bar\rho_n$ defined in the following way are a solution to* (MMSEP).

$$\bar\rho_0 := 0 \;\; \textit{and} \;\; \bar\rho_k := \inf\left\{t > \bar\rho_{k-1} : (t, W_t) \in \bar R_k\right\} \wedge Z_k \;\; \textit{for} \;\; k \ge 1$$

where

$$Z_k := \begin{cases} 0 & \textit{with probability } \mathbb{P}^\lambda\left[Z_k = 0 \middle| W_0, \bar\rho_{k-1} = 0\right] = \frac{\mathrm{d}\left(\mu^0_{k-1}\wedge\mu_k\right)}{\mathrm{d}\mu^0_{k-1}}(W_0) \\ \infty & \textit{otherwise} \end{cases}$$

for $\mu^0_{k-1} := \mathbb{P}^\lambda\left[\bar\rho_{k-1} = 0, W_0 \in \cdot\right]$.

Proof. Note that for $k = 1$ this exact structure of the Rost solution $\bar\rho_1$ has already been established in [36]. Observing that $\bar\rho_1 \le \cdots \le \bar\rho_n$ are hitting times of inverse barriers, we can establish that $\mathcal{L}^\lambda(\bar\rho_1)$ and consequently any $\mathcal{L}^\lambda(\bar\rho_k)$, $k \ge 2$, do not possess atoms beyond time 0.

Inductively recovering the multi-marginal Rost solutions from Theorem 2.4 becomes feasible by considering the single atom $t_0 = 0$. Thus, for $k \ge 2$, set $\eta = \bar\rho_{k-1}$ and $\mu = \mu_k$, with $\nu_0 = \mu_k$ and $\alpha_0 = \mu^0_{k-1}$. Subsequently, invoking Theorem 2.4, we establish the existence of a Rost barrier $\bar R := \bar R_k$, ensuring that $\bar\rho_k \ge \bar\rho_{k-1}$ is of the desired structure with $Z_k := Z_0$. □

2.2 Optimal stopping representation of (dSEP) and (MMSEP)

We will derive the multi marginal Root optimal stopping problem (1.7) - (1.8) resp. the multi marginal Rost optimal stopping problem (1.9) - (1.10) as a special case of a *delayed* Root resp. Rost representation.

For this purpose we consider a starting distribution λ for a Brownian motion and an integrable delay stopping time η. Recall the space time distribution induced by the delay

$$\alpha := \mathcal{L}^\lambda((\eta, W_\eta)) \in \mathcal{P}([0,\infty)\times\mathbb{R}).$$

We denote the $\mathbb{P}^\lambda$-potential of $W_{\eta\wedge T}$ by

$$V^\alpha_T(x) = -\mathbb{E}^\lambda\left[|W_{\eta\wedge T} - x|\right]$$

and the spatial marginal of α by

$$\alpha_X(A) := \alpha([0,\infty) \times A) = \mathcal{L}^\lambda(W_\eta)(A) \quad \text{for } A \subseteq \mathbb{R} \text{ measurable.}$$

Let now $R \subseteq [0,\infty) \times \mathbb{R}$ denote a Root or Rost barrier respectively, then we consider the (delayed) Root stopping time

$$\rho_\eta := \inf\{t \geq \eta : (t, W_t) \in R\},$$

while a (delayed) Rost stopping time will be defined according to Definition 2.3.

These stopping times induce the following measures

$$\beta^\alpha := \mathcal{L}^\lambda\left(W_{\rho^\eta}\right) \text{ as well as } \beta_T^\alpha := \mathcal{L}^\lambda\left(W_{\rho^\eta \wedge T}\right) \text{ for } T \geq 0$$

and the (stopped) potential

$$U_{\beta_T^\alpha}(x) = -\mathbb{E}^\lambda\left[|W_{\rho^\eta \wedge T} - x|\right].$$

The extension of the Root optimal stopping problem (1.2)-(1.3) will then be given in the following theorem.

Theorem 2.7. *Let R denote a* Root *barrier inducing the potentials U_{β^α} and $U_{\beta_T^\alpha}$. Then for every $(T,x) \in [0,\infty) \times \mathbb{R}$ we have the representation*

$$U_{\beta_T^\alpha}(x) = \mathbb{E}^x\left[V_{T-\tau^*}^\alpha(W_{\tau^*}) + (U_{\beta^\alpha} - U_{\alpha_X})(W_{\tau^*})\mathbb{1}_{\tau^*<T}\right] \tag{2.3}$$

$$= \sup_{\tau \leq T} \mathbb{E}_y\left[V_{T-\tau}^\alpha(W_\tau) + (U_{\beta^\alpha} - U_{\alpha_X})(W_\tau)\mathbb{1}_{\tau<T}\right]. \tag{2.4}$$

where the optimiser is given by $\tau^* := \inf\{t \geq 0 : (T-t, W_t) \in R\} \wedge T$.

Moreover, given the interpolating potentials $U_{\beta_T^\alpha}(x)$ for all $(T,x) \in [0,\infty) \times \mathbb{R}$ we can recover the Root barrier R embedding the measure β^α in the following way

$$R = \left\{(T,x) : \left(U_{\beta_T^\alpha} - V_T^\alpha\right)(x) = \left(U_{\beta^\alpha} - U_{\alpha_X}\right)(x)\right\}.$$

Analogously we can give an optimal stopping representation for delayed Rost embeddings extending (1.4)-(1.5) in the following way.

Theorem 2.8. *Let $\bar{R}$ denote a* Rost *barrier inducing the potentials U_{β^α} and $U_{\beta_T^\alpha}$. Then for every $(T,x) \in [0,\infty) \times \mathbb{R}$ we have the representation*

$$U_{\beta^\alpha}(x) - U_{\beta_T^\alpha}(x) = \mathbb{E}^x\left[U_{\beta^\alpha}(W_{\sigma^*}) - V_{T-\sigma^*}^\alpha(W_{\sigma_*})\right] \tag{2.5}$$

$$= \sup_{\sigma \leq T} \mathbb{E}^x\left[U_{\beta^\alpha}(W_\sigma) - V_{T-\sigma}^\alpha(W_\sigma)\right], \tag{2.6}$$

where the optimiser is given by $\sigma^* := \inf\{t \geq 0 : (T-t, W_t) \in \bar{D}\} \wedge T$.

Moreover, given the interpolating potentials $U_{\beta_T^\alpha}(x)$ for all $(T,x) \in [0,\infty) \times \mathbb{R}$ we can recover the Rost barrier $\bar{R}$ embedding the measure β^α in the following way

$$\bar{R} = \left\{(T,x) : U_{\beta_T^\alpha}(x) = V_T^\alpha(x)\right\}.$$

Note that, while the interpolating potentials were used to recover the Rost barrier, the probability densities of the external randomisations that might be associated with the Rost stopping time can be recovered in an inductive manner as described in Section 2.1, precisely by Equation (2.1).

We now want to observe that, for the choice of $\eta = 0$, the single marginal identities can easily be recovered. Indeed,

- $\beta^\alpha = \mu$ and $\beta_T^\alpha = \mu_T$ as in Section 1,

- $\alpha = \delta_0 \times \lambda$ thus, $\alpha_X = \lambda$,
- $V_T^\alpha(y) = -\mathbb{E}^\lambda[|W_0 - y|] = U_\lambda(y)$ for any $T \geq 0$.

Hence

$$\begin{aligned}\mathbb{E}_y\left[V_{T-\tau}^\alpha(W_\tau) + (U_\mu - U_{\alpha_X})(W_\tau)\mathbb{1}_{\tau<T}\right] &= \mathbb{E}_y\left[U_\lambda(W_\tau) + (U_\mu - U_\lambda)(W_\tau)\mathbb{1}_{\tau<T}\right] \\ &= \mathbb{E}_y\left[U_\mu(W_\tau)\mathbb{1}_{\tau<T} + U_\lambda(W_\tau)\mathbb{1}_{\tau=T}\right],\end{aligned}$$

in the Root case, and

$$\mathbb{E}^x\left[U_{\beta^\alpha}(W_\sigma) - V_{T-\sigma}^\alpha(W_\sigma)\right] = \mathbb{E}^x\left[U_\mu(W_\sigma) - U_\lambda(W_\sigma)\right]$$

in the Rost case respectively.

Furthermore, the multi-marginal extensions now follow as an corollary as they can be deduced inductively from the delayed optimal stopping problem (2.3)-(2.4) (resp. (2.5)-(2.6)) in the following way.

Corollary 2.9. *Let $R_1, R_2, \dots, R_n$ denote a sequence of Root (resp. Rost) barriers inducing the potentials $U_{\mu_1}, \dots, U_{\mu_n}$ as well as $U_{\mu_1,T}, \dots, U_{\mu_n,T}$. Then the optimal stopping representations* (1.7)-(1.8) *(resp.* (1.9)-(1.10)*) hold.*

Proof. For the first marginal induced by R_1 the optimal stopping representation is simply the single marginal representation (1.2)-(1.3) (resp. (1.4)-(1.5)) already known (and also recovered above). Let now $k \in \{2, \dots, n\}$ and consider $\eta = \rho_k$ and $R = R_{k+1}$. Then

- $\alpha = \mathcal{L}^\lambda(\rho_k, W_{\rho_k})$, thus $\alpha_X = \mu_k$ and $U_{\alpha_X}(y) = U_{\mu_k}(y)$,
- $\beta^\alpha = \mu_{k+1}$, thus $U_{\beta^\alpha}(y) = U_{\mu_{k+1}}(y)$ and $U_{\beta_T^\alpha}(y) = U_{\mu_{k+1},T}(y)$,
- $V_T^\alpha(y) = -\mathbb{E}^\lambda[|W_{\rho_k\wedge T} - y|] = U_{\mu_k,T}(y)$.

Hence we recover the multi marginal Root (resp. Rost) optimal stopping problems (1.7)-(1.8) (resp. (1.9)-(1.10)). □

3 A discrete delayed (SEP) and its optimal stopping representation

We take on the idea of [1] to cast our problem in a discrete setting in order to use arguments for SSRWs to derive a discrete version of the desired result and then recover the continuous result through a Donsker type limiting argument. In this section we define a discrete version of (dSEP). We furthermore define discrete Root and Rost solutions and give an explicit construction.

Let X denote a SSRWs on some probability space $(\Omega, \mathbb{P})$, started at some possibly random initial position. Given $x \in \mathbb{Z}$ we write $\mathbb{P}^x$ for the conditional distribution when $X_0 = x$. Hence, we can consider a probability measure λ on $\mathbb{Z}$ a "starting" distribution for X by setting $\mathbb{P}^\lambda := \sum_{x\in\mathbb{Z}} \mathbb{P}^x \lambda(\{x\})$.

3.1 A discrete (dSEP)

Consider a SSRW $(X_t)_{t\in\mathbb{N}}$ started in $X_0 \sim \lambda$. Given a discrete measure μ the *discrete Skorokhod embedding problem* is to find a (discrete) stopping time ρ such that

$$X_\rho \sim \mu \text{ and } (X_{t\wedge\rho})_{t\in\mathbb{N}} \text{ is uniformly integrable.} \tag{3.1}$$

Given a delay stopping time η, the *delayed* discrete Skorokhod embedding problem (or *delayed* (dSEP)) is to find a (discrete) stopping time $\rho \geq \eta$ such that (3.1) is satisfied.

Just as in the continuous case we are interested in the existence of barrier type solutions to the discrete (dSEP). These barriers can be defined analogously to the continuous setting as subsets $R \subseteq \mathbb{N} \times \mathbb{Z}$ satisfying the following respective property.

- R will denote a (discrete) *Root* barrier if for $(t, x) \in R$ and $s > t$ we also have $(s, x) \in R$.
- $\bar{R}$ will denote a (discrete) *Rost* barrier if for $(t, x) \in \bar{R}$ and $s < t$ we also have $(s, x) \in \bar{R}$.

While in the continuous setting it is known that no external randomisation is needed in order to give a Root solution to the (SEP) and external randomisation is needed only in $t = 0$ for the Rost solution (and only if the initial measure λ and the terminal measure μ share mass) this is no longer the case in the discrete setting. We will see that in order to give a discrete Root (resp. Rost) solution external randomisation will be necessary in both cases also beyond time $t = 0$, however, only at the boundary of the respective barrier.

Example 3.1. Consider the simple example

$$\lambda = \delta_0$$
$$\mu = \frac{1}{3}\delta_{-2} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_2.$$

First note that all paths of a SSRW X that reach ± 2 at any point in time will be absorbed there. As all paths originate from $(0, 0)$, the probability of being in site $x = 0$ at time $t = 2$ is given by $\frac{1}{2} > \frac{1}{3} = \mu(\{0\})$. Consequently, we cannot stop all paths in $(2, 0)$ as this would embed too much mass in $\{0\}$. However, if we allow all paths to proceed from $(2, 0)$, then the probability of being in site $x = 0$ at time $t = 4$ without having hit ± 2 before is given by $\frac{1}{4} < \frac{1}{3} = \mu(\{0\})$. In other words, this would result in insufficient mass being embedded in 0. Hence, it becomes necessary to stop *some* paths in $(2, 0)$, but not all of them. Precisely, we will need to stop $\frac{1}{3}$ of paths in $(2, 0)$ while allowing the remainder to reach either of $\{-2, 0, 2\}$ at time $t = 4$ in order to correctly embed the distribution μ. Refer to Figure 2 for an illustration on possible realisations of such paths.

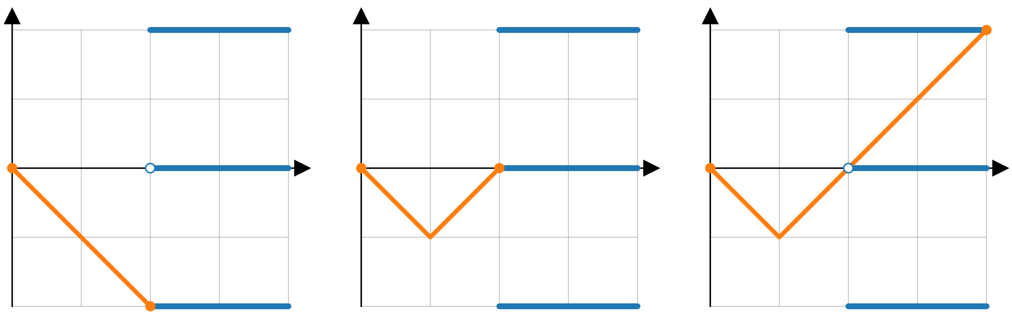


Figure 2: Possible realisations of paths hitting a Root barrier that solves the discrete (SEP) given by $\lambda = \delta_0$ and $\mu = \frac{1}{3}\delta_{-2} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_2$.

We will formalise this concept of randomised stopping in the following way. If ρ denotes a stopping time for the process X then we can for each point $(t, x) \in \mathbb{N} \times \mathbb{Z}$ consider the probability of ρ stopping in site x at time t after η by

$$r_t(x) := \mathbb{P}^\lambda \left[\rho = t \middle| (\rho \wedge t, X_{\rho \wedge t}) = (t, x), \eta \leq t\right]. \tag{3.2}$$

If now R denotes a barrier and ρ its hitting time then in the classic (deterministic) case we should have $r_t(x) \in \{0, 1\}$ and $r_t(x) = 1$ if and only if $(t, x) \in R$, i.e. we stop in the barrier immediately after arrival. However, if we allow for $r_t(x) \in (0, 1)$ but agree on the convention that $(t, x) \in R$ when $r_t(x) > 0$, then this can be understood as stopping inside of the barrier only with probability

$r_t(x)$. We can now give an equivalent definition of discrete Root (resp. Rost) barriers on the basis of these stopping probabilities. A field of stopping probabilities denoted as $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ represents a

- Root barrier if for all $(t,x) \in \mathbb{N}\times\mathbb{Z}$ such that $r_t(x) > 0$ we have $r_{t+s}(x) = 1\,\forall s \geq 1$, (3.3)
- Rost barrier if for all $(t,x) \in \mathbb{N}\times\mathbb{Z}$ such that $r_t(x) > 0$ we have $r_{t-s}(x) = 1\,\forall s = 1,\dots,t$. (3.4)

This definition guarantees the respective barrier structure and furthermore assures that randomisation only happens at the boundary points of the respective barriers. We will also refer to $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ as Root and Rost field of stopping probabilities, respectively.

Let $u = (u_0, u_1, \dots)$ denote a sequence of iid uniform random variable, i.e. $u_i \sim U[0,1]$. Then our Root (resp. Rost) stopping times subject to external randomisation at the endpoints can be represented as follows

$$\rho = \rho_u = \inf\{t \geq \eta : r_t(X_t) > u_t\}. \tag{3.5}$$

We could think of flipping a coin with success probability $r_t(X_t)$ at every time point t. Was the coin flip successful, the process stops. If not, it moves onto the next point. Note that if $r_t(x) \in \{0,1\}$ for all $(t,x) \in \mathbb{N}\times\mathbb{Z}$, then this definition will coincide with the usual, non-randomized definition

$$\rho = \inf\{t \geq \eta : (t, X_t) \in R\}. \tag{3.6}$$

Our stopping time ρ is now a *randomised* stopping time taking the two arguments ω and u, thus $\rho : \Omega \times [0,1]^{\mathbb{N}} \to \mathbb{N}$. While this is a slightly unconventional definition of an external randomisation used for its more intuitive explanation in this context, we can also cast this randomised stopping time into the more conventional framework of using only a single variable of external randomisation in the following way. For each $\omega \in \Omega$ we consider the sequence $(\zeta_\omega(t))_{t\in\mathbb{N}}$ of the probability of the process $X_{\rho\wedge t}(\omega)$ being "alive" at time t given via

$$\begin{aligned}\zeta_\omega(0) &:= (1 - r_0(X_0(\omega))) \vee \mathbb{1}_{\eta(\omega)=0} \quad \text{and}\\ \zeta_\omega(t) &:= \zeta_\omega(t-1)\cdot(1 - r_t(X_t(\omega))) \vee \mathbb{1}_{t<\eta(\omega)}.\end{aligned}$$

Note that $(\zeta_\omega(t))_{t\in\mathbb{N}}$ is a decreasing sequence and $\xi_\omega(\{0,\dots,t\}) := 1-\zeta_\omega(t)$ will define a measure on $\mathbb{N}$. Given a uniformly distributed random variable $u \sim U[0,1]$ independent of X the (randomised) stopping time

$$\tilde\rho(\omega,u) := \inf\{t \geq 0 : \xi_\omega(\{0,\dots,t\}) \geq u\}$$

now gives a more traditional representation of our randomised Root (resp. Rost) stopping time.

In the following two subsections we will give an explicit construction of the stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ solving the discrete (dSEP) both in the Root and Rost case. To guarantee uniform integrability of all solutions constructed we will henceforth always assume μ to denote a finite sum of atoms. Let x_* and x^* denote the smallest and largest site respectively at which we have an atom. Then in both the Root and the Rost case our solution should be bounded from above by the (delayed) hitting time

$$H_{x_*,x^*} := \inf\{t \geq \eta : X_t \in \{x_*, x^*\}\}.$$

Since η is a stopping time with finite expectation then so is H_{x_*,x^*}, which in turn implies uniform integrability of $(X_{t\wedge\rho})_{t\geq 0}$.

3.1.1 Existence of discrete delayed Root solutions

An explicit Root solution to the discrete (dSEP) can be established with the help of expected local times. For a SSRW random walk X the local time at site x up to time t is defined via

$$L_t^x(X) = \#\{s : 0 \leq s < t, X_s = x\} = \sum_{s=0}^{t-1} \mathbb{1}_{X_s=x}.$$

We note the following discrete version of Tanaka's Formula for local times.

Theorem 3.2. *Consider a random walk X defined by $X_t = X_0 + \sum_{s=1}^{t} \xi_s$ where ξ_s are iid Rademacher random variables. Then for $x \in \mathbb{Z}$ we can represent the local time $L_t^x(X)$ of X until time t as*

$$L_t^x(X) = |X_t - x| - |X_0 - x| - \sum_{s=0}^{t-1} \operatorname{sgn}(X_s - x)\xi_{s+1}. \tag{3.7}$$

We refer to [16] for a proof.

Let

$$L_{s,t}^x(X) = \#\{u : s \le u < t \text{ and } X_u = x\} = \sum_{u=s}^{t-1} \mathbb{1}_{X_u = x} = L_t^x(X) - L_s^x(X)$$

denote the local time of the process X at site x *between* time s and t (for $s \le t$).

For our delay stopping time η write $\alpha_X = \mathcal{L}^\lambda(X_\eta)$. Consider now $X_0 \sim \lambda$ and let $\rho \ge \eta$ be a stopping time for the random walk X such that $X_\rho \sim \mu$. Then due to (3.7) we have the following representation of the *expected* local time between η and ρ

$$\mathbb{E}^\lambda \left[L_{\eta,\rho}^x(X) \right] = U_{\alpha_X}(x) - U_\mu(x) =: L(x). \tag{3.8}$$

Conversely, for any solution $\rho \ge \eta$ to (dSEP) Equation (3.8) must be satisfied.

Let $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ denote a field of stopping probabilities and ρ its corresponding stopping time delayed by the stopping time η.

Then we define

$$\tilde{\alpha}_t(x) := \mathbb{P}^\lambda \left[X_t = x, \eta \le t, \rho \ge t \right],$$

the free mass available in site x at time t.

Define $\alpha(t,x) := \mathbb{P}^\lambda [X_\eta = x, \eta = t]$, then we can compute $\tilde{\alpha}_t$ inductively using only information up to time $t-1$ in the following way

$$\tilde{\alpha}_t(x) = \frac{1}{2} \left((1 - r_{t-1}(x+1))\tilde{\alpha}_{t-1}(x+1) + (1 - r_{t-1}(x-1))\tilde{\alpha}_{t-1}(x-1) \right) + \alpha(t,x).$$

Precisely, paths can only end up in (t,x) from the neighboring points $(t-1, x+1)$ and $(t-1, x+1)$. They will with probability $1 - r_{t-1}(x+1)$ resp. $1 - r_{t-1}(x-1)$ leave the respective site and then with probability $\frac{1}{2}$ end up in (t,x). Moreover, we only consider mass after the stopping time η, this is represented by $\alpha(t,x)$, the probability of particles *appearing* in site x at time t. Note that at time 0 we thus only have $\tilde{\alpha}_0(x) = \alpha(0,x)$.

With the help of the quantities $\tilde{\alpha}_t(x)$ the expected number of visits in site x up until time t of the process X can then be written as

$$l_t^x := \sum_{s=0}^{t-1} (1 - r_s(x))\tilde{\alpha}_s(x). \tag{3.9}$$

In particular we will have $l_0^x = 0$ for all $x \in \mathbb{Z}$.

We recall $L(x) := U_{\alpha_X}(x) - U_\mu(x)$ and define

$$\tilde{L}_t(x) := L(x) - l_t^x,$$

the "missing local time" after time t. Note that $\tilde{L}_t(x)$ (resp. l_t^x) can be computed only using information up to time $t-1$.

We are now equipped to give a precise construction of a Root solution to the discrete (dSEP).

Theorem 3.3. *Consider a field of stopping probabilities* $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ *inductively defined via the formula*

$$r_t(x) := 1 - \frac{L(x) - l_t(x)}{\tilde{\alpha}_t(x)} \wedge 1 = 1 - \frac{\left(\tilde{L}_t \wedge \tilde{\alpha}_t\right)(x)}{\tilde{\alpha}_t(x)} \quad \textit{when } \tilde{\alpha}_t(x) > 0$$

and chosen to comply with (3.3) *otherwise.*

Then this field represents a Root field of stopping probabilities and the associated stopping time

$$\rho = \inf\{t \geq \eta : r_t(X_t) > u_t\}.$$

will be a Root solution to the discrete (dSEP).

Proof. Note the following

- As both $\tilde{L}_t$ and $\tilde{\alpha}_t$ only use information that is either given or acquired up to time $t-1$, it is clear that r_t can be computed inductively to give a stopping time.
- When $\tilde{\alpha}_t(x) = 0$ the point (t,x) cannot be reached by the stopped random walk after time η. The specific value of $r_t(x)$ is hence irrelevant for all the respective quantities and can be chosen to comply with the definition of a Root field of stopping probabilities.
- $r_t(x) = 1 \Leftrightarrow l_t^x = L(x)$. In other words, we must have already accrued enough local time at site x. Moreover we will then have $l_{t+s}^x = l_t^x$ and $r_{t+s}(x) = 1$ for all $s \geq 0$, hence the Root structure of our barrier is guaranteed.
- $r_t(x) = 0 \Leftrightarrow \tilde{L}_t(x) = L(x) - l_t^x \geq \tilde{\alpha}_t(x)$. So no mass will be stopped in (t,x) since we still need to accrue more local time and adding all available mass to l_t^x will make us remain in our local time budget, i.e. $l_{t+1}^x = l_t^x + \tilde{\alpha}_t(x) \leq L(x)$.
- $r_t(x) \in (0,1) \Leftrightarrow 0 < \frac{L(x)-l_t^x}{\tilde{\alpha}_t(x)} < 1$. As letting all mass run would add too much to our expected local time only a fraction should be added and we have $l_{t+1}^x = l_t^x + \frac{L(x)-l_t^x}{\tilde{\alpha}_t(x)}\tilde{\alpha}_t(x) = L(x)$. This will imply $r_{t+s} = 1$ for all $s \geq 1$ again showing that we will recover the desired Root barrier structure.

Thus for each time point t this construction will reveal a new layer of stopping probabilities $(r_t(x))_{x\in\mathbb{Z}}$ which will be of Root structure. Let ρ_∞ denote the stopping time corresponding to the field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ constructed as above. It remains to show that $\mathcal{L}^\lambda(X_{\rho_\infty}) = \mu$. As l_t^x is increasing and bounded from above by $L(x)$ the limit $l_\infty^x := \lim_{t\to\infty} l_t^x$ exists and equals the expected local time of X_{ρ_∞} after the stopping time η, $\mathbb{E}^\lambda\left[L_{\eta,\rho_\infty}^x(X)\right] = l_\infty^t$. Let U_{μ_∞} denote the potential of $\mathcal{L}^\lambda\left(X_{\rho_\infty}\right)$, then due to Tanaka's formula we have

$$U_{\mu_\infty}(x) = U_{\alpha_x} - l_\infty^t,$$

hence our stopping time will embed the right measure if and only if $l_\infty^t = L(x)$. Consider the set

$$I := \{x \in \mathbb{Z} : l_\infty^x < L(x)\} = \{x \in \mathbb{Z} : U_{\mu_\infty}(x) > U_\mu(x)\}.$$

To conclude the proof of the theorem we must show that I is empty. A first step is to show that for every $x \in I$ we have

$$\mathbb{P}^\lambda\left[X_{\rho_\infty} = x\right] = 0. \tag{3.10}$$

Note that this is equivalent to $\tilde{\alpha}_t(x) \wedge r_t(x) = 0$ for all $t \in \mathbb{N}$. However, we cannot have $\tilde{\alpha}(x) = 0$ for all $t \in \mathbb{N}$ as this would imply that site x is never reached by the process X with positive probability after time η, furthermore implying $L(x) = 0$, which is a contradiction. Thus there must exist at least one $T \in \mathbb{N}$ such that $\tilde{\alpha}_T(x) > 0$. Assume now that also $r_T(x) > 0$. Then

$l^x_\infty \geq l^x_{T+1} = l^x_T + \left(\frac{L(x)-l^x_T}{\tilde{\alpha}_T(x)}\right)\tilde{\alpha}_T(x) = L(x)$ which is, again, a contradiction, allowing us to conclude (3.10).

Now we proceed to show that I must be empty. Consider a maximal interval $J \subseteq I$, i.e. either $J = \{\dots, x_+ - 1, x_+\}$ and $x_+ + 1 \notin I$, $J = \{x_-, x_- + 1, \dots\}$ and $x_- - 1 \notin I$ or $J = \{x_-, x_- + 1, \dots, x_+ - 1, x_+\}$ and both $x_- - 1, x_+ + 1 \notin I$. Then we must have $U_{\mu_\infty}(x_- - 1) = U_\mu(x_- - 1)$ resp. $U_{\mu_\infty}(x_+ + 1) = U_\mu(x_+ + 1)$. Note that U_{μ_∞} is bounded from above by the concave function U_{α_X} and bounded from below by the concave function U_μ. Hence we cannot have that $U_{\mu_\infty}|_J$ is a linear function and there must exists $x \in J$ such that

$$U_{\mu_\infty}(x-1) - U_{\mu_\infty}(x) < U_{\mu_\infty}(x) - U_{\mu_\infty}(x+1).$$

However, this is equivalent to

$$\mathbb{P}^\lambda\left[X_{\rho_\infty} = x\right] = -\frac{1}{2}\left(U_{\mu_\infty}(x-1) + U_{\mu_\infty}(x+1)\right) + U_{\mu_\infty}(x) > 0$$

which contradicts Equation (3.10). Hence $J = \emptyset$, thus also $I = \emptyset$ and our claim is proven. □

3.1.2 Existence of discrete delayed Rost solutions

Given a delayed stopping time ρ corresponding to a field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ we consider the quantity

$$\tilde{\mu}_t(x) := \mathbb{P}^\lambda\left[X_\rho = x, \rho \leq t\right],$$

the "mass embedded until time t". Then

$$\tilde{\mu}_t(x) = \sum_{s=0}^{t} \tilde{\alpha}_s(x) r_s(x)$$

for $\tilde{\alpha}_t$ as defined in the previous section. Furthermore consider

$$\nu_t(x) := \mu\left(\{x\}\right) - \tilde{\mu}_{t-1}(x),$$

the "mass not yet embedded" and note that it can be computed inductively using information only up to time $t-1$. We can then also write

$$\tilde{\mu}_t(x) = \mu\left(\{x\}\right) - \nu_t(x) + \tilde{\alpha}_t(x) r_t(x).$$

Note that if the field of stopping probabilities embeds the measure μ, then

$$\lim_{t\to\infty} \tilde{\mu}_t(x) = \mu\left(\{x\}\right) \text{ resp. } \lim_{t\to\infty} \nu_t(x) = 0.$$

With the help of these quantities we can give an explicit construction of Rost solution to the discrete (dSEP).

Theorem 3.4. *Consider a field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ inductively defined via the formula*

$$r_t(x) = \frac{\nu_t(x)}{\tilde{\alpha}_t(x)} \wedge 1 = \frac{\left(\tilde{\alpha}_t \wedge \nu_t\right)(x)}{\tilde{\alpha}_t\,(x)} \quad \textit{when } \tilde{\alpha}_t(x) > 0$$

and chosen to comply with (3.3) *otherwise.*

Then this field represents a Rost field of stopping probabilities and the associated stopping time

$$\rho = \inf\left\{t \geq \eta : r_t(X_t) > u_t\right\}.$$

will be a Rost solution to the discrete (dSEP).

Proof. As both ν_t and $\tilde{\alpha}_t$ only use information that is either given or acquired up to time $t-1$, it is clear that r_t can be computed iteratively through time, thus in each time step t we identify the t-layer $r_t(x)_{x\in\mathbb{Z}}$ of the probability field. Let us justify that $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ defined this way both embeds the right measure μ and is of Rost structure.

Firstly, if $\mu(\{x\}) = 0$ then $r_t(x) = 0$ for all $t \in \mathbb{N}$, thus consider only those x such that $\mu(\{x\}) > 0$. Then consider the following cases.

$t = 0$: Only consider those x such that $\tilde{\alpha}_0(x) = \alpha(0, x) > 0$ as otherwise the point x cannot be reached by the process X at time 0 after time η.

- If $\tilde{\alpha}_0(x) \leq \mu(\{x\})$, then all mass available in $(0, x)$ should be stopped, hence $r_0(x) = 1$ and $\tilde{\mu}_0(\{x\}) = \tilde{\alpha}_0(x) \leq \mu(\{x\})$.
- If $\tilde{\alpha}_0(x) > \mu(\{x\})$, then some mass needs to be stopped in $(0, x)$, precisely $r_0(x) = \frac{\mu(\{x\})}{\tilde{\alpha}_t(x)}$ as then we have $\tilde{\mu}_0(\{x\}) = \mu(\{x\})$.

$t > 0$: Only consider those x such that $\tilde{\alpha}_t(x) > 0$ as otherwise the point x cannot be reached by the process X at time t after time η.

- If $\nu_t(x) = 0$ then all necessary mass has already been embedded before time t, hence $r_t(x) = 0$.
- If $0 < \tilde{\alpha}_t(x) \leq \nu_t(x)$ then the available mass in site x at time t does not exceed the mass still needed at site x, hence $r_t(x) = 1$ as all mass can be added.
- If $0 < \nu_t(x) < \tilde{\alpha}_t(x)$ then we have mass missing in site x but the available mass at time t exceeds the necessary mass, hence we should only add the fraction $r_t(x) = \frac{\nu_t(x)}{\tilde{\alpha}_t(x)}$ as then we have $\tilde{\mu}_t(\{x\}) = \mu(\{x\}) - \nu_t(x) + \tilde{\alpha}_t(x)\frac{\nu_t(x)}{\tilde{\alpha}_t(x)} = \mu(\{x\})$.

It is crucial to note that the quantity $\nu_t(x)$ is decreasing and remains at 0 once it has reached there, hence clearly the field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ defined as above is of Rost structure.

We are left to formally show that the stopping time

$$\rho_\infty := \inf\{t \geq \eta : r_t(X_t) > u_t\}$$

embeds the right measure into the SSRW, i.e. $\mathbb{P}^\lambda[X_{\rho_\infty} = x] = \mu(\{x\})$ for all $x \in \mathbb{Z}$.

We consider the measure

$$\mu_\infty(\{x\}) := \mathbb{P}^\lambda[X_{\rho_\infty} = x] = \lim_{t\to\infty} \tilde{\mu}_t(x) = \sum_{s=0}^{\infty} \tilde{\alpha}_s(x)p_s(x).$$

Fist assume $\mu_\infty(\{x\}) > \mu(\{x\})$. As this would be equivalent to $\lim_{t\to\infty} \nu_t(x) < 0$ there would need to exist a first $T \in \mathbb{N}$ such that $\nu_T(x) < 0$. However this contradicts the construction of $r_t(x)$ as described above.
Hence we must have $\mu_\infty(\{x\}) \leq \mu(\{x\})$ for all $x \in \mathbb{Z}$.

Consider the set $I := \{x \in \mathbb{Z} : \mu_\infty(\{x\}) < \mu(\{x\})\}$. If we assume that I is not empty then

$$0 \leq \sum_{x\in\mathbb{Z}} \mu_\infty(\{x\}) < \sum_{x\in\mathbb{Z}} \mu(\{x\}) = 1.$$

Note that $\mu_\infty(\{x\}) = \lim_{t\to\infty} \tilde{\mu}_t(x) < \mu(\{x\})$ is equivalent to $\nu_t(x) > 0$ for every $t \in \mathbb{N}$, which in turn implies $r_t(x) = 1$ for all $t \in \mathbb{N}$ by the construction above. In particular, this implies $\rho_\infty \leq H_x^\eta := \inf\{t \geq \eta : X_t = x\}$ which is a finite stopping time, hence also $\mathbb{P}^\lambda[\rho_\infty < \infty] = 1$. So X_ρ is a well defined random variable and we must have $\sum_{x\in\mathbb{Z}} \mu_\infty(\{x\}) = 1$ which is a contradiction. This implies $I = \emptyset$ and $\mu_\infty(\{x\}) = \mu(\{x\})$, hence our constructed field of stopping probabilities and consequently the associated stopping time ρ_∞ embeds the right measure. □

3.2 A discrete optimal stopping representation

In this section we will consider a second simple symmetric second random walk Y on the probability space $(\Omega, \mathbb{P})$, mutually independent of the SSRW X introduced in previous section. For given $x, y \in \mathbb{Z}$ we recall $\mathbb{P}^x$ as the conditional distribution given $X_0 = x$ and analogously define $\mathbb{P}_y$ as the conditional distribution given $Y_0 = y$. Similarly, conditioning on both events simultaneously is denoted by $\mathbb{P}^x_y$. We recall the starting distribution λ for the process X and introduce a corresponding starting distribution ν for Y analogously, leading to $\mathbb{P}_\nu$ as well as $\mathbb{P}^\lambda_\nu$.

From this point onward we envision the process X evolving "rightwards" from some lattice point $(0, x)$ (where possibly $x = X_0 \sim \lambda$) and the process Y evolving "leftwards" from the lattice point (T, y) at time zero. In this context a (stopping) time τ for the "backward" process Y will be measured as $T - \tau$ for the "forward" process X.

Let us define the discrete time equivalents of the objects introduced in Section 2. We consider a discrete delay stopping time $\eta \in \mathbb{N} = \{0, 1, \dots\}$ for the process X. Then analogous to the continuous case we define the space time measure

$$\alpha := \mathcal{L}^\lambda((\eta, X_\eta)) \in \mathcal{P}(\mathbb{N} \times \mathbb{Z}),$$

its spatial marginal

$$\alpha_X(A) := \alpha(\mathbb{N} \times A) = \mathcal{L}^\lambda(X_\eta)(A) \quad \text{for } A \subseteq \mathbb{Z} \text{ measurable}$$

and the $\mathbb{P}^\lambda$-potential of $X_{\eta\wedge T}$ by

$$V^\alpha_T(y) := -\mathbb{E}^\lambda\left[|X_{\eta\wedge T} - y|\right].$$

For a Root (resp. Rost) Barrier $R \subseteq \mathbb{Z} \times \mathbb{Z}$ consider the discrete, delayed Root (resp. Rost) stopping time

$$\rho_\eta := \inf\{t \geq \eta : (t, X_t) \in R\}.$$

Similarly for a Root (resp. Rost) field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ and a vector $U = (u_0, u_1, \dots)$ of iid $U[0,1]$-distributed random variables we consider the discrete, delayed Root (resp. Rost) stopping time

$$\rho_\eta := \inf\{t \geq \eta : r_t(X_t) > u_t\}.$$

We will throughout this paper assume that $(X_{\rho^\eta\wedge t})_{t\in\mathbb{N}}$ is uniformly integrable. The stopping time ρ_η induces the measures

$$\beta^\alpha := \mathcal{L}^\lambda\left(X_{\rho_\eta}\right) \text{ and } \beta^\alpha_T := \mathcal{L}^\lambda\left(X_{\rho_\eta\wedge T}\right) \text{ for } T \in \mathbb{N},$$

as well as the (stopped) potential

$$U_{\beta^\alpha_T}(y) = -\mathbb{E}^\lambda\left[\left|X_{\rho_\eta\wedge T} - y\right|\right].$$

Moreover, given a field $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ of Root (resp. Rost) stopping probabilities we denote by $\mathcal{R}_r$ the set of all fields of Root (resp. Rost) stopping probabilities that agree with r everywhere except on the boundary, precisely

$$\mathcal{R}_r := \left\{ (s_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}} : \begin{array}{l} (s_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}} \text{ is a Root (resp. Rost) field of stopping} \\ \text{probabilities s.t. } s_t(x) = r_t(x) \text{ given } r_t(x) \in \{0,1\}. \end{array} \right\}.$$

We will sometimes use the shorthand notation $s \in \mathcal{R}_r$ for $s = (s_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ when the context makes it obvious. Furthermore, we define the set of all associated stopping times in the following way

$$\mathcal{T}_r := \left\{ \inf\{t \geq 0 : s_{T-t}(Y_t) > v_t\} : \begin{array}{l} (s_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}} \in \mathcal{R}_r \text{ and } V = (v_0, \dots, v_T) \text{ denotes} \\ \text{a vector of iid } U[0,1]\text{-distributed random variables.} \end{array} \right\}.$$

Then we can then give the following discrete time analogue of (2.3)-(2.4),

Theorem 3.5. *Let $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ denote a Root field of stopping probabilities, then we have the representation*

$$U_{\beta_T^\alpha}(y) = \mathbb{E}_y\left[V^\alpha_{T-\tau^*}(Y_{\tau^*}) + (U_{\beta^\alpha} - U_{\alpha_X})(Y_{\tau^*})\mathbb{1}_{\tau^*<T}\right] \tag{3.11}$$
$$= \sup_{\tau\le T}\mathbb{E}_y\left[V^\alpha_{T-\tau}(Y_\tau) + (U_{\beta^\alpha} - U_{\alpha_X})(Y_\tau)\mathbb{1}_{\tau<T}\right], \tag{3.12}$$

where all optimisers will be of the form $\tau^* := \tau\wedge T$ for $\tau\in\mathcal{T}_r$.

When ρ_η corresponds to a discrete delayed Rost stopping the discrete version of (2.5)-(2.6) reads the following.

Theorem 3.6. *Let $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ denote a Rost field of stopping probabilities, then we have the representation*

$$U_{\beta^\alpha}(x) - U_{\beta_T^\alpha}(x) = \mathbb{E}^x\left[U_{\beta^\alpha}(X_{\sigma^*}) - V^\alpha_{T-\sigma^*}(X_{\sigma^*})\right] \tag{3.13}$$
$$= \sup_{\sigma\le T}\mathbb{E}^x\left[U_{\beta^\alpha}(X_\sigma) - V^\alpha_{T-\sigma}(X_\sigma)\right], \tag{3.14}$$

where all optimisers will be of the form $\sigma^* := \sigma\wedge T$ for $\sigma\in\mathcal{T}_r$.

It will become clear in Section 3.4 why the processes X and Y appear differently in the Root and Rost optimal stopping problem.

3.2.1 The core argument

We present the core argument repeatedly used in [1] and crucial for further results, an equality in expectation for differences of independent random walks.

$$\mathbb{E}^x_y\left[|X_{T-s} - Y_s|\right] = \mathbb{E}^x_y\left[|X_{T-(s-1)} - Y_{s-1}|\right]. \tag{3.15}$$

The following two remarks break down the key ingredients needed for proving the core argument.

Remark 3.7. Let Z be a Rademacher random variable. Then for any $a\in\mathbb{Z}$ we have

$$\mathbb{E}\left[|a+Z|\right] = |a| + \mathbb{1}_{a=0}.$$

Remark 3.8. For a sigma algebra $\mathcal{H}$ let A be a $\mathcal{H}$-measurable random variable taking values in $\mathbb{Z}$ and let Z be a Rademacher random variable independent of $\mathcal{H}$. Then by Remark 3.7 and the Independence Lemma we can conclude

$$\mathbb{E}\left[|A+Z|\big|\mathcal{H}\right] = |A| + \mathbb{1}_{A=0}.$$

We can now prove the core argument (3.15). On the one hand we have

$$\begin{aligned}
\mathbb{E}^x_y\left[|X_{T-s} - Y_s|\right] &= \mathbb{E}^x_y\left[\mathbb{E}^x_y\left[|X_{T-s} - Y_s|\big|X_{T-s}, Y_{s-1}\right]\right]\\
&= \mathbb{E}^x_y\left[\mathbb{E}^x_y\left[|(X_{T-s} - Y_{s-1}) - (Y_s - Y_{s-1})|\big|X_{T-s}, Y_{s-1}\right]\right]\\
&= \mathbb{E}^x_y\left[|X_{T-s} - Y_{s-1}| + \mathbb{1}_{X_{T-s}=Y_{s-1}}\right]
\end{aligned}$$

where we applied Remark 3.8 for the sigma algebra $\mathcal{H} := \sigma(X_{T-s}, Y_{s-1})$, the $\mathcal{G}$-measurable random variables $A := X_{T-s} - Y_{s-1}$ and the Rademacher random variable $Z := Y_s - Y_{s-1}$ which is independent of $\mathcal{H}$. Analogously, we can show

$$\begin{aligned}
\mathbb{E}^x_y\left[|X_{T-(s-1)} - Y_{s-1}|\right] &= \mathbb{E}^x_y\left[\mathbb{E}^x_y\left[|X_{T-s+1} - Y_{s-1}|\big|X_{T-s}, Y_{s-1}\right]\right]\\
&= \mathbb{E}^x_y\left[\mathbb{E}^x_y\left[|(X_{T-s} - Y_{s-1}) + (X_{T-s+1} - X_{T-s})|\big|X_{T-s}, Y_{s-1}\right]\right]\\
&= \mathbb{E}^x_y\left[|X_{T-s} - Y_{s-1}| + \mathbb{1}_{X_{T-s}=Y_{s-1}}\right],
\end{aligned}$$

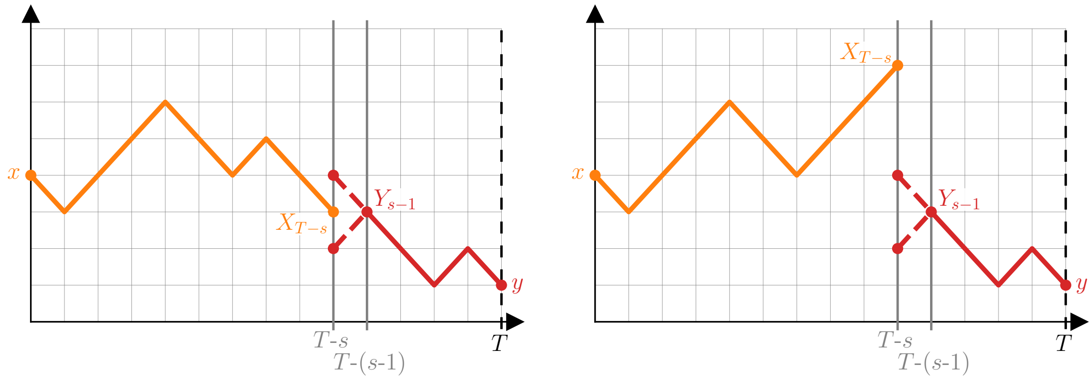


Figure 3: Illustrating the appearance of the indicator function in the core argument.

so altogether we have (3.15). See also Figure 3 for an illustration of the appearance of the indicator function. In particular, Equation (3.15) implies an independence of s, more precisely

$$\mathbb{E}_y^x\left[|X_{T-s} - Y_s|\right] = \mathbb{E}_y^x\left[|X_T - Y_0|\right] = \mathbb{E}_y^x\left[|X_0 - Y_T|\right].$$

Note that we can replace s by a Y stopping time $\tau < T$ to receive

$$\mathbb{E}_y^x\left[|X_{T-\tau} - Y_\tau|\right] = \mathbb{E}_y^x\left[|X_T - Y_0|\right].$$

Analogously, for an X stopping time $\sigma < T$ we have

$$\mathbb{E}_y^x\left[|X_\sigma - Y_{T-\sigma}|\right] = \mathbb{E}_y^x\left[|X_0 - Y_T|\right],$$

collectively this leads to the identity

$$\mathbb{E}_y^x\left[|X_{T-\tau} - Y_\tau|\right] = \mathbb{E}_y^x\left[|X_\sigma - Y_{T-\sigma}|\right] \tag{3.16}$$

which will serve as a key identity for subsequent arguments.

3.2.2 The interpolation function

In [1] it was crucial to define an interpolating function $F(s)$ which connects between the LHS of Equation (1.2) and the RHS of Equations (1.2) and (1.3). With only slight modification of the proofs in [1] we can define a more general interpolation function and state the following lemma.

Lemma 3.9. *Let λ be a starting measure for the process X and ν be a starting measure for the process Y. For an X stopping time σ, a Y stopping time τ and any $T \geq 0$ we consider the following interpolation function*

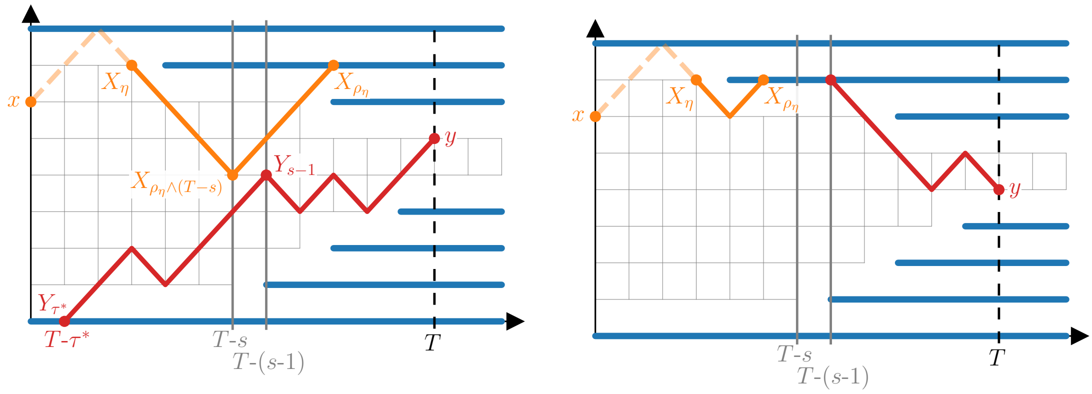
$$F_{\sigma,\tau}^{\lambda,\nu}(s) := \mathbb{E}_\nu^\lambda\left[\left|X_{\sigma\wedge(T-\tau\wedge s)} - Y_{\tau\wedge s}\right|\right]. \tag{3.17}$$

Then we have the following:

(i) The function $F_{\sigma,\tau}^{\lambda,\nu}(s)$ is increasing in s, more precisely

$$\begin{aligned} F_{\sigma,\tau}^{\lambda,\nu}(s) &= \mathbb{E}_\nu^\lambda\left[\left|X_{\sigma\wedge(T-\tau\wedge s)} - Y_{\tau\wedge(s-1)}\right| + \mathbb{1}_{X_{\sigma\wedge(T-s)}=Y_{s-1},\tau\geq s}\right] &(3.18)\\ &\geq \mathbb{E}_\nu^\lambda\left[\left|X_{\sigma\wedge(T-\tau\wedge s)} - Y_{\tau\wedge(s-1)}\right| + \mathbb{1}_{X_{\sigma\wedge(T-s)}=Y_{s-1},\tau\geq s,\sigma>T-s}\right] = F_{\sigma,\tau}^{\lambda,\nu}(s-1). &(3.19)\end{aligned}$$

(ii) If σ is a (possibly randomised, possibly delayed) forward Root stopping time for the process X and τ is a (possibly randomised) backwards Root stopping time for the process Y then $F_{\sigma,\tau}^{\lambda,\nu}(s)$ is constant in s.



(A) Here $\tau^* \geq s$ and $X_{\rho_\eta \wedge (T-s)} = Y_{s-1}$. We see that in this case $\rho_\eta \wedge (T-s) = T-s$ must hold as otherwise Y_{τ^*} would have stopped earlier.

(B) Here $\rho_\eta \wedge (T-s) = \rho_\eta$. We see how $X_{\rho_\eta \wedge (T-s)} = Y_{s-1}$ implies that we must have $\tau^* \leq s-1$ due to the Root barrier structure.

Figure 4: An illustration of some different cases that appear when investigating equality of the interpolation function for forward and backward Root stopping times.

Proof. The essence of the proof of (i) is an appropriate application of the Core Argument presented above, and (i) follows in complete analogy to the proofs of [1, Lemma 3.3, Lemma 3.4]. For (ii) we have to make some modifications.

Let $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ denote a Root field of stopping probabilities, then for $(s_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}} \in \mathcal{R}_r$ consider the stopping times

$$\begin{aligned} \rho^\eta &:= \inf\{t \geq \eta : r_t(X_t) > u_t\}, \quad \text{and} \\ \tau^* &:= \inf\{t \geq 0 : s_{T-t}(Y_t) > v_t\} \wedge T. \end{aligned}$$

In order to show equality in (i) for these stopping times, it suffices to show equality of the indicator functions appearing in (3.18) and (3.19), equivalently

$$\{\rho^\eta > T-s\} \supseteq \{X_{\rho^\eta \wedge (T-s)} = Y_{s-1}, \tau^* \geq s\}$$

up to nullsets. Indeed, on $\{\tau^* \geq s\}$ we deduce $\tau^* > s-1$. Hence we must have $s_{T-(s-1)}(Y_{s-1}) < 1$ as otherwise $\tau^* = s-1$ almost surely. By the definition of the Root stopping probabilities we thus have $s_u(Y_{s-1}) = r_u(Y_{s-1}) = 0$ for all $u < T-(s-1)$. Moreover, on the set $\{X_{\rho^\eta \wedge (T-s)} = Y_{s-1}\}$ we have $r_u\left(X_{\rho^\eta \wedge (T-s)}\right) = 0$ for all $u < T-(s-1)$. In particular, since $\rho^\eta \wedge (T-s) < T-(s-1)$, on the set $\{X_{\rho^\eta \wedge (T-s)} = Y_{s-1}\}$ we have $r_{\rho^\eta \wedge (T-s)}\left(X_{\rho^\eta \wedge (T-s)}\right) = 0$. Hence $\rho^\eta > T-s$ needs to be satisfied as otherwise the Root properties of the stopping probabilities would be violated. See Figure 4 for an illustration of these arguments. □

3.3 The Root optimal stopping representation

Let us introduce some frequently used notations. For the choice of $\nu = \delta_y$ we simplify the notation of the interpolation function between the forward and backward Root stopping times to

$$F^\alpha(s) := F^{\lambda,y}_{\rho_\eta,\tau^*}(s).$$

Here α denotes the space-time law induced by the delay stopping time η and β^α is the corresponding delayed Root law as defined in Section 3.2. For a given stopping time τ consider the function

$$u^\alpha_\tau(T,y) := \mathbb{E}_y\left[V^\alpha_{T-\tau}(Y_\tau) + (U_{\beta^\alpha} - U_{\alpha_X})(Y_\tau)\mathbb{1}_{\tau<T}\right]. \tag{3.20}$$

We also introduce
$$u^\alpha(T,y) := u^\alpha_{\tau^*}(T,y)$$
for a backward Root stopping time $\tau^* = \inf\{t \geq 0 : s_{T-t}(Y_t) > v_t\} \wedge T$. Considering this refined notation Equations (3.11)-(3.12) can be written in the following way
$$U_{\beta^\alpha_T}(y) = u^\alpha(T,y) \tag{3.21}$$
$$= \sup_{\tau\leq T} u^\alpha_\tau(T,y). \tag{3.22}$$

Furthermore, it will be of importance that the function u^α_τ has the following alternative representation
$$u^\alpha_\tau(T,y) = \mathbb{E}_y\left[V^\alpha_{T-\tau}(Y_\tau) + (U_{\beta^\alpha} - U_{\alpha_X})(Y_\tau)\mathbb{1}_{\tau<T}\right] \tag{3.23}$$
$$= -\mathbb{E}^\lambda_y\left[\left|X_{\eta\wedge(T-\tau)} - Y_\tau\right| + \left(\left|X_{\rho_\eta} - Y_\tau\right| - |X_\eta - Y_\tau|\right)\mathbb{1}_{\tau<T}\right]. \tag{3.24}$$

To further simplify and rewrite these expressions we observe the following lemma.

Lemma 3.10. *Let η and θ be delay stopping times for the processes X and Y respectively. By $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ we denote a Root field of stopping probabilities. For mutually independent $U[0,1]$-distributed random variables $u_0, u_1, \ldots, v_0, v_1, \ldots$ we consider the forward Root stopping time $\rho := \inf\{t \geq \eta : r_t(X_t) > u_t\}$ and for $s \in \mathcal{R}_r$ the backward Root stopping time $\tau := \inf\{t \geq 0 : s_{T-t}(Y_t) > v_t\} \in \mathcal{T}_r$.*

Furthermore we introduce the filtration $\mathcal{G}_t := \sigma\left((X_s)_{0\leq s\leq t}, (Y_s)_{0\leq s\leq T}, u_1, \ldots, u_t, v_1, \ldots, v_T\right)$. Then for all $A \in \mathcal{H} := \mathcal{G}_{\rho\wedge(\eta\vee(T-\tau))}$ with $\{\tau < T\} \subseteq A$ we have
$$\mathbb{E}^\lambda_\nu\left[|X_\rho - Y_\tau|\,\mathbb{1}_A\right] = \mathbb{E}^\lambda_\nu\left[\left|X_{\rho\wedge(\eta\vee(T-\tau))} - Y_\tau\right|\mathbb{1}_A\right]. \tag{3.25}$$

Proof. As X is independent of Y, we have that X is a $(\mathcal{G}_t)$ martingale and furthermore η, ρ as well as $\tilde\sigma := \eta \vee (T-\tau)$ are $(\mathcal{G}_t)$ stopping times. Hence, by the optional sampling theorem we can conclude that for every $A \in \mathcal{H} = \mathcal{G}_{\rho\wedge\tilde\sigma}$ the following holds
$$\mathbb{E}^\lambda_\nu\left[X_\rho \mathbb{1}_A\right] = \mathbb{E}^\lambda_\nu\left[X_{\rho\wedge\tilde\sigma}\mathbb{1}_A\right]. \tag{3.26}$$

By definition of the Root stopping probabilities and the definition of τ, on the set $\{\tau < T\}$ we have $s_{T-\tau}(Y_\tau) > 0$. Furthermore $s_{T-\tau+u}(Y_\tau) = r_{T-\tau+u}(Y_\tau) = 1$ for all $u \geq 1$. Moreover, on the set $\{\rho > \tilde\sigma, X_{\tilde\sigma} < Y_\tau\}$ we must have that $X_\rho \leq Y_\tau$ as otherwise the Root property of the stopping probabilities would be violated. We can argue analogously that on $\{\rho > \tilde\sigma, X_{\tilde\sigma} \geq Y_\tau\}$ we must have $X_\rho \geq Y_\tau$. Then
$$\begin{aligned}
&\mathbb{E}^\lambda_\nu\left[|X_\rho - Y_\tau|\,\mathbb{1}_A\right]\\
&= \mathbb{E}^\lambda_\nu\left[|X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho\leq\tilde\sigma} + |X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho>\tilde\sigma}\right]\\
&= \mathbb{E}^\lambda_\nu\left[|X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho\leq\tilde\sigma} + |X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho>\tilde\sigma,X_{\tilde\sigma}<Y_\tau} + |X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho>\tilde\sigma,X_{\tilde\sigma}\geq Y_\tau}\right]\\
&= \mathbb{E}^\lambda_\nu\left[|X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho\leq\tilde\sigma} - (X_\rho - Y_\tau)\,\mathbb{1}_{A,\rho>\tilde\sigma,X_{\tilde\sigma}<Y_\tau} + (X_\rho - Y_\tau)\,\mathbb{1}_{A,\rho>\tilde\sigma,X_{\tilde\sigma}\geq Y_\tau}\right]\\
&= \mathbb{E}^\lambda_\nu\left[|X_\rho - Y_\tau|\,\mathbb{1}_{A,\rho\leq\tilde\sigma} - (X_{\rho\wedge\tilde\sigma} - Y_\tau)\,\mathbb{1}_{A,\rho>\tilde\sigma,X_{\tilde\sigma}<Y_\tau} + (X_{\rho\wedge\tilde\sigma} - Y_\tau)\,\mathbb{1}_{A,\rho>\tilde\sigma,X_{\tilde\sigma}\geq Y_\tau}\right]\\
&= \mathbb{E}^\lambda_\nu\left[|X_{\rho\wedge\tilde\sigma} - Y_\tau|\,\mathbb{1}_A\right].
\end{aligned}$$
where we use Equation (3.26) and the fact that $\{\rho > \tilde\sigma, X_{\tilde\sigma} < Y_\tau\}, \{\rho > \tilde\sigma, X_{\tilde\sigma} \geq Y_\tau\} \in \mathcal{H}$. □

Choosing $\nu = \delta_y$, $\rho = \rho_\eta$ and $\theta = 0$ in the lemma above enables us to prove the following.

Lemma 3.11. *The RHS of Equation* (3.11) *equals the terminal value of the interpolation function, that is*
$$-u^\alpha(T,y) = F^\alpha(T). \tag{3.27}$$

Proof. Recall the representation (3.24) and consider the following decomposition

$$\begin{aligned}
-u^{\alpha}(T,y) =& \mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau^*)} - Y_{\tau^*}\right| + \left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T}\right] \\
=& \mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau^*)} - Y_{\tau^*}\right|\mathbb{1}_{\eta>T} + \left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T,\eta>T}\right] && (3.28) \\
+& \mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau^*)} - Y_{\tau^*}\right|\mathbb{1}_{\eta\leq T} + \left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T,\eta\leq T}\right]. && (3.29)
\end{aligned}$$

We will investigate the two lines (3.28) and (3.29) separately.

Let us first consider (3.28). As $\rho_\eta \geq \eta$, we observe that on the event $\{\eta > T\}$ we have $\eta \wedge (T-\tau^*) = T - \tau^* = \rho_\eta \wedge (T-\tau^*)$, hence for the first term of (3.28) we can conclude

$$\mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau^*)} - Y_{\tau^*}\right|\mathbb{1}_{\eta>T}\right] = \mathbb{E}_y^{\lambda}\left[\left|X_{\rho_\eta\wedge(T-\tau^*)} - Y_{\tau^*}\right|\mathbb{1}_{\eta>T}\right]. \tag{3.30}$$

Furthermore, note the following:

- On the event $\{\tau^* < T\}$ the stopping time τ^* equals a backward Root stopping time.
- On the event $\{\eta > T\}$ we have $\eta \vee (T-\tau^*) = \eta = \rho_\eta \wedge (\eta \vee (T-\tau^*))$.
- We have $\{\tau^* < T, \eta > T\} \in \mathcal{H}$ (for $\mathcal{H}$ defined in Lemma 3.10).

Therefore we can apply Lemma 3.10 to the second term of (3.28) to arrive at

$$\begin{aligned}
&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T,\eta>T}\right] \\
=&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - \left|X_{\rho_\eta\wedge(\eta\vee(T-\tau^*))} - Y_{\tau^*}\right|\right)\mathbb{1}_{\tau^*<T,\eta>T}\right] = 0. && (3.31)
\end{aligned}$$

To summarise, line (3.28) reduces to

$$\mathbb{E}_y^{\lambda}\left[\left|X_{\rho_\eta\wedge(T-\tau^*)} - Y_{\tau^*}\right|\mathbb{1}_{\eta>T}\right]. \tag{3.32}$$

Next we consider (3.29). Let us start by examining the first term of (3.29) which can be decomposed further into

$$\begin{aligned}
&\mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau)} - Y_{\tau^*}\right|\mathbb{1}_{\eta\leq T}\right] \\
=&\mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau)} - Y_{\tau^*}\right|\mathbb{1}_{\eta\leq T,T-\eta\leq\tau^*}\right] + \mathbb{E}_y^{\lambda}\left[\left|X_{\eta\wedge(T-\tau)} - Y_{\tau^*}\right|\mathbb{1}_{\eta\leq T,\tau^*<T-\eta}\right] \\
=&\mathbb{E}_y^{\lambda}\left[\left|X_{\rho_\eta\wedge(T-\tau)} - Y_{\tau^*}\right|\mathbb{1}_{\eta\leq T,T-\eta\leq\tau^*}\right] + \mathbb{E}_y^{\lambda}\left[|X_\eta - Y_{\tau^*}|\mathbb{1}_{\eta\leq T,\tau^*<T-\eta}\right]. && (3.33)
\end{aligned}$$

Similarly, the second term of (3.29) decomposes into

$$\begin{aligned}
&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T,\eta\leq T}\right] \\
=&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T,\eta\leq T,T-\eta\leq\tau^*}\right] && (3.34) \\
+&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\eta\leq T,\tau^*<T-\eta}\right]. && (3.35)
\end{aligned}$$

Importantly, the second term of (3.33) cancels out the second term of (3.35).

We will proceed to conclude that the terms depicted in (3.34) will equate to 0. For this purpose observe that:

- On the event $\{\tau^* < T\}$ the stopping time τ^* equals a backward Root stopping time.
- On the event $\{T-\eta \leq \tau^*\}$ we have $\eta \vee (T-\tau^*) = \eta = \rho_\eta \wedge (\eta \vee (T-\tau^*))$.
- We have $\{\tau^* < T, \eta \leq T, T-\eta \leq \tau^*\} \in \mathcal{H}$.

Hence, we can apply Lemma 3.10 to deduce

$$\begin{aligned}
&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - |X_\eta - Y_{\tau^*}|\right)\mathbb{1}_{\tau^*<T,\eta\leq T,T-\eta\leq\tau^*}\right] \\
=&\mathbb{E}_y^{\lambda}\left[\left(\left|X_{\rho_\eta} - Y_{\tau^*}\right| - \left|X_{\rho_\eta\wedge(\eta\vee(T-\tau^*))} - Y_{\tau^*}\right|\right)\mathbb{1}_{\tau^*<T,\eta\leq T,T-\eta\leq\tau^*}\right] \\
=&0. && (3.36)
\end{aligned}$$

The final term left to consideration is the first term in (3.35). For this purpose, note that

- on the event $\{\tau^* < T-\eta\}$ we have $\eta \vee (T-\tau^*) = (T-\tau^*)$, and
- $\{\eta \le T, \tau^* < T-\eta\} \in \mathcal{H}$.

Hence, application of Lemma 3.10 to the first term of (3.35) yields

$$\mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta} - Y_{\tau^*} \right| \mathbb{1}_{\eta \le T, \tau^* < T-\eta} \right] = \mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau^*)} - Y_{\tau^*} \right| \mathbb{1}_{\eta \le T, \tau^* < T-\eta} \right]. \tag{3.37}$$

To summarise, (3.29) reduces to

$$\mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau^*)} - Y_{\tau^*} \right| \mathbb{1}_{\eta \le T} \right]. \tag{3.38}$$

Combining the observations (3.32) and (3.38) the desired identity now becomes easy to see:

$$-u^\alpha(T,y) = \mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau^*)} - Y_{\tau^*} \right| \right] = F^\alpha(T).$$

□

We are ready to prove Theorem 3.5, the optimal stopping representation of Root solutions to (dSEP).

Proof of Theorem 3.5. We begin by proving (3.11) through consideration of the interpolation function

$$F^\alpha(s) = F^{\lambda,y}_{\rho_\eta,\tau^*}(s) = \mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau^* \wedge s)} - Y_{\tau^* \wedge s} \right| \right].$$

Then $-F^\alpha(0) = \mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge T} - Y_0 \right| \right] = U_{\beta^\alpha_T}(y)$ which equals the LHS of (3.11), equivalently of (3.21). It was the purpose of Lemma 3.11 to show that $-F^\alpha(T)$ equals the RHS of (3.11), equivalently of (3.21). The equality in (3.11) now follows by Lemma 3.9 (*ii*), which states that the interpolation function $F^\alpha(T)$ is constant given our consideration of Root stopping times.

Let us proceed to establish the optimal stopping problem (3.12). We keep the Root stopping time ρ_η but allow for an arbitrary Y stopping time $\tau \le T$. It still remains true that $F^{\lambda,y}_{\rho_\eta,\tau}(0) = -U_{\beta^\alpha}(y)$, and due to Lemma 3.9 (*i*) we have the inequality $-U_{\beta^\alpha}(y) \le F^{\lambda,y}_{\rho_\eta,\tau}(T)$. Therefore, we are left to show that

$$F^{\lambda,y}_{\rho_\eta,\tau}(T) \le u^\alpha_\tau(T,y) = -\mathbb{E}_y \left[V^\alpha_{T-\tau}(Y_\tau) + (U_{\beta^\alpha} - U_{\alpha_X})(Y_\tau) \mathbb{1}_{\tau<T} \right],$$

or equivalently

$$\mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau)} - Y_\tau \right| \right] \le \mathbb{E}_y^\lambda \left[\left| X_{\eta \wedge (T-\tau)} - Y_\tau \right| + \left(\left| X_{\rho_\eta} - Y_\tau \right| - \left| X_\eta - Y_\tau \right| \right) \mathbb{1}_{\tau<T} \right]. \tag{3.39}$$

We will consider a decomposition into the events $\{\eta > T-\tau\}$ and $\{\eta \le T-\tau\}$.

Due to Jensen's inequality and optional sampling we have the following

$$\mathbb{E}_y^\lambda \left[\left(\left| X_{\rho_\eta} - Y_\tau \right| - \left| X_\eta - Y_\tau \right| \right) \mathbb{1}_{\tau<T, \eta>T-\tau} \right] \ge 0$$

and can conclude

$$\begin{aligned}
&\mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau)} - Y_\tau \right| \mathbb{1}_{\eta > T-\tau} \right] \\
&= \mathbb{E}_y^\lambda \left[\left| X_{\eta \wedge (T-\tau)} - Y_\tau \right| \mathbb{1}_{\eta > T-\tau} \right] \\
&\le \mathbb{E}_y^\lambda \left[\left| X_{\eta \wedge (T-\tau)} - Y_\tau \right| \mathbb{1}_{\eta > T-\tau} + \left(\left| X_{\rho_\eta} - Y_\tau \right| - \left| X_\eta - Y_\tau \right| \right) \mathbb{1}_{\tau<T, \eta>T-\tau} \right].
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
&\mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau)} - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau} \right] \\
&= \mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta \wedge (T-\tau)} - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau, \tau<T} + \left| X_0 - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau, \tau=T} \right] \\
&\le \mathbb{E}_y^\lambda \left[\left| X_{\rho_\eta} - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau, \tau<T} + \left| X_\eta - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau, \tau=T} \right] \\
&= \mathbb{E}_y^\lambda \left[\left| X_\eta - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau} + \left(\left| X_{\rho_\eta} - Y_\tau \right| - \left| X_\eta - Y_\tau \right| \right) \mathbb{1}_{\eta \le T-\tau, \tau<T} \right] \\
&= \mathbb{E}_y^\lambda \left[\left| X_{\eta \wedge (T-\tau)} - Y_\tau \right| \mathbb{1}_{\eta \le T-\tau} + \left(\left| X_{\rho_\eta} - Y_\tau \right| - \left| X_\eta - Y_\tau \right| \right) \mathbb{1}_{\eta \le T-\tau, \tau<T} \right]
\end{aligned}$$

where the inequality again follows by Jensen's inequality and optional sampling. Together this proves (3.39). Thus we have shown that

$$U_{\beta^\alpha}(y) \geq u^\alpha_\tau(T, y)$$

which concludes the proof of (3.12). □

3.4 The Root-Rost-symmetry

To establish the optimal stopping representation of Rost solution to (dSEP) we could simply repeat the appropriate analogues of the arguments in the section above. Instead however we will derive the Rost representation from the Root one by observing a symmetry between the two similar to the approach in [1].

We will first give a brief summary of the results in [1] which were stated in non-randomised terms.

Recall the definition of a discrete Rost barrier $\bar{R} \subseteq \mathbb{N} \times \mathbb{Z}$, that is

- If $(t, m) \in \bar{R}$ then for all $s < t$ also $(s, m) \in \bar{R}$.

Now observe that for any $T \in \mathbb{N}$ we can transform this Rost barrier into a Root Barrier R_T in the following way

$$R_T := \{(T - t, x) : (t, x) \in R\}. \tag{3.40}$$

To more conveniently apply results from the sections above we will reverse the roles of X and Y for Rost stopping times.

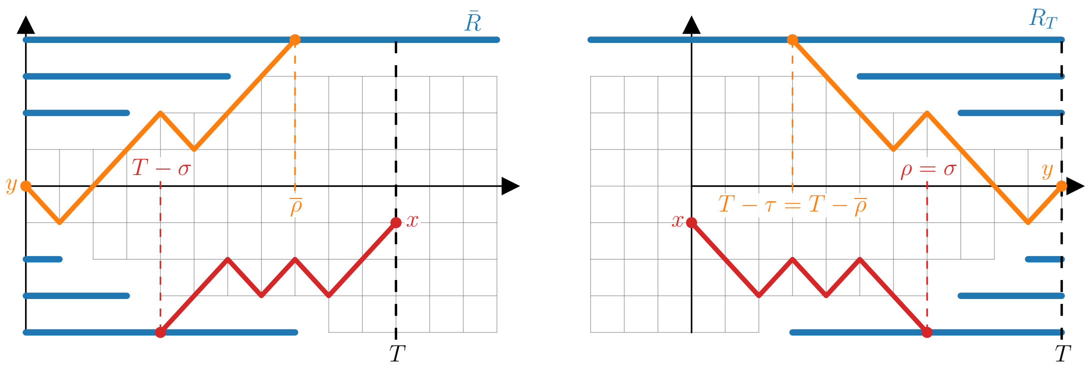


Figure 5: An illustration of the connection between $\bar{R}$ and R_T, resp. between $\bar{\rho}$ and τ as well as σ and ρ without delay.

Consider a non-delayed (forward) Rost stopping time $\bar{\rho} = \inf\left\{t \geq 0 : (t, Y_t) \in \bar{R}\right\}$. We can now observe as illustrated in Figure 5 that the (forward) Rost stopping time $\bar{\rho}$ can be represented as a backward Root stopping time $\tau := \inf\{t \geq 0 : (T - t, Y_t) \in R_T\}$, and that the backwards Rost stopping time $\sigma := \inf\left\{t \geq 0 : (T - t, X_t) \in \bar{R}\right\}$ can be represented as a (non-delayed, forward) Root stopping time $\rho := \rho_0 = \inf\{t \geq 0 : (t, X_t) \in R_T\}$, to summarise

$$\begin{aligned} \bar{\rho} &= \inf\left\{t \geq 0 : (t, Y_t) \in \bar{R}\right\} = \inf\{t \geq 0 : (T - t, Y_t) \in R_T\} = \tau \\ \sigma &= \inf\left\{t \geq 0 : (T - t, X_t) \in \bar{R}\right\} = \inf\{t \geq 0 : (t, X_t) \in R_T\} = \rho. \end{aligned} \tag{3.41}$$

Using this symmetry, in [1] the Rost optimal stopping problem was then derived with the methods established for the Root optimal stopping problem using the following proposition.

Proposition 3.12 ([1, Proposition 3.7])**.** *For each x, y, T, any Y-stopping time τ such that $\mathbb{E}_y[|Y_\tau|] < \infty$, and every $\{0, \dots, T\}$-valued X-stopping time σ, we have*

$$\mathbb{E}_y\left[|x - Y_\tau| - |x - Y_{\tau\wedge T}|\right] \leq \mathbb{E}_y^x\left[|X_\sigma - Y_\tau| - |X_\sigma - y|\right]. \tag{3.42}$$

Suppose furthermore that

$$\tau = \inf\{t \in \mathbb{N} : (T - t, Y_t) \in R\}, \tag{3.43}$$

for a Root barrier R and that $\sigma = \rho^{Root} \wedge T$. Then there is equality in (3.42).

The proof given for [1, Proposition 3.7] relied heavily on the interpolation function $F_{\sigma,\tau}(s) = F_{\sigma,\tau}^{x,y}(s) = \mathbb{E}_y^x\left[\left|X_{\sigma\wedge(T-\tau\wedge s)} - Y_{\tau\wedge s}\right|\right]$ in the following way. To see inequality we observe

$$\begin{aligned}
\mathbb{E}_y\left[|x - Y_\tau| - |x - Y_{\tau\wedge T}|\right] &\leq \mathbb{E}_y^x\left[|X_\sigma - Y_\tau|\right] - F_{\sigma,\tau}(T)\\
&\leq \mathbb{E}_y^x\left[|X_\sigma - Y_\tau|\right] - F_{\sigma,\tau}(0)\\
&= \mathbb{E}_y^x\left[|X_\sigma - Y_\tau| - |X_\sigma - y|\right].
\end{aligned}$$

Equality for Root forward and backward stopping times follows from Lemma 3.9 (ii). In the delayed regimes however, these arguments cannot be repeated verbatim as Lemma 3.9 (ii) does not hold for delayed backwards Root stopping times.

We will observe an analogous symmetry to (3.41) in the delayed regime and give the appropriate generalisation of the symmetries (3.41) as well as Proposition 3.12.

A delayed randomised Root-Rost-symmetry

Let $(\bar{r}_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ denote a field of Rost stopping probabilities. Then for any $T \in \mathbb{N}$ we can transform these Rost stopping probabilities into a field $(r_t^T(x))_{(t,x)\in\mathbb{Z}\times\mathbb{Z}}$ of Root stopping probabilities in the following way

$$r_t^T(x) := \bar{r}_{T-t}(x). \tag{3.44}$$

Note that the definitions of Root and Rost fields of stopping probabilities naturally extend to $(t, x) \in \mathbb{Z}\times\mathbb{Z}$. Consider a delayed (forward) Rost stopping time

$$\bar{\rho}_\theta = \inf\left\{t \geq \theta : \bar{r}_t(Y_t) > v_t\right\}.$$

We can now observe as illustrated in Figure 5 that the (forward) Rost stopping time $\bar{\rho}$ can be represented as a delayed backward Root stopping time

$$\tau_\theta := \inf\left\{t \geq \theta : r_{T-t}^T(Y_t) > v_t\right\},$$

and that the (undelayed) backwards Rost stopping time

$$\sigma := \inf\left\{t \geq 0 : \bar{r}_{T-t}(X_t) > u_t\right\}$$

can be represented as a (undelayed, forward) Root stopping time

$$\rho := \rho_0 = \inf\left\{t \geq 0 : r_t^T(X_t) > u_t\right\},$$

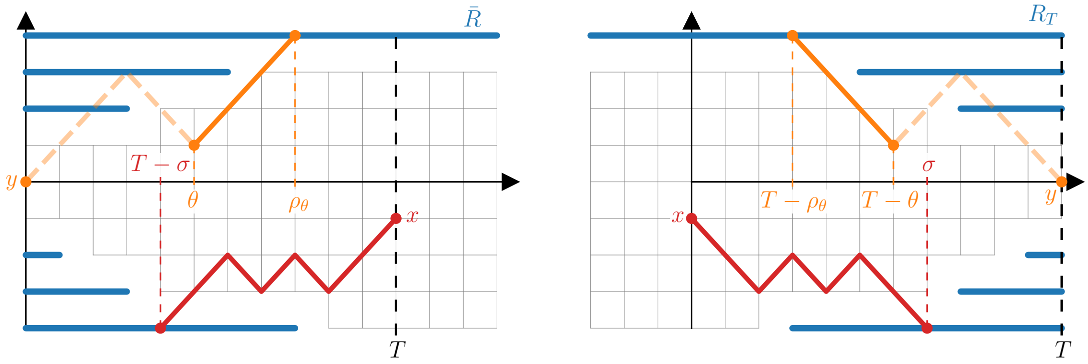
in sum

$$\bar{\rho}_\theta := \inf\left\{t \geq \theta : \bar{r}_t(Y_t) > v_t\right\} = \inf\left\{t \geq \theta : r_{T-t}^T(Y_t) > v_t\right\} =: \tau_\theta \tag{3.45}$$

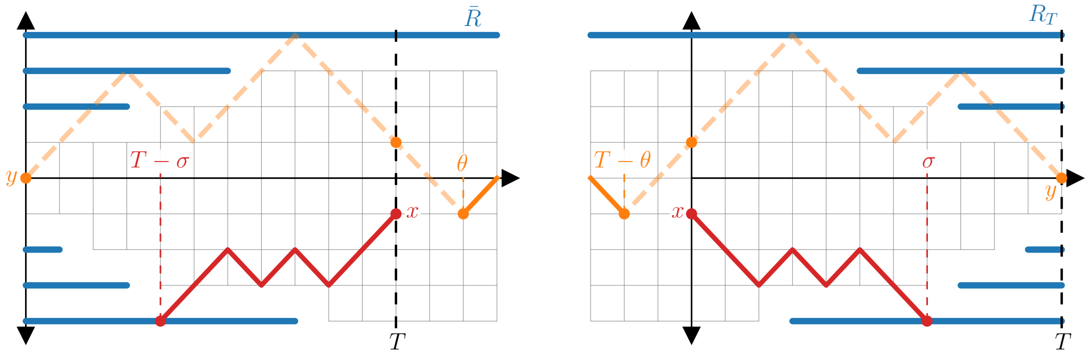
$$\sigma := \inf\left\{t \geq 0 : \bar{r}_{T-t}(X_t) > u_t\right\} = \inf\left\{t \geq 0 : r_t^T(X_t) > u_t\right\} =: \rho. \tag{3.46}$$

See also Figure 6 for an illustration of these stopping times.

We see that Proposition 3.12 is valid for *any* stopping times σ and τ, hence also for the choices $\sigma = \rho_0$ and $\tau = \tau_\theta$. However, in this case is was already observed in Section 3.2.2 that the



(A) An illustration of the connection between $\bar{R}$ and R_T in the delayed regime for $\theta < T$.



(B) An illustration of the connection between $\bar{R}$ and R_T in the delayed regime for $\theta \geq T$.

Figure 6

interpolation function $F_{\rho_0,\tau_\theta}(s)$ will in general no longer be constant when τ is subject to a non-zero delay θ. Thus we will no longer recover an equality in Equation (3.42) for a *delayed* backward Root stopping time. Moreover the RHS of Equation (3.42) will in general also not equal the RHS of Equation (3.13), thus cannot be applied to prove our desired generalisation of the delayed Rost optimal stopping representation.

The following generalisation of Proposition 3.12 is needed.

Proposition 3.13. *Let $\sigma \leq T$ be a stopping time for X and let τ be a stopping time for Y such that $\mathbb{E}_\nu[|Y_\tau|]$. Then*

(i)

$$\mathbb{E}_\nu\left[|x - Y_\tau| - |x - Y_{\tau\wedge T}|\right] \leq \mathbb{E}_\nu^x\left[|X_\sigma - Y_\tau| - \left|X_\sigma - Y_{\tau\wedge(T-\sigma)}\right|\right]. \tag{3.47}$$

(ii) For any Y-stopping time θ such that $\theta \leq \tau$ we have

$$\mathbb{E}_\nu\left[|x - Y_\tau| - |x - Y_{\tau\wedge T}|\right] \leq \mathbb{E}_\nu^x\left[|X_\sigma - Y_\tau| - \left|X_\sigma - Y_{\theta\wedge(T-\sigma)}\right|\right]. \tag{3.48}$$

(iii) Furthermore, let $(r_t(x))_{(t,x)\in\mathbb{Z}\times\mathbb{Z}}$ denotes a field of Root stopping probabilities, then consider a field $(s_t(x))_{(t,x)\in\mathbb{Z}\times\mathbb{Z}} \in \mathcal{R}_r$ and let θ denote a delay stopping time for Y. Then there is equality in (3.47) and (3.48) when considering the stopping times

$$\sigma = \inf\{t \geq 0 : r_t(X_t) > u_t\} \quad \textit{and}$$
$$\tau = \inf\{t \geq \theta : s_{T-t}(Y_t) > v_t\} \wedge T.$$

Proof. As observed in the proof of [1, Proposition 3.7] we can conclude the following with the help of Jensen's inequality and optional sampling

$$\mathbb{E}_\nu\left[|x-Y_\tau|-|x-Y_{\tau\wedge T}|\right] \leq \mathbb{E}_\nu^x\left[|X_\sigma-Y_\tau|-\left|X_{\sigma\wedge(T-\tau\wedge T)}-Y_{\tau\wedge T}\right|\right].$$

It is also a straightforward application of the Core Argument (3.15) together with optional sampling to obtain

$$\mathbb{E}_\nu^x\left[\left|X_{\sigma\wedge(T-\tau\wedge T)}-Y_{\tau\wedge T}\right|\right]=\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\tau\wedge(T-\sigma)}\right|\right]$$

which concludes the proof of (i).

Note that (ii) easily follows from (i) via Jensen's inequality and optional sampling by observing that when $\theta\leq\tau$ we clearly also have $\theta\wedge(T-\sigma)\leq\tau\wedge(T-\sigma)$.

It now remains to show (iii), the equality in the Root case. We will apply Lemma 3.10 for the choices $\lambda=\delta_x$, $\rho=\sigma$ and $\eta=0$. We will consider a decomposition into the events $\{\tau<T\}$ and $\{\tau\geq T\}$. First note that on $\{\tau<T\}$ we have $\eta\vee(T-\tau)=T-\tau$, hence by Lemma 3.10 we obtain

$$\mathbb{E}_\nu^x\left[|X_\sigma-Y_\tau|\,\mathbb{1}_{\tau<T}\right]=\mathbb{E}_\nu^x\left[\left|X_{\sigma\wedge(T-\tau)}-Y_\tau\right|\mathbb{1}_{\tau<T}\right],$$

thus

$$\begin{aligned}\mathbb{E}_\nu\left[(|x-Y_\tau|-|x-Y_{\tau\wedge T}|)\,\mathbb{1}_{\tau<T}\right]=0&=\mathbb{E}_\nu^x\left[\left(|X_\sigma-Y_\tau|-\left|X_{\sigma\wedge(T-\tau)}-Y_\tau\right|\right)\mathbb{1}_{\tau<T}\right]\\&=\mathbb{E}_\nu^x\left[\left(|X_\sigma-Y_\tau|-\left|X_{\sigma\wedge(T-\tau\wedge T)}-Y_{\tau\wedge T}\right|\right)\mathbb{1}_{\tau<T}\right].\end{aligned}$$

On the other hand, on $\{\tau\geq T\}$ we have $\eta\vee(T-\tau)=0$, hence by Lemma 3.10

$$\mathbb{E}_\nu^x\left[|X_\sigma-Y_\tau|\,\mathbb{1}_{\tau\geq T}\right]=\mathbb{E}_\nu^x\left[|X_{\sigma\wedge 0}-Y_\tau|\,\mathbb{1}_{\tau\geq T}\right]=\mathbb{E}_\nu^x\left[|X_0-Y_\tau|\,\mathbb{1}_{\tau\geq T}\right]=\mathbb{E}_\nu\left[|x-Y_\tau|\,\mathbb{1}_{\tau\geq T}\right]$$

which gives

$$\begin{aligned}\mathbb{E}_\nu\left[(|x-Y_\tau|-|x-Y_{\tau\wedge T}|)\,\mathbb{1}_{\tau\geq T}\right]&=\mathbb{E}_\nu^x\left[(|X_\sigma-Y_\tau|-|X_0-Y_{\tau\wedge T}|)\,\mathbb{1}_{\tau\geq T}\right]\\&=\mathbb{E}_\nu^x\left[\left(|X_\sigma-Y_\tau|-\left|X_{\sigma\wedge(T-\tau\wedge T)}-Y_{\tau\wedge T}\right|\right)\mathbb{1}_{\tau\geq T}\right].\end{aligned}$$

Altogether equality in (3.47) follows. Note that $\eta=0$ was essential here in order to gain this equality.

It now remains to establish equality in (3.48), that is we need to show

$$\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\tau\wedge(T-\sigma)}\right|\right]=\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\theta\wedge(T-\sigma)}\right|\right].$$

On the one hand we trivially have

$$\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\tau\wedge(T-\sigma)}\right|\mathbb{1}_{\sigma=T}\right]=\mathbb{E}_\nu^x\left[|X_\sigma-Y_0|\,\mathbb{1}_{\sigma=T}\right]=\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\theta\wedge(T-\sigma)}\right|\mathbb{1}_{\sigma=T}\right],$$

whereas

$$\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\tau\wedge(T-\sigma)}\right|\mathbb{1}_{\sigma<T}\right]=\mathbb{E}_\nu^x\left[\left|X_\sigma-Y_{\theta\wedge(T-\sigma)}\right|\mathbb{1}_{\sigma<T}\right]$$

can be argued verbatim as in Lemma 3.10, switching the roles of the two stopping times therein. □

Remark 3.14. Note that Proposition 3.12 can easily be recovered from Proposition 3.13 (ii) via the choices $\nu=\delta_y$ and $\theta=0$.

Equipped with this proposition and the (delayed) Root-Rost symmetries (3.45) and (3.46) we can now give a proof of Theorem 3.6, the optimal stopping representation of Rost solutions to (dSEP)

Proof of Theorem 3.6. We want to apply Proposition 3.13. Consider $\tau = \tau_\theta$ and note that for the LHS of (3.47) respectively (3.48) with the help of the Root-Rost symmetry (3.45) we have

$$\mathbb{E}_\nu\left[|x - Y_\tau| - |x - Y_{\tau\wedge T}|\right] = U_{\beta^\alpha}(x) - U_{\beta^\alpha_T}(x).$$

Furthermore for any X-stopping time σ we have

$$\mathbb{E}^x_\nu\left[|X_\sigma - Y_\tau| - \left|X_\sigma - Y_{\theta\wedge(T-\sigma)}\right|\right] = \mathbb{E}^x\left[U_{\beta^\alpha}(X_\sigma) - V^\alpha_{T-\sigma}(X_\sigma)\right]. \tag{3.49}$$

It now becomes obvious that Equation (3.13) follows by Proposition 3.13 (*iii*) by considering $\sigma = \sigma^*$ and the Root-Rost symmetry (3.46), while Equation (3.14) follows from Proposition 3.13 (*ii*) considering general X- stopping times $\sigma \leq T$. $\square$

4 Recovering the continuous optimal stopping representation as a limit of the discrete optimal stopping representation

The aim of this section is to recover Theorem 2.7 (resp. Theorem 2.8) from its discrete counterpart, Theorem 3.5 (resp. Theorem 3.6) through taking limits of appropriately scaled SSRWs.

This passage to continuous time essentially relies on the application of Donsker-type results, similar to the limiting procedure carried out in [1, Section 5] which in turn heavily built on arguments established in [13]. For this purpose it is important to point out that the results and arguments presented in Section 3 remain invariant under appropriate scaling of the space-time grid.

In Section 4.1 we recall the appropriate space-time grid and an associated discretisation of a Brownian motion into a scaled SSRW as proposed in [13]. We proceed to introducing a discretisation of a continuous (dSEP) to a scaled discrete delayed problem in Section 4.2 and furthermore explain how our results of Section 3 apply in this case.

It is subject of Section 4.3 to recover continuous Root and Rost solutions to (dSEP) as limits of solutions to a appropriately scaled and discretised (dSEP). This will give us convergence of the LHS of 3.11 to the LHS of 2.3 (resp. convergence of the LHS of 3.13 to the LHS of 2.5).

In Section 4.4 these limiting results are furthermore applied to the recovery of the optimal stopping problem and its value, namely establishing convergences of the RHS of 3.11 to the LHS of 3.11 as well as 3.12 to 3.12 (analogously for Rost).

4.1 Discretisation of a Brownian motion and its stopping times

We begin by recalling the discretisation of a Brownian motion $(W_t)_{t\geq 0}$ into a scaled SSRW as proposed in [13]. For a given $N \in \mathbb{N}$ consider a discrete set of spatial points $\mathcal{X}^N := \{x_1^N, \ldots, x_{L^N}^N\}$ such that

- $|x_{k+1}^N - x_k^N| = \frac{1}{\sqrt{N}}$ for $j = 1, \ldots, L^N - 1$.
- $L^N \sim \sqrt{N}$.

We define a random walk $(Y_k^N)_{k\in\mathbb{N}}$ in the following way. Let $Y_k^N := W_{\tau_k^N}$ where the τ_k^N are given as those times when a Brownian motion hits a new grid point in $\mathcal{X}^N$. More precisely, let $(W_t)_{t\geq 0}$ be a Brownian motion started in $W_0 \sim \lambda$. Then we define

- $\tau_0^N = \inf\{t \geq 0 : W_t \in \mathcal{X}^N\}$, and
- given $W_{\tau_k^N} = x_j^N$ we define $\tau_{k+1}^N := \inf\{t \geq \tau_k^N : W_t \in \{x_{j-1}^N, x_{j+1}^N\}\}$.

Note, this definition implies that the random walk $(Y_k^N)_{k\in\mathbb{N}}$ is started according to a discretisation λ^N of the starting law λ given by

$$\lambda^N := \mathcal{L}(Y_0^N) = \mathcal{L}^\lambda(W_{\tau_0^N}).$$

Now let τ be a stopping time for the Brownian Motion W_t. Then we can define a stopping rule $\tilde\sigma^N$ for the discretised process Y_k^N via

$$\tilde\sigma^N := \inf\{t \in \mathbb{N} : \tau_{t-1}^N < \tau \le \tau_t^N\}.$$

For such a discretised stopping time we have the following convergence results.

Lemma 4.1 (Cf.[13, Lemma 5.2])**.** *Let $\tilde\sigma^N$ be the discretisation of a stopping time τ of a Brownian motion W_t as defined above. Then*

$$Y^N_{\tilde\sigma^N} \to W_\tau \quad \textit{almost surely, and} \tag{4.1}$$

$$\frac{\tilde\sigma^N}{N} \to \tau \quad \textit{in probability, as } N \to \infty. \tag{4.2}$$

In particular, this implies that for $\mu^N := \mathcal{L}(Y^N_{\tilde\sigma^N})$ and for $\mu := \mathcal{L}^\lambda(W_\tau)$ we have that $\mu^N \to \mu$. Moreover, for every $T \ge 0$ we have

$$\begin{aligned} Y^N_{\tilde\sigma^N \wedge NT} &\to W_{\tau\wedge T} \quad \textit{almost surely, and} \\ \frac{\tilde\sigma^N \wedge NT}{N} &\to \tau \wedge T \quad \textit{in probability, as } N \to \infty. \end{aligned}$$

Note that while in [13] this lemma was stated for stopping times τ that are optimisers of a given (OptSEP), this characterisation is not essential to the proof given, thus we state the lemma in more generality. Given the convergence of $\frac{\tilde\sigma^N}{N} \to \tau$ in probability, the convergence of $\frac{\tilde\sigma^N \wedge NT}{N} \to \tau \wedge T$ in probability is clear. Regarding the convergence of the processes, it is important to observe that by construction of the discretisation we have $\tau^N_{NT} \to T$ a.s. by the strong law of large numbers. Note furthermore that $Y^N_{\tilde\sigma^N\wedge NT} = W_{\tau^N_{\tilde\sigma^N} \wedge \tau^N_{NT}}$, hence the almost sure convergence $Y^N_{\tilde\sigma^N\wedge NT} \to W_{\tau\wedge T}$ is clear.

4.2 A suitable discretisation of Root and Rost solutions to (dSEP)

Consider a continuous (dSEP) given by the triple (λ, η, μ). Theorem 2.2 (resp. Theorem 2.4) gives us existence of a Root (resp. Rost) solution ρ embedding this measure μ after the stopping time η.

We propose the following discretisation $(\lambda^N, \tilde\eta^N, \mu^N)$ of (λ, η, μ).

$$\begin{aligned} \lambda^N &:= \mathcal{L}(Y_0^N), \\ \tilde\eta^N &:= \inf\{t \in \mathbb{N} : \tau_{t-1}^N < \eta \le \tau_t^N\} \text{ and} \\ \mu^N &:= \mathcal{L}^{\lambda^N}(Y^N_{\tilde\sigma^N}) \text{ where} \\ \tilde\sigma^N &:= \inf\{t \in \mathbb{N} : \tau_{t-1}^N < \rho \le \tau_t^N\}. \end{aligned}$$

(Note that since $\rho \ge \eta$, it is clear that we also have $\tilde\sigma^N \ge \tilde\eta^N$.) If the measure μ is supported also beyond the grid $\mathcal{X}^N$ given at the discretisation step N we make the boundaries of the grid, x_1^N resp. $x^N_{L^N}$ absorbing barriers for the Brownian motion to ensure that the random walk only moves on the given grid.

The following convergences are now clear by the definitions in Section 4.1 and Lemma 4.1

$$\begin{aligned} \lambda^N &\Rightarrow \lambda \\ \eta^{(N)} := \frac{\tilde\eta^N}{N} &\xrightarrow{p} \eta \\ \mu^N &\Rightarrow \mu. \end{aligned}$$

Moreover, it is important to observe that the conditions $\lambda^N \leq_c \mathcal{L}^{\lambda^N}\left(Y^N_{\tilde{\eta}^N}\right) \leq_c \mu^N$ and $\mathbb{E}\left[\tilde{\eta}^N\right] < \infty$ are satisfied. Therefore, by Theorem 3.3 (resp. Theorem 3.4) and rescaling there exists a field of Root (resp. Rost) stopping probabilities $(r^N_t(x))_{(t,x)\in\mathbb{N}\times\mathcal{X}^N}$ where $\mathcal{X}^N \subseteq \frac{1}{\sqrt{N}}\mathbb{Z}$ such that

$$\tilde{\rho}^N := \inf\left\{t \geq \tilde{\eta}^N : r^N_t(Y^N_t) > u_t\right\}$$

defines a discrete Root (resp. Rost) stopping time embedding the measure μ^N into the SSRW Y^N, solving the discrete (and rescaled) (dSEP) given by $(\lambda^N, \tilde{\eta}^N, \mu^N)$.

By $\left(W^{(N)}_t\right)_{t\geq 0}$ we denote the rescaled continuous version of the random walk Y^N defined via

$$W^{(N)}_t := Y^N_{\lfloor Nt\rfloor} + (Nt - \lfloor Nt\rfloor)\left(Y^N_{\lfloor Nt\rfloor+1} - Y^N_{\lfloor Nt\rfloor}\right).$$

Note that we have $\left(\frac{\tilde{\eta}^N}{N}, Y^N_{\tilde{\eta}^N}\right) = \left(\eta^{(N)}, W^{(N)}_{\eta^{(N)}}\right)$ almost surely.

We now want to define Root and Rost stopping times for $\left(W^{(N)}_t\right)_{t\geq 0}$. By definition of the discretisation we have $W^{(N)}_t \in \mathcal{X}^N$ if and only if $Nt \in \mathbb{N}$. Hence we can extend a Root field of stopping probabilities $(r^N_t(x))_{(t,x)\in\mathbb{N}\times\mathcal{X}^N}$ to $[0,\infty)\times\mathbb{R}$ by setting

$$r^N_t(x) := \begin{cases} r^N_{\lfloor t\rfloor}(x) & \text{for } x \in \mathcal{X}^N, \\ 0 & \text{for all } t \text{ when } x \in \mathbb{R}\setminus\mathcal{X}^N. \end{cases}$$

Then we define

$$\rho^{(N)} := \inf\left\{t \geq \eta^{(N)} : r^N_{\lfloor Nt\rfloor}\left(W^{(N)}_t\right) > u_{\lfloor Nt\rfloor}\right\} \tag{4.3}$$

in order to have

$$\left(\tilde{\rho}^N, Y^N_{\tilde{\rho}^N}\right) = \left(\rho^{(N)}, W^{(N)}_{\rho^{(N)}}\right) \text{ almost surely.}$$

Furthermore $(r^N_t(x))_{(t,x)\in\mathbb{N}\times\mathcal{X}^N}$ induce a Root barrier $\tilde{R}^N \subseteq \mathbb{N}\times\frac{1}{\sqrt{N}}\mathbb{Z}$ via

$$\tilde{R}^N := \left\{(t,x) : r^N_t(x) > 0\right\},$$

and a corresponding Root barrier in $[0,\infty)\times\mathbb{R}$ can be defined in the following way

$$R^{(N)} := \left\{(t,x) \in [0,\infty)\times\mathbb{R} : (\lfloor Nt\rfloor, x) \in \tilde{R}^N\right\}. \tag{4.4}$$

We will sometimes use the notation $\tilde{R}^N =: \tilde{R}^N_+$ (resp. $R^{(N)} =: R^{(N)}_+$) and analogously define $\tilde{R}^N_-$ (resp. $R^{(N)}_-$) via

$$\tilde{R}^N_- := \left\{(t,x) : r^N_t(x) = 1\right\}.$$

Note that $\tilde{R}^N_- \subseteq \tilde{R}^N_+$ and $R^{(N)}_- \subseteq R^{(N)}_+$. We make the same definitions for Rost fields of stopping probabilities replacing the floor function by a ceiling function.

4.3 Taking limits

The main objective of this section is to show the following convergence of the discrete time objects defined above to their continuous conterparts. This will be carried out individually for Root and Rost respectively.

We follow [34] for the definition of a metric on the space of barriers. Let $\mathscr{R}$ denote the space of all Root and Rost barriers in $[0,\infty]\times[-\infty,\infty]$. Then a complete metric on this space can be defined in the following way. Consider the map $f : [0,\infty]\times[-\infty,\infty] \to [0,1]\times[-1,1]$, $f(t,x) := \left(\frac{t}{1+t}, \frac{x}{1+|x|}\right)$. Let d denote the ordinary Euclidean metric on $\mathbb{R}$. Then for two barriers $R, S \subseteq [0,\infty]\times[-\infty,\infty]$ *Root's metric* $d_{\mathscr{R}}$ on the space of all barriers $\mathscr{R}$ is defined as follows

$$d_{\mathscr{R}}(R,S) := \max\left\{\sup_{(t,x)\in R} d\left(f(t,x), f(S)\right), \sup_{(s,y)\in S} d\left(f(R), f(s,y)\right)\right\}.$$

4.3.1 A limit of Root stopping times

Lemma 4.2. *Let $\left(\rho^{(N)}\right)_{n\in\mathbb{N}}$ be a sequence of discrete Root stopping times as defined in* (4.3)*. Then there exists a Root barrier $R \subseteq [0,\infty)\times\mathbb{R}$ such that we have the following convergence*

$$\left(\rho^{(N)}, W^{(N)}_{\rho^{(N)}}\right) \xrightarrow{d} (\rho, W_\rho) \tag{4.5}$$

where

$$\rho := \inf\left\{t \geq \eta : (t, W_t) \in R\right\}.$$

Proof. The Root barriers $(R^{(N)})_{N\in\mathbb{N}}$ converge (possibly along a subsequence) to a Root barrier R in Root's metric, details can be found in [34].

We define the following auxiliary hitting time

$$\rho^N := \inf\left\{t \geq \eta : (t, W_t) \in R^{(N)}\right\}$$

and show convergence in two steps.

(i) We have $\rho^N \xrightarrow{p} \rho$. This is basically a consequence of a delayed version of Root's original convergence lemma, [34, Lemma 2.4], which we provide in Lemma 4.3.

(ii) We furthermore have
$$\left|\left(\rho^{(N)}, W^{(N)}_{\rho^{(N)}}\right) - \left(\rho^N, W_{\rho^N}\right)\right| \xrightarrow{p} 0,$$
which is a delayed version of [13, Lemma 5.6] which we give in Lemma 4.5.

□

We formulate a delayed version of Root's convergence lemma, [34, Lemma 2.4]

Lemma 4.3. *Let η be a delay stopping time and R be a Root barrier such that for the corresponding delayed hitting time*

$$\tau = \inf\left\{t \geq \eta : (t, W_t) \in R\right\}$$

we have $\mathbb{E}[\tau] < \infty$.

Then for any $\varepsilon > 0$ there exist $\delta, \tilde{\delta} > 0$ such that if for another Root barrier $\bar{R}$ we have $d_{\mathscr{R}}(R, \bar{R}) < \delta$ and for another delay stopping time $\bar{\eta}$ we have $\mathbb{P}[|\eta - \bar{\eta}| > \tilde{\delta}] < \tilde{\delta}$, then for the corresponding delayed hitting time $\bar{\tau}$ defined as

$$\bar{\tau} = \inf\left\{t \geq \bar{\eta} : (t, W_t) \in \bar{R}\right\},$$

it holds that

$$\mathbb{P}[\bar{\tau} > \tau + \varepsilon] < \varepsilon.$$

Proof. Similar to [34] we make the following choices for $\varepsilon > 0$.

- Choose $\tilde{\varepsilon} > 0$ such that $\tilde{\varepsilon} < \frac{\varepsilon}{4}$ and
$$\mathbb{P}\left[\sup_{\tilde{\varepsilon}<t<\varepsilon} W_t > \tilde{\varepsilon} \text{ and } \inf_{\tilde{\varepsilon}<t<\varepsilon} W_t < -\tilde{\varepsilon}\right] > 1 - \frac{\varepsilon}{4}.$$
- Choose $T > \frac{3\mathbb{E}[\tau]}{\varepsilon}$, then $\mathbb{P}[\tau \geq T] < \frac{\varepsilon}{4}$ (due to Markov's inequality).
- Choose M and $\delta > 0$ such that if $(t,x) \in R$, $t \leq T$, $|x| \leq M$ and $d_{\mathscr{R}}(R, \bar{R}) < \delta$, then $d((t,x), \bar{R}) \leq \tilde{\varepsilon}$.
- Choose $\tilde{\delta} = \tilde{\varepsilon}$, then $\mathbb{P}[|\eta - \bar{\eta}| > \tilde{\varepsilon}] < \tilde{\varepsilon}$.

Then we consider the set

$$A := \{\omega \in \Omega : (i) \sup_{\tau(\omega)+\tilde{\varepsilon}<t<\tau(\omega)+\varepsilon} (W_t(\omega) - W_\tau(\omega)) > \tilde{\varepsilon} \text{ and}$$
$$\inf_{\tau(\omega)+\tilde{\varepsilon}<t<\tau(\omega)+\varepsilon} (W_t(\omega) - W_\tau(\omega)) < -\tilde{\varepsilon},$$
$$(ii)\ \tau(\omega) < T,$$
$$(iii)\ |W_\tau(\omega)| < M,$$
$$(vi)\ \tilde{\eta} < \eta + \tilde{\varepsilon}.\}$$

We then see from the definitions made above that for $\omega \in A$ follows $\overline{\tau}(\omega) < \tau(\omega) + \varepsilon$. Moreover, these choices and the strong Markov property give us $\mathbb{P}[A] > 1 - \varepsilon$. ☐

In order to complete step (ii) in the proof of Lemma 4.2, we need to provide the following auxiliary convergence result, which is basically a delayed version of [13, Lemma 5.7].

Lemma 4.4. *Consider the stopping times*

$$\rho^{(M,N)} := \inf\left\{t \geq \eta^{(N)} : \left(t, W_t^{(N)}\right) \in R^{(M)}\right\},$$
$$\rho^M := \inf\left\{t \geq \eta : (t, W_t) \in R^{(M)}\right\},$$

where $\eta^{(N)} \to \eta$ in probability. Then we have the following convergence

$$\left(W^{(N)}_{\rho^{(M,N)}}, \rho^{(M,N)}\right) \xrightarrow[N\to\infty]{d} \left(W_{\rho^M}, \rho^M\right). \tag{4.6}$$

Proof. We aim to apply Donsker's theorem to the continuous scaled symmetric random walk $\left(W_t^{(N)}\right)_{t\in[\eta^{(N)},T]}$. Let $(B_t)_{t\in[\eta,T]}$ denote a Brownian motion, then for each $\omega \in \Omega$ we have $\left(W_t^{(N)}\right)_{t\in[\eta^{(N)},T]}, (B_t)_{t\in[\eta,T]} \in \mathcal{D}_T$ where $\mathcal{D}_T := \bigcup_{a\leq T} C([a,T])$.

We define an appropriate metric on $\mathcal{D}_T$. More generally, consider the function space $\mathcal{D} := \bigcup_{a\leq b} C([a,b])$. Thus for each $f \in \mathcal{D}$ there exist $l_f, r_f \in \mathbb{R}$, $l_f \leq r_f$ such that $f \in C([l_f, r_f])$. For $f \in \mathcal{D}$ and $t \geq 0$ we consider $\overline{f}(t) := f(l_f \vee (t \wedge r_f)) \in C([0,\infty))$ and define the following distance on $\mathcal{D}$.

$$d_{\mathcal{D}}(f,g) := \left(\sup_{t\in[0,\infty)} |\overline{f}(t) - \overline{g}(t)|\right) \vee |l_f - l_g| \vee |r_f - r_g|.$$

Considering this, metric we can apply Donsker's theorem to get

$$\left(W_t^{(N)}\right)_{t\in[\eta^{(N)},T]} \Rightarrow (B_t)_{t\in[\eta,T]}$$

and we are able to use the Portmanteau Theorem in the following way

$$\lim_{N\to\infty} \mathbb{P}\left[\left(W_t^{(N)}\right)_{t\in[\eta^{(N)},T]} \in \mathcal{K}\right] = \mathbb{P}\left[(B_t)_{t\in[\eta,T]} \in \mathcal{K}\right]$$
$$\text{for all } \mathcal{K} \subseteq \mathcal{D} \text{ Borel with } \mathbb{P}^\lambda\left[(B_t)_{t\in[\eta,T]} \in \partial\mathcal{K}\right] = 0.$$

It is now left to consider cleverly chosen sets $\mathcal{K} \in \mathcal{B}(\mathcal{D})$ to support our convergence claim (4.6). For M fixed consider $W^{(N)}_{\rho^{(M,N)}}$ and B_{ρ^M} as well as the spatial grid $\mathcal{X} := \{x_1^M, \ldots, x_{L^M}^M\}$.

Choose an arbitrary point $x_i^M \in \mathcal{X}$, $t \in (0,T]$ such that (t, x_i^M) lies somewhere in the interior of $R^{(M)}$ and some $\gamma < \big(t - \inf\big\{s : \big(s, x_i^M\big) \in R^{(M)}\big\}\big) \wedge (T-t)$ to consider the set

$$\bar{R}(\gamma) := [t-\gamma, t+\gamma] \times \big\{x_i^M\big\} \subseteq R^{(M)}.$$

Since the barrier $R^{(M)}$ will be a subset of $[0,\infty) \times \mathcal{X}$ we can identify a smallest $z \in \mathcal{X}$ such that $z > x_i^M$ and $([t-\gamma, t+\gamma] \times \{z\}) \subseteq R^{(M)}$. Analogously, we identify a largest $y \in \mathcal{X}$, $y < x_i^M$.

Moreover, let $y < w_1 < \dots < w_n < z$, $w_i \in \mathcal{X} \setminus \big\{x_i^M\big\}$ denote all those points at which a piece of the barrier $R^{(M)}$ sticks into our area of interest, precisely $([t-\gamma, t+\gamma] \times \{w_i\}) \cap R^{(M)} \neq \emptyset$. It is clear that there can only be finitely may such points. Consider $\bar{R}_{w_i} := ([t-\gamma, t+\gamma] \times \{w_i\}) \cap R^{(M)}$ for $i = 1, \dots, n$.

We proceed to define the following subsets of $\mathcal{D}_T$.

$$\begin{aligned}
\hat{\mathcal{K}} &:= \bigcup_{\substack{\hat\gamma > 0\\ \hat\gamma \in \mathbb{Q}}} \Big\{f \in \mathcal{D}_T : \|(s, f(s)) - R^{(M)}\| > \hat\gamma \ \forall\, s \in [l_f, (t-\gamma) \vee l_f]\Big\}\\
\hat{\mathcal{K}}^w &:= \bigcap_{i=1,\dots,n} \bigcup_{\substack{\hat\gamma > 0\\ \hat\gamma \in \mathbb{Q}}} \{f \in \mathcal{D}_T : \|(s, f(s)) - \bar{R}_{w_i}\| > \hat\gamma \ \forall\, s \in [(t-\gamma) \vee l_f, t+\gamma]\}\\
\hat{\mathcal{K}}_z &:= \{f \in \hat{\mathcal{K}} \cap \hat{\mathcal{K}}^w : f((t-\gamma) \vee l_f) \in (x_i^M, z)\}\\
\mathcal{K}_q^\varepsilon &:= \{f \in \mathcal{D}_T : f(s) < z - \varepsilon \ \forall\, s \in [(t-\gamma) \vee l_f, q] \cap \mathbb{Q}\}\\
\mathcal{K}_q^{\varepsilon,\delta} &:= \{f \in \mathcal{D}_T : f(q) < x_i^M - \delta\} \cap \mathcal{K}_q^\varepsilon\\
\mathcal{K}^{\varepsilon,\delta} &:= \bigcup_{\substack{q \in [t-\gamma, t+\gamma]\\ q \in \mathbb{Q}}} \mathcal{K}_q^{\varepsilon,\delta}\\
\mathcal{K}^z &:= \bigcap_{\substack{\delta > 0\\ \delta \in \mathbb{Q}}} \bigcup_{\substack{\varepsilon > 0\\ \varepsilon \in \mathbb{Q}}} \mathcal{K}^{\varepsilon,\delta} \cap \hat{\mathcal{K}}_z
\end{aligned}$$

Then

- $\hat{\mathcal{K}}$ is the set of all paths which stay away from the barrier $R^{(M)}$ at any time point before $(t-\gamma) \vee l_f$. Since $R^{(M)}$ is a finite barrier, $\hat{\mathcal{K}}$ is a Borel set.
- $\hat{\mathcal{K}}^w$ is the set of all paths which stay away from the barrier pieces $\bar{R}_{w_1}, \dots, \bar{R}_{w_n}$ before time $t+\gamma$.
- $\hat{\mathcal{K}}_z$ is the set of all paths in $\hat{\mathcal{K}} \cap \hat{\mathcal{K}}^w$ that are between x_i^M and z at time $(t-\gamma) \vee l_f$.
- $\mathcal{K}_q^\varepsilon$ is the set of all paths that stay below $z - \varepsilon$ after time $(t-\gamma) \vee l_f$ and before time q.
- $\mathcal{K}_q^{\varepsilon,\delta}$ is the set of all paths that are below $x_i^M - \delta$ at time q and stay below $z - \varepsilon$ after time $(t-\gamma) \vee l_f$ and before time q.
- $\mathcal{K}^{\varepsilon,\delta}$ is the set of all paths that such that there exist a $q \in [t-\gamma, t+\gamma] \cap \mathbb{Q}$ fulfilling the properties of set $\mathcal{K}_q^{\varepsilon,\delta}$.
- $\mathcal{K}^z$ is the set of all paths such that for all $\delta > 0$ there exists an $\varepsilon > 0$ such that the properties of set $\mathcal{K}^{\varepsilon,\delta}$ are fulfilled.

Analogously, we define $\mathcal{K}^y$ with the opposite inequalities.

Note that $\mathcal{K}^z$ will be a Borel set consisting of all those paths that start somewhere away from the barrier $R^{(M)}$ but hit the barrier for the first time in $\bar{R}(\gamma)$ coming from above, while $\mathcal{K}^y$ will be a Borel set of those paths that hit the barrier $R^{(M)}$ for the first time in $\bar{R}(\gamma)$ coming from below. Altogether, $\mathcal{K} := \mathcal{K}^y \cup \mathcal{K}^z$ will be a Borel set consisting of exactly those paths that hit the barrier $R^{(M)}$ for the first time in $\bar{R}(\gamma)$ after starting in some initial time l_f away from the barrier $R^{(M)}$.

We need to identify $\partial\mathcal{K}$ and to argue that $\mathbb{P}^\lambda\left[(B_t)_{t\in[\eta,T]}\in\partial\mathcal{K}\right]=0$. The set $\partial\mathcal{K}$ will consist of exactly those paths that start somewhere away from the boundary and either

- hit $\bar{R}(\gamma)$ exactly at time $t\pm\gamma$, or
- hit $\bar{R}(\gamma)$ elsewhere, but also *touch* the barrier $R^{(M)}$ before, without passing *through* the barrier.
- are born somewhere inside the barrier $R^{(M)}$ but immediately leave it without ever touching the barrier again.
- are born somewhere inside the barrier and remain flat.

By standard properties of a Brownian motion it is clear that these events must have zero probability.

With appropriate choices of x_i^N, t and γ it now becomes clear that we can cover the whole barrier $R^{(M)}$ with such sets $\bar{R}(\gamma)$.

We are left to consider the possibility of *appearing* exactly *on* the barrier. As a combination of the convergence of η^N to η and Donsker's theorem gives (should give?)

$$\left(\eta^{(N)},W^{(N)}_{\eta^{(N)}}\right)\xrightarrow[N\to\infty]{d}(\eta,B_\eta).$$

Now note that similarly to the arguments given in the proof of Lemma 4.3, a Brownian motion that is very close to a Root barrier will proceed to hit this barrier almost immediately. It is then a consequence of this regularity of Root barriers which gives us

$$\mathbb{P}^{\lambda^N}\left[\left(\eta^{(N)},W^{(N)}_{\eta^{(N)}}\right)\in R^{(M)}\right]\xrightarrow[N\to\infty]{}\mathbb{P}^\lambda\left[(\eta,B_\eta)\in R^{(M)}\right],$$

concluding the proof of the convergence (4.6). □

We are now ready to finish step (ii) in the proof of Lemma 4.2.

Lemma 4.5 (Cf.[13, Lemma 5.6]). *We have the following convergence*

$$\left|\left(\rho^{(N)},W^{(N)}_{\rho^{(N)}}\right)-\left(\rho^N,W_{\rho^N}\right)\right|\xrightarrow{p}0 \text{ for } N\to\infty.$$

Proof. Define the following hitting times for the scaled continuous random walk $W^{(N)}$

$$\tau^{(M,N)}:=\inf\left\{t\geq\eta^{(N)}:\left(t,W^{(N)}_t\right)\in R^{(M)}\right\},\quad\text{and}$$
$$\tau^{(N)}:=\inf\left\{t\geq\eta^{(N)}:\left(t,W^{(N)}_t\right)\in R^{(N)}\right\}.$$

Then analogous to the proof of [13, Lemma 5.6], but replacing Root's original convergence lemma with its delayed version Lemma 4.3, we can deduce the convergence

$$\left|\left(\tau^{(M,N)},W^{(N)}_{\tau^{(M,N)}}\right)-\left(\tau^{(N)},W^{(N)}_{\tau^{(N)}}\right)\right|\xrightarrow[M,N\to\infty]{p}0.$$

It is a consequence of Lemma 4.4 that

$$\left(\tau^{(M,N)},W^{(N)}_{\tau^{(M,N)}}\right)\xrightarrow[N\to\infty]{d}\left(\rho^M,W_{\rho^M}\right).$$

Thus it remains to show that

$$\left|\left(\rho^{(N)},W^{(N)}_{\rho^{(N)}}\right)-\left(\tau^{(N)},W^{(N)}_{\tau^{(N)}}\right)\right|\xrightarrow[N\to\infty]{p}0. \tag{4.7}$$

For this purpose recall the Root barriers $R_+^{(N)}$ and $R_-^{(N)}$ defined above and consider

$$\tau_+^{(N)} := \inf\left\{t \geq \eta : \left(t, W_t^{(N)}\right) \in R_+^{(N)}\right\} = \tau^{(N)},$$
$$\tau_-^{(N)} := \inf\left\{t \geq \eta : \left(t, W_t^{(N)}\right) \in R_-^{(N)}\right\}.$$

Note that by Root's barrier structure and the definition of the discretisation we have

$$d_{\mathscr{R}}\left(R_+^{(N)}, R_-^{(N)}\right) \leq \frac{1}{N} \to 0 \text{ for } N \to \infty,$$

hence both $R_+^{(N)}$ and $R_-^{(N)}$ converge to the same Root barrier R in Root's metric. Moreover, it is a consequence of Lemma 4.3 together with Donsker's Theorem that

$$\left|\tau_+^{(N)} - \tau_-^{(N)}\right| \xrightarrow[N\to\infty]{p} 0.$$

Now since $\tau_+^{(N)} \leq \rho^{(N)} \leq \tau_-^{(N)}$ the convergence (4.7) follows, concluding the proof. □

4.3.2 A limit of Rost stopping times

We are left to show convergence of discrete Rost solutions of (dSEP) to continuous ones. For a measurable set A and $t \geq 0$ recall the following measures defined in Section 2

$$\nu_t(A) = \mu(A) - \mathbb{P}^\lambda\left[W_\rho \in A, \rho < t\right] \quad \text{and}$$
$$\alpha_t(A) = \alpha\left(\{t\} \times A\right) = \mathbb{P}^\lambda\left[W_t \in A, \eta = t\right].$$

Lemma 4.6. *Let $\left(\rho^{(N)}\right)_{n\in\mathbb{N}}$ be a sequence of discrete Rost stopping times. Then there exists a Rost barrier $\bar{R} \subseteq [0,\infty) \times \mathbb{R}$ such that we have the following convergence*

$$\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right) \xrightarrow{d} (\rho, W_\rho) \tag{4.8}$$

where

$$\rho := \inf\left\{t > \eta : (t, W_t) \in \bar{R}\right\} \wedge Z$$

for

$$Z = \min_{k\in\mathbb{N}} Z_k, \quad Z_k := \begin{cases} t_k & \text{with probability } \mathbb{P}^\lambda\left[Z = t_k \middle| W_\eta, \eta = t_k\right] = \frac{\mathrm{d}\left(\alpha_{t_k} \wedge \nu_{t_k}\right)}{\mathrm{d}\left(\alpha_{t_k}\right)}(W_\eta) \\ \infty & \text{else.} \end{cases}$$

(Here $\{t_0, t_1, \dots\}$ denote the set of (at most countably many) atoms of $\mathcal{L}^\lambda(\eta)$.)

Proof. Let $\bar{R}^{(N)}$ denote the Rost barrier associated to $\rho^{(N)}$ via Definition (4.4). Then the sequence $(\bar{R}^{(N)})_{N\in\mathbb{N}}$ of these Rost barriers converges (possibly along a subsequence) to a Rost barrier $\bar{R}$ in Root's metric.

Consider the following decomposition

$$\begin{aligned}\left|\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right) - (\rho, W_\rho)\right| \leq &\left|\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)} > \eta^{(N)}+\varepsilon} - (\rho, W_\rho)\,\mathbb{1}_{\rho>\eta}\right| \\ &+ \left|\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)} \leq \eta^{(N)}+\varepsilon} - (\rho, W_\rho)\,\mathbb{1}_{\rho=\eta}\right|,\end{aligned}$$

hence (4.8) is equivalent to both

$$\left|\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)} > \eta^{(N)}+\varepsilon} - (\rho, W_\rho)\,\mathbb{1}_{\rho>\eta}\right| \xrightarrow[\varepsilon\searrow 0, N\to\infty]{p} 0 \quad \text{and} \tag{4.9}$$

$$\left|\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)} \leq \eta^{(N)}+\varepsilon} - (\rho, W_\rho)\,\mathbb{1}_{\rho=\eta}\right| \xrightarrow[\varepsilon\searrow 0, N\to\infty]{p} 0. \tag{4.10}$$

Let us first consider (4.9). Define the stopping time

$$\tau := \inf \left\{ t > \eta : (t, W_t) \in \bar{R} \right\},$$

thus $\rho = \tau \wedge Z$ and on $\{\rho > \eta\}$ we have $\rho = \tau$. Furthermore, for $\bar{R}^{(N)}$ consider the following (auxiliary) stopping times

$$\tau^N := \inf \left\{ t > \eta : (t, W_t) \in \bar{R}^{(N)} \right\} \quad \text{and}$$
$$\tau^{(N)} := \inf \left\{ t > \eta^{(N)} : (t, W_t^{(N)}) \in \bar{R}^{(N)} \right\}.$$

We will then show convergence in three steps.

(i) $\tau^N \xrightarrow{p} \tau$, see Lemma 4.7.

(ii) Convergence for non-randomised hitting times, precisely

$$\left| \left(\tau^{(N)}, W^{(N)}_{\tau^{(N)}} \right) - \left(\tau^N, W_{\tau^N} \right) \right| \xrightarrow{d} 0,$$

see Lemma 4.8.

(iii) Convergence for randomised Rost stopping times after $\eta^{(N)}$, precisely

$$\left| \left(\rho^{(N)}, W^{(N)}_{\rho^{(N)}} \right) \mathbb{1}_{\rho^{(N)} > \eta^{(N)} + \varepsilon} - \left(\tau^N, W_{\tau^N} \right) \mathbb{1}_{\tau^N > \eta} \right| \xrightarrow[\varepsilon \searrow 0, N \to \infty]{p} 0,$$

see Lemma 4.9.

It remains to show (4.10). For $\varepsilon > 0$ define

$$\hat{\mu}^N_\varepsilon := \mathcal{L}^{\lambda^N} \left(W^{(N)}_{\rho^{(N)}}; \rho^{(N)} > \eta^{(N)} + \varepsilon \right)$$
$$\overline{\mu}^N_\varepsilon := \mu^N - \hat{\mu}^N_\varepsilon$$

as well as

$$\hat{\mu} := \mathcal{L}^\lambda \left(W_\rho; \rho > \eta \right) \quad \text{and}$$
$$\overline{\mu} := \mu - \hat{\mu}.$$

Then by (4.9) we have

$$\hat{\mu}^N_\varepsilon = \mathcal{L}^{\lambda^N} \left(W^{(N)}_{\rho^{(N)}}; \rho^{(N)} > \eta^{(N)} + \varepsilon \right) \Rightarrow \hat{\mu} \quad \text{for } N \to \infty, \text{ then } \varepsilon \searrow 0.$$

Thus recall $\mu^N \Rightarrow \mu$ for $N \to \infty$ to conclude that we must have

$$\overline{\mu}^N_\varepsilon \Rightarrow \overline{\mu} := \mu - \hat{\mu} \quad \text{for } N \to \infty, \text{ then } \varepsilon \searrow 0$$

where $\overline{\mu}$ corresponds to the mass that is acquired by stopping instantly. In other words

$$\mathcal{L}^{\lambda^N} \left(W^{(N)}_{\rho^{(N)}}; \rho^{(N)} \leq \eta^{(N)} + \varepsilon \right) \Rightarrow \mathcal{L}^\lambda \left(W_\rho; \rho = \eta \right) \quad \text{for } N \to \infty, \text{ then } \varepsilon \searrow 0. \tag{4.11}$$

The convergence (4.11) is equivalent to

$$\mathbb{P}^{\lambda^N} \left[W^{(N)}_{\rho^{(N)}} \in A, \rho^{(N)} \leq \eta^{(N)} + \varepsilon \right] \xrightarrow[\varepsilon \searrow 0, N \to \infty]{} \mathbb{P}^\lambda \left[W_\rho \in A, \rho = \eta \right] \tag{4.12}$$

for all measurable A such that $\mathbb{P}^\lambda \left[W_\rho \in \partial A, \rho = \eta \right] = 0$. As we have the convergence $\eta^{(N)} \to \eta$ in probability we can furthermore conclude

$$\mathbb{P}^{\lambda^N} \left[W^{(N)}_{\rho^{(N)}} \in A, \rho^{(N)} \leq \eta^{(N)} + \varepsilon, |\eta^{(N)} - t| < \varepsilon \right] \xrightarrow[\varepsilon \searrow 0, N \to \infty]{} \mathbb{P}^\lambda \left[W_\rho \in A, \rho = \eta, \eta = t \right] \tag{4.13}$$

for all measurable A such that $\mathbb{P}^\lambda \left[W_\rho \in \partial A, \rho = \eta, \eta = t \right] = 0$.

That ρ must now share the desired representation on $\{\rho = \eta\}$ is a consequence of Lemma 2.5 and concludes the proof. $\square$

In order to conclude the proof above we show a delayed Version of [13, Lemma 5.5].

Lemma 4.7. *The Rost barriers $\bar{R}^{(N)}$ converge (possibly along a subsequence) to another Rost barrier $\bar{R}^\infty$ and*

$$\rho^N := \inf\left\{t > \eta : (t, W_t) \in \bar{R}^{(N)}\right\} \xrightarrow{p} \inf\left\{t > \eta : (t, W_t) \in \bar{R}^\infty\right\} =: \rho^\infty$$

Proof. Note that both stopping times are subject to the *same* delay η, hence we can repeat the proof of Lemma 5.5 in [13] with minimal adaptations. The arguments used therein rely on the existence of boundary functions $b : (0,\infty) \to \mathbb{R} \cup \{\infty\}$ and $c : (0,\infty) \to \mathbb{R} \cup \{-\infty\}$ such that the Rost hitting time can be represented as the first time the Brownian motions (W_t) either rises above $b(t)$ or falls below $c(t)$.

While in [13, Lemma 5.5] the Brownian motion was assumed to start in $(0, W_0)$ it was possible to give this description using only a single upper boundary function b and lower boundary function c. Due to the possibly delayed starting in (η, W_η) however, the processes might see different parts of the barrier that remain unseen for other starting positions. Hence we will no longer be able to describe the Rost hitting time with a single boundary function but as at most countably many such areas of the boundary can exists we can instead consider a family of boundary function $\{b_n : (0,\infty) \to \mathbb{R} \cup \{\infty\}, n \in \mathbb{N}\}$ and $\{c_n : (0,\infty) \to \mathbb{R} \cup \{-\infty\}, n \in \mathbb{N}\}$. Then for each starting position (η, W_η) we can identify the *closest* boundary functions b_η and c_η. Precisely, for each realisation of η we can define b_η as the boundary function b_n such that $W_\eta \le b_n(\eta)$ and $|(\eta, W_\eta) - (\eta, b_n(\eta))|$ is minimal. There might be several boundary functions fulfilling these properties, however as they all have to coincide for $t \ge \eta$ the specific choice is irrelevant. Analogously we define c_η to conclude the proof. □

We now give a delayed Version of [13, Lemma 5.6].

Lemma 4.8 (Cf.[13, Lemma 5.6])**.** *For $N \in \mathbb{N}$ consider a Rost barrier $\bar{R}^{(N)} \subseteq [0,\infty) \times \frac{1}{\sqrt{N}}\mathbb{Z}$ and let $\eta^{(N)} \in \frac{1}{N}\mathbb{N}$ and $\eta \in [0,\infty)$ be delay stopping times such that $\eta^{(N)} \xrightarrow{p} \eta$. Then for the stopping times*

$$\tau^{(N)} := \inf\left\{t > \eta^{(N)} : (t, W_t^{(N)}) \in \bar{R}^{(N)}\right\} \quad \text{and}$$
$$\tau^N := \inf\left\{t > \eta : (t, W_t) \in \bar{R}^{(N)}\right\}$$

we have the following convergence

$$\left|\left(\tau^{(N)}, W^{(N)}_{\tau^{(N)}}\right) - \left(\tau^N, W_{\tau^N}\right)\right| \xrightarrow{p} 0 \text{ for } N \to \infty.$$

Proof. First we give a slight refinement of the proof in [13, Lemma 5.6] for $\eta = 0$.

As in [13, Lemma 5.6] consider the Brownian motions $W^{\pm\varepsilon} = W_t \pm \varepsilon t$ with drift. Then we define the following stopping times

$$\rho^{+,N,\pm\varepsilon} := \inf\left\{t > 0 : W_t \ge W_0, \left(t, W_t^{\pm\varepsilon}\right) \in \bar{R}^{(N)}\right\},$$
$$\rho^{-,N,\pm\varepsilon} := \inf\left\{t > 0 : W_t \le W_0, \left(t, W_t^{\pm\varepsilon}\right) \in \bar{R}^{(N)}\right\},$$

and

$$\rho^{+,N} := \inf\left\{t > 0 : W_t \ge W_0, (t, W_t) \in \bar{R}^{(N)}\right\},$$
$$\rho^{-,N} := \inf\left\{t > 0 : W_t \le W_0, (t, W_t) \in \bar{R}^{(N)}\right\},$$

as well as

$$\rho^{+,(N)} := \inf\left\{t > 0 : W_t^{(N)} \geq W_0^{(N)}, \left(t, W_t^{(N)}\right) \in \bar{R}^{(N)}\right\},$$
$$\rho^{-,(N)} := \inf\left\{t > 0 : W_t^{(N)} \leq W_0^{(N)}, \left(t, W_t^{(N)}\right) \in \bar{R}^{(N)}\right\}.$$

Furthermore for $T > 0$ and $\varepsilon > 0$ consider the set

$$A^{N,\varepsilon} = \left\{\omega \in \Omega : \sup_{0 \leq t \leq T}\left|W_t - W_t^{(N)}\right| \leq \varepsilon\right\}$$

and note that by Donsker's Theorem the probability of $\left(A^{N,\varepsilon}\right)^c$ goes to zero for $N \to \infty$.

We have the following properties of the stopping times defined above.

- On the set $A^{N,\varepsilon}$ we have that $|\rho^{+,(N)} - \rho^{+,N}| \leq |\rho^{+,N,+\varepsilon} - \rho^{+,N,-\varepsilon}|$ as well as $|\rho^{-,(N)} - \rho^{-,N}| \leq |\rho^{-,N,+\varepsilon} - \rho^{-,N,-\varepsilon}|$.
- As demonstrated in the proof of [13, Lemma 5.6] it follows from Girsanov's Theorem that $|\rho^{+,N,+\varepsilon} - \rho^{+,N,-\varepsilon}| \xrightarrow{p} 0$ and $|\rho^{-,N,+\varepsilon} - \rho^{-,N,-\varepsilon}| \xrightarrow{p} 0$ for $\varepsilon \searrow 0$.
- Now note that $\rho^{(N)} = \rho^{+,(N)} \wedge \rho^{-,(N)}$ and $\rho^N = \rho^{+,N} \wedge \rho^{-,N}$, thus on the set $A^{N,\varepsilon}$ we have

$$\begin{aligned}|\rho^{(N)} - \rho^N| &= |\rho^{+,N} \wedge \rho^{-,N} - \rho^{+,(N)} \wedge \rho^{-,(N)}| \\ &\leq |\rho^{+,N} - \rho^{+,(N)}| + |\rho^{-,N} - \rho^{-,(N)}| \\ &\leq |\rho^{+,N,+\varepsilon} - \rho^{+,N,-\varepsilon}| + |\rho^{-,N,+\varepsilon} - \rho^{-,N,-\varepsilon}|\end{aligned}$$

- Combining all these ingredients we can conclude that $|\rho^{(N)} - \rho^N| \xrightarrow{p} 0$ just as in [13, Lemma 5.6].

We now add a delay.

Since $\eta^{(N)} \xrightarrow{p} \eta$ we can for ever $\varepsilon, \delta > 0$ find an $N_0 \in \mathbb{N}$ such that for

$$B^{N,\varepsilon} := \left\{\left|\eta^{(N)} - \eta\right| \leq \varepsilon\right\}$$

we have $\mathbb{P}\left[\left(B^{N,\varepsilon}\right)^c\right] \leq \delta$ for all $N \geq N_0$. Define the set $C^{N,\varepsilon} := A^{N,\varepsilon} \cap B^{N,\varepsilon}$ for $A^{N,\varepsilon}$ defined above, then similarly $\mathbb{P}\left[\left(C^{N,\varepsilon}\right)^c\right] \leq \tilde{\delta}$ for any $\tilde{\delta} > 0$ in N is chosen big enough.

Define the following stopping times.

$$\rho^{+,N,\pm\varepsilon} := \inf\left\{t > \eta : W_t \geq W_0, \left(t, W_t^{\pm\varepsilon}\right) \in \bar{R}^{(N)}\right\},$$
$$\rho^{-,N,\pm\varepsilon} := \inf\left\{t > \eta : W_t \leq W_0, \left(t, W_t^{\pm\varepsilon}\right) \in \bar{R}^{(N)}\right\},$$

and

$$\rho^{+,N} := \inf\left\{t > \eta : W_t \geq W_0, (t, W_t) \in \bar{R}^{(N)}\right\},$$
$$\rho^{-,N} := \inf\left\{t > \eta : W_t \leq W_0, (t, W_t) \in \bar{R}^{(N)}\right\},$$

as well as

$$\rho^{+,(N)} := \inf\left\{t > \eta^{(N)} : W_t^{(N)} \geq W_0^{(N)}, \left(t, W_t^{(N)}\right) \in \bar{R}^{(N)}\right\},$$
$$\rho^{-,(N)} := \inf\left\{t > \eta^{(N)} : W_t^{(N)} \leq W_0^{(N)}, \left(t, W_t^{(N)}\right) \in \bar{R}^{(N)}\right\}.$$

Replacing the stopping times as well as $A^{N,\varepsilon}$ with $C^{N,\varepsilon}$ in the arguments above resp. in [13, Lemma 5.6] we can conclude the desired convergence. □

We show the final convergence via a sandwiching argument.

For a given field of stopping probabilities $(r_t^N(x))_{(t,x)\in\mathbb{N}\times\frac{1}{\sqrt{N}}\mathbb{Z}}$ define

$$\begin{aligned}\tilde{R}_+^N &:= \left\{(t,x) : r_t^N(x) > 0\right\},\\ \tilde{R}_-^N &:= \left\{(t,x) : r_t^N(x) = 1\right\} =: \tilde{R}^N\end{aligned}$$

and let $\bar{R}_+^{(N)}$ and $\bar{R}_-^{(N)}$ respectively denote their (Rost) continuifications as defined in (4.4). Note that $\bar{R}^{(N)} = \bar{R}_-^{(N)} \subseteq \bar{R}_+^{(N)}$ and the two barriers coincide on all points (t,x) such that $r_t^N(x) \in \{0,1\}$. Furthermore define $\bar{S}_+^{(N)} := \bar{R}_+^{(N)} \setminus (\{0\}\times\mathbb{R})$ as well as $\bar{S}_-^{(N)} := \bar{R}_-^{(N)} \setminus (\{0\}\times\mathbb{R})$ and recall $d_{\mathscr{R}}$, Roots metric on barriers. While $\bar{R}_+^{(N)}$ and $\bar{R}_-^{(N)}$ might differ substantially on $\{0\}\times\mathbb{R}$, for $\bar{S}_+^{(N)}$ and $\bar{S}_-^{(N)}$ by Rost's barrier structure and definition of the discretisation we have

$$d_{\mathscr{R}}\left(\bar{S}_+^{(N)}, \bar{S}_-^{(N)}\right) \leq \frac{1}{N} \to 0 \quad \text{for } N\to\infty. \tag{4.14}$$

Hence both $\bar{S}_+^{(N)}$ and $\bar{S}_-^{(N)}$ converge to the same object $\bar{S}$ in Roots metric.

Define the hitting times

$$\begin{aligned}\tau_+^{(N)} &:= \inf\left\{t > \eta^{(N)} : \left(t, W_t^{(N)}\right) \in \bar{R}_+^{(N)}\right\},\\ \tau_-^{(N)} &:= \inf\left\{t > \eta^{(N)} : \left(t, W_t^{(N)}\right) \in \bar{R}_-^{(N)}\right\}.\end{aligned}$$

Then both $\tau_+^{(N)}$ and $\tau_-^{(N)}$ are non-randomised hitting times and we will almost surely have

$$\tau_+^{(N)} \leq \rho^{(N)} \leq \tau_-^{(N)}.$$

We will also consider the hitting times

$$\begin{aligned}\tau_+^N &:= \inf\left\{t > \eta : (t, W_t) \in \bar{R}_+^{(N)}\right\},\\ \tau_-^N &:= \inf\left\{t > \eta : (t, W_t) \in \bar{R}_-^{(N)}\right\}.\end{aligned}$$

These auxiliary stopping times will help us prove the following lemma.

Lemma 4.9. *We have the convergence*

$$\left|\left(\rho^{(N)}, W_{\rho^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)}>\eta^{(N)}+\varepsilon} - \left(\tau^N, W_{\tau^N}\right)\mathbb{1}_{\tau^N>\eta}\right| \xrightarrow[\varepsilon\searrow 0, N\to\infty]{p} 0. \tag{4.15}$$

Proof. Since $\tau_+^{(N)} \leq \rho^{(N)} \leq \tau_-^{(N)}$ almost surely it suffices to show the two convergences

$$\left|\left(\tau_+^{(N)}, W_{\tau_+^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)}>\eta^{(N)}+\varepsilon} - \left(\tau^N, W_{\tau^N}\right)\mathbb{1}_{\tau^N>\eta}\right| \xrightarrow[\varepsilon\searrow 0, N\to\infty]{p} 0 \quad \text{and} \tag{4.16}$$

$$\left|\left(\tau_-^{(N)}, W_{\tau_-^{(N)}}^{(N)}\right)\mathbb{1}_{\rho^{(N)}>\eta^{(N)}+\varepsilon} - \left(\tau^N, W_{\tau^N}\right)\mathbb{1}_{\tau^N>\eta}\right| \xrightarrow[\varepsilon\searrow 0, N\to\infty]{p} 0. \tag{4.17}$$

Lemma 4.8 gives us the following convergence

$$\left|\left(\tau_+^{(N)}, W_{\tau_+^{(N)}}^{(N)}\right) - \left(\tau_+^N, W_{\tau_+^N}\right)\right| \xrightarrow[N\to\infty]{d} 0 \tag{4.18}$$

as well as

$$\left|\left(\tau_-^{(N)}, W_{\tau_-^{(N)}}^{(N)}\right) - \left(\tau_-^N, W_{\tau_-^N}\right)\right| \xrightarrow[N\to\infty]{d} 0. \tag{4.19}$$

Note that (4.18) already implies (4.16) as $\tau^N = \tau_+^N$.

To see the convergence (4.17), observe that on $\{\rho^{(N)} > \eta^{(N)} + \varepsilon\}$ we will also have $\tau_-^{(N)} > \eta^{(N)} + \varepsilon$. Furthermore, since $\tau^N = \tau_+^N \leq \tau_-^N$ on $\{\tau^N > \eta\}$ we thus also have $\tau_-^N > \eta$. Hence it is a consequence of (4.14) and Lemma 4.7 that

$$\left|\tau_+^N \mathbb{1}_{\tau^N > \eta} - \tau_-^N \mathbb{1}_{\tau^N > \eta}\right| \xrightarrow{p} 0.$$

This together with (4.19) implies convergence (4.17), concluding the proof. $\square$

Lemma 4.2 (resp. 4.6) now enables the first step of recovering the continuous optimal stopping representation of Theorem 2.7 (resp. 2.8) from the discrete counterpart Theorem 3.5 (resp. 3.6).

Corollary 4.10. *Let $(\lambda^N, \tilde{\eta}^N, \mu^N)$ denote a (rescaled) discrete* (dSEP) *approximating the continuous* (dSEP) *given by (λ, η, μ) as defined in Section 4.2. Let $\rho^{(N)}$ denote a Root (resp. Rost) solution to $(\lambda^N, \tilde{\eta}^N, \mu^N)$ while ρ denotes a Root (resp. Rost) solution to (λ, η, μ). Then for every $T \in [0, \infty)$ we have the convergence*

$$\mu_T^N = \mathcal{L}^{\lambda^N}\left(W^{(N)}_{\rho^{(N)} \wedge T}\right) \Rightarrow \mathcal{L}^{\lambda}\left(W_{\rho \wedge T}\right) = \mu_T. \tag{4.20}$$

Moreoever this gives convergence of the LHS of 3.11 (resp. 3.13) to the LHS of 2.3 (resp. 2.5).

4.4 A discretisation of the optimal stopping problem and its convergence

It remains to show convergence of the discrete optimal stopping problem to the continuous one.

Les us consider a continuous (dSEP) determined by (λ, η, μ). We will first prove convergence of the Root case, the Rost case can then be argued analogously.

To avoid heavy usage of floor functions, we will assume from now on for the cutoff time $T \in I := \left\{\frac{m}{2^n} : m, n \in \mathbb{N}\right\}$. Then taking limits along the subsequence $\left(Y^{2^n}\right)_{n\in\mathbb{N}}$ (resp. $\left(W^{(2^n)}\right)_{n\in\mathbb{N}}$) ensures that there exists an $N_0 \in \mathbb{N}$ such that T will always be a multiple of the step size $\frac{1}{2^n}$ for all $n \geq N_0$. For arbitrary $T > 0$ the results can be recovered via density arguments.

With the help of the function

$$G_T(x, t) := V^{\alpha}_{T-t}(x) + \left(U_\mu(x) - U_{\alpha_X}(x)\right)\mathbb{1}_{t<T}$$

we can recall the continuous identities of interest

$$U_{\mu_T}(x) = \mathbb{E}^x\left[G_T\left(W_{\tau^*}, \tau^*\right)\right] \tag{2.3}$$

$$= \sup_{\tau \leq T} \mathbb{E}^x\left[G_T\left(W_\tau, \tau\right)\right] \tag{2.4}$$

where the optimiser is given by

$$\tau^* := \tau_T \wedge T \text{ for } \tau_T := \inf\{t \geq 0 : (T-t, W_t) \notin R\}$$

for a Root barrier R.

Let now $(\lambda^N, \tilde{\eta}^N, \mu^N)$ denote a (rescaled) discrete (dSEP) approximating the continuous (dSEP) as defined in Section 4.2.

Note that Lemma 4.1 gives us the following additional convergence results

$$\alpha_X^N := \mathcal{L}^{\lambda^N}(Y^N_{\tilde{\eta}^N}) \to \alpha_X \quad \text{and} \tag{4.21}$$

$$\mathcal{L}^{\lambda^N}(Y^N_{\tilde{\eta}^N \wedge TN}) \to \mathcal{L}^{\lambda}(W_{\eta \wedge T}). \tag{4.22}$$

Thus we are able to define a discrete version of G_T in the following way

$$G_T^N(x, t) := V^N_{T-t}(x) + \left(U_{\mu^N}(x) - U_{\alpha_X^N}(x)\right)\mathbb{1}_{t<T}$$

where

$$V_T^N(x) := \mathbb{E}^{\lambda^N}\left[\left|Y^N_{\tilde{\eta}^N \wedge NT} - x\right|\right].$$

Here V_T^N (resp. V_T^α) is the potential function with respect to the LHS (resp. RHS) of (4.22), hence we have uniform convergence of G_T^N to G_T due to the uniform convergence of all the potential functions involved.

Theorem 3.3 gives existence of a Root field of stopping probabilities $(r_t^N(x))_{(t,x)\in\mathbb{N}\times\mathcal{X}^N}$ such that the corresponding discrete delayed Root stopping time $\tilde{\rho}^N$ embeds the measure μ^N. Let

$$\tilde{R}^N := \left\{(t,x)\in\mathbb{N}\times\mathcal{X}^N : r_t^N(x) > 0\right\}$$

denote a Root barrier induced by the field of stopping probabilities and let $R^{(N)}$ denote its continuification as defined in (4.4).

The rescaled version of the results in Theorem 3.5 then read

$$U_{\mu_T^N}(x) = \mathbb{E}^x\left[G_T^N\left(Y^N_{\tilde{\tau}^{N*}}, \frac{\tilde{\tau}^{N*}}{N}\right)\right] \tag{3.11*}$$

$$= \sup_{\frac{\tau}{N}\le T} \mathbb{E}^x\left[G_T^N\left(Y^N_\tau, \frac{\tau}{N}\right)\right], \tag{3.12*}$$

and an optimiser is given by

$$\tilde{\tau}^{N*} := \tilde{\tau}^N_T \wedge NT \quad \text{for} \quad \tilde{\tau}^N_T := \inf\{t\in\mathbb{N} : (NT - t, Y_t^N) \in \tilde{R}^N\}.$$

To conclude the proof of Theorem 2.7, it remains to establish the following two convergence results.

Lemma 4.11.

(i) We have convergence of the RHS of (3.11*) *to the RHS of* (2.3)*, precisely*

$$\mathbb{E}^x\left[G_T^N\left(Y^N_{\tilde{\tau}^{N*}}, \frac{\tilde{\tau}^{N*}}{N}\right)\right] \xrightarrow{N\to\infty} \mathbb{E}^x\left[G_T\left(W_{\tau^*}, \tau^*\right)\right].$$

(ii) We furthermore have convergence of the RHS of (3.12*) *to the RHS of* (2.4)*, precisely*

$$\sup_{\frac{\tau}{N}\le T} \mathbb{E}^x\left[G_T^N\left(Y^N_\tau, \frac{\tau}{N}\right)\right] \xrightarrow{N\to\infty} \sup_{\tau\le T} \mathbb{E}^x\left[G_T\left(W_\tau, \tau\right)\right].$$

Proof. To see *(i)*, recall the continuous rescaled version $W^{(N)}$ of Y^N. Then due to [13], Section 5, alternatively also due to Lemma 4.6 for the trivial delay $\eta^{(N)} = \eta = 0$ and additionally considering the Root-Rost symmetry we have the following Donsker type convergence

$$\left(\tau^{(N)*}, W^{(N)}_{\tau^{(N)*}}\right) \xrightarrow{d} (\tau^*, W_{\tau^*}) \ \text{ as } N\to\infty, \tag{4.23}$$

where

$$\tau^{(N)*} := \inf\left\{t\ge 0 : \left(T - t, W_t^{(N)}\right) \in R^{(N)}\right\} \wedge T$$

is the backwards Root stopping time of $W^{(N)}$ and

$$\tau^* := \inf\left\{t\ge 0 : (T - t, W_t) \in R\right\} \wedge T$$

is the backwards Root stopping time of the Brownian motion. Furthermore, note that by the definition of the discretisation we have

$$\left(\tau^{(N)*}, W^{(N)}_{\tau^{(N)*}}\right) = \left(\frac{\tilde{\tau}^{N*}}{N}, Y^N_{\tilde{\tau}^{N*}}\right).$$

Since all other crucial properties - that G_T is a lower semi-continuous function and that G_T^N converges uniformly to G_T - are also given here, we can repeat the convergence proof given in [1], Section 5 verbatim.

It is left to show *(ii)*, convergence of the optimal stopping problems. Trivially, we have

$$\mathbb{E}^x\left[G^T(W_{\tau^*},\tau^*)\right] \le \sup_{\tau\le T}\mathbb{E}^x\left[G^T(W_\tau,\tau)\right].$$

Hence it remains to show that

$$\sup_{\tau\le T}\mathbb{E}^x\left[G^T(W_\tau,\tau)\right] \le \mathbb{E}^x\left[G^T(W_{\tau^*},\tau^*)\right].$$

Let $\bar\tau$ be an optimiser of the continuous optimal stopping problem (2.4). Consider the auxiliary functions

$$\begin{aligned}\bar G_T^\varepsilon(x,t) &:= V_{T-t}^\alpha(x) + \left(U_\mu(x) - U_{\alpha_X}(x)\right)\mathbb{1}_{t\le T-\varepsilon} \quad\text{and}\\ \tilde G_T^{N,\varepsilon}(x,t) &:= V_{T-t}^N(x) + \left(U_{\mu^N}(x) - U_{\alpha_X^N}(x)\right)\mathbb{1}_{t\le T-\varepsilon}.\end{aligned}$$

Then for any $\delta>0$ we have the following

$$\begin{aligned}\sup_{\tau\le T}\mathbb{E}^x\left[G_T(W_\tau,\tau)\right]-\delta &< \mathbb{E}^x\left[G_T\left(W_{\bar\tau},\bar\tau\right)\right] &(4.24)\\
&= \lim_{\varepsilon\searrow 0}\mathbb{E}^x\left[\bar G_T^\varepsilon(W_{\bar\tau},\bar\tau)\right] &(4.25)\\
&\le \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\mathbb{E}^x\left[\tilde G_T^{N,\varepsilon}\left(Y_{\tilde\sigma^N}^N,\frac{\tilde\sigma^N}{N}\right)\right] &(4.26)\\
&\le \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\sup_{\frac{\tau}{N}\le T}\mathbb{E}^x\left[\tilde G_T^{N,\varepsilon}\left(Y_\tau^N,\frac{\tau}{N}\right)\right] &(4.27)\\
&\le \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\sup_{\frac{\tau}{N}\le T-\varepsilon}\mathbb{E}^x\left[G_{T-\varepsilon}^N\left(Y_\tau^N,\frac{\tau}{N}\right)\right] &(4.28)\\
&= \lim_{\varepsilon\searrow 0}\liminf_{N\to\infty}\mathbb{E}^x\left[G_{T-\varepsilon}^N\left(Y_{\tilde\tau_{T-\varepsilon}^{N*}}^N,\frac{\tilde\tau_{T-\varepsilon}^{N*}}{N}\right)\right] &(4.29)\\
&= \lim_{\varepsilon\searrow 0}\mathbb{E}^x\left[G_{T-\varepsilon}\left(W_{\tau_{T-\varepsilon}^*},\tau_{T-\varepsilon}^*\right)\right] &(4.30)\\
&= \lim_{\varepsilon\searrow 0}U_{\mu_{T-\varepsilon}}(x) &(4.31)\\
&= U_{\mu_T}(x) = \mathbb{E}^x\left[G\left(W_{\tau^*},\tau^*\right)\right]. &(4.32)\end{aligned}$$

We justify this chain of inequalities and equalities step by step.

The inequality in (4.24) is clear as $\bar\tau$ was assumed to be an optimiser. For the difference between G_T and $\bar G_T^\varepsilon$ we have

$$|G_T(x,t)-\bar G_T^\varepsilon(x,t)| = \underbrace{|U_\mu(x)-U_{\alpha_X}(x)|}_{\le c<\infty}\mathbb{1}_{T-\varepsilon<t<T},$$

hence for a random variable X and stopping time τ we have

$$\mathbb{E}^x\left[|G_T(X,\tau)-\bar G_T^\varepsilon(X,\tau)|\right] \le c\cdot\mathbb{P}\left[\tau\in(T-\varepsilon,T)\right].$$

Now since $\lim_{\varepsilon\searrow 0}\mathbb{P}\left[\tau\in(T-\varepsilon,T)\right]=0$ we can conclude

$$\lim_{\varepsilon\searrow 0}\mathbb{E}^x\left[|G_T(W_{\bar\tau},\bar\tau)-\bar G_T^\varepsilon(W_{\bar\tau},\bar\tau)|\right]=0,$$

which gives (4.25). The inequality in (4.26) follows from uniform convergence of $\tilde{G}_T^{N,\varepsilon}$ to $\bar{G}_T^{\varepsilon}$ and the fact that $\tilde{G}_T^{N,\varepsilon}$ is lower semi-continuous. In (4.27) we simply take a supremum. To see (4.28), let us extend the function $V_T^N(x)$ to $T<0$ in the following way.

$$V_T^N(x) := \begin{cases} -\mathbb{E}^{\lambda^N}\left[\left|Y^N_{\tilde{\eta}^N \wedge NT} - x\right|\right] & \text{for } T \geq 0 \\ U_{\lambda^N}(x) & \text{for } T < 0. \end{cases} \tag{4.33}$$

Consider $t \leq u$, then $T-u \leq T-t$, thus also $\tilde{\eta}^N \wedge (T-u) \leq \tilde{\eta}^N \wedge (T-t)$. Due to Jensen and Optional Sampling we then have

$$-V^N_{T-u}(x) = \mathbb{E}^{\lambda^N}\left[\left|Y^N_{\tilde{\eta}^N\wedge(T-u)} - x\right|\right] \leq \mathbb{E}^{\lambda^N}\left[\left|Y^N_{\tilde{\eta}^N\wedge(T-t)} - x\right|\right] = -V^N_{T-t}(x),$$

or equivalently

$$V^N_{T-t}(x) \leq V^N_{T-u}(x).$$

Especially, we have for $\varepsilon > 0$ that

$$V^N_{T-t}(x) \leq V^N_{T-\varepsilon-t}(x),$$

hence the function $v(t,x) := V^N_{T-t}(x)$ is *increasing* in t and convex in x. Now

$$\begin{aligned}
\sup_{\frac{\tau}{N}\leq T} \mathbb{E}^x\left[\tilde{G}_T^{N,\varepsilon}\left(Y_\tau^N, \frac{\tau}{N}\right)\right] &= \sup_{\frac{\tau}{N}\leq T} \mathbb{E}^x\Big[\underbrace{V^N_{T-\tau}\left(Y^N_\tau\right)}_{\leq V^N_{T-\varepsilon-\tau}(Y^N_\tau)} + \underbrace{\left(U_{\mu^N} - U_{\alpha^N_X}\right)\left(Y^N_\tau\right)\mathbb{1}_{\frac{\tau}{N}\leq T-\varepsilon}}_{\leq \left(U_{\mu^N} - U_{\alpha^N_X}\right)(Y^N_\tau)\mathbb{1}_{\frac{\tau}{N}<T-\varepsilon}}\Big] \\
&\leq \sup_{\frac{\tau}{N}\leq T}\left[V^N_{T-\varepsilon-\tau}\left(Y^N_\tau\right) + \left(U_{\mu^N} - U_{\alpha^N_X}\right)\left(Y^N_\tau\right)\mathbb{1}_{\frac{\tau}{N}<T-\varepsilon}\right] && (4.34) \\
&\leq \sup_{\frac{\tau}{N}\leq T} \mathbb{E}^x\left[\tilde{G}^{N,\varepsilon}_{T-\varepsilon}\left(Y^N_\tau, \frac{\tau}{N}\right)\right] && (4.35)
\end{aligned}$$

It remains to show that no optimiser of (4.34) resp. (4.35) will stop after time $T-\varepsilon$. By definition (4.33) we have $V^N_{T-\varepsilon-\tau}\left(Y^N_\tau\right) = V^N_{T-\varepsilon-\tau\wedge(T-\varepsilon)}\left(Y^N_\tau\right)$ since we cannot do better than $V^N_0\left(Y^N_\tau\right) = U_{\lambda^N}\left(Y^N_\tau\right)$. Now, let $(Z_t)_{t\geq 0}$ be a martingale. Then we consider $\left(\tilde{G}^{N,\varepsilon}_{T-\varepsilon}(Z_t,t)\right)_{t\in[T-\varepsilon,T]} = \left(U_{\lambda^N}(Z_t)\right)_{t\in[T-\varepsilon,T]}$. As U_{λ^N} is a concave function, $\left(U_{\lambda^N}(Z_t)\right)_{t\in[T-\varepsilon,T]}$ is a supermartingale and for any stopping time τ we have

$$\mathbb{E}^x\left[\tilde{G}^{N,\varepsilon}_{T-\varepsilon}\left(Z_{\tau\wedge(T-\varepsilon)}, \tau\wedge(T-\varepsilon)\right)\right] \geq \mathbb{E}^x\left[\tilde{G}^{N,\varepsilon}_{T-\varepsilon}\left(Z_{\tau\wedge T}, \tau\wedge T\right)\right].$$

Hence we see that no optimiser of (4.35) will stop after time $T-\varepsilon$ which concludes (4.28). Due to Theorem 3.5 we know that

$$\tilde{\tau}^{N*}_{T-\varepsilon} := \inf\{t\in\mathbb{N} : (N(T-\varepsilon)-t, Y^N_t) \in \tilde{R}^N\} \wedge N(T-\varepsilon)$$

is an optimiser of the optimal stopping problem

$$\sup_{\frac{\tau}{N}\leq T-\varepsilon} \mathbb{E}^x\left[G^N_{T-\varepsilon}\left(Y^N_\tau, \frac{\tau}{N}\right)\right]$$

which gives equality in (4.29). The convergence *(i)* together with Corollary 4.10 gives equality in (4.30) and (4.31). The convergence (4.32) follows by considering the barrier R such that the corresponding Root stopping time ρ embeds the measure μ. Then the measure μ_T will be embedded by $\rho_T := \inf\{t \geq \eta : (t, W_t) \in R \cup ([T,\infty)\times\mathbb{R})\}$. Obviously, we have that $R\cup([T,\infty)\times\mathbb{R})$ converges to $R\cup([T,\infty)\times\mathbb{R})$ in Root's barrier distance for $\varepsilon \searrow 0$ and convergence of the corresponding measures follows from Lemma 4.3. □

This concludes the convergence of the discrete optimal stopping representation to the continuous optimal stopping representation, hence completing the proof of Theorem 2.7.

The convergence in the Rost case now follows analogously, choosing the function G_T and G_T^N as

$$G_T(x,t) := U_\mu(x) - V_{T-t}^\alpha(x) \quad \text{and}$$
$$G_T^N(x,t) := U_{\mu^N}(x) - V_{T-t}^N(x),$$

concluding the proof of Theorem 2.8.

References

[1] J. Backoff, A. M. G. Cox, A. Grass, and M. Huesmann. Switching Identities by Probabilistic Means. *Séminaire de Probabilités IX Université de Strasbourg* (2021).

[2] E. Bayraktar and T. Bernhardt. On the continuity of the Root barrier. *Proc. Amer. Math. Soc.* 150.7 (2022), pp. 3133–3145.

[3] E. Bayraktar and X. Zhang. Embedding of Walsh Brownian motion. *Stochastic Process. Appl.* 134 (2021), pp. 1–28.

[4] M. Beiglböck, M. Huesmann, and F. Stebegg. Root to Kellerer. *Séminaire de Probabilités XLVIII* (2016), pp. 1–12.

[5] M. Beiglböck, A. M. G. Cox, and M. Huesmann. Optimal transport and Skorokhod embedding. *Invent. Math.* 208.2 (2017), pp. 327–400.

[6] M. Beiglböck, A. M. G. Cox, and M. Huesmann. The geometry of multi-marginal Skorokhod Embedding. *Probab. Theory Related Fields* 176.3-4 (2020), pp. 1045–1096.

[7] M. Beiglböck, M. Nutz, and F. Stebegg. Fine properties of the optimal Skorokhod embedding problem. *J. Eur. Math. Soc. (JEMS)* 24.4 (2022), pp. 1389–1429.

[8] M. Brückerhoff and M. Huesmann. Shadows and Barriers. *ArXiv e-prints* (2021).

[9] P. Carr and R. Lee. Hedging variance options on continuous semimartingales. *Finance Stoch.* 14.2 (2010), pp. 179–207.

[10] A. M. G. Cox and D. Hobson. A Unifying Class of Skorokhod Embeddings: Connecting the Azéma: Yor and Vallois Embeddings. *Bernoulli* 13.1 (2007), pp. 114–130.

[11] A. M. G. Cox and J. Wang. Optimal robust bounds for variance options. *ArXiv e-prints* (Aug. 2013).

[12] A. M. G. Cox and A. Grass. Robust option pricing with volatility term structure – An empirical study for variance options. *ArXiv e-prints* (2023).

[13] A. M. G. Cox and S. M. Kinsley. Discretisation and duality of optimal Skorokhod embedding problems. *Stochastic Process. Appl.* 129.7 (2019), pp. 2376–2405.

[14] A. M. G. Cox, J. Obłój, and N. Touzi. The Root solution to the multi-marginal embedding problem: an optimal stopping and time-reversal approach. *Probab. Theory Related Fields* 173.1-2 (2019), pp. 211–259.

[15] A. M. G. Cox and J. Wang. Root's barrier: construction, optimality and applications to variance options. *Ann. Appl. Probab.* 23.3 (2013), pp. 859–894.

[16] M. Csörgő and P. Révész. On strong invariance for local time of partial sums. *Stochastic Process. Appl.* 20.1 (1985), pp. 59–84.

[17] T. De Angelis. From optimal stopping boundaries to Rost's reversed barriers and the Skorokhod embedding. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.2 (2018), pp. 1098–1133.

[18] P. Gassiat, A. Mijatović, and H. Oberhauser. An integral equation for Root's barrier and the generation of Brownian increments. *Ann. Appl. Probab.* 25.4 (2015), pp. 2039–2065.

[19] P. Gassiat, H. Oberhauser, and G. dos Reis. Root's barrier, viscosity solutions of obstacle problems and reflected FBSDEs. *Stochastic Processes and their Applications* 125.12 (2015), pp. 4601–4631.

[20] P. Gassiat, H. Oberhauser, and C. Z. Zou. A free boundary characterisation of the Root barrier for Markov processes. *Probab. Theory Related Fields* 180.1-2 (2021), pp. 33–69.

[21] N. Ghoussoub, Y.-H. Kim, and T. Lim. Optimal Brownian stopping when the source and target are radially symmetric distributions. *SIAM J. Control Optim.* 58.5 (2020), pp. 2765–2789.

[22] N. Ghoussoub, Y.-H. Kim, and A. Z. Palmer. PDE methods for optimal Skorokhod embeddings. *Calc. Var. Partial Differential Equations* 58.3 (2019), Paper No. 113, 31.

[23] A. Grass. Perkins Embedding for General Starting Laws. *ArXiv e-prints* (2023).

[24] P. Henry-Labordère, J. Obłój, P. Spoida, and N. Touzi. The maximum maximum of a martingale with given n marginals. *Ann. Appl. Probab.* 26.1 (2016), pp. 1–44.

[25] D. Hobson. Robust hedging of the lookback option. *Finance and Stochastics* 2 (4 1998), pp. 329–347.

[26] D. Hobson. The Skorokhod embedding problem and model-independent bounds for option prices. *Paris-Princeton Lectures on Mathematical Finance 2010*. Vol. 2003. Lecture Notes in Math. Springer, Berlin, 2011, pp. 267–318.

[27] D. Hobson and M. Klimmek. Model independent hedging strategies for variance swaps. *ArXiv e-prints* (Apr. 2011).

[28] R. M. Loynes. Stopping times on Brownian motion: Some properties of Root's construction. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 16 (1970), pp. 211–218.

[29] M. Beiglböck, A. Cox, and M. Huesmann. Optimal Transport and Skorokhod Embedding. *Invent. Math.* 208.2 (2017), pp. 327–400.

[30] A. Neuberger. The Log Contract. *Journal of Portfolio Managment,* 20(2) (1994), pp. 74–80.

[31] J. Obłój. The Skorokhod embedding problem and its offspring. *Probab. Surv.* 1 (2004), pp. 321–390.

[32] J. Obłój and P. Spoida. An iterated Azéma-Yor type embedding for finitely many marginals. *Ann. Probab.* 45.4 (2017), pp. 2210–2247.

[33] A. Richard, X. Tan, and N. Touzi. On the Root solution to the Skorokhod embedding problem given full marginals. *SIAM J. Control Optim.* 58.4 (2020), pp. 1874–1892.

[34] D. H. Root. The existence of certain stopping times on Brownian motion. *Ann. Math. Statist.* 40 (1969), pp. 715–718.

[35] H. Rost. Skorokhod stopping times of minimal variance. *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975)*. Berlin: Springer, 1976, 194–208. Lecture Notes in Math., Vol. 511.

[36] H. Rost. Skorokhod's theorem for general Markov processes. *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Tech. Univ. Prague, Prague, 1971; dedicated to the memory of Antonın Špaček)*. Academia [Publishing House of the Czechoslovak Academy of Sciences], Prague, 1973, pp. 755–764.

[37] A. V. Skorohod. Issledovaniya po teorii sluchainykh protsessov (Stokhasticheskie differentsialnye uravneniya i predelnye teoremy dlya protsessov Markova). Izdat. Kiev. Univ., Kiev, 1961, p. 216.

[38] A. V. Skorokhod. Studies in the theory of random processes. Translated from the Russian by Scripta Technica, Inc. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1965, pp. viii+199.

[39] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.* 36 (1965), pp. 423–439.

Supplemental material: Examples of Root and Rost Solutions to discrete and continuous (dSEP)

In this supplemental material we give detailed examples of the computations involved in determining concrete Root and Rost solutions to the discrete, possibly delayed (SEP) as proposed in Theorems 3.3 and 3.4 as well as concrete example of continuous time Root and Rost solutions to the (dSEP).

The structure of this supplemental material is as follows. In Section 1 we choose an initial and terminal measure and follow the algorithm proposed in Theorem 3.3 to first solve the *undelayed* discrete (SEP) with a discrete Root solution. We provide detailed computations along with discussions of resulting barriers and show potential realisations of the random walk hitting these barriers. Subsequently, we introduce a non-trivial delay and revisit the computations to illustrate the modifications required when accommodating such a delay. In Section 2 we proceed in exactly the same manner to illustrate the algorithm proposed in Theorem 3.4 to provide a discrete Rost solution.

Section 3 gives discussions for concrete solutions to continuous time Root and Rost solutions to the (dSEP) and Section 4 does the same for the (MMSEP).

1 Examples of Root solutions to the discrete (SEP) and (dSEP)

Consider the following example (with random starting but without delay)

$$\lambda = \frac{1}{4}\delta_{-1} + \frac{1}{2}\delta_0 + \frac{1}{4}\delta_1,$$
$$\mu = \frac{1}{8}\delta_{-2} + \frac{7}{32}\delta_{-1} + \frac{5}{16}\delta_0 + \frac{7}{32}\delta_1 + \frac{1}{8}\delta_2$$

and $\eta = 0$.

Then $\alpha(0,x) = \lambda(\{x\})$ and $\alpha(t,x) = 0$ for every $t > 0$. Moreover, $\alpha_X = \mathcal{L}^\lambda(X_\eta) = \lambda$ and the measures are in convex order as seen in Figure 1.

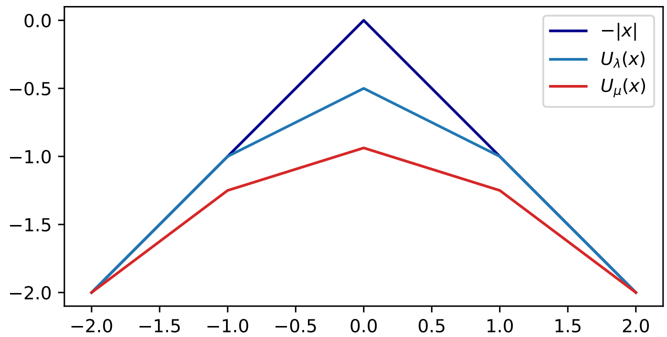


Figure 1: Potential Functions.

The correct Root solution to this problem would be the barrier shown in Figure 2 where the probability of staying in the point $(0,0)$ is given by $\frac{1}{8}$. We will recover the field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ by following the algorithm proposed in Theorem 3.3.

First note that we will only have to consider $x \in \{-2,-1,0,1,2\}$, as all other spatial points will never be reached by the random walk.

For the expected local time $\mathbb{E}^\lambda\left[L^x_{\eta,\rho}(X)\right] = L(x)$ we have

$$L(2) = 0,\ L(1) = \frac{1}{4},\ L(0) = \frac{7}{16},\ L(-1) = \frac{1}{4} \text{ and } L(-2) = 0.$$

Consider first time $t = 0$, then we have $\tilde{\alpha}_0(x) = \lambda(\{x\})$, thus

$\tilde{\alpha}_t(x)$	$t = 0$
$x = 2$	$\tilde{\alpha}_0(2) = 0$
$x = 1$	$\tilde{\alpha}_0(1) = \frac{1}{4}$
$x = 0$	$\tilde{\alpha}_0(0) = \frac{1}{2}$
$x = -1$	$\tilde{\alpha}_0(-1) = \frac{1}{4}$
$x = -2$	$\tilde{\alpha}_0(-2) = 0$

and

l^x_t	$t = 0$
$x = 2$	$l^2_0 = 0$
$x = 1$	$l^1_0 = 0$
$x = 0$	$l^0_0 = 0$
$x = -1$	$l^{-1}_0 = 0$
$x = -2$	$l^{-2}_0 = 0$

.

The quantity $\frac{L(x)-l^x_t}{\tilde{\alpha}_t(x)} \wedge 1$ yields

	$t = 0$
$x = 2$	
$x = 1$	1
$x = 0$	$\frac{7}{8}$
$x = -1$	1
$x = -2$	

, hence

$r_t(x)$	$t = 0$
$x = 2$	
$x = 1$	0
$x = 0$	$\frac{1}{8}$
$x = -1$	0
$x = -2$	

.

For the next time step $t = 1$ we have

$\tilde{\alpha}_t(x)$	$t = 1$
$x = 2$	$\tilde{\alpha}_1(2) = \frac{1}{8}$
$x = 1$	$\tilde{\alpha}_1(1) = \frac{7}{32}$
$x = 0$	$\tilde{\alpha}_1(0) = \frac{1}{4}$
$x = -1$	$\tilde{\alpha}_1(-1) = \frac{7}{32}$
$x = -2$	$\tilde{\alpha}_1(-2) = \frac{1}{8}$

and

l_t^x	$t = 1$
$x = 2$	$l_1^2 = 0$
$x = 1$	$l_1^1 = \frac{1}{4}$
$x = 0$	$l_1^0 = \frac{7}{16}$
$x = -1$	$l_1^{-1} = \frac{1}{4}$
$x = -2$	$l_1^{-2} = 0$

.

The quantity $\frac{L(x) - l_t^x}{\tilde{\alpha}_t(x)} \wedge 1$ yields

	$t = 1$
$x = 2$	0
$x = 1$	0
$x = 0$	0
$x = -1$	0
$x = -2$	0

hence

$r_t(x)$	$t = 1$
$x = 2$	1
$x = 1$	1
$x = 0$	1
$x = -1$	1
$x = -2$	1

.

Repeating the computation for $t > 2$ we again recieve $r_t(x) = 1$ for all $x \in \{-2, -1, 0, 1, 2\}$.

Altogehter, we recover a barrier R described by the following stopping probabilities

	$t = 0$	$t = 1$	$t = 2$
$x = 2$		$r_1(2) = 1$	$r_2(2) = 1$
$x = 1$	$r_0(1) = 0$	$r_1(1) = 1$	$r_2(1) = 1$
$x = 0$	$r_0(0) = \frac{1}{8}$	$r_1(0) = 1$	$r_2(0) = 1$
$x = -1$	$r_0(-1) = 0$	$r_1(-1) = 1$	$r_2(-1) = 1$
$x = -2$		$r_1(-2) = 1$	$r_2(-2) = 1$

.

Check Figure 2 for an illustration of this barrier and several different possible realisations of a random walk embedding the measure μ.

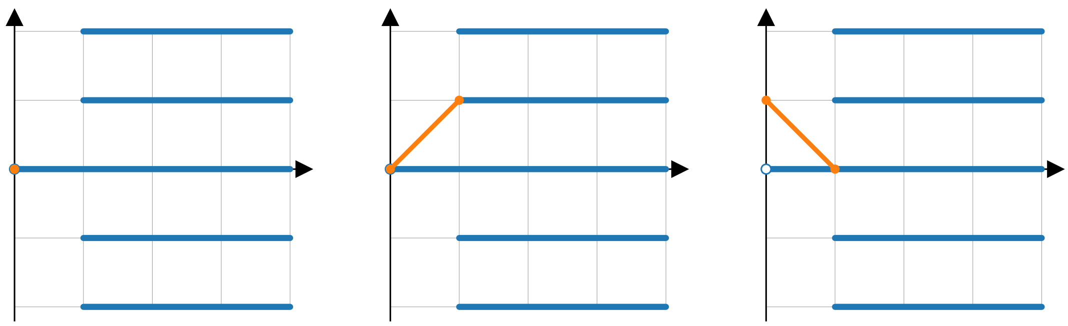


Figure 2: Discrete Root barriers solving the discrete (SEP) and possible realisations of the stopped random walk.

Adding a delay. Let us now consider a non-trivial delay η such that $\mathbb{P}[\eta = 0] = \frac{1}{3}$ and $\mathbb{P}[\eta = 1] = \frac{2}{3}$ and η is independent of X. Then

$$\alpha_X = \frac{1}{12}\delta_{-2} + \frac{1}{4}\delta_{-1} + \frac{1}{3}\delta_0 + \frac{1}{4}\delta_1 + \frac{1}{12}\delta_2$$

and $\alpha_X \leq_c \mu$ will be satisfied as seen in Figure 3.

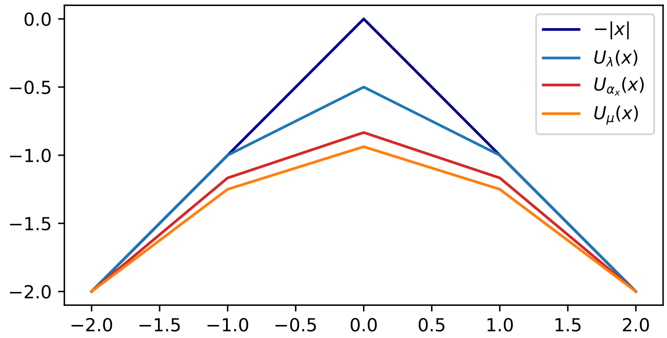


Figure 3: Potential Functions.

Considering this delay, the correct Root solution will be the same as in the undelayed case above (Figure 2), however, the probability of staying in the point $(0,0)$ increases to $\frac{3}{8}$.

For the expected local time $\mathbb{E}^\lambda\left[L^x_{\eta,\rho}(X)\right] = L(x)$ we have

$$L(2) = 0,\ L(1) = \frac{1}{12},\ L(0) = \frac{5}{48},\ L(-1) = \frac{1}{12} \text{ and } L(-2) = 0.$$

For $\alpha(t,x) := \mathbb{P}^\lambda\left[X_\eta = x, \eta = t\right]$ we have the following.

$\alpha(t,x)$	$t=0$	$t=1$	$t=2$
$x=2$	0	$\frac{1}{12}$	0
$x=1$	$\frac{1}{12}$	$\frac{1}{6}$	0
$x=0$	$\frac{1}{6}$	$\frac{1}{6}$	0
$x=-1$	$\frac{1}{12}$	$\frac{1}{6}$	0
$x=-2$	0	$\frac{1}{12}$	0

.

At time $t=0$ we have $\tilde{\alpha}_0(x) = \alpha(0,x)$, thus

$\tilde{\alpha}_t(x)$	$t=0$
$x=2$	$\tilde{\alpha}_0(2) = 0$
$x=1$	$\tilde{\alpha}_0(1) = \frac{1}{12}$
$x=0$	$\tilde{\alpha}_0(0) = \frac{1}{6}$
$x=-1$	$\tilde{\alpha}_0(-1) = \frac{1}{12}$
$x=-2$	$\tilde{\alpha}_0(-2) = 0$

and

l^x_t	$t=0$
$x=2$	$l^2_0 = 0$
$x=1$	$l^1_0 = 0$
$x=0$	$l^0_0 = 0$
$x=-1$	$l^{-1}_0 = 0$
$x=-2$	$l^{-2}_0 = 0$

.

The quantity $\frac{L(x)-l^x_0}{\tilde{\alpha}_0(x)} \wedge 1$ yields

	$t=0$
$x=2$	
$x=1$	1
$x=0$	$\frac{5}{8}$
$x=-1$	1
$x=-2$	

hence

$r_t(x)$	$t=0$
$x=2$	
$x=1$	0
$x=0$	$\frac{3}{8}$
$x=-1$	0
$x=-2$	

,

which is precisely what we expect to recover. For the next time step we have

$\tilde{\alpha}_t(x)$	$t=1$
$x=2$	$\tilde{\alpha}_1(2) = \frac{1}{8}$
$x=1$	$\tilde{\alpha}_1(1) = \frac{7}{32}$
$x=0$	$\tilde{\alpha}_1(0) = \frac{1}{4}$
$x=-1$	$\tilde{\alpha}_1(-1) = \frac{7}{32}$
$x=-2$	$\tilde{\alpha}_1(-2) = \frac{1}{8}$

and

l^x_t	$t=1$
$x=2$	$l^2_1 = 0$
$x=1$	$l^1_1 = \frac{1}{12}$
$x=0$	$l^0_1 = \frac{5}{48}$
$x=-1$	$l^{-1}_1 = \frac{1}{12}$
$x=-2$	$l^{-2}_1 = 0$

.

The quantity $\frac{L(x)-l_1^x}{\tilde{\alpha}_1(x)} \wedge 1$ yields

$$\begin{array}{c|c} & t=1 \\ \hline x=2 & 0 \\ x=1 & 0 \\ x=0 & 0 \\ x=-1 & 0 \\ x=-2 & 0 \end{array} \quad \text{hence} \quad \begin{array}{c|c} r_t(x) & t=1 \\ \hline x=2 & 1 \\ x=1 & 1 \\ x=0 & 1 \\ x=-1 & 1 \\ x=-2 & 1 \end{array},$$

as desired. Check Figure 4 for in illustration of this barrier and several different possible realisations of a random walk embedding the measure μ after the stopping time η.

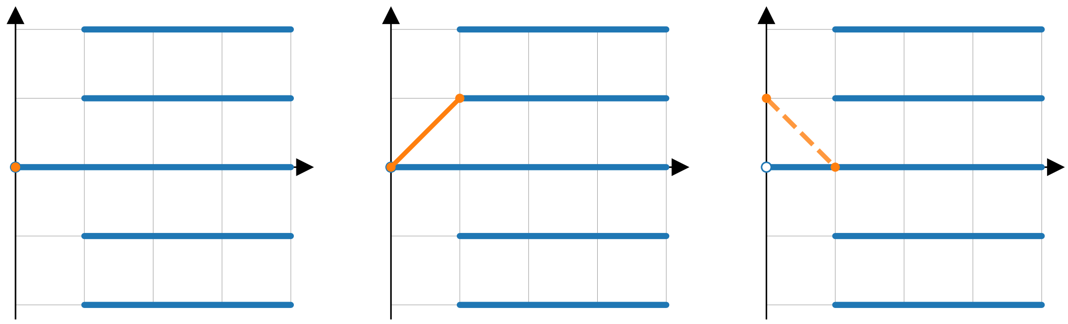


Figure 4: Discrete Root barriers solving the discrete (dSEP) and possible realisations of the stopped random walk.

2 Examples of Rost solutions to the discrete (SEP) and (dSEP)

Let us now construct a Rost solution to the discrete delayed (SEP). Consider the following example (with random starting but without delay).

$$\lambda = \frac{1}{4}\delta_{-1} + \frac{1}{2}\delta_0 + \frac{1}{4}\delta_1 \quad \text{and}$$
$$\mu = \frac{1}{8}\delta_{-2} + \frac{5}{16}\delta_{-1} + \frac{1}{8}\delta_0 + \frac{5}{16}\delta_1 + \frac{1}{8}\delta_2.$$

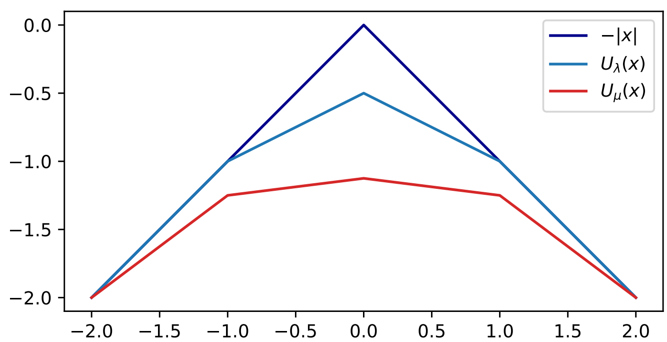


Figure 5: Potential Functions.

We want to follow the algorithm proposed in Theorem 3.4 in order to recover a Rost field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ solving the discrete (SEP).

We need to compute the quantities

$$\nu_t(x) := \mu(x) - \mathbb{P}^\lambda\left[X_\rho = x, \rho \le 0\right]$$

and

$$\tilde{\alpha}_t(x) := \mathbb{P}^\lambda\left[X_t = x, \eta \le t, \rho \ge t\right].$$

Also here we only have to consider $x \in \{-2,-1,0,1,2\}$ as all other spatial points will never be reached by the random walk.

Let us first consider $t = 0$. Then $\nu_t(x) = \mu(\{x\})$ for all $t \in \mathbb{N}$ and $\tilde{\alpha}_0(x) = \lambda(\{x\})$, hence

$\nu_t(x)$	$t=0$
$x=2$	$\nu_0(2)=\frac{1}{8}$
$x=1$	$\nu_0(1)=\frac{5}{16}$
$x=0$	$\nu_0(0)=\frac{1}{8}$
$x=-1$	$\nu_0(-1)=\frac{5}{16}$
$x=-2$	$\nu_0(-2)=\frac{1}{8}$

and

$\tilde{\alpha}_t(x)$	$t=0$
$x=2$	$\tilde{\alpha}_0(2)=0$
$x=1$	$\tilde{\alpha}_0(1)=\frac{1}{4}$
$x=0$	$\tilde{\alpha}_0(0)=\frac{1}{2}$
$x=-1$	$\tilde{\alpha}_0(-1)=\frac{1}{4}$
$x=-2$	$\tilde{\alpha}_0(-2)=0$

.

Computing the quantity $r_0(x) = \frac{\nu_0(x)}{\tilde{\alpha}_0(x)} \wedge 1$ yields

$r_0(x)$	$t=0$
$x=2$	
$x=1$	1
$x=0$	$\frac{1}{4}$
$x=-1$	1
$x=-2$	

.

For the next time $t = 1$ step we have

$\nu_1(x)$	$t=1$
$x=2$	$\nu_1(2)=\frac{1}{8}$
$x=1$	$\nu_1(1)=\frac{1}{16}$
$x=0$	$\nu_1(0)=0$
$x=-1$	$\nu_1(-1)=\frac{1}{16}$
$x=-2$	$\nu_1(-2)=\frac{1}{8}$

and

$\tilde{\alpha}_1(x)$	$t=1$
$x=2$	$\tilde{\alpha}_1(2)=0$
$x=1$	$\tilde{\alpha}_1(1)=\frac{3}{16}$
$x=0$	$\tilde{\alpha}_1(0)=0$
$x=-1$	$\tilde{\alpha}_1(-1)=\frac{3}{16}$
$x=-2$	$\tilde{\alpha}_1(-2)=0$

.

Computing the quantity $r_1(x) = \frac{\nu_1(x)}{\tilde{\alpha}_1(x)} \wedge 1$ yields

$r_1(x)$	$t=1$
$x=2$	
$x=1$	$\frac{1}{3}$
$x=0$	0
$x=-1$	$\frac{1}{3}$
$x=-2$	

.

Now for $t = 2$ we have

$\nu_2(x)$	$t=2$
$x=2$	$\nu_2(2)=\frac{1}{8}$
$x=1$	$\nu_2(1)=0$
$x=0$	$\nu_2(0)=0$
$x=-1$	$\nu_2(-1)=0$
$x=-2$	$\nu_2(-2)=\frac{1}{8}$

and

$\tilde{\alpha}_2(x)$	$t=2$
$x=2$	$\tilde{\alpha}_2(2)=\frac{1}{16}$
$x=1$	$\tilde{\alpha}_2(1)=0$
$x=0$	$\tilde{\alpha}_2(0)=\frac{1}{8}$
$x=-1$	$\tilde{\alpha}_2(-1)=0$
$x=-2$	$\tilde{\alpha}_2(-2)=\frac{1}{16}$

.

Computing the quantity $r_2(x) = \frac{\nu_2(x)}{\tilde{\alpha}_2(x)} \wedge 1$ and using the convention that $r_t(x) = 0$ when $\nu_t(x) = 0$ yields

$r_2(x)$	$t=2$
$x=2$	1
$x=1$	0
$x=0$	0
$x=-1$	0
$x=-2$	1

For all $t > 2$ a similar computation to above will show that $r_t(2) = r_t(-2) = 1$ and $r_t(x) = 0$ for $x \in \{-1, 0, 1\}$. Check Figure 6 for in illustration of this barrier and several different possible realisations of a random walk embedding the measure μ.

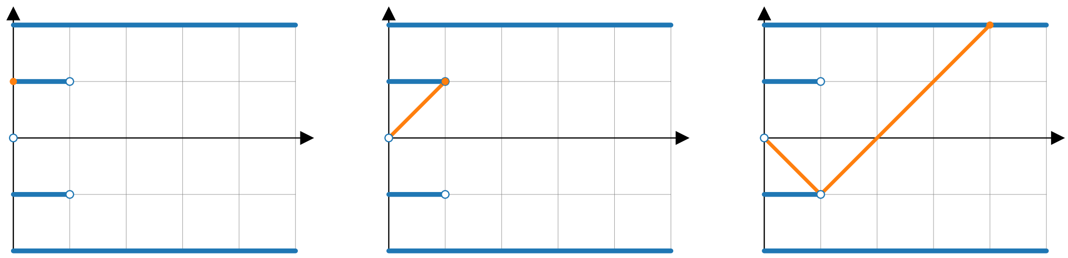


Figure 6: Discrete Rost barriers solving the discrete (SEP) and possible realisations of the stopped random walk.

Adding a delay. Let us now consider a non-trivial delay η such that $\mathbb{P}[\eta = 0] = \frac{1}{6}$ and $\mathbb{P}[\eta = 1] = \frac{5}{6}$ and η is independent of X. Then

$$\alpha_X = \frac{5}{48}\delta_{-2} + \frac{6}{24}\delta_{-1} + \frac{7}{24}\delta_0 + \frac{6}{24}\delta_1 + \frac{5}{48}\delta_2.$$

and $\alpha_X \leq_c \mu$ will be satisfied as seen in Figure 7.

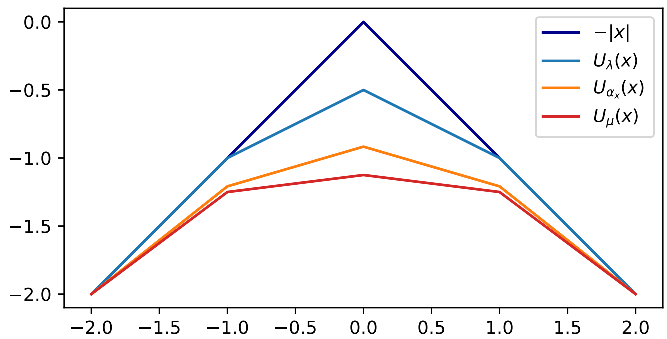


Figure 7: Potential Functions.

Let us again follow the algorithm proposed in Theorem 3.4 in order to recover a Rost field of stopping probabilities $(r_t(x))_{(t,x)\in\mathbb{N}\times\mathbb{Z}}$ solving the discrete (dSEP).

First consider $t = 0$, then $\nu_t(x) = \mu(\{x\})$ for all $t \in \mathbb{N}$ and $\tilde{\alpha}_0(x) = \alpha(\{0\}, \{x\})$, hence

$\nu_t(x)$	$t=0$
$x=2$	$\nu_0(2) = \frac{1}{8}$
$x=1$	$\nu_0(1) = \frac{5}{16}$
$x=0$	$\nu_0(0) = \frac{1}{8}$
$x=-1$	$\nu_0(-1) = \frac{5}{16}$
$x=-2$	$\nu_0(-2) = \frac{1}{8}$

and

$\tilde{\alpha}_t(x)$	$t=0$
$x=2$	$\tilde{\alpha}_0(2) = 0$
$x=1$	$\tilde{\alpha}_0(1) = \frac{1}{24}$
$x=0$	$\tilde{\alpha}_0(0) = \frac{1}{12}$
$x=-1$	$\tilde{\alpha}_0(-1) = \frac{1}{24}$
$x=-2$	$\tilde{\alpha}_0(-2) = 0$

Computing the quantity $r_0(x) = \frac{\nu_0(x)}{\tilde{\alpha}_0(x)} \wedge 1$ yields

$r_0(x)$	$t=0$
$x=2$	
$x=1$	1
$x=0$	1
$x=-1$	1
$x=-2$	

For the next time $t=1$ step we have

$\nu_1(x)$	$t=1$
$x=2$	$\nu_1(2)=\frac{1}{8}$
$x=1$	$\nu_1(1)=\frac{13}{48}$
$x=0$	$\nu_1(0)=\frac{1}{24}$
$x=-1$	$\nu_1(-1)=\frac{13}{48}$
$x=-2$	$\nu_1(-2)=\frac{1}{8}$

and

$\tilde{\alpha}_1(x)$	$t=1$
$x=2$	$\tilde{\alpha}_1(2)=\frac{5}{48}$
$x=1$	$\tilde{\alpha}_1(1)=\frac{5}{24}$
$x=0$	$\tilde{\alpha}_1(0)=\frac{5}{24}$
$x=-1$	$\tilde{\alpha}_1(-1)=\frac{5}{24}$
$x=-2$	$\tilde{\alpha}_1(-2)=\frac{5}{48}$

Computing the quantity $r_1(x) = \frac{\nu_1(x)}{\tilde{\alpha}_1(x)} \wedge 1$ yields

$r_1(x)$	$t=1$
$x=2$	1
$x=1$	1
$x=0$	$\frac{1}{5}$
$x=-1$	1
$x=-2$	1

Now for $t=2$ we have

$\nu_2(x)$	$t=2$
$x=2$	$\nu_2(2)=\frac{1}{48}$
$x=1$	$\nu_2(1)=\frac{3}{48}$
$x=0$	$\nu_2(0)=0$
$x=-1$	$\nu_2(-1)=\frac{3}{48}$
$x=-2$	$\nu_2(-2)=\frac{1}{48}$

and

$\tilde{\alpha}_2(x)$	$t=2$
$x=2$	$\tilde{\alpha}_2(2)=0$
$x=1$	$\tilde{\alpha}_2(1)=\frac{1}{12}$
$x=0$	$\tilde{\alpha}_2(0)=0$
$x=-1$	$\tilde{\alpha}_2(-1)=\frac{1}{12}$
$x=-2$	$\tilde{\alpha}_2(-2)=0$

Computing the quantity $r_2(x) = \frac{\nu_2(x)}{\tilde{\alpha}_2(x)} \wedge 1$ and using the convention that $r_t(x) = 0$ when $\nu_t(x) = 0$ yields

$r_2(x)$	$t=2$
$x=2$	
$x=1$	$\frac{3}{4}$
$x=0$	0
$x=-1$	$\frac{3}{4}$
$x=-2$	

Now for $t=3$ we have

$\nu_3(x)$	$t=3$
$x=2$	$\nu_2(2)=\frac{1}{48}$
$x=1$	$\nu_2(1)=0$
$x=0$	$\nu_2(0)=0$
$x=-1$	$\nu_2(-1)=0$
$x=-2$	$\nu_2(-2)=\frac{1}{48}$

and

$\tilde{\alpha}_3(x)$	$t=3$
$x=2$	$\tilde{\alpha}_2(2)=\frac{1}{96}$
$x=1$	$\tilde{\alpha}_2(1)=0$
$x=0$	$\tilde{\alpha}_2(0)=\frac{1}{96}$
$x=-1$	$\tilde{\alpha}_2(-1)=0$
$x=-2$	$\tilde{\alpha}_2(-2)=\frac{1}{96}$

Computing the quantity $r_3(x) = \frac{\nu_3(x)}{\tilde{\alpha}_3(x)} \wedge 1$ and using the convention that $r_t(x) = 0$ when $\nu_t(x) = 0$ yields

$$\begin{array}{c|c} r_3(x) & t=3 \\ \hline x=2 & 1 \\ x=1 & 0 \\ x=0 & 0 \\ x=-1 & 0 \\ x=-2 & 1 \end{array}.$$

For all $t > 3$ a similar computation to above will show that $r_t(3) = r_t(-3) = 1$ and $r_t(x) = 0$ for $x \in \{-1, 0, 1\}$. Check Figure 8 for in illustration of this barrier and several different possible realisations of a random walk embedding the measure μ.

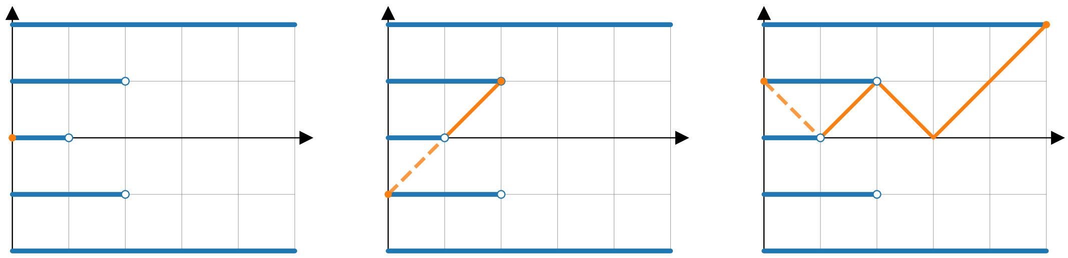


Figure 8: Discrete Rost barriers solving the discrete (dSEP) and possible realisations of the stopped random walk.

3 Continuous time Root and Rost solutions to (dSEP)

Let μ and ν be two measures with the respective densities φ_μ and φ_ν. Then note that for the infimum of the two measures μ and ν

$$(\mu \wedge \nu)(A) := \inf_{B \subseteq A,\, B \text{ measurable}} \big(\mu(B) + \nu(A \setminus B)\big)$$

we have that

$$(\mu \wedge \nu)(A) = \int_A \varphi_\mu(x) \wedge \varphi_\nu(x) dx.$$

Given a probability density function φ we will write

$$U_\varphi(x) = -\int |x-y|\varphi(y)dy$$

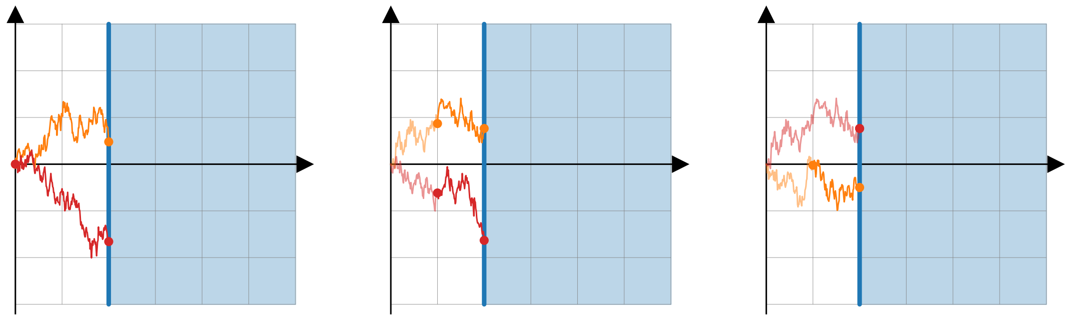
for the associated potential.

From now on φ_t will denote the probability density function associated to $\mathcal{N}(0,t)$.

Let us consider the following example

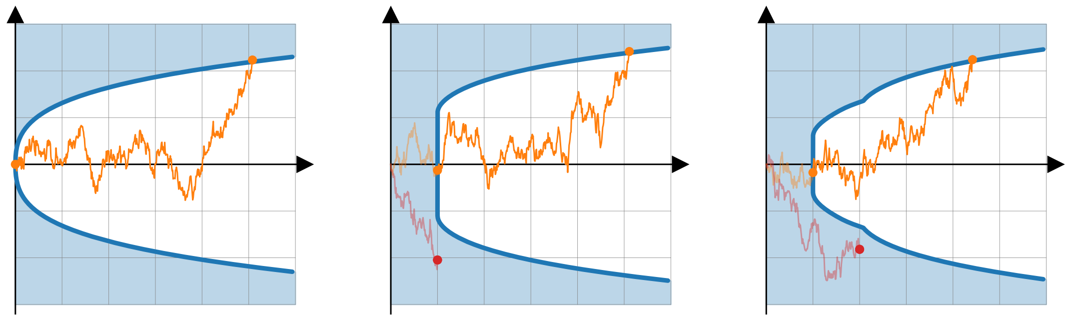
$$\lambda = \delta_0 \quad \text{and} \quad \mu = \mathcal{N}(0,2).$$

We will consider three different delay scenarios.



(A) No delay. (B) Constant delay. (C) Stochastic delay.

Figure 9: Root solutions to (dSEP) for $\lambda = \delta_0$, $\mu = \mathcal{N}(0,2)$ and different delays.



(A) No delay. (B) Constant delay. (C) Stochastic delay.

Figure 10: Rost solutions to (dSEP) for $\lambda = \delta_0$, $\mu = \mathcal{N}(0,2)$ and different delays.

3.1 No delay.

First consider the delay $\eta \equiv 0$, i.e. a conventional (SEP). Then

$$\mathcal{L}^\lambda(W_\eta) = \mathcal{L}^\lambda(W_0) = \delta_0 \leq_c \mathcal{N}(0,2) = \mu.$$

The Root case

The unique Root solution in this case is given by $R = [2,\infty) \times \mathbb{R}$, thus $\rho \equiv 2$ and trivially $W_2 \sim \mathcal{N}(0,2)$ as seen in Figure 9 (a).

The Rost case

Note that $\mathcal{L}^\lambda(\eta) = \delta_0$, hence we only need to consider the single atom in $t = 0$ of η. Then

$$\nu_0 = \mu = \mathcal{N}(0,2)$$

and

$$\alpha_0(A) = \mathbb{P}^\lambda[W_0 \in A, \eta = 0] = \mathbb{P}^\lambda[W_0 \in A] = \delta_0(A).$$

Thus as the measures ν_0 and α_0 are mutually singular we have $(\nu_0 \wedge \alpha_0)(A) = 0$ for any measurable set A and for the random variable Z of Theorem 2.4 we have $Z = \infty$ almost surely.

Given the potential $U_{\mu_t}(x)$ for every $(t,x) \in [0,\infty)\times\mathbb{R}$ obtained via solving the optimal stopping problem of Theorem 2.8 we can identify the set

$$\bar{R} := \{(t,x) : U_{\mu_t}(x) = U_\lambda(x)\},$$

which will be a Rost barrier and the stopping time

$$\rho = \inf\left\{t > 0 : (t, W_t) \in \bar{R}\right\}$$

will embed the measure μ. A numeric approximation of this solution can be seen in Figure 10 (a).

3.2 Constant delay.

Consider now a constant but non-zero delay $\eta \equiv 1$. Then

$$\mathcal{L}^\lambda(W_\eta) = \mathcal{L}^\lambda(W_1) = \mathcal{N}(0,1) \leq_c \mathcal{N}(0,2) = \mu.$$

The Root case

The unique Root solution in this case is still given by $R = [2,\infty)\times\mathbb{R}$ respectively $\rho \equiv 2$ as seen in Figure 9 (b).

The Rost case

Note that $\mathcal{L}^\lambda(\eta) = \delta_1$, hence η has a single atom in 1 and we only need to consider $t = 1$. Then

$$\nu_1 = \mu = \mathcal{N}(0,2)$$

and

$$\alpha_1(A) = \mathbb{P}^\lambda\left[W_1 \in A, \eta = 1\right] = \mathbb{P}^\lambda\left[W_1 \in A\right] = \mathcal{N}(0,1).$$

For x in the interval $\overline{I} := \left[-2\sqrt{\log\left(\sqrt{2}\right)}, 2\sqrt{\log\left(\sqrt{2}\right)}\right]$ we have $\varphi_2(x) \leq \varphi_1(x)$, hence

$$(\nu_1 \wedge \alpha_1)(A) = \int_A \varphi_1(x) \wedge \varphi_2(x)dx = \int_{A\cap\overline{I}} \varphi_2(x)dx + \int_{A\cap\overline{I}^c} \varphi_1(x)dx.$$

This implies that inside of the interval $\overline{I}$ we stop with the probability given via the density φ_2 and outside of the interval we stop immediately.

Thus for the random variable $Z = Z_1$ of Theorem 2.4 we have

$$\mathbb{P}^\lambda\left[Z = 1 \middle| W_1 \in A\right] = \frac{1}{\mathbb{P}^\lambda\left[W_1 \in A\right]} \int_A \varphi_1(x) \wedge \varphi_2(x)dx.$$

Especially, this implies $Z = 1$ almost surely given that $W_1 \in A \subseteq \overline{I}^c$.

Note furthermore that

$$V_t^\alpha(x) = \begin{cases} U_{\varphi_t}(x) & \text{for } t < 1 \text{ and} \\ U_{\varphi_1}(x) & \text{for } t \geq 1. \end{cases}$$

Given the potential $U_{\mu_t}(x)$ for every $(t,x) \in [0,\infty)\times\mathbb{R}$ obtained via solving the optimal stopping problem of Theorem 2.8 we can identify the set

$$\bar{R} := \{(t,x) : U_{\mu_t}(x) = V_t^\alpha(x)\},$$

which will be a Rost barrier and the stopping time

$$\rho = \inf\left\{t > 1 : (t, W_t) \in \bar{R}\right\}$$

will embed the measure μ. See Figure 10 (b) for a numeric approximation of this solution.

3.3 Stochastic delay.

Consider now a delay stopping time η such that $\mathbb{P}^\lambda[\eta = 1] = \gamma$ and $\mathbb{P}^\lambda[\eta = 2] = 1 - \gamma$ for $\gamma \in (0, 1)$. Then

$$\mathcal{L}^\lambda(W_\eta) \sim \gamma\varphi_1 + (1-\gamma)\varphi_2,$$

i.e. the random variable W_η is distributed according to the probability density function

$$\varphi_\eta(x) := \gamma\varphi_1(x) + (1-\gamma)\varphi_2(x).$$

The Root case

For any choice of γ the unique Root solution will still be given by $R = [2, \infty) \times \mathbb{R}$ respectively $\rho \equiv 2$ as seen in Figure 9 (c).

The Rost case

We have $\mathcal{L}^\lambda(\eta) = \gamma\delta_1 + (1-\gamma)\delta_2$, thus we need to investigate the two atoms $t = 1$ and $t = 2$. But the Rost solution will depend heavily on the choice of γ.

Consider $\gamma = \frac{3}{4}$
At time $t = 1$ no mass has yet been embedded, hence

$$\nu_1(A) = \mu(A) = \int_A \varphi_2(x)dx.$$

Furthermore, for the free running mass at time $t = 1$ we have

$$\alpha_1(A) = \mathbb{P}^\lambda\left[W_1 \in A, \eta = 1\right] = \int_A \frac{3}{4}\varphi_1(x)dx.$$

Then for x in the interval $\overline{I}_1 := \left[-2\sqrt{\log\left(\frac{3}{4}\sqrt{2}\right)}, 2\sqrt{\log\left(\frac{3}{4}\sqrt{2}\right)}\right]$ we have that $\varphi_2(x) \leq \frac{3}{4}\varphi_1(x)$, hence

$$(\nu_1 \wedge \alpha_1)(A) = \int_A \varphi_2(x) \wedge \left(\frac{3}{4}\varphi_1(x)\right) dx = \int_{A\cap\overline{I}_1} \varphi_2(x)dx + \int_{A\cap\overline{I}_1^c} \frac{3}{4}\varphi_1(x)dx.$$

For the random variable Z_1 of Theorem 2.4 we have

$$\mathbb{P}^\lambda\left[Z_1 = 1 \middle| W_1 \in A, \eta = 1\right] = \frac{1}{\mathbb{P}^\lambda\left[W_1 \in A, \eta = 1\right]} \int_A \varphi_2(x) \wedge \left(\frac{3}{4}\varphi_1(x)\right) dx.$$

Especially, this implies $Z_1 = 1$ almost surely given that $\eta = 1$ and $W_1 \in A \subseteq \overline{I}_1^c$.
Consider now $t = 2$.
Then for $A \subseteq \overline{I}_1$ we will already have embedded all necessary mass, hence

$$\nu_2(A) = 0 \quad \text{for} \quad A \subseteq \overline{I}_1.$$

However, how much mass is left to embed in $\overline{I}_1^c$ is not to straight forward to compute any more.
We turn to the results of a numeric approximation of this solution given in Figure 10 (b). From this figure we can infer existence of a second interval $\overline{I}_2$ such that $\overline{I}_1 \subseteq \overline{I}_2$ and for the conditional density ψ_2 of Z_2 we have $\psi_2(x) = 0$ for $x \in \overline{I}_2$ while $\psi_2(x) = 1$ for $x \in \overline{I}_2^c$.
For the potential V_t^α of the delay we have

$$V_t^\alpha(x) = \begin{cases} U_{\varphi_t}(x) & \text{for } t \leq 1, \\ U_{\frac{3}{4}\varphi_1 + \frac{1}{4}\varphi_t}(x) & \text{for } 1 < t < 2 \text{ and} \\ U_{\varphi_\eta}(x) & \text{for } t \geq 2. \end{cases}$$

Given the potential $U_{\mu_t}(x)$ for every $(t,x) \in [0,\infty) \times \mathbb{R}$ obtained via solving the optimal stopping problem of Theorem 2.8 we can identify the set

$$\bar{R} := \{(t,x) : U_{\mu_t}(x) = V_t^{\alpha}(x)\},$$

which will be a Rost barrier and the stopping time

$$\rho = \inf\left\{t > \eta : (t, W_t) \in \bar{R}\right\}$$

will embed the measure μ. See Figure 10 (c) for a numeric approximation of this solution.

4 Root and Rost solutions to the (MMSEP)

Instead of a delay stopping time let us consider a (MMSEP) given by the initial distribution $\lambda := \delta_0$ and the two marginals

$$\mu_1 := \mathcal{N}(0,1) \quad \text{and} \quad \mu_2 := \mathcal{N}(0,2).$$

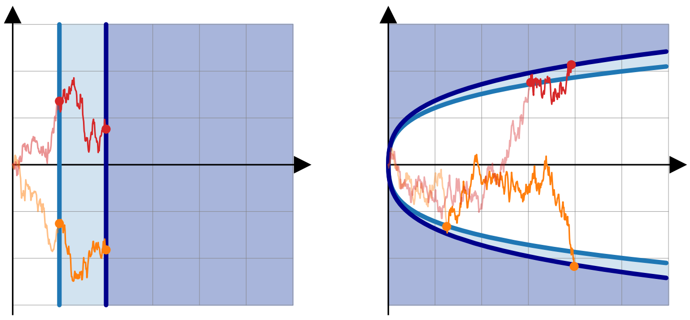


Figure 11: Root and Rost solutions to (MMSEP) for $\lambda = \delta_0$, $\mu_1 = \mathcal{N}(0,1)$ and $\mu_2 = \mathcal{N}(0,2)$.

The Root case

The (MMSEP) will be solved by $R_1 := [1,\infty) \times \mathbb{R}$, respectively $\rho_1 \equiv 2$ and $R_2 = [2,\infty) \times \mathbb{R}$, respectively $\rho_2 \equiv 2 \geq 1 = \rho_1$, see Figure 11 (a).

The Rost case

Consider the first marginal. As the measures λ and μ_1 are mutually singular no external randomisation is needed in this case. Given the potential $U_{\mu_{1,t}}(x)$ for every $(t,x) \in [0,\infty) \times \mathbb{R}$ obtained via solving the optimal stopping problem of Theorem 2.8 we can identify the set

$$\bar{R}_1 := \left\{(t,x) : U_{\mu_{1,t}}(x) = U_\lambda(x)\right\},$$

which will be a Rost barrier and the stopping time

$$\rho_1 = \inf\left\{t > 0 : (t, W_t) \in \bar{R}_1\right\}$$

will embed the measure μ_1.

Let us now consider the second marginal. Note that $\eta = \rho_1$ is an atomless stopping time, hence no external randomisation is needed. Then given the potential $U_{\mu_{2,t}}(x)$ for every $(t,x) \in [0,\infty) \times \mathbb{R}$ obtained via solving the optimal stopping problem of Theorem 2.8 we can identify the set

$$\bar{R}_2 := \left\{(t,x) : U_{\mu_{2,t}}(x) = U_{\mu_{1,t}}(x)\right\},$$

which will be a Rost barrier and the stopping time

$$\rho_2 = \inf\left\{t > \rho_1 : (t, W_t) \in \bar{R}_2\right\}$$

will embed the measure μ_2. See Figure 11 (b) for a numeric approximation.

Robust Option Pricing with Volatility Term Structure — An Empirical Study for Variance Options

A. M. G. Cox and A. M. Grass

Abstract

The robust option pricing problem is to find upper and lower bounds on fair prices of financial claims using only the most minimal assumptions. It contrasts with the classical, model-based approach and gained prominence in the wake of the 2008 financial crisis, and can be used to understand the extent to which a model-based price is sensitive to the underlying model assumptions. Common approaches involve pricing exotic derivatives such as variance options by incorporating market data through implied volatility. The existing literature focuses largely on incorporating implied volatility information corresponding to the maturity of the exotic option. In this paper, we aim to explain how intermediate data can and should be incorporated.

It is natural to expect that this additional information will improve the robust pricing bounds. To investigate this question, we consider variance options, where the bounds of the informed robust pricing problem are known. We proceed to conduct an empirical study uncovering a surprising finding: Contrary to common belief, the incorporation of more information does not lead to an improvement of the robust pricing bounds.

Keywords: variance option; robust pricing; Skorokhod embedding problem; Root barrier; Rost barrier.

MSC 2020 Primary: 91G20; 91G80; Secondary: 60J70; 60G40.

1 Introduction

It is well understood by market participants and regulators that the problem of pricing financial derivatives is highly susceptible to model risk — that is, the risk associated with using incorrect pricing models. One aspect of reducing model risk is to ensure that pricing models accurately reflect quoted market prices (calibration), however by itself, this is insufficient to remove all model risk, since there are typically multiple models which can fit the same set of market prices. It is therefore necessary to understand the set of all models which calibrate to a given set of market prices. By increasing the set of prices that we require a model to calibrate to, we reduce the class of models which can calibrate to the data, and we would therefore expect to get a smaller class of models that are consistent with increased market data, and therefore a better understanding of possible behaviour implied by market prices.

A natural question that arises in practice is: given all relevant prices, can I significantly reduce the range of models which are consistent with observed market prices? In this paper, we will consider this question in the case of market prices for options on variance, given market information in the form of the implied variance term-structure, with information about all available maturities up to the maturity date of the variance option.

Classical approaches to derivative pricing rely on calibrating a class of parametric models to market prices of vanilla options (for example, in the equity context, European Call options). In practice, this process leads to inconsistencies, since parametric models are generally insufficient to capture all possible market features, across multiple maturities for both long- and short-dated maturities. On the other hand, more flexible models (such as Dupire's Local Volatility model [36]) offer greater versatility, but are known not to match market dynamics well, [43], and as a result, are not always reliable models for pricing exotic options.

An alternative approach lies in methods generally called robust, or model-independent pricing. In this approach, dating back to the seminal paper of Hobson [50], one looks to find prices corresponding to models which are extremal, while imposing as few bounds on the price of options as possible. In this context, one is typically able to identify structural properties of extremal models, as well as potential super- or sub-hedging strategies which are robust to model-misspecification. Since the original paper of Hobson, a substantial literature has been developed on this topic, see for example, [15, 19, 20, 22, 25, 46, 54] among others.

In the robust setting, one looks for hedging strategies which will work in any market scenario which is possible, given the possibility of trading in any liquid options. If one increases the set of liquid options that can be used to super-replicate a given exotic option, clearly one has a larger class of possible trading strategies, and therefore a lower model-independent super-replication price. Hence increasing the class of traded options both reduces the minimal super-replication price, and increases the maximal sub-replication price. As a consequence, including more options into the set of liquid options should tighten the bounds on the set of prices which are admissible without (model-independent) arbitrage. By a duality argument (and as a consequence of e.g. [17]), one can also see this as a reduction in the set of pricing models which are consistent with observed market prices.

In general, this perspective allows us to understand inclusion of additional liquid options as a method of potentially generating tighter pricing bounds, and therefore getting a tighter constraint on the extent of model-risk inherent in a given price of an exotic option.

In this paper, we consider the case of options on variance, and in particular, variance calls, that is, options which pay the holder $(V_T - K)_+$, where V is the integrated quadratic variation of the log-asset price, that is, if σ_t is the observed volatility process of the stock price $(P_t)_{t\in[0,T]}$, then $V_T = \int_0^T \sigma_t^2 \, dt$. These derivatives are of interest due in part to their close connection to other options on variance/volatility such as the VIX index, and also in the model-independent context because the variance swap $(V_T - K)$, is well known to have a model-free replication strategy (see [29]). As a result, we might expect model-independent approaches to be informative about (although not completely fix) prices of, for example, variance calls.

We consider specifically whether the model-independent pricing bounds narrow with the incorporation of additional information contained in European call options with maturity earlier than the maturity date of the variance call options. The essence of our approach relies on a construction which allows us to identify the optimal models with the construction of certain space-time boundaries. We are able to informally equate the presence of additional (useful) information of earlier marginals with geometric properties of the underlying boundary structure. Moreover, by fitting market data, we are able to compute the space-time boundaries of market data, and conduct an empirical study which indicates that, at least at the times tested, the additional information content coming from earlier marginal information appears to be negligible, indicating that incorporating information contained in the full implied volatility surface is of little benefit for understanding the model risk inherent in variance call options.

One implication of our results is that market participants who look to robustly hedge variance call options using vanilla options to mitigate their risk (see for example [19]) will not see substantial benefit through the use of call options from earlier maturities. Since model-independent trading strategies of variance call options are known for single maturity options due to [21], this suggests that these trading strategies are sufficient for many practical situations.

We proceed as follows. We first summarise the model-independent, or robust approach to determining worst case bounds for option prices subject to model risk. Then we explain how this connects the case of options on variance to the solutions of Root and Rost of the Skorokhod embedding problem in the single and multiple-marginal settings, and how these solutions can be numerically computed to reveal their geometric structure. In Section 3, we describe the market data which we will use to assess our findings. In particular, in order to get reliable numerical results for the underlying geometric structure, we need to smooth market prices, which are inherently noisy, as well as renormalisation to identify appropriate estimates for interest rate and dividend parameters. Our approach to this relies on calibrating market data to well known models, where we can reliable interpolate complete distributions. This is explained in Section 3.2. We conclude by presenting our numerical results, and showing that the resulting models do not appear to offer any evidence that the multi-marginal setting offers any benefit over calibration for the single marginal setting.

1.1 Worst case bounds

Throughout this article we denote the price process of our underlying asset by $(P_t)_{t\in[0,T]}$. We consider a pay-off function $F : C[0,T] \to \mathbb{R}$ and are interested in determining a fair price F_T of the claim $F\left((P_t)_{t\in[0,T]}\right)$ with *maturity* $T > 0$.

Focusing on pricing these derivatives with maximum caution we could consider *all* possible models for $(P_t)_{t\in[0,T]}$ satisfying only the most minimal assumptions. That is, under a pricing measure a price process is a continuous, non-negative martingale process which agrees with the present-day market value P_0. Let us denote the set of all such martingales by $\mathcal{M}^{P_0}$. It is important to note here, for the sake of simplicity of this discussion, we currently assume zero interest rates and dividend yields as otherwise $(P_t)_{t\in[0,T]}$ will only be a martingale after discounting. We will, however, give all necessary details for non-zero interest rates and dividend yields in Section 3.

Worst Case Bound or *robust bound* pricing entails computing an *upper* and *lower* limit of all possible prices. We then find the following bounds for the price of the derivative F_T:

$$\inf_{M\in\mathcal{M}^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] \leq F_T \leq \sup_{M\in\mathcal{M}^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right]. \tag{1}$$

It, however, quickly becomes apparent that this class $\mathcal{M}^{P_0}$ is too extensive to make much sense of such bounds. Numerous authors have suggested and contributed various restrictions of $\mathcal{M}^{P_0}$ to some sensible subclasses. For instance [1, 13, 28] consider the case where information on a finite number of financial derivatives is available. If a continuum of European call options at multiple time points is liquidly traded, market data yields information about the marginal distribution of the price process by the Breeden-Litzenbeger theorem [14]. The corresponding pricing problem has been studied from a martingale transport perspective, see [7, 11, 17, 39, 53] among many others. It is well known that under mild technical conditions, the upper limit of prices is equal to the minimal model-independent super-hedging price, and vice versa, the lower limit is equal to the maximal model-independent sub-hedging price, see for example [9, 32].

In this article we adopt an approach first proposed by Hobson in [50] to address this challenge. Said method uses market given European call option data to inform on our set of plausible models, to then establish connections to specific solutions to the so called *Skorokhod embedding problem* to find concrete optimizer of (1).

We give more details in the next section and also refer readers to [51] for an extensive survey. Furthermore, we will give an example of a class of claims, namely variance options, which can be examined very rigorously with methods developed in the recent years. While incorporating time T market data to refine the set $\mathcal{M}^{P_0}$ is a well studied problem, only recently the necessary theory was developed which allows to extend such results to considering all available European call option data up to time T.

Our aim is to investigate if, in fact, this additional information would indeed results in tighter robust bounds, as one would expect. In Section 3 we carry out an in-depth empirical study on variance options using real market data quoted for multiple maturities, examining this very question.

2 Market-informed robust pricing and the Skorokhod embedding problem

It is a well-established perspective, as articulated in early works like [35, 36] by Dupire, to no longer perceive European call options merely as derivatives of underlying assets to be priced and instead to view them as independent assets with their prices given by market dynamics. Adopting this sentiment we can invoke the Breeden-Litzenberger Theorem [14], stating that under the assumption that European call options are computed as the discounted payoff under a pricing measure μ,

$$\mathcal{C}(K,T) = \mathbb{E}_\mu\left[(P_T - K)^+\right]$$

for every strike price $K \in (0,\infty)$ and maturity $T > 0$, and furthermore assuming that the pricing function $\mathcal{C}$ is twice differentiable we can recover the pricing measure μ via

$$\mu(P_T \in \mathrm{d}K) = \frac{\partial^2}{\partial K^2}\mathcal{C}(K,T).$$

We can hence refine our set of plausible models $\mathcal{M}^{P_0}$ to $\mathcal{M}^{P_0}_\mu$, the set of those continuous, non-negative martingale models that adhere to the time-T market-given marginal distribution μ, precisely

$$\mathcal{M}^{P_0}_\mu := \left\{(M_t)_{t\in[0,T]} \text{ is a continuous, non-negative martingale such that } M_0 = P_0 \text{ and } M_T \sim \mu\right\}.$$

This allows us to refine the uninformed worst case bounds (1) to the following market-informed robust bounds

$$\underline{F} := \inf_{M\in\mathcal{M}^{P_0}_\mu} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] \le F_T \le \sup_{M\in\mathcal{M}^{P_0}_\mu} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] =: \overline{F}. \tag{2}$$

Let now $(X_t)_{t\geq 0}$ denote a continuous, non-negative martingale subject to $X_0 = P_0$. Consider a stopping time τ and define the time-changed martingale $(M_t)_{t\geq 0}$ as follows

$$M_t := X_{\frac{t}{T-t}\wedge\tau}. \tag{3}$$

Note that $M_T = X_\tau$, thus we have the equivalence

$$(M_t)_{t\geq 0} \in \mathcal{M}_\mu^{P_0} \;\Leftrightarrow\; X_\tau \sim \mu. \tag{4}$$

Such stopping times are precisely the solutions to the (generalized) *Skorokhod embedding problem* ($\mathrm{SEP}_{X,\mu}$), where given a process $(X_t)_{t\geq 0}$ and a measure μ the task is to find a stopping time τ such that

$$X_\tau \sim \mu. \tag{$\mathrm{SEP}_{X,\mu}$}$$

This problem was initially proposed and solved by Skorokhod himself [65, 66] in the early 1960s. However, it continues to be a subject of extensive research, with numerous distinct solutions being established in recent years. More than 20 different solutions are surveyed in [58], and substantial contributions have emerged thereafter, notably in [5, 6, 8, 10, 18, 21, 26, 27, 40, 41, 42, 44, 46, 55, 59]. Most of these solutions to ($\mathrm{SEP}_{X,\mu}$) distinguish themselves through unique additional extremal properties, particular examples will be presented in the next section.

The application of these solutions to the robust pricing problem (2) has been pioneered by Hobson in [50]. The idea, initially focused on lookback options, is to cleverly select $(X_t)_{t\geq 0}$ and a corresponding specific τ solutions to ($\mathrm{SEP}_{X,\mu}$) such that the bounds in (2) are attained by the martingale $(M_t)_{t\geq 0}$ defined in (3).

This idea has been extended in subsequent works, such as [46]. Other types of options have been considered, Barrier options were treated in [15, 19], forward-start digital options in [54], double touch/no-touch options in [19, 20] and options on leverage exchange traded funds in [24, 25].

Of particular interest to our study are the extensions to options on variance, as explored in [16, 21, 27, 52, 56] among others. Further details on this topic will be provided in the subsequent section.

2.1 Options on variance

We now turn our focus to a special class of financial derivatives, namely *variance options*. Variance options are contracts written on the quadratic variation of the log price process $\langle \log P\rangle_T$, more precisely for some function $f : [0,\infty) \to \mathbb{R}$, a variance option has a payoff of the form

$$f\left(\langle \log P\rangle_T\right).$$

Frequently considered examples are

- $F = f(\langle \log P\rangle_T) = \langle \log P\rangle_T - K$, a *variance swap* or
- $F = f(\langle \log P\rangle_T) = (\langle \log P\rangle_T - K)^+$, a *variance call.*

It is worth noting that the quadratic variation of the log price process is invariant under discounting.

Intriguingly, in [34] and more recently in [16, 21, 27] the authors connected the robust pricing problem of variance options to two early and illustrative solutions to ($\mathrm{SEP}_{X,\mu}$), namely the Root [62] and Rost [64] solutions. We state an existence theorem of these solutions and refer to Figure 1 for a graphical illustration.

Theorem 1 ([21, 27, 40, 41, 42, 62, 64]). *Let $(X_t)_{t\geq 0}$ denote a continuous Markov martingale. Consider a measure μ and an increasing convex function $f : [0,\infty) \to [0,\infty)$. Then*

- *there exists a* (Root) barrier $R \subseteq \mathbb{R}^2$ *such that*

$$\tau_{Root} = \inf\left\{t \geq 0 : (t, X_t) \in R\right\}$$

is a stopping time solving the ($\mathrm{SEP}_{X,\mu}$). *A* (Root) barrier *is a set R such that $(t,x) \in R$ implies $(s,x) \in R$ for any $s \geq t$.*

Moreover τ_{Root} minimizes $\mathbb{E}[f(\tau)]$ *among all solutions τ to* ($\mathrm{SEP}_{X,\mu}$).

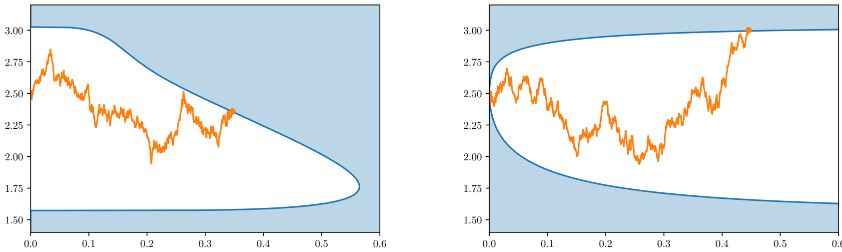


Figure 1: Illustration of a Root (left) resp. Rost (right) solution to (SEP$_{X,\mu}$).

- *there exists an* inverse (Rost) barrier $\bar{R} \subseteq \mathbb{R}^2$ *such that*
$$\tau_{Rost} = \inf \left\{ t > 0 : (t, X_t) \in \bar{R} \right\}$$
is a stopping time solving the (SEP$_{X,\mu}$). *An* inverse (Rost) barrier *is a set* $\bar{R}$ *such that* $(t, x) \in \bar{R}$ *implies* $(s, x) \in \bar{R}$ *for any* $s \leq t$.

  Moreover τ_{Rost} maximizes $\mathbb{E}[f(\tau)]$ *among all solutions* τ *to* (SEP$_{X,\mu}$).

We will use these two concrete solutions to (SEP$_{X,\mu}$) and use observation (4) to construct two concrete candidate models contained in $\mathcal{M}_\mu^{P_0}$.

Let $(X_t)_{t\geq 0}$ denote a *geometric Brownian motion* started in $X_0 = P_0$. Precisely, let $(B_t)_{t\geq 0}$ denote a Brownian motion, then X_t will satisfy the SDE

$$\mathrm{d}X_t = X_t \mathrm{d}B_t, \quad X_0 = P_0. \tag{GBM}$$

It is explained in great detail in [51, Section 5.2] as well as [27, Section 1] and [21, Section 2] that any candidate price process can be represented as a geometric Brownian motion run at a different speed, using a clever time-change argument. Conversely, we can construct candidate price processes as time-changed geometric Brownian motions. A martingales defined in the following way

$$M_t := X_{\frac{t}{T-t} \wedge \tau}, t \in [0, T] \tag{5}$$

where τ is a solution to (SEP$_{X,\mu}$) satisfies the following

- $M_0 = P_0$,
- $M_T = X_\tau \sim \mu$,
- and most importantly, the quadratic variation of the log price process has the following representation $\langle \log M \rangle_T = \tau$.

Given the Root solution τ_{Root} (resp. Rost solution τ_{Rost}) to (SEP$_{X,\mu}$) provided by Theorem 1 we can define the two candidate prices processes

- $M_t^{\text{Root}} := X_{\frac{t}{T-t} \wedge \tau_{\text{Root}}}$, $t \in [0, T]$,
- $M_t^{\text{Rost}} := X_{\frac{t}{T-t} \wedge \tau_{\text{Rost}}}$, $t \in [0, T]$,

and call them Root and Rost model respectively. We can now state that the market-informed robust bounds (2) are attained by these models.

Theorem 2 ([16, 21, 27, 56])**.** *For an increasing convex function* $f : [0, \infty) \to [0, \infty)$ *consider the variance option* $F\left((P_t)_{t\in[0,T]}\right) = f(\langle \log P \rangle_T)$. *Then for any price* F_T *of this claim we have the following robust bounds*

$$\mathbb{E}\left[F\left(M^{Root}\right)\right] = \inf_{M \in \mathcal{M}_\mu^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] \leq F_T \leq \sup_{M \in \mathcal{M}_\mu^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] = \mathbb{E}\left[F\left(M^{Rost}\right)\right] \tag{6}$$

More precisely, the lower and upper bound on these options on variance are attained by a Root and Rost model respectively.

Given the nice graphical properties of Root and Rost solutions to ($\mathrm{SEP}_{X,\mu}$), Theorem 2 offers a possibility of gaining deeper insights into the behaviour of such extremal models attaining the robust pricing bounds.

While the early works on Root and Rost stopping times were purely existential, various more recent works like [21, 27, 30, 40, 41] offer the necessary tools to analyse such models more concretely and in-depth.

Incorporating the market-given marginal at the expiration time T is a well studied problem with established results. But before we apply this theory to market data we aim to generalize the results to not only incorporate the time T marginal at expiration, but to furthermore incorporate *all* maturities available on the market before said time T.

This extension will be explored in the following section.

2.2 Incorporating multiple maturities

The market informed robust pricing problem has been well studied for the class of plausible models $\mathcal{M}_{\mu}^{P_0}$, making the asset adhere to the market-informed time T marginal distribution at times of expiration.

However, European call options are traded liquidly and frequently with maturities on numerous stocks and indices expiring several times per week.

Given this rich supply of data it becomes apparent that we are very likely able to use more data for our robust pricing problem than the time T market informed marginals, namely the marginal distributions for all expiries quoted in between. Consequently, it is reasonable to assert that our plausible model should conform to all these intermediate marginal constraints, thereby tightening the robust bounds on the pricing problem.

To formalize this, let $T_1, \ldots, T_n = T$ denote the available maturities leading up to the expiration time T. Correspondingly, we denote the associated marginals recovered through the Breeden-Litzenberger theorem by $(\mu_1, \ldots, \mu_n)$.

Then we can define the set of all plausible martingale models adhering to the market given marginals by

$$\begin{aligned}\mathcal{M}_{(\mu_1,\ldots,\mu_n)}^{P_0} := \{&(M_t)_{t\in[0,T]} \text{ is a continuous, non-negative martingale} \\ &\text{such that } M_0 = P_0 \text{ and } M_k \sim \mu_k, k = 1, \ldots, n\}\end{aligned}$$

and consider the refined robust bounds

$$\underline{F} := \inf_{M\in\mathcal{M}_{(\mu_1,\ldots,\mu_n)}^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] \le F_T \le \sup_{M\in\mathcal{M}_{(\mu_1,\ldots,\mu_n)}^{P_0}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] =: \overline{F}. \tag{7}$$

Similarly to the single-marginal case, we can again establish a connection between models in $\mathcal{M}_{(\mu_1,\ldots,\mu_n)}^{P_0}$ and solutions to a Skorokhod embedding problem. An analogous observation to the equivalence (4) can be made if instead of the ($\mathrm{SEP}_{X,\mu}$) we consider its multi-marginal extension:

For a sequence of measures $(\mu_1, \ldots \mu_n)$ and a process $(X_t)_{t\ge0}$ the *multi-marginal Skorokhod embedding problem* ($\mathrm{MMSEP}_{X,(\mu_1,\ldots\mu_n)}$) is to find a sequence of stopping times $\tau_1 \le \cdots \le \tau_n$ such that

$$X_{\tau_k} \sim \mu_k \quad \text{for} \quad k = 1, \ldots, n. \tag{$\mathrm{MMSEP}_{X,(\mu_1,\ldots\mu_n)}$}$$

Let $(X_t)_{t\ge0}$ denote a continuous martingale subject to $X_0 = P_0$. Given a sequence of market-given marginals $(\mu_1, \ldots \mu_n)$, denote by $(\tau_1, \ldots, \tau_n)$ a solution to ($\mathrm{MMSEP}_{X,(\mu_1,\ldots\mu_n)}$). With such solution, analogously to the single marginal case we can construct a martingale $(M_t)_{t\ge0} \in \mathcal{M}_{(\mu_1,\ldots,\mu_n)}^{P_0}$

in the following way. Set $T_0 = \tau_0 = 0$ and define

$$M_t := \sum_{i=1}^{n} X_{\tau_{i-1}+\frac{t-T_{i-1}}{T_i-t}\wedge(\tau_i-\tau_{i-1})} \mathbb{1}_{t\in(T_{i-1},T_i]} + P_0. \tag{8}$$

Then analogously to the single marginal case, a model defined in this way will satisfy the following

- $M_0 = P_0$,
- $M_{T_k} = X_{\tau_k} \sim \mu_k$ for $k = 1, \ldots, n$,
- $\langle \log M \rangle_T = \tau_n$.

The existence of multi-marginal Root and Rost solution to ($\text{MMSEP}_{X,(\mu_1,\ldots\mu_n)}$) are ensured by [8, 23, 26]. It is furthermore established in [8] that these solutions still exhibit the same extremal properties necessary to identify Root and Rost solutions as optimizers of the right kind of optimization problem.

More precisely, for an increasing convex function $f : [0,\infty) \to [0,\infty)$, the sequence of Root solutions $(\tau_1^{\text{Root}}, \ldots, \tau_n^{\text{Root}})$ minimizes $\mathbb{E}[f(\tau_k)]$ simultaneously for all $k = 1, \ldots, n$ among all other sequences of stopping times $(\tau_1, \ldots, \tau_n)$ solving the ($\text{MMSEP}_{X,(\mu_1,\ldots\mu_n)}$). Similarly, a sequence of Rost solutions $(\tau_1^{\text{Rost}}, \ldots, \tau_n^{\text{Rost}})$ maximizes $\mathbb{E}[f(\tau_k)]$ simultaneously for all $k = 1, \ldots, n$.

We can now define a multi-marginal Root (resp. Rost) model. Let $(X_t)_{t\geq 0}$ denote a geometric Brownian motion (GBM) started in $X_0 = P_0$. Given the multi-marginal Root solution $\tau^{\text{Root}_1}, \ldots, \tau^{\text{Root}_n}$ (resp. multi-marginal Rost solution $\tau_1^{\text{Rost}}, \ldots, \tau_n^{\text{Rost}}$) to ($\text{MMSEP}_{X,(\mu_1,\ldots\mu_n)}$) we can utilize Definition (8) to propose the two candidate prices processes

- $M_t^{\text{Root}} := \sum_{i=1}^{n} X_{\tau_{i-1}^{\text{Root}}+\frac{t-T_{i-1}}{T_i-t}\wedge(\tau_i^{\text{Root}}-\tau_{i-1}^{\text{Root}})} \mathbb{1}_{t\in(T_{i-1},T_i]} + P_0$, $t \in [0,T]$,
- $M_t^{\text{Rost}} := \sum_{i=1}^{n} X_{\tau_{i-1}^{\text{Rost}}+\frac{t-T_{i-1}}{T_i-t}\wedge(\tau_i^{\text{Rost}}-\tau_{i-1}^{\text{Rost}})} \mathbb{1}_{t\in(T_{i-1},T_i]} + P_0$, $t \in [0,T]$,

and call them multi-marginal Root and multi-marginal Rost model respectively.

For an increasing convex function $f : [0,\infty) \to [0,\infty)$ let us now consider the variance option $F\left((P_t)_{t\in[0,T]}\right) = f(\langle \log P \rangle_T)$. Then any fair price F_T of this claim will adhere to the following robust bounds informed by the market-given marginals $(\mu_1, \ldots, \mu_n)$ as follows

$$\begin{aligned}\mathbb{E}\left[F\left(M^{\text{MM-Root}}\right)\right] &= \inf_{M\in\mathcal{M}^{P_0}_{(\mu_1,\ldots,\mu_n)}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] \\ &\leq F_T \leq \sup_{M\in\mathcal{M}^{P_0}_{(\mu_1,\ldots,\mu_n)}} \mathbb{E}\left[F\left((M_t)_{t\in[0,T]}\right)\right] = \mathbb{E}\left[F\left(M^{\text{MM-Rost}}\right)\right].\end{aligned} \tag{9}$$

It is noteworthy that the inclusion $\mathcal{M}^{P_0}_{(\mu_1,\ldots,\mu_n)} \subseteq \mathcal{M}^{P_0}_{\mu}$ implies that we have $\mathbb{E}\left[F\left(M^{\text{MM-Root}}\right)\right] \geq \mathbb{E}\left[F\left(M^{\text{Root}}\right)\right]$ as well as $\mathbb{E}\left[F\left(M^{\text{MM-Rost}}\right)\right] \leq \mathbb{E}\left[F\left(M^{\text{Rost}}\right)\right]$.

Therefore, incorporating more market given marginal into the robust pricing problem should possibly refine the bounds.

In the next section we summarize the results required for numerical examination of such extremal models achieving the robust pricing bounds informed by the market-given marginals $(\mu_1, \ldots, \mu_n)$.

2.3 Analysis of Root and Rost models

We come to the main objective of this article, to investigate the behaviour of those models for which the bounds of the robust pricing problem (7), incorporating multiple market given marginals are attained.

These models, namely the multi-marginal Root and Rost models defined in Section 2.2 are characterized by the respective sequences of barriers which need to be passed by the candidate price processes consecutively. Hence, we believe that analysing these barriers might offer valuable insights into the underlying models.

While the original results of Root [62] and Rost [63] are purely of existential nature and concrete barriers were not known save for a few very simple examples, these solutions later became computationally tractable thanks to [21, 27, 30, 40, 41] among others.

In this single-marginal case, where the robust bounds are only informed by the time T marginal distribution at time of expiration as outlined in Section 2.1, a thorough numerical investigation of the Root and Rost models was conducted in [21] using simulated market data. Below, we extend these results to the multi-marginal case. In order to perform a similar meaningful analysis of the multi-marginal Root and Rost models attaining the bounds of the robust pricing problem (7) informed by multiple maturities, tractable representations of multi-marginal Root and Rost solutions are essential. A computable multi-marginal Root solution is provided in [26], while a computable multi-marginal Rost solution, to the best of our knowledge, has only been developed recently in [23].

We will give a brief summary of the well-established single-marginal results and proceed to state their extensions to multiple marginals.

Single-marginal results

Let $(X_t)_{t\geq 0}$ represent a geometric Brownian motion (GBM). Given a measure μ let τ_{Root} be a Root solution to ($\text{SEP}_{X,\mu}$), induced by the Root barrier R. Throughout this paper we consistently assume the process $(X_t)_{t\geq 0}$ to start according to the starting measure $\lambda = \delta_{P_0}$, however more general starting measure can be considered. We refer to the original papers for details.

Define the function $u(t,x) := -\mathbb{E}\left[|X_{\tau_{\text{Root}}\wedge t} - e^x|\right]$, and for a measure m, let $U_m(x) := \int |x - y| \mathrm{d}m(y)$ denote its *potential*. The Root barrier R now also has the following representation

$$R = \{(t, e^x) : u(t,x) = U_\mu(e^x)\} . \tag{10}$$

However, it is crucial to know and the main result in [27] that the function $u(t,x)$ can be recovered without any knowledge of τ_{Root} and R as a solution to the following PDE in variational form

$$\begin{gathered}\max\left\{\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(t,x) - \frac{1}{2}\frac{\partial u}{\partial x}(t,x) - \frac{\partial u}{\partial t}(t,x), U_\mu(e^x) - u(t,x)\right\} = 0,\\ u(0,x) = U_\lambda(e^x).\end{gathered} \tag{11}$$

This gives a concrete method for the numerically computing Root barriers by solving Equation (11) and recovering the Root barrier R via Equation (10).

Similarly, let τ_{Rost} denote a Rost solution to ($\text{SEP}_{X,\mu}$), induced by the Rost barrier $\bar{R}$ and consider the function $v(t,x) := U_\mu(e^x) + \mathbb{E}\left[|X_{\tau_{\text{Rost}}\wedge t} - e^x|\right]$. Then the Rost barrier $\bar{R}$ also has the following representation

$$\bar{R} = \{(t, e^x) : v(t,x) = U_\mu(e^x) - U_\lambda(e^x)\} . \tag{12}$$

Moreover, similar to the Root case and as discovered in [21, 30] the function $v(t,x)$ can also be recovered as the solution to the following PDE in variational form

$$\begin{gathered}\max\left\{\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(t,x) - \frac{1}{2}\frac{\partial u}{\partial x}(t,x) - \frac{\partial u}{\partial t}(t,x), (U_\mu - U_\lambda)(e^x) - v(t,x)\right\} = 0,\\ v(0,x) = (U_\mu - U_\lambda)(e^x).\end{gathered} \tag{13}$$

Multi-marginal results

The preceding results allow for analysis of Root and Rost models associated to options on variance when considering the market given marginal at time of maturity T. However, extending this analysis to a multi-marginal context, to the best of our knowledge, has remained unexplored.

In Section 2.2 it was explained how the bounds of the robust pricing problem informed by multiple maturities as formulated in (7) are attained by a multi-marginal Root and Rost model respectively. In order to perform a similarly meaningful investigation of such models as in the single-marginal case it remains to establish a method of numerical analysis analogous to the above. For the multi-marginal Root case a computable representation is know due to [26], while [23] recently introduced a similar representation for multi-marginal Rost solutions. Both these works give an optimal stopping characterization of multi-marginal Root and Rost solutions and state results for embedding into a Brownian motion.

It remains to discuss the application of these results to a geometric Brownian motion (GBM) and how to establish at a PDE representation similar to (11) and (13). The process of deriving a PDE representation in variational form starting from an optimal stopping problem is well known due to [12], we also refer to Section 5.2 in [61] for a more accessible explanation.

To extend such results form a Brownian motion to a geometric Brownian motion we utilize the same methods as described in [27, Section 4.3] where the authors propose the change of variables $(t, x) \mapsto (t, e^x)$. While the transitions densities of a Brownian motion must satisfy the heat equation, the transition densities of the (GBM) will in turn, after said change of variables satisfy a drift-diffusion equation as seen on the left hand side of the PDEs in variational form (11), (13) (14) and (15). For more details we refer to the proof of [27, Theorem 4.6] and the preceding discussion therein.

Altogether, this leads us to the following characterizations of multi-marginal Root and Rost solutions to $(\mathrm{MMSEP}_{X,(\mu_1,\ldots\mu_n)})$ as solutions to PDEs in variational form.

Consider now a geometric Brownian motion (GBM) $(X_t)_{t\geq 0}$ and a sequence of measures $\mu_1, \ldots, \mu_n$.

We first consider the multi-marginal Root case, hence let $(\tau_1^{\mathrm{Root}}, \ldots, \tau_n^{\mathrm{Root}})$ be a multi-marginal Root solution to $(\mathrm{MMSEP}_{X,(\mu_1,\ldots\mu_n)})$ induced by the sequence of Root barriers $R_1, \ldots, R_n$.

For $k = 1, \ldots, n$ define the function $u_k(t, x) = -\mathbb{E}\left[|X_{\tau_k^{\mathrm{Root}}\wedge t} - e^x|\right]$ as well as $u_0(t, x) = U_\lambda(e^x) = U_{\delta_{P_0}}(e^x)$.

Then the k-th Root barrier R_k corresponding to the k-th Root stopping time τ_k^{Root} has the following representation

$$R_k = \left\{(t, e^x) : u_{k-1}(t, x) - u_k(t, x) = \left(U_{\mu_{k-1}} - U_{\mu_k}\right)(e^x)\right\}.$$

Similar to the single-marginal case, the functions $u_k(t, x)$ for $k = 1, \ldots, n$ can be recovered inductively as the solution of the following PDE in variational form

$$\begin{gathered}\max\left\{\frac{1}{2}\frac{\partial^2 u_k}{\partial x^2}(t,x) - \frac{1}{2}\frac{\partial u_k}{\partial x}(t,x) - \frac{\partial u_k}{\partial t}(t,x),\, u_{k-1}(t,x) - u_k(t,x) - \left(U_{\mu_{k-1}} - U_{\mu_k}\right)(e^x)\right\} = 0,\\ u_k(0,x) = U_\lambda(e^x).\end{gathered} \tag{14}$$

A similar result holds in the multi-marginal Rost case. For $k = 1, \ldots, n$ define the function $v_k(t, x) = U_{\mu_k}(e^x) - \mathbb{E}\left[|X_{\tau_k^{\mathrm{Rost}}\wedge t} - e^x|\right]$ as well as $v_0(t, x) = (U_{\mu_k} - U_\lambda)(e^x) = (U_{\mu_k} - U_{\delta_{P_0}})(e^x)$. Then the k-th Rost barrier $\bar{R}_k$ corresponding to the k-th Rost stopping time τ_k^{Rost} also has the following representation

$$\bar{R}_k = \left\{(t, e^x) : v_{k-1}(t, x) - v_k(t, x) = \left(U_{\mu_{k-1}} - U_{\mu_k}\right)(e^x)\right\}.$$

Again, the functions $v_k(t, x)$ for $k = 1, \ldots, n$ can be recovered inductively as the solution to the following PDE in variational form

$$\max\left\{\frac{1}{2}\frac{\partial^2 v_k}{\partial x^2}(t,x)-\frac{1}{2}\frac{\partial v_k}{\partial x}(t,x)-\frac{\partial u}{\partial t}(t,x),\, v_{k-1}(t,x)-v_k(t,x)-\left(U_{\mu_{k-1}}-U_{\mu_k}\right)(e^x)\right\}=0,$$
$$v_k(0,x)=(U_{\mu_k}-U_\lambda)(e^x). \tag{15}$$

Hence, provided market-given marginal $(\mu_1,\ldots,\mu_n)$ the above allows us to compute the concrete barriers the Root and Rost models must adhere to. As these barriers each have to be passed through consecutively, this possibly gives us valuable insight into the behaviour of these extremal models.

Computation of robust bounds

We conclude this section by explaining the computation of prices for variance options given Root and Rost barriers. For a function $f:[0,\infty)\to[0,\infty)$ we consider the variance option $F\left((M_t)_{t\in[0,T]}\right)=f(\langle\log M\rangle_T)$.

For a Root barrier R we define a corresponding *barrier function* $R(x)$ as follows

$$R(x):=\inf\{t\geq 0:(t,x)\in R\},$$

while for a Rost barrier $\bar{R}$ such a function will be defined via

$$\bar{R}(x):=\sup\{t\geq 0:(t,x)\in \bar{R}\}.$$

Let us now assume that μ is atomless and the barrier functions $R(x)$ and $\bar{R}(x)$ are well defined and continuous. In this case, we particularly note that

$$\tau=R(X_\tau).$$

Recall that for our martingale model $(M_t)_{t\in[0,T]}$ defined via (5) or (8) we have $\langle\log M\rangle_T=\tau$ where τ denotes the Root (resp. Rost) barrier corresponding to the terminal maturity T.

Thus, under such a Root (resp. Rost) model, a fair price F_T of our variance option can be easily computed as

$$F_T=\mathbb{E}\left[f(\tau)\right]=\int f(R(x))\mathrm{d}\mu_T(x). \tag{16}$$

We carry out an empirical study on real market data in the next section.

3 Empirical study

We conduct an empirical study utilizing European put and call options data quoted on the S&P500 index as of March 19, 2019, acquired from the CBOE Datashop.

We opt for using closing quotes from the dataset for our study, investigating the 12 available (non-weekly) expiries which mature every third Friday of the month.

The shortest expiry within our dataset is 31 days to maturity (0.085 years to maturity), while the longest maturity extends to 1005 days (2.75 years to maturity). Furthermore, there are between 83 and 296 prices quoted per maturity, ranging between 100\$ and 4200\$ in strike value while the closing spot price of the S&P500 was provided by CBOE as $P_0=2833\$$.

The prices sourced from CBOE are provided as bid-ask quotes, for our further analysis we compute and work with the mid-point prices. These quoted (mid-point) call and put options prices are illustrated in Figure 2.

While the exact interest rates and dividend yield used for option pricing were not disclosed by CBOE, we estimate a constant implied interest rate, denoted as r, at $r=0.0265$, and a constant implied dividend yield, denoted as q, at $q=0.0185$. Details on our estimation methods can be

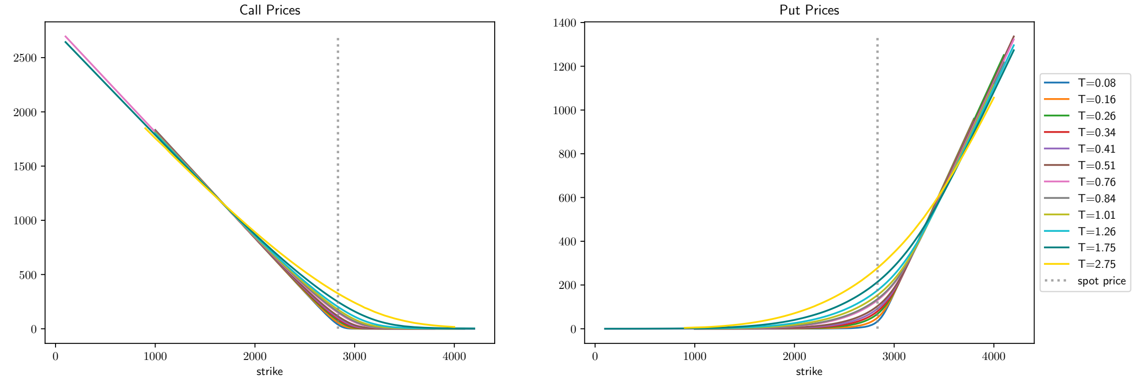


Figure 2: Market given European call and put option prices on the S&P500.

found in Appendix A. It is noteworthy, however, that these estimated values are in alignment with the T-bill rates quoted by the U.S. Department of the Treasury, as well as reported annual dividend yields of the S&P500, promoting confidence in the validity of our empirical setup.

In Section 3.1 we take the necessary preparations in order to compute multi-marginal Root and Rost barriers as detailed in Section 2.3. The numerical results will be presented in Section 3.2.

3.1 Preparation of the data

Market-quoted data is inherently noisy, occasionally disrupted by reporting delays or inaccuracies. These irregularities can introduce problems like artificial arbitrage opportunities within the data, which may not accurately represent real market conditions. Moreover, while the prices themselves may appear orderly, closer examination reveals increasing levels of noise in their first and second derivatives, as depicted in Figure 3.

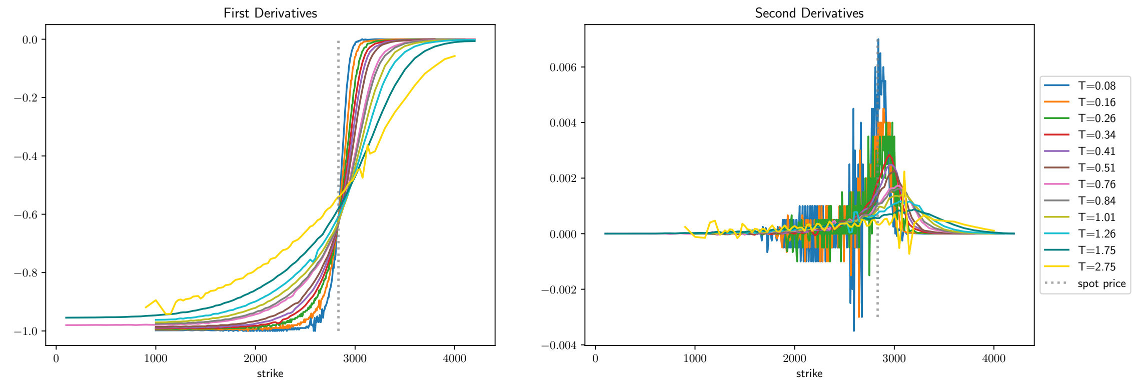


Figure 3: First and second derivatives of market given call prices.

In the context of numerical solutions to partial differential equation, as introduced in Section 2.3, these noisy derivatives pose a great challenge, usually resulting in failure of the algorithm.

There are several possibilities to tackling this problem, we choose to investigate two possible calibration scenarios, both calibrating a Black-Scholes model and a Heston model to the market given data. Our rationale for this choice is as follows: Despite its recognized limitations, the Black-Scholes model remains a widely used method for pricing European call options by numerous financial institutions.

The Heston model [49] is know to overcome some of the limitations of the Black-Scholes model while remaining tractable, hence being a favored alternative to the Black-Scholes model by practitioners. Heston models can in practice demonstrate exact fits to market data in part due to its widespread use by market participators.

In sum, rather than considering market prices as direct indicators, we view the market as providing the underlying pricing parameters for widely used pricing models.

This approach is not necessary, provided alternative approaches to smoothing the market data can be used. Other potential methods include [31] or arbitrage free spline smoothing as suggested in [38], but we have not implemented these methods to date.

Forward prices.

Our dataset provides European call option quoted on the spot price process $(P_t)_{t\geq 0}$. However, as presented in Section 2, it is essential to consider call option prices quoted on the forward price process. Therefore, the final step in our data preparation requires appropriate modification of the quotes.

Given a strike price x and a maturity T we will by $\mathcal{C}(x,T)$ denote the price of a European call option on the asset $(P_t)_{t\geq 0}$. We furthermore assume this price to be given under a risk neutral (martingale) measure via

$$\mathcal{C}(x,T) = e^{-rT}\mathbb{E}\left[(P_T - x)^+\right]$$

where r denotes the constant and deterministic risk-free interest rate.

As the S&P500 is a stock index consisting of 500 different stocks, each paying dividends at individual times throughout the year we choose to model a continuously paid dividend which can effectively be seen as an average of all the different dividends paid by the different companies at various times.

Assuming such a continuously compounded dividend yield q, we consider the forward price process $(S_t)_{t\geq 0}$

$$S_t := e^{(q-r)t}P_t,$$

which is a martingale under the pricing measure.

By $c(x,T)$ we will denote the time-zero price of a call option on the discounted process $(S_t)_{t\geq 0}$ with strike x and maturity T,

$$c(x,T) := \mathbb{E}\left[(S_T - x)^+\right].$$

We then observe the following relationship between $\mathcal{C}(x,T)$ and $c(x,T)$

$$c(x,T) = \mathbb{E}\left[\left(e^{(q-r)T}P_T - x\right)^+\right] = e^{qT}\left(e^{-rT}\mathbb{E}\left[\left(P_T - e^{-(q-r)T}x\right)^+\right]\right) = e^{qT}\mathcal{C}\left(e^{-(q-r)T}x, T\right). \tag{17}$$

Hence, we use Equation (17) to compute prices quoted on the forward price process given a pricing function $\mathcal{C}$.

The forward prices associated to our market given data can be seen in Figure 4 (c) resp. Figure 8 (c) using a pricing function obtained via calibration of a Black-Scholes resp. Heston pricing model to the market given data.

As presented in Section 2.3, the necessary input for the computation of Root and Rost barriers are the potentials $U_{\mu^k}(x)$ corresponding to market implied measures μ^k, $k \in \{1,\dots,n\}$. Note that the potential of a measure can easily be recovered from the call prices (17) in the following way

$$U_{\mu^k}(x) = P_0 - 2c(x,T_k) - x,$$

we show the potentials for our data in Figure 4 (c) resp. Figure 8 (c).

3.2 Numerical results

In this section we calibrate both a Black-Scholes and a Heston model to our market-given data. This calibration process enables us to use the respective call prices as a smoothed approximation of our market prices to inform on the robust pricing problem (7).

The effectiveness of the respective calibration can be seen in Figure 4 (a) and Figure 8 (a) where we show a comparison of the calibrated prices to the market-given prices. Additionally, we observe the indeed very well-behaved first and second derivatives of the calibrated prices in Figure 4 (b) and Figure 8 (b).

Using a classic Euler numeric differentiation scheme we first solve the single-marginal embedding problem given in (11) resp. (13) for each maturity. Here the stopping time embedding the k-th marginal is not required to wait for the stopping time embedding the $(k-1)$-th marginal. The results are depicted on the left side in Figure 5 (a) and (c) resp. Figure 9 (a) and (c). On the right these same figures show the numerical solution to the multi-marginal problem given in (14) resp. (15), namely in Figure 5 (b) and (d) resp. Figure 9 (b) and (d).

We conduct a more extensive discussion of the resulting barriers in Section 3.3.

3.2.1 Black-Scholes data

Calibration results. Calibrating a *Black-Scholes model* to the market data yields the pricing parameter $\sigma = 0.153$.

We want to point out that the fit of the calibrated model to the market given data as seen in Figure 4 (a) seems unfavourable and not ideal. The Black-Scholes model does however provide very smooth derivatives, depicted in Figure 4 (b) that allow for easy application of the Root and Rost barrier computation algorithm. The Black-Scholes forward prices and the associated potential functions used for the algorithm can be seen in Figure 4 (c).

Root and Rost barriers.

We give the numerical results obtained by solving both the PDE in variational form (11) resp. (13) as well as (14) resp. (15) using an explicit Euler scheme.

It is important to note that the algorithm exhibits certain challenges when dealing with situations where the potential associated with the time T_k marginal closely approaches the initial potential $-|x - P_0|$.

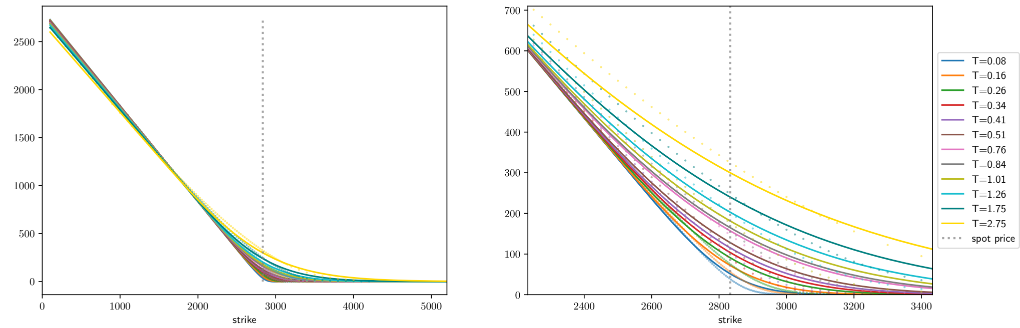
Consequently, we depict the region where these potentials remain a minimum distance of $\epsilon = 0.01$ away from the initial potential using solid lines. The remaining portions of the barrier are shown transparently. However, it is worth mentioning that we believe this boundary behavior to be an artifact of the numerical scheme and not indicative of the true behavior of the barriers.

In the case of the Root barriers, this is very easy to see. It is straightforward to verify that the case of constant barrier functions will recover exactly the Black-Scholes model. Hence when trying to compute the barrier function corresponding to the Black-Scholes model, we would expect to recover a constant barrier. This is exactly the behaviour of the barriers shown in Figure 5, at least in the likely region, corresponding to the 'thick' lines in the figures.

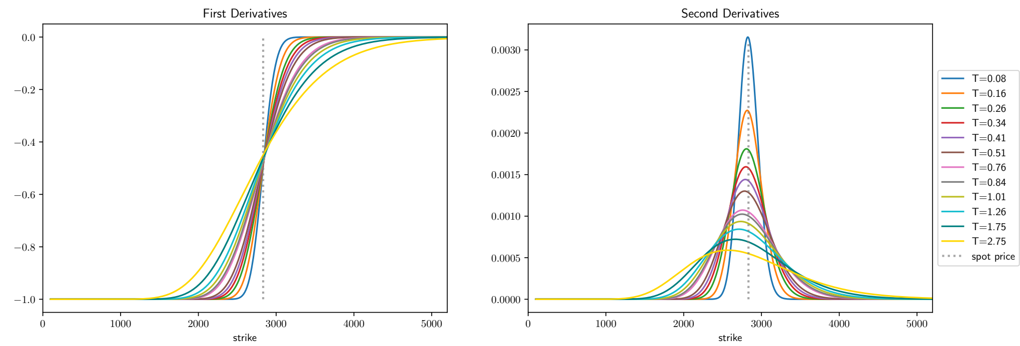
Prices.

We use the barriers obtained above to establish robust bounds for a variance call, as detailed in Section 2.3 and Equation (16). Given a maturity T and a strike price K, our objective is to price the *variance call* defined as

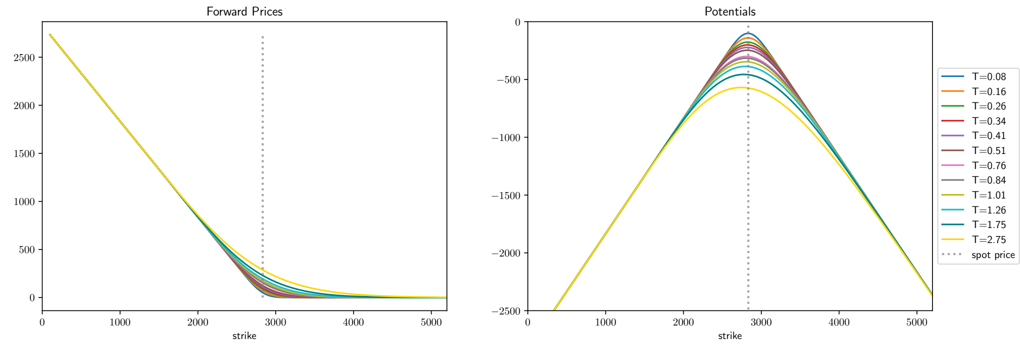
$$f(\langle \log M \rangle_T) = (\langle \log M \rangle_T - K)^+ .$$



(A) Comparing calibrated prices to market given prices.

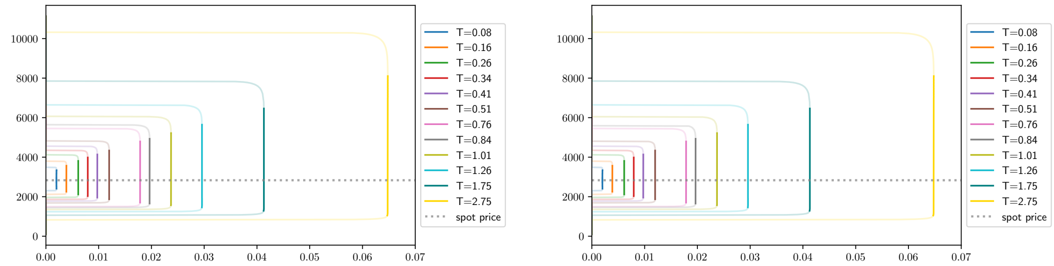


(B) First and second derivatives of calibrated prices.

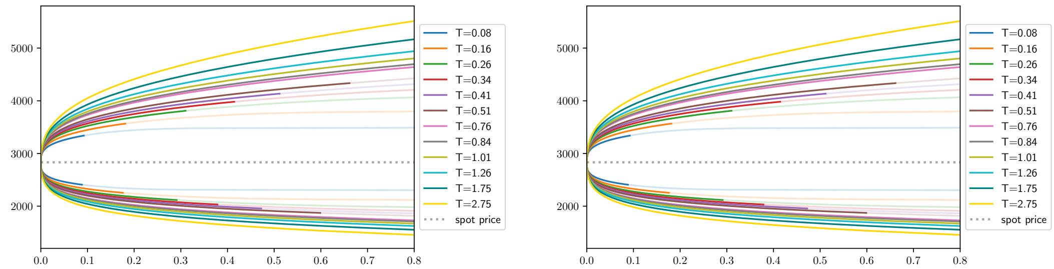


(C) Forward prices and the associated potentials.

Figure 4: Black-Scholes model calibration results.



(A) Single marginal Root barriers (B) Multi marginal Root barriers



(C) Single marginal Rost barriers (D) Multi marginal Rost barriers

Figure 5: Root (top) and Rost (bottom) barriers embedding the market implied marginals of the S&P500.

We consider both an intermediate maturity, $T = 1.01$ as well as the terminal maturity $T = 2.75$. As the single-marginal and the multi-marginal barriers coincide, the resulting bounds are identical, regardless of whether we employ the single-marginal or multi-marginal model.

In Figure 6 we illustrate the option prices for these variance calls for a wide range of possible strike prices. Notably, while the upper and lower bounds coincide for the strike price 0, the gap between the upper and lower robust pricing bounds widens notably for increasing strike prices.

3.2.2 Heston data

Calibration results. Calibrating a *Heston model* to the market data yields the following pricing parameters.

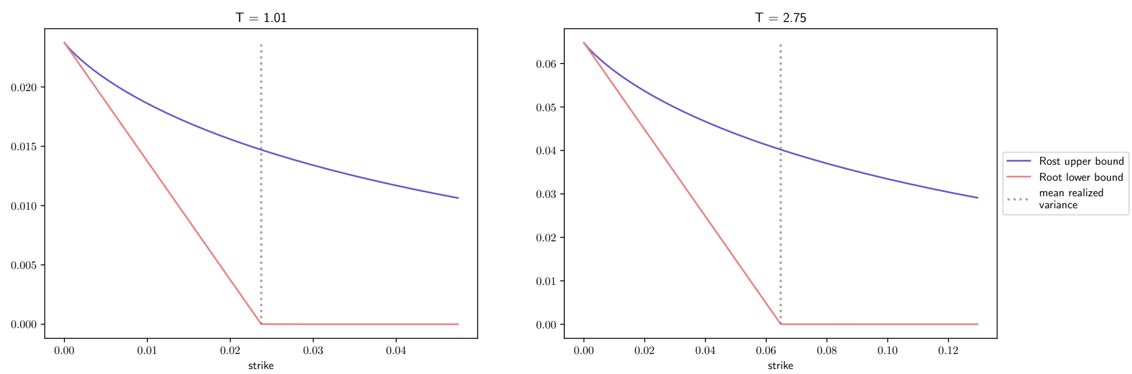
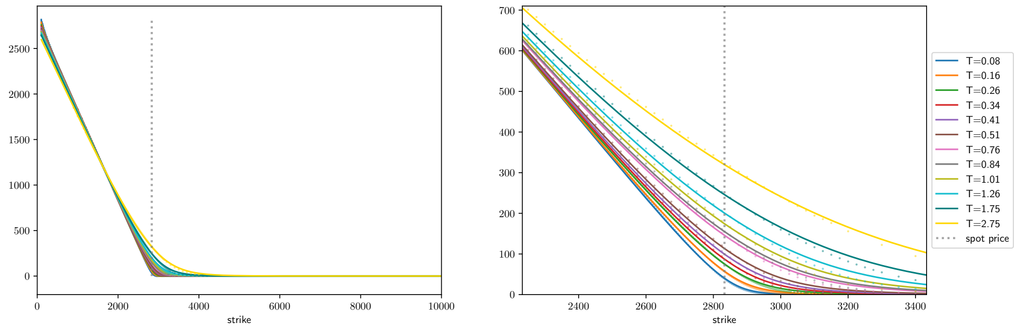
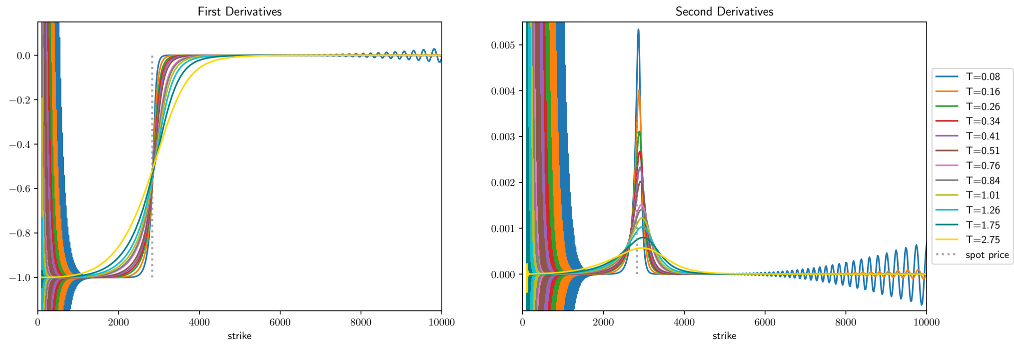


Figure 6: Upper and lower robust bounds on variance call prices for two different maturities.



(A) Comparing calibrated prices to market given prices.



(B) First and second derivatives of calibrated prices.

Figure 7: Heston model calibration results before tail correction.

Parameter	Calibration Result	Parameter Meaning
v_0	0.0158	initial volantility
$\bar{v}$	0.0361	long-run mean variance
ρ	-0.5199	corellation of asset BM and volatility BM
κ	2.6280	speed of mean reversion
σ	0.7902	vol of vol

Table 1: Calibration Results

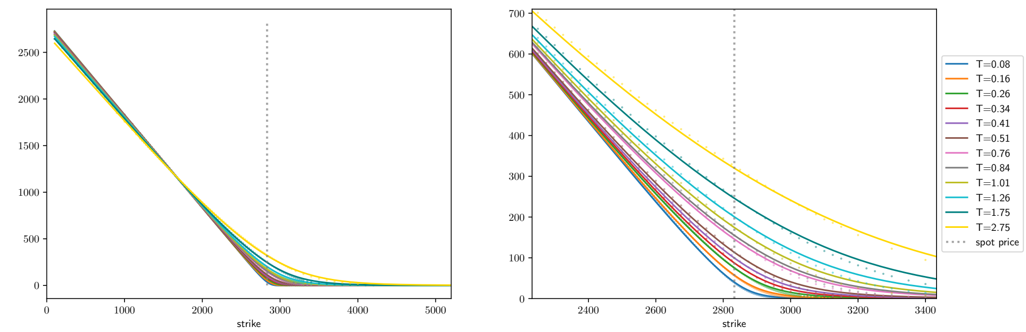
The fit of the Heston prices to the market given data has improved substantially over the Black-Scholes model as seen in Figure 8 (a), but can still not be considered ideal.

However, most common implementation of the Heston pricing functionals such as the python package `QuantLib` seem to exhibit numerical instability in the tails, especially prominent in the derivatives as depicted in Figure 7. To mitigate this problem it was necessary to smooth out the tails in a sensible manner. Just before the convexity assumption of the Heston prices breaks down, we replace the Heston tails with exponential tails fitted to the function values as well as the first derivatives. This approach seems to provide a surprisingly nice fit.

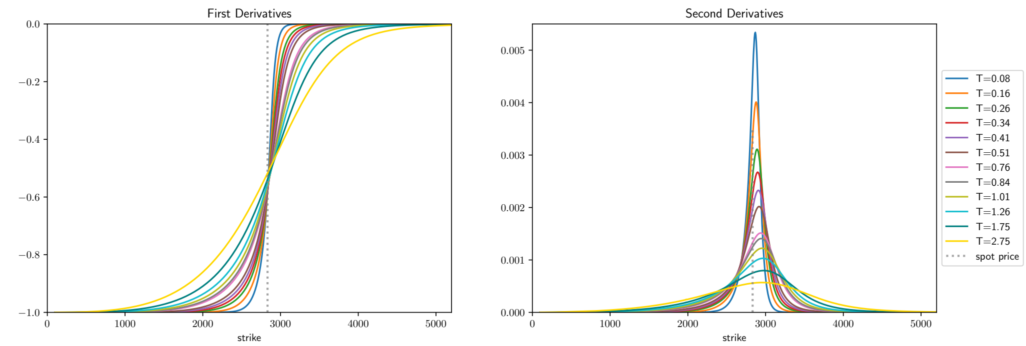
Nevertheless, this truncation of course leads to a loss of confidence in the behavior of the barriers at the boundaries regions. We will later on depict these replaced sections with transparency to highlight the uncertainty.

In Figure 8 (b), we present the first and second derivatives following the tail correction. Figure 8 (c) illustrates the associated forward prices and potentials.

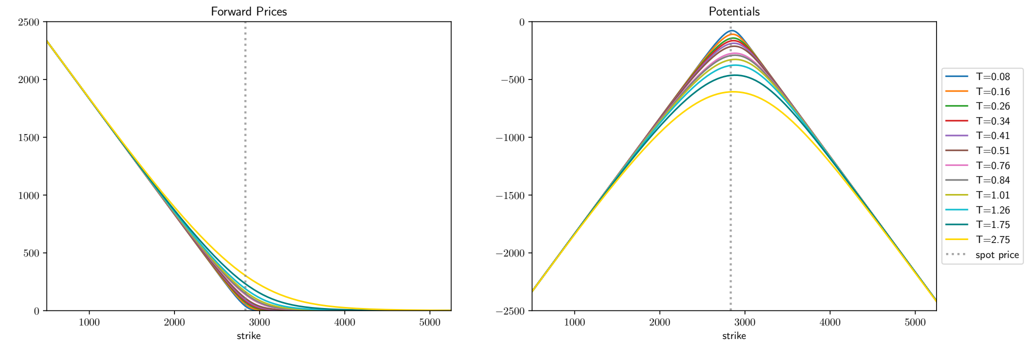
Root and Rost barriers. We give the numerical results obtained by solving both the PDE in variational form (11) resp. (13) as well as (14) resp. (15) using an explicit Euler scheme and the tail-corrected Heston data.



(A) Comparing calibrated prices to market given prices.

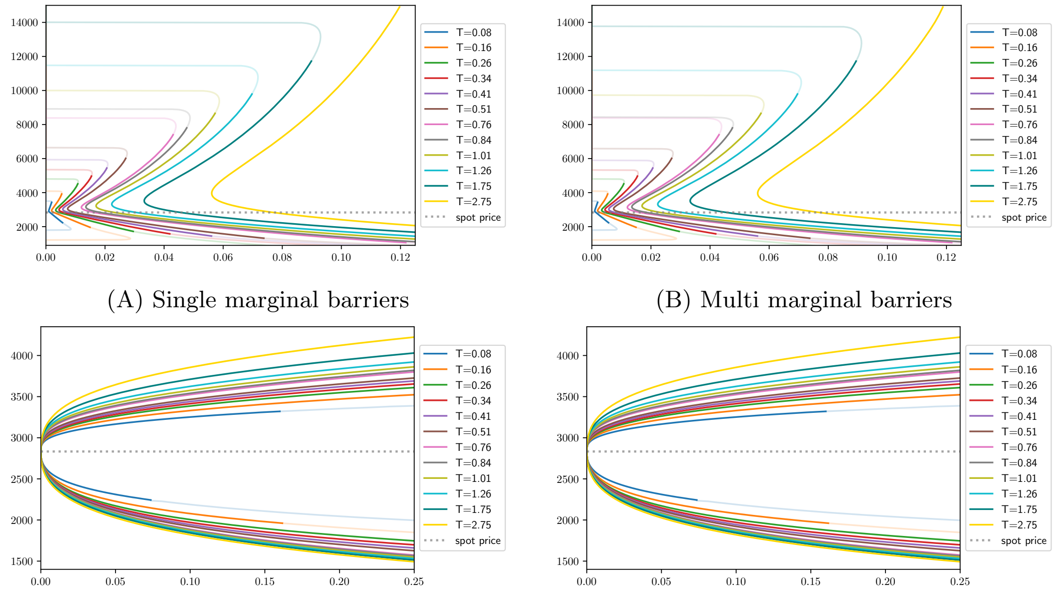


(B) First and second derivatives of calibrated prices.



(C) Forward prices and the associated potentials.

Figure 8: Heston model calibration results.



(A) Single marginal barriers (B) Multi marginal barriers

(C) Single marginal barriers (D) Multi marginal barriers

Figure 9: Root (top) and Rost (bottom) barriers embedding the market implied marginals of the S&P500.

As already pointed out in the discussion of the Black-Scholes model, the Euler scheme does not seem to return the result we would expect when the potential associated with the time T_k marginal too closely approaches the initial potential $-|x - P_0|$.

The algorithm does, however, seem to remain stable for a little while longer than in the Black-Scholes case, we depict the region where these potentials remain a minimum distance of $\epsilon = 0.0002$ away from the initial potential and no tail correction has taken place in solid lines. The remaining portions of the barrier are shown transparently.

Prices. We use the barriers obtained above to establish robust bounds for a variance call, as detailed in Section 2.3 and Equation (16). Given a maturity T and a strike price K, our objective is to price the *variance call* defined as

$$f(\langle \log M \rangle_T) = (\langle \log M \rangle_T - K)^+ .$$

We consider both an intermediate maturity, $T = 1.01$ as well as the terminal maturity $T = 2.75$. As the single-marginal and the multi-marginal barriers coincide, the resulting bounds are identical, regardless of whether we employ the single-marginal or multi-marginal model.

In Figure 10 we illustrate the option prices for these variance calls for a wide range of possible strike prices. Notably, similar to the Black-Scholes case, while the upper and lower bounds coincide for the strike price 0, the gap between the upper and lower robust pricing bounds widens notably for increasing strike prices.

Additionally, it is evident that the bounds computed here differ greatly from those computed in the Black-Scholes section, demonstrating a certain sensitivity of the robust bounds to variations in the data inputs.

3.3 Discussion

In both calibration regimes, the Root and Rost barriers adhering to the market given data which emerge from our numerical analysis appear to be in perfect consecutive order. In the case of the Root construction for the Black-Scholes model, this should not be surprising since (as noted

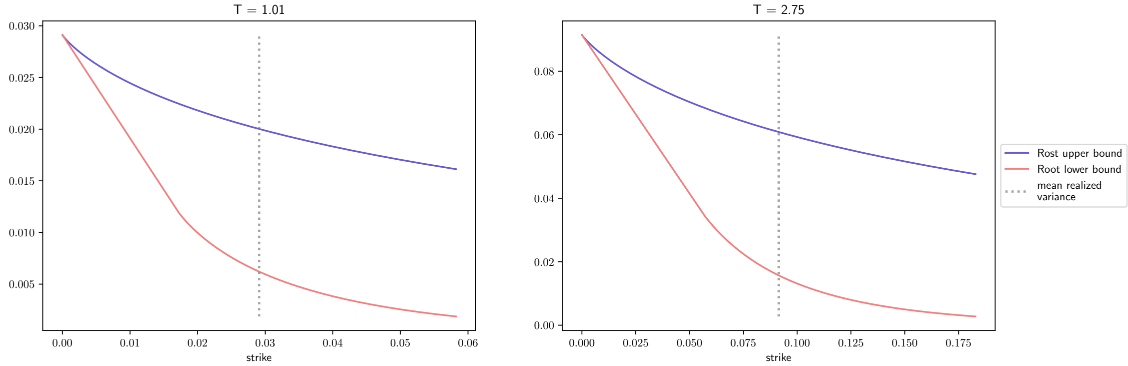


Figure 10: Upper and lower robust bounds on variance call prices for two different maturities.

above), the optimal barriers correspond to barriers which are constant in time for the time-changed model, and therefore convex ordering will imply ordering of the corresponding barriers. For the barriers based on the Heston model, it would be natural to conjecture that the barriers are ordered when the call prices come from the same parameters at different maturities (as indicated here), however it does not appear to be possible to prove this easily.

The fact that our multi-marginal barriers align precisely with their single-marginal counterparts indicates that the incorporation of earlier maturities yields no additional information for future maturities and the requirement for these barriers to be traversed chronologically imposes no additional constraints on the extremal models. As a consequence, the evidence we have collected suggests that under normal market regimes, it would be natural to assume that it would be sufficient to compute model-independent bounds based only on the terminal marginal information, and there is limited useful information contained in earlier marginals. It would be interesting to establish whether this behaviour is preserved in extreme market scenarios.

We have carried out similar analysis on prices collected during the 2020 Covid crash. Although our results were similar to those presented, we encountered the following problems during the investigation. Firstly, while the estimation of the option-implied interest rate yielded results somewhat aligned with the T-bill rates reported by the U.S. Department of the Treasury, the corresponding implied dividend yields deviated significantly and, notably, presented as negative. This suggests a potential flaw in our underlying setup for this scenario. Furthermore, the fit of the resultant calibrated models is regrettably poor, rendering them unsuitable for serious consideration. A different smoothing procedure may reveal a different answer and a more systematic review of abnormal market conditions may also reveal some periods where this behaviour breaks down.

The primary result from our study is that the robust pricing paradigm, when only considering the inclusion of marginal information about price-processes, do not substantially reduce price bounds beyond the information contained in the marginal law at the maturity time. This suggests that robust pricing is inherently limited in the extent to which the upper or lower bounds can be used for pricing, without the addition of further information which can reduce the class of models, for example through the knowledge of prices of more complex derivatives. Another alternative is to reduce the class of models under which pricing can happen. However, this approach also has potentially limited scope for reducing pricing bounds. Work of [33] showed that provided the class of models which are included in the "robust" set includes some stochastic volatility model with additional noise relative to the asset, and under mild technical conditions, then the class of possible pricing measures is essentially the set of all calibrated measures, that is, we cannot significantly reduce the class $\mathcal{M}^{P_0}$ without making strong assumptions on the dynamics of the true model.

In practice, and as is well established, this means that robust methods are of limited practical application for pricing, a fact which has been well known for many years. Rather, the main applications of such methods should come through more sophisticated approaches to penalisation.

For a completely different approach to these problems, employing modern machine learning techniques see for example [45] using neural SDE models, while e.g. in [57] a machine learning approch to the discrete time problem was proposed.

Another recent methodology, often referred to as "sensitivity analysis", consists of restricting to models within a small "distance" to some reasonable reference model, see, e.g., [3, 47, 48] for results in continuous time and [2, 4, 37, 60] for results in discrete time, as well as the references therein.

We note, however that such methods are again subject to (more subtle notions of) model-risk, and may again expose institutions adopting these metrics to unanticipated risks. More generally, it seems that model-risk is a phenomena which is unavoidable for accurate pricing and hedging.

4 Conclusion

In this paper we considered robust pricing bounds for options on variance. We considered the classical solutions to the robust pricing of variance options given vanilla options with the same maturity as the variance options, and (through the application of appropriate smoothing of market prices), we were able to identify the underlying geometric structure which identifies the optimal model.

Further, by incorporating market data on vanilla options with earlier maturity dates, we were able to show that, contrary to some expectations, and through explicit calculation of the geometric structure relating to the extremal models, that the extremal prices of variance options do not appear to change when we include this additional information. As a consequence, we conclude that there is some evidence that robust pricing bounds for variance options are not significantly changed when additional market information coming from vanilla options is incorporated. We conclude that there is no simple approach to eliminating model risk when pricing and hedging exotic options.

A Estimating implied interest rate and dividend yield.

In instances where interest rates and dividend yields used in pricing are unavailable, as is the situation with the data obtained from the CBOE Datashop, there exist several methods for approximating these quantities. A commonly used technique is utilizing the put-call parity

$$\mathcal{C}(K,T) - \mathcal{P}(K,T) = P_0 e^{-qT} - Ke^{-rT}. \tag{18}$$

Then taking the first derivative in the strike price K yields the following formula for the *implied* interest rate given (K,T)

$$\frac{\partial}{\partial K}\left(\mathcal{C}(K,T) - \mathcal{P}(K,T)\right) = -e^{-rT} \quad \Leftrightarrow \quad r = r(K,T) = -\frac{1}{T}\frac{\partial}{\partial K}\left(\mathcal{P}(K,T) - \mathcal{C}(K,T)\right).$$

Let $\mathcal{X} \subseteq [0,\infty)^2$ denote the set of market given strike-maturity pairs (K,T) and let N denote its cardinality. Then we estimate the *implied* interest rate $\hat{r}$ as

$$\hat{r} := \frac{1}{N}\sum_{(K,T)\in\mathcal{X}} r(K,T). \tag{19}$$

Given an interest rate r (or an estimate thereof) we estimate the *implied* dividend yield given (K,T) analogously

$$q(K,T) = -\frac{1}{T}\log\left(\frac{1}{P_0}\left(\mathcal{C}(K,T) - \mathcal{P}(K,T) + Ke^{-rT}\right)\right), \tag{20}$$

hence we can define $\hat{q}$ by

$$\hat{q} := \frac{1}{N}\sum_{(K,T)\in\mathcal{X}} q(K,T). \tag{21}$$

References

[1] B. Acciaio, M. Beiglböck, F. Penkner, and W. Schachermayer. A model-free version of the fundamental theorem of asset pricing and the super-replication theorem. *Math. Finance* 26.2 (2016), pp. 233–251.

[2] D. Bartl, S. Drapeau, J. Obłój, and J. Wiesel. Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proc. A.* 477.2256 (2021), Paper No. 20210176, 19.

[3] D. Bartl, A. Neufeld, and K. Park. Sensitivity of robust optimization problems under drift and volatility uncertainty. *ArXiv e-prints* (2023).

[4] D. Bartl and J. Wiesel. Sensitivity of Multiperiod Optimization Problems with Respect to the Adapted Wasserstein Distance. *SIAM J. Financ. Math.* 14.2 (2023), pp. 704–720.

[5] E. Bayraktar and T. Bernhardt. On the continuity of the Root barrier. *Proc. Amer. Math. Soc.* 150.7 (2022), pp. 3133–3145.

[6] E. Bayraktar and X. Zhang. Embedding of Walsh Brownian motion. *Stochastic Process. Appl.* 134 (2021), pp. 1–28.

[7] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.* 44.1 (2016), pp. 42–106.

[8] M. Beiglböck, A. M. G. Cox, and M. Huesmann. The geometry of multi-marginal Skorokhod Embedding. *Probab. Theory Related Fields* 176.3-4 (2020), pp. 1045–1096.

[9] M. Beiglböck, A. M. G. Cox, M. Huesmann, N. Perkowski, and D. J. Prömel. Pathwise superreplication via Vovk's outer measure. *Finance Stoch.* 21.4 (2017), pp. 1141–1166.

[10] M. Beiglböck, M. Nutz, and F. Stebegg. Fine properties of the optimal Skorokhod embedding problem. *J. Eur. Math. Soc. (JEMS)* 24.4 (2022), pp. 1389–1429.

[11] M. Beiglböck, M. Nutz, and N. Touzi. Complete duality for martingale optimal transport on the line. *Ann. Probab.* 45.5 (2017), pp. 3038–3074.

[12] A. Bensoussan and J. Lions. Applications of variational inequalities in stochastic control. eng. Studies in mathematics and its applications ; 12. Amsterdam [u.a.]: North-Holland, 1982. ISBN: 0444863583.

[13] B. Bouchard and M. Nutz. Arbitrage and duality in nondominated discrete-time models. *The Annals of Applied Probability* 25.2 (2015), pp. 823–859.

[14] D. T. Breeden and R. H. Litzenberger. Prices of State-contingent Claims Implicit in Option Prices. *The Journal of Business* 51.4 (1978), pp. 621–51.

[15] H. Brown, D. Hobson, and L. Rogers. Robust hedging of barrier options. *Math. Finance* 11.3 (2001), pp. 285–314.

[16] P. Carr and R. Lee. Hedging variance options on continuous semimartingales. *Finance Stoch.* 14.2 (2010), pp. 179–207.

[17] P. Cheridito, M. Kiiski, D. J. Prömel, and H. M. Soner. Martingale optimal transport duality. *Math. Ann.* 379.3-4 (2021), pp. 1685–1712.

[18] A. M. G. Cox and D. Hobson. A Unifying Class of Skorokhod Embeddings: Connecting the Azéma: Yor and Vallois Embeddings. *Bernoulli* 13.1 (2007), pp. 114–130.

[19] A. M. G. Cox and J. Obłój. Robust hedging of double touch barrier options. *SIAM J. Financial Math.* 2 (2011), pp. 141–182.

[20] A. M. G. Cox and J. Obłój. Robust pricing and hedging of double no-touch options. *Finance Stoch.* 15.3 (2011), pp. 573–605.

[21] A. M. G. Cox and J. Wang. Optimal robust bounds for variance options. *ArXiv e-prints* (Aug. 2013).

[22] A. Cox and J. Obłój. Robust pricing and hedging of double no-touch options. *ArXiv e-prints* (Jan. 2009).

[23] A. M. G. Cox and A. Grass. Delayed switching identities and multi-marginal solutions to the Skorokhod embedding problem. *ArXiv e-prints* (2023).

[24] A. M. G. Cox and S. M. Kinsley. Discretisation and duality of optimal Skorokhod embedding problems. *Stochastic Process. Appl.* 129.7 (2019), pp. 2376–2405.

[25] A. M. G. Cox and S. M. Kinsley. Robust hedging of options on a leveraged exchange traded fund. *Ann. Appl. Probab.* 29.1 (2019), pp. 531–576.

[26] A. M. G. Cox, J. Obłój, and N. Touzi. The Root solution to the multi-marginal embedding problem: an optimal stopping and time-reversal approach. *Probab. Theory Related Fields* 173.1-2 (2019), pp. 211–259.

[27] A. M. G. Cox and J. Wang. Root's barrier: construction, optimality and applications to variance options. *Ann. Appl. Probab.* 23.3 (2013), pp. 859–894.

[28] M. Davis and D. Hobson. The range of traded option prices. *Math. Finance* 17.1 (2007), pp. 1–14.

[29] M. Davis, J. Obłój, and V. Raval. Arbitrage Bounds for Prices of Options on Realized Variance. *ArXiv e-prints* (Jan. 2010).

[30] T. De Angelis. From optimal stopping boundaries to Rost's reversed barriers and the Skorokhod embedding. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.2 (2018), pp. 1098–1133.

[31] H. De March and P. Henry-Labordère. Building arbitrage-free implied volatility: Sinkhorn's algorithm and variants. *ArXiv e-prints* (2023).

[32] Y. Dolinsky and H. M. Soner. Martingale optimal transport in the Skorokhod space. *Stochastic Processes and their Applications* 125.10 (2015), pp. 3893–3931.

[33] Y. Dolinsky and A. Neufeld. Super-replication in fully incomplete markets. *Math. Finance* 28.2 (2018), pp. 483–515.

[34] B. Dupire. Arbitrage bounds for volatility derivatives. *Presentation at PDE and Mathematical Finance KTH, Stockholm* (2005).

[35] B. Dupire. Model art. *Risk* 6 (1993), pp. 118–124.

[36] B. Dupire. Pricing with a smile. *Risk* 7.1 (1994), pp. 18–20.

[37] S. Eckstein, G. Guo, T. Lim, and J. Obłój. Robust pricing and hedging of options on multiple assets and its numerics. *SIAM J. Financial Math.* 12.1 (2021), pp. 158–188.

[38] M. R. Fengler. Arbitrage-free smoothing of the implied volatility surface. *Quant. Finance* 9.4 (2009), pp. 417–428.

[39] A. Galichon, P. Henry-Labordère, and N. Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to Lookback options. *Ann. Appl. Probab.* 24.1 (2014), pp. 312–336.

[40] P. Gassiat, A. Mijatović, and H. Oberhauser. An integral equation for Root's barrier and the generation of Brownian increments. *Ann. Appl. Probab.* 25.4 (2015), pp. 2039–2065.

[41] P. Gassiat, H. Oberhauser, and G. dos Reis. Root's barrier, viscosity solutions of obstacle problems and reflected FBSDEs. *Stochastic Processes and their Applications* 125.12 (2015), pp. 4601–4631.

[42] P. Gassiat, H. Oberhauser, and C. Z. Zou. A free boundary characterisation of the Root barrier for Markov processes. *Probab. Theory Related Fields* 180.1-2 (2021), pp. 33–69.

[43] J. Gatheral. The volatility surface: a practitioner's guide. *WILEY* (2006).

[44] N. Ghoussoub, Y.-H. Kim, and A. Z. Palmer. PDE methods for optimal Skorokhod embeddings. *Calc. Var. Partial Differential Equations* 58.3 (2019), Paper No. 113, 31.

[45] P. Gierjatowicz, M. Sabate-Vidales, D. Šiška, Ł. Szpruch, and Ž. Žurič. Robust pricing and hedging via neural stochastic differential equations. en. *Journal of Computational Finance* (Feb. 2023).

[46] P. Henry-Labordère, J. Obłój, P. Spoida, and N. Touzi. The maximum maximum of a martingale with given n marginals. *Ann. Appl. Probab.* 26.1 (2016), pp. 1–44.

[47] S. Herrmann and J. Muhle-Karbe. Model uncertainty, recalibration, and the emergence of delta–vega hedging. *Finance and Stochastics* 21.4 (Oct. 2017), pp. 873–930.

[48] S. Herrmann, J. Muhle-Karbe, and F. T. Seifried. Hedging with small uncertainty aversion. *Finance and Stochastics* 21.1 (Jan. 2017), pp. 1–64.

[49] S. L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financ. Stud.* 6.2 (1993), pp. 327–343.

[50] D. Hobson. Robust hedging of the lookback option. *Finance and Stochastics* 2 (4 1998), pp. 329–347.

[51] D. Hobson. The Skorokhod embedding problem and model-independent bounds for option prices. *Paris-Princeton Lectures on Mathematical Finance 2010.* Vol. 2003. Lecture Notes in Math. Springer, Berlin, 2011, pp. 267–318.

[52] D. Hobson and M. Klimmek. Model independent hedging strategies for variance swaps. *ArXiv e-prints* (Apr. 2011).

[53] D. Hobson and A. Neuberger. Robust bounds for forward start options. *Math. Finance* 22.1 (2012). Ed. by I. Wiley-Blackwell Publishing, pp. 31–56.

[54] D. Hobson and J. Pedersen. The minimum maximum of a continuous martingale with given initial and terminal laws. *Ann. Probab.* 30.2 (2002), pp. 978–999.

[55] M. Beiglböck, A. Cox, and M. Huesmann. Optimal Transport and Skorokhod Embedding. *Invent. Math.* 208.2 (2017), pp. 327–400.

[56] A. Neuberger. The Log Contract. *Journal of Portfolio Managment,* 20(2) (1994), pp. 74–80.

[57] A. Neufeld and J. Sester. Robust Q-learning Algorithm for Markov Decision Processes under Wasserstein Uncertainty. *ArXiv e-prints* (2023).

[58] J. Obłój. The Skorokhod embedding problem and its offspring. *Probab. Surv.* 1 (2004), pp. 321–390.

[59] J. Obłój and P. Spoida. An iterated Azéma-Yor type embedding for finitely many marginals. *Ann. Probab.* 45.4 (2017), pp. 2210–2247.

[60] J. Obłój and J. Wiesel. Distributionally robust portfolio maximization and marginal utility pricing in one period financial markets. *Math. Finance* 31.4 (2021), pp. 1454–1493.

[61] H. Pham. Continuous-time stochastic control and optimization with financial applications. Vol. 61. Springer Science & Business Media, 2009.

[62] D. H. Root. The existence of certain stopping times on Brownian motion. *Ann. Math. Statist.* 40 (1969), pp. 715–718.

[63] H. Rost. Skorokhod stopping times of minimal variance. *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975).* Berlin: Springer, 1976, 194–208. Lecture Notes in Math., Vol. 511.

[64] H. Rost. Skorokhod's theorem for general Markov processes. *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Tech. Univ. Prague, Prague, 1971; dedicated to the memory of Antonın Špaček).* Academia [Publishing House of the Czechoslovak Academy of Sciences], Prague, 1973, pp. 755–764.

[65] A. V. Skorohod. Issledovaniya po teorii sluchainykh protsessov (Stokhasticheskie differentsialnye uravneniya i predelnye teoremy dlya protsessov Markova). Izdat. Kiev. Univ., Kiev, 1961, p. 216.

[66] A. V. Skorokhod. Studies in the theory of random processes. Translated from the Russian by Scripta Technica, Inc. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1965, pp. viii+199.

Perkins Embedding for General Starting Laws

A. M. Grass

Abstract

The *Skorokhod embedding problem* (SEP) is to represent a given probability measure as a Brownian motion B at a particular stopping time. In recent years particular attention has gone to solutions which exhibit additional optimality properties due to applications to martingale inequalities and robust pricing in mathematical finance.

Among these solutions, the *Perkins embedding* sticks out through its distinct geometric properties. Moreover, it is the only optimal solution to the SEP which so far has been limited to the case of Brownian motion started in a dirac distribution.

In this paper we provide, for the first time, an optimal solution to the Skorokhod embedding problem for the general SEP which leads to the Perkins solution when applied to Brownian motion started in a Dirac. This solution to the SEP also suggests a new geometric interpretation of the Perkins solution which better clarifies the relation to other optimal solutions of the SEP.

Keywords: Skorokhod embedding problem; Perkins embedding; barrier type solutions; inverse barriers.

MSC 2020 Primary 60G40, 60G44; Secondary 91G20.

1 Introduction

Let μ be a probability measure on the real line with finite second moment and write $(B_t)_{t\geq 0}$ for a Brownian motion started according to a probability measure λ, i.e. $B_0 \sim \lambda$. The Skorokhod embedding problem ([33, 34]) is to find a stopping time τ such that the measure μ can be represented in the sense that

$$B_\tau \sim \mu, \quad \text{and } \mathbb{E}[\tau] < \infty. \tag{SEP$_{\lambda,\mu}$}$$

Skorokhod [33, 34] gave a solution in the case where Brownian motion is started in an atom. A necessary and sufficient condition for the SEP$_{\lambda,\mu}$ to admit a solution is for λ to be prior to μ in convex order, i.e. $\int f(x)\lambda(dx) \leq \int f(x)\mu(dx)$ for any convex function $f : \mathbb{R} \to \mathbb{R}$. This follows by combining Skorokhod's results with Strassen's theorem [35] on the existence of martingales with given marginals. We will tacitly assume this condition throughout the paper. The condition $\mathbb{E}[\tau] < \infty$ is imposed to exclude trivial solutions.

Skorokhod's work initiated an active field of research. Obłój's survey [27] provides a comprehensive overview of the developments up to 2004 and describes more than 20 different solutions to the Skorokhod problem given by different authors.

During the 2000's a particular stimulus for the field has come from the connection with robust finance which was discovered in Hobson's seminal paper [19], see also [21]. In view of these applications, it is of particular importance to construct stopping times which optimize certain functionals subject to satisfying the embedding constraint. This question is known as the *optimal Skorokhod embedding problem* and has been considered extensively, see [2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 26, 28] among others.

Perkins embedding

In this article we focus on a particular solution to (SEP$_{\lambda,\mu}$) established by Perkins [29] for the *deterministic start case* where $B_0 = 0$, i.e. $\lambda = \delta_0$. The Perkins embedding solves an optimal Skorokhod embedding problem: It has the characteristic optimality property of minimizing the law of the running maximum $\overline{B}_\tau := \max_{t\leq\tau} B_t$ of the underlying Brownian motion while simultaneously maximizing the law of the running minimum $\underline{B}_\tau := \min_{t\leq\tau} B_t$ among all solution to (SEP$_{\lambda,\mu}$). Here (and below) maximization / minimization of laws is understood with respect to first order stochastic dominance.

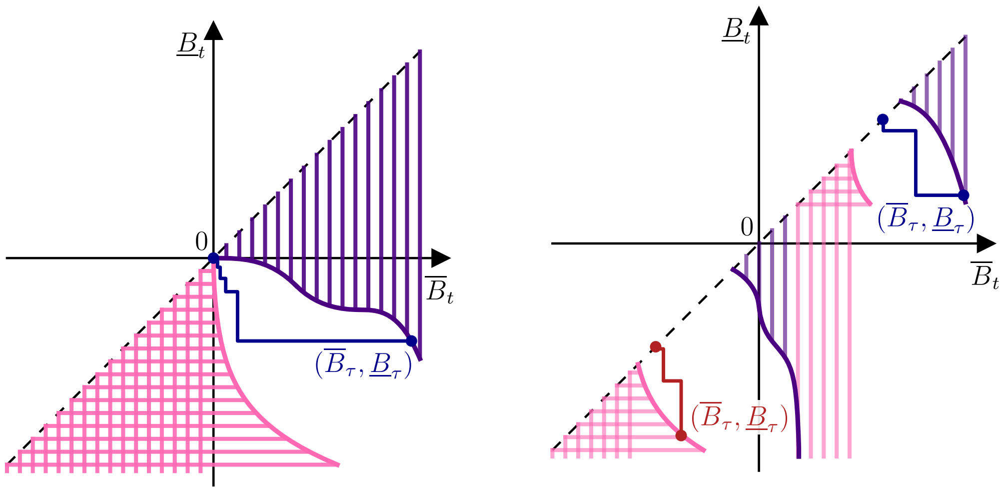
The Perkins embedding is of importance for robust pricing of barrier and lookback options, see [6, 19, 21]. While other solutions to the optimal Skorokhod embedding problem have been given in the case of a general starting distribution $B_0 \sim \lambda$ or admit direct extensions to this case, there is no known solution in the general starting case which extends the Perkins embedding.

In [22] Hobson and Pedersen propose a solution to (SEP$_{\lambda,\mu}$) allowing for a general starting law which minimizes the law of the running maximum. Remarkably, the Hobson-Pederson embedding bears some resemblance to the Azéma-Yor [1] embedding. As pointed out in [22], the Hobson-Pederson embedding does not maximize the running minimum in the $B_0 = 0$ case and specifically differs from the Perkins embedding also in this case. A further difference between the Hobson-Pederson embedding and the Perkins embedding, is that the latter does not require external randomization (unless $\mu(0) > 0$, see Section 3.2).

Of specific interest is the geometric structure of Perkins solution. In the $\lambda = \delta_0$ case it can be identified as a hitting time of the process $(\overline{B}, \underline{B})$ of a specifically structured target set $R \subseteq \mathbb{R}^2$, see Figure 1 (A). The set R is a union of 'lines' of two types:

(v) *vertical lines* which start below the diagonal and terminate in the diagonal. These lines will be depicted in '*v*iolett'.

(h) *lines* which start on the right of the diagonal. They move horizontally until being reflected by the diagonal and continue to move vertically towards $-\infty$. These lines will be depicted in '*h*ot pink'.



(A) Perkins solution in the $\lambda = \delta_0$ case.

(B) Perkins solution with general starting.

Figure 1: Examples of Perkins solutions to (SEP$_{\lambda,\mu}$).

We will say that R has *vh-barrier structure*. Note that traditionally the Perkins picture is drawn without the horizontal lines being reflected downwards. These downward lines are irrelevant in the deterministic starting case however become crucial when allowing for general starting as we will see below.

Our main contribution is to extend the Perkins solution to the case of general starting law. Specifically, setting $(\lambda \wedge \mu)(A) := \inf_{B \subseteq A,\, B \text{ measurable}} \big(\lambda(B) + \mu(A \setminus B)\big)$ we have:

Theorem 1. There exists a vh-barrier $R \subseteq \mathbb{R}^2$ such that the stopping time τ_P defined by $P[\tau_{\mathrm{P}} = 0, B_0 \in A] = (\lambda \wedge \mu)(A)$ for Borel $A \subseteq \mathbb{R}$ and on $\{\tau_P > 0\}$ by

$$\tau_P = \inf \big\{ t \geq 0 : \big(\overline{B}_t, \underline{B}_t\big) \in R \big\} \tag{1.1}$$

is a solution to (SEP$_{\lambda,\mu}$). Moreover τ_P minimizes the law of the running maximum $\overline{B}$ and further maximizes the law of the running minimum $\underline{B}$ among all these solutions.

If λ and μ are mutually singular measures we have $\tau_P > 0$ almost surely and τ_P is adapted to the filtration generated by the underlying Brownian motion.

The solution τ_P will be unique in the sense that if $\tilde{\tau}$ is another stopping time of the form (1.1) then we have $\tau_P = \tilde{\tau}$ almost surely.

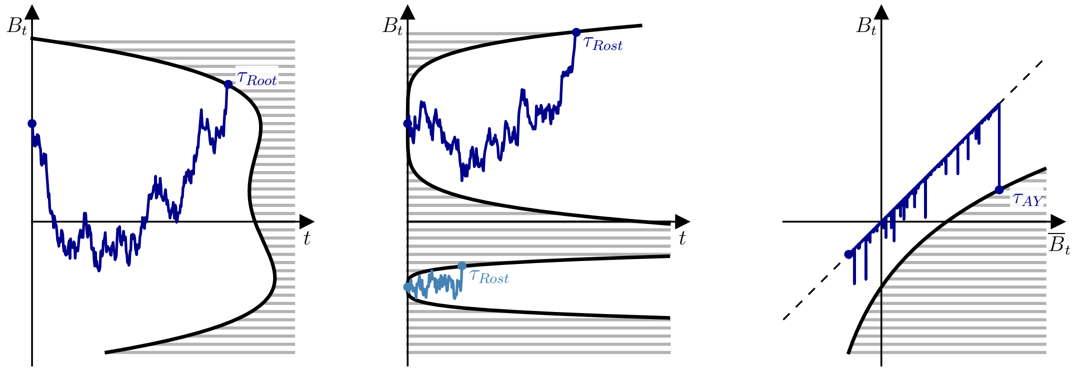
A possible realization of a general starting Perkins solution can be seen in Figure 1 (B). We point out two important properties of this solution. Firstly external randomization is only needed at time 0 and only if the two measure λ and μ share mass. This answers a question raised in [22, Remark 2.3] concerning the existence of adapted solutions in this setting. Secondly this solutions features the same representation as the hitting time of the process $(\overline{B}, \underline{B})$ as the original Perkins solution and recovers it in the $\lambda = \delta_0$ case as illustrated in Figure 1.

We will focus on discussing the Perkins solution from a *barrier type solution* viewpoint.

Barrier type solutions

A *barrier type solution* to (SEP$_{\lambda,\mu}$) is traditionally of the form

$$\tau_R = \inf \{ t \geq 0 : (A_t, B_t) \in R \} \tag{1.2}$$



(A) Root solution. (B) Rost solution. (C) Azéma-Yor solution.

Figure 2: Barrier type solutions to (SEP$_{\lambda,\mu}$).

for a sufficiently regular processes A_t and a Borel set $R \subseteq \mathbb{R}^2$ featuring some additional *barrier* structure.

Barrier type solutions not only come with natural geometric interpretations, we also see that all these solution are adapted to the filtration of the underlying Brownian motion (apart from potential additional randomization at time 0 if λ and μ share mass). Furthermore, barrier type solutions feature the following intrinsic uniqueness property highlighted by Loynes [25]: if $S \subseteq \mathbb{R}^2$ is another barrier such that τ_S solves (SEP$_{\lambda,\mu}$) then we have $\tau_R = \tau_S$ almost surely. Prototypical barrier type solutions and also the solutions to originally coin the notion of a barrier resp. inverse barrier are the solutions by Root [30] and Rost [31, 32]. A (Root) *barrier* is a set $R \subseteq \mathbb{R}_+ \times \mathbb{R}$ such that $(s,x) \in R$ implies $(t,x) \in R$ for all $t > s$ while for an *inverse barrier* $S \subseteq \mathbb{R}_+ \times \mathbb{R}$ we require that $(s,x) \in S$ implies $(t,x) \in S$ for all $t < s$. The Root resp. Rost solution is then given as the hitting time

$$\tau_{\text{Root}} = \inf\{t \geq 0 : (t, B_t) \in R\} \quad \text{resp.} \quad \tau_{\text{Rost}} = \inf\{t \geq 0 : (t, B_t) \in S\}$$

and is known for *minimizing* resp. *maximizing* $\mathbb{E}[\tau^2]$ among all solutions to (SEP$_{\lambda,\mu}$) (see [24] and [31]). To be precise, in the Rost case, if the initial and the terminal law share mass, this shared mass might need to be stopped immediately, i.e. $P[\tau_{\text{Rost}} = 0, B_0 \in A] = (\lambda \wedge \mu)(A)$.

Another barrier type solution to (SEP$_{\lambda,\mu}$) worth mentioning here is the Azéma-Yor solution [1] which is known for maximizing the law of the running maximum of the underlying Brownian motion. This solution is also given as the hitting time of a barrier in the Root sense, however of the joint process $(\max_{s\leq t} B_s, B_t)$, thus

$$\tau_{AY} = \inf\left\{t \geq 0 : \left(\max_{s\leq t} B_s, B_t\right) \in R\right\}.$$

A general starting law version was found by Hobson in [20] and - as in the Root case - the barrier type representation of the solution did not change. See Figure 2 for examples of Root, Rost and Azéma-Yor barrier type solution. As both the Perkins solution and the Hobson-Pedersen solution *minimize* the law of the running maximum of the underlying Brownian motion, they can be seen as a pendant to the Azéma-Yor embedding in a similar way as the Rost embedding can be seen as a pendant to the Root embedding.

Further barrier type solutions to (SEP$_{\lambda,\mu}$) are e.g. the Vallois embedding [36], the Jacka embedding [23] or the cave embedding [26]. We also refer to [26] for a unifying framework for barrier type solutions.

Notably it is not possible to find a process (A_t) such that the Perkins solution can be represented as a barrier type solution in the classical sense (1.2), that is as a hitting time of the joint process

(A_t, B_t). However, it is possible to identify it as a hitting time of the process $(\overline{B}, \underline{B})$ of a vh-barrier as described above.

In [26] the authors use ideas and concepts of optimal mass transport to give existence results for solutions to $(\mathrm{SEP}_{\lambda,\mu})$ featuring additional extremal properties and moreover provide techniques on how connections between these additional properties of the solutions and their geometric characterizations as barrier type solutions can be made. Minor variations of these techniques will allow us prove Theorem 1 and moreover to identify the right kind of phase space to interpret the Perkins embedding as the hitting time of a Rost inverse barrier, completing the Azéma-Yor↔Perkins - Root↔Rost analogy described earlier. Furthermore this representation will enable us to prove a Loynes-type uniqueness result for this solution.

Outline of the article

In Section 2 we recall notation and results of [26] and extend them to a multifold optimization problem over solutions to $(\mathrm{SEP}_{\lambda,\mu})$. These results will be applied in Section 3 to obtain existence of a Perkins embedding with general starting law as a solution to $(\mathrm{SEP}_{\lambda,\mu})$ with the desired additional extremal properties. In Section 4 we discuss uniqueness of barrier type solutions and propose a setting in which these results can be applied to the Perkins embedding. This will conclude the proof of Theorem 1.

2 Multifold optimal Skorokhod embedding

For this article we will adopt the underlying assumptions of [26], that is, we will work on a stochastic basis $\Omega = (\Omega, \mathcal{G}, (\mathcal{G}_t)_{t\geq 0}, P)$ which is rich enough to support a Brownian motion B and a uniformly distributed $\mathcal{G}_0$-random variable independent of B.

2.1 Optimal Skorokhod embedding

The theory of finding solutions to the Skorokhod embedding problem featuring some optimality properties is presented from a systematic viewpoint in [26], see also [17]. Historically it was also common to find a new solution to $(\mathrm{SEP}_{\lambda,\mu})$ and to only later establish its additional optimality property. In [26] this was turned around by imposing an optimization problem to the Skorokhod embedding problem and then to derive structural properties of the embedding problem from a 'first order condition'.

For this we consider the set of *stopped paths*.

Definition 2.1 (stopped paths). We define the set of *stopped paths*

$$S := \{(f, s) : s \geq 0, f \in C([0, s], \mathbb{R})\}.$$

The optimal Skorokhod embedding is formulated in the following way:

Definition 2.2 (optimal Skorokhod embedding problem). For a function

$$\gamma : S \to \mathbb{R}$$

the *optimal Skorokhod embedding problem* is to find a stopping time τ on Ω solving the following optimization problem

$$\inf \{\mathbb{E}[\gamma((B_t)_{t\leq\tau}, \tau)] : \tau \text{ solves } (\mathrm{SEP}_{\lambda,\mu})\}. \tag{$\mathrm{OptSEP}_{\lambda,\mu}$}$$

We will always assume $(\mathrm{OptSEP}_{\lambda,\mu})$ to be *well posed* in the sense that for all stopping times solving $(\mathrm{SEP}_{\lambda,\mu})$, we have that $\mathbb{E}[\gamma((B_t)_{t\leq\tau}, \tau)]$ exists, that $\mathbb{E}[\gamma((B_t)_{t\leq\tau}, \tau)] \in (-\infty, \infty]$ and that there is at least one stopping time τ such that $\mathbb{E}[\gamma((B_t)_{t\leq\tau}, \tau)] < \infty$.

The solutions presented in the introduction can be recovered as solutions to the optimal Skorokhod embedding problem in the following way. The Root solution can be obtained by choosing $\gamma((f,s)) := s^2$, the Rost solution by $\gamma((f,s)) := -s^2$ and the Azéma-Yor solution by $\gamma((f,s)) := -\overline{f} = \max_{t\leq s} f(t)$. Many more classic solutions to the Skorokhod embedding problem can be represented this way, see [26].

As pointed out in the introduction, the Perkins solution stands out due to its two fold optimality property. The optimal Skorokhod embedding problem as described above will not suffice to describe this special embedding. Also, in the construction of optimal solutions it is often useful to impose additional auxiliary optimality conditions in order to guarantee uniqueness of solutions. Thus we consider the following multifold optimal Skorokhod embedding problem.

Definition 2.3 (multifold optimal Skorokhod embedding problem). Let $n \in \mathbb{N}$ and consider a function

$$\gamma = (\gamma_1, \dots, \gamma_n) : S \to \mathbb{R}^n.$$

Let $\mathrm{Opt}_\gamma^{(1)}$ be the set of all $\mathcal{G}$-stopping times τ on Ω solving the following optimization problem

$$\inf \{\mathbb{E}[\gamma_1((B_t)_{t\leq\tau}, \tau)] : \tau \text{ solves } (\mathrm{SEP}_{\lambda,\mu})\}. \tag{$\mathrm{OptSEP}^{(1)}_{\lambda,\mu}$}$$

For $j \in \{2,\dots,n\}$ define $\mathrm{Opt}_\gamma^{(j)}$ as the set of all stopping times $\tau \in \mathrm{Opt}_\gamma^{(j-1)}$ that solve the optimization problem

$$\inf \left\{\mathbb{E}[\gamma_j((B_t)_{t\leq\tau}, \tau)] : \tau \in \mathrm{Opt}_\gamma^{(j-1)}\right\}. \tag{$\mathrm{OptSEP}^{(j)}_{\lambda,\mu}$}$$

We call $(\mathrm{OptSEP}_{\lambda,\mu}) := (\mathrm{OptSEP}^{(n)}_{\lambda,\mu})$ the *multifold optimal Skorokhod embedding problem.*

The optimization problem $(\mathrm{OptSEP}^{(1)}_{\lambda,\mu})$ is *well posed* if for all stopping times solving $(\mathrm{SEP}_{\lambda,\mu})$, we have that $\mathbb{E}[\gamma_1((B_t)_{t\leq\tau}, \tau)]$ exists, that $\mathbb{E}[\gamma_1((B_t)_{t\leq\tau}, \tau)] \in (-\infty, \infty]$ and that there is at least one stopping time τ so that $\mathbb{E}[\gamma_1((B_t)_{t\leq\tau}, \tau)] < \infty$. We call $(\mathrm{OptSEP}_{\lambda,\mu})$ *well posed*, if $(\mathrm{OptSEP}^{(1)}_{\lambda,\mu})$ is well posed and for all $j \in \{2,\dots,n\}$ the problem $(\mathrm{OptSEP}^{(j)}_{\lambda,\mu})$ is well posed in the above sense (considering stopping times in $\mathrm{Opt}_\gamma^{(j-1)}$).

2.2 Randomized stopping times

The requirement of solutions to $(\mathrm{SEP}_{\lambda,\mu})$ to be stopping times with respect to the filtration generated by Brownian motion is often too restrictive (as seen for the Hobson-Pedersen solution in the next section).

The key idea of linking the Skorokhod embedding problem to optimal transport is to think of a stopping time τ as a transport plan, mapping the mass of a trajectory $(B_t(\tilde\omega))_{t\geq 0}$ in the space $(C(\mathbb{R}_+), \mathbb{W}_\lambda)$ (where $\mathbb{W}_\lambda$ denotes the law of a Brownian motion started according to the distribution λ) to the endpoint $B_{\tau(\tilde\omega)}(\tilde\omega)$ in $\mathbb{R}$. Even tough optimal transport theory cannot be applied directly, the appropriate analogues for this setup are developed in [26].

The necessary relaxation of this problem is to consider so called *randomized stopping times* which can be seen as stopping times in the usual sense but on a probability space that is possibly enlarged.

The idea of randomized stopping times is made precise by defining them as a subset of subprobability measures on $C(\mathbb{R}_+) \times \mathbb{R}_+$, for details we refer to [26, Chapter 3.2].

Definition 2.4 (randomized stopping times). A subprobability measure ξ on $C(\mathbb{R}_+) \times \mathbb{R}_+$ is called a *randomized stopping time* (of Brownian motion with initial distribution λ) if

(i) $\mathrm{proj}_{C(\mathbb{R}_+)} \leq \mathbb{W}_\lambda$.

(ii) Given the disintegration $(\xi_\omega)_{\omega\in C(\mathbb{R}_+)}$ of ξ w.r.t. the first coordinate $\omega \in C(\mathbb{R}_+)$, $\tilde A_t(\omega) = \xi_\omega([0,t])$ is $\mathcal{F}^a_t$-measurable for all $t \in \mathbb{R}_+$ (Here $(\mathcal{F}^a)$ denotes the natural augmented filtration on $C(\mathbb{R}_+)$).

For the probability measure μ on $\mathbb{R}$ we define the subset $\mathrm{RST}_\lambda(\mu)$ which consists of those $\xi \in \mathrm{RST}_\lambda$ such that

(i) $\mathrm{proj}_{C(\mathbb{R}_+)} = \mathbb{W}_\lambda$,

(ii) For all $A \in \mathcal{G}$ we have $\xi(\{(\omega, t) \in C(\mathbb{R}_+) \times \mathbb{R}_+ : \omega(t) \in A\}) = \mu(A)$.

2.3 Existence of an optimizer

One important result of [26] is to establish existence of optimizers under fairly general conditions. Even though the main proofs were carried out for $n = 1$ (Theorem 4.1) and generalizations are provided for $n = 2$ (Theorem 6.1), we can generalize in precisely the same way to arbitrary $n \in \mathbb{N}$. Let us formulate the optimal Skorokhod embedding problem for randomized stopping times and then give the existence results for this generalized problem.

Definition 2.5 ((OptSEP$\star_{\lambda,\gamma}$)). Let $n \in \mathbb{N}$ and consider a function

$$\gamma = (\gamma_1, \ldots, \gamma_n) : S \to \mathbb{R}^n.$$

Let $\mathrm{Opt}\star_\gamma^{(1)}$ be the set of all randomized stopping times solving the following optimization problem

$$\inf \left\{ \int_{C(\mathbb{R}_+)\times\mathbb{R}_+} \gamma_1((\omega|_{[0,t]}, t)) \mathrm{d}\xi((\omega, t)) : \xi \in \mathrm{RST}_\lambda(\mu) \right\}. \tag{OptSEP$\star_{\lambda,\mu}^{(1)}$}$$

For $j \in \{2, \ldots, n\}$ define $\mathrm{Opt}\star_\gamma^{(j)}$ as the set of all stopping times $\tau \in \mathrm{Opt}\star_\gamma^{(j-1)}$ wich solve the optimization problem

$$\inf \left\{ \int_{C(\mathbb{R}_+)\times\mathbb{R}_+} \gamma_j((\omega|_{[0,t]}, t)) \mathrm{d}\xi((\omega, t)) : \xi \in \mathrm{Opt}\star_\gamma^{(j-1)} \right\}. \tag{OptSEP$\star_{\lambda,\mu}^{(j)}$}$$

We define (OptSEP$\star_{\lambda,\mu}$):= (OptSEP$\star_{\lambda,\mu}^{(n)}$).

Theorem 2 (Existence of a minimizer, cf. [26], Theorem 4.1). Let $\gamma : S \to \mathbb{R}^n$ be lsc and bounded from below in the following sense:

For all $j \in \{1, \ldots, n\}$ there exist constants $a_j, b_j, c_j \in \mathbb{R}_+$ such that

$$-\left(a_j + b_j t + c_j \max_{s\le t} B_s^2\right) \le \gamma_j((B_s)_{s\le t}, t)). \tag{2.1}$$

holds on $C(\mathbb{R}_+) \times \mathbb{R}_+$.
Then (OptSEP$\star_{\lambda,\mu}$) admits a minmizer $\xi \in \mathrm{RST}_\lambda(\mu)$.

The following lemma provides equivalence between optimization over stopping times on our enlarged probability space and optimization over randomized stopping times.

Lemma 2.6 (cf. [26], Lemma 3.11). Let τ be a $(G_t)_{t\ge0}$-stopping time and consider the map

$$\Phi : \Omega \to C(\mathbb{R}_+) \times \mathbb{R}_+, \quad \overline{\omega} \mapsto ((B_t(\overline{\omega}))_{t\ge0}, \tau(\overline{\omega})).$$

Then $\xi := \Phi_\# P \in \mathrm{RST}_\lambda$ and for any measurable map $\gamma : S \to \mathbb{R}$ we have

$$\int \gamma((f, s)) dr_\# \xi((f, s)) = \mathbb{E}_P \left[\gamma((B_t)_{t\le\tau}, \tau)\right] \tag{$\star$}$$

for the function

$$r : C(\mathbb{R}_+) \times \mathbb{R}_+ \to S, \quad (\omega, t) \mapsto (\omega|_{[0,t]}, t).$$

For any randomized stopping time $\xi \in \mathrm{RST}_\lambda$ we can find a $(G_t)_{t\ge0}$-stopping time τ such that $\xi := \Phi_\# P$ and ($\star$) holds.

2.4 Monotonicity principle

In optimal transport theory it was an important development to be able to identify optimal transport plans by the geometry of its support - see *c-cyclical monotonicity*. A monotonicity principle ('first order condition') similar to c-cyclical monotonicity for optimal Skorokhod embeddings is given in [26]. We define the *support* of a stopping time τ as a subset $\Gamma \subseteq S$ such that

$$P\left[((B_t)_{t\leq\tau}, \tau) \in \Gamma\right] = 1.$$

Specific properties of this set Γ will be crucial to our geometric approach to the Skorokhod embedding problem.

We will consider possible continuations of a given path and therefore define an operation of concatenation.

Definition 2.7 (concatenation of paths). For two paths $(f,s),(g,t) \in S$ we define an operation of *concatenation* $\oplus$ by

$$(f \oplus g)(r) := \begin{cases} f(r) & r \in [0,s] \\ f(s) - g(0) + g(r-s) & r \in (s, s+t]. \end{cases}$$

Definition 2.8 (going paths). For any set of stopped paths $\Gamma \subseteq S$ we define the set of initial segments of this paths as

$$\Gamma^< := \left\{ (f,s) \in S : \exists (\tilde{f}, \tilde{s}) \in \Gamma \text{ such that } s < \tilde{s} \text{ and } f|_{[0,s]} = \tilde{f}|_{[0,s]} \right\}$$

and call it the set of *going paths*.

In the multifold optimal Skorokhod embedding problem we consecutively optimize over functions of stopped paths. We will soon learn that it is interesting and useful to derive structural arguments about the sets of stopped paths satisfying these optimality conditions. In order to do this, we would like to have a strategy for dealing with different paths stopping at the same value. Among those paths we would like to identify those paths that should be stopped and those paths that should be allowed to continue - keeping in mind our optimization problem. This leads us to the notion of *stop-go* pairs, which by considering possible continuations of the paths gives a rule on how to decide on which one to stop and which one to allow to continue.

Definition 2.9 (Stop-go pair). A pair of paths $((f,s),(g,t)) \in S \times S$ is called a *stop-go pair* with respect to γ (short: SG-pair) if

(i) $f(s) = g(t)$

(ii) $\mathbb{E}\left[\gamma(f \oplus (B_u)_{u\leq\sigma}, s+\sigma)\right] + \gamma(g,t) > \gamma(f,s) + \mathbb{E}\left[\gamma(g \oplus (B_u)_{u\leq\sigma}, t+\sigma)\right]$
in the lexicographic ordering of $\mathbb{R}^n$ for every $(\mathcal{F}_t^B)_{t\geq 0}$-stopping time σ such that $\mathbb{E}[\sigma] \in (0,\infty)$, both sides are well defined and the left-hand side is finite in every component.

For the set of all SG-pairs we will write

$$\mathsf{SG}_\gamma := \left\{ ((f,s),(g,t)) \in S \times S : ((f,s),(g,t)) \text{ is a stop-go pair with respect to } \gamma \right\},$$

and for $j \in \{1,\dots,n\}$ we will call

$$\mathbb{E}\left[\gamma_j(f \oplus (B_u)_{u\leq\sigma}, s+\sigma)\right] + \gamma_j(g,t) \geq \gamma_j(f,s) + \mathbb{E}\left[\gamma_j(g \oplus (B_u)_{u\leq\sigma}, t+\sigma)\right]$$

the *j-th stop-go condition* (SGC_j).

We now want to identify sets of stopped paths, such that it is not advantageous to stop any of these paths earlier in comparison to the other paths in this set. In other words, the stopping rule cannot be improved within this set.

Definition 2.10 (γ-monotonicity). A set of stopped paths $\Gamma \subseteq S$ is called *γ-monotone* if

$$\mathsf{SG}_\gamma \cap (\Gamma^< \times \Gamma) = \emptyset$$

Theorem 3 (Monotonicity principle). Let $\gamma : S \to \mathbb{R}^n$ be Borel measurable and let τ be a minimizer of (OptSEP$_{\lambda,\mu}$). Then there exists a γ-monotone Borel set $\Gamma \subseteq S$ such that

$$P\left[((B_t)_{t\leq\tau}, \tau) \in \Gamma\right] = 1.$$

Similar to the existence result also the monotonicity principle can be formulated for randomized stopping times. As these technicalities are not important for the rest of this paper we will refer the reader to [26] for further details.

We will see that the monotonicity principle is the key to the geometric approach to optimal Skorokhod embedding problems.

3 The Perkins embedding

3.1 Perkins embedding with deterministic starting

We will first give a precise formulation of the original Perkins solution in a barrier type formulation. Finding a geometric interpretation of this solution in the case of $\lambda = \delta_0$ is feasible with the methods established in [26], see Theorem 6.8 therein.
We will give the following slight reformulation of this theorem in order to stress our interest in the specific structure of the target set. In addition to the barriers defined in the introduction we will furthermore consider an *upwards barrier*, that is a set $R \subseteq \mathbb{R}^2$ such that $(s, x) \in R$ implies $(s, y) \in R$ for all $y > x$.

Theorem 4 (The Perkins embedding, cf. [29]). Let $\lambda = \delta_0$ and assume $\mu(\{0\}) = 0$. Let $\varphi : \mathbb{R}^2_+ \to \mathbb{R}$ be a bounded continuous function which is strictly increasing in both arguments. Then there exists a stopping time τ_{P_0} which minimizes

$$\mathbb{E}\left[\varphi\left(\overline{B}_\tau, -\underline{B}_\tau\right)\right]$$

over all solution to (SEP$_{\lambda,\mu}$) and which is of the form

$$\tau_{P_0} = \inf\left\{t \geq 0 : (\overline{B}_t, \underline{B}_t) \in R\right\}.$$

Here $R \subseteq \mathbb{R} \times \mathbb{R}_-$ will be a vh-barrier which can be represented as $R = R_1 \cup R_2$ with R_1 being an upwards barrier induced by (v)-lines and R_2 being an inverse barrier induced by (h)-lines. Moreover the boundaries of R_1 and R_2 are both given by decreasing functions $\mathbb{R}_- \to \mathbb{R}$ (see Figure 1 A).

The solution τ_{P_0} will in fact coincide almost surely for each choice of the auxiliary function φ thus giving first order stochastic dominance.

Unfortunately, some of the arguments in the proof of this theorem do not extend to the random starting case. Foremost it is no longer possible to optimize over the running minimum and the running maximum *simultaneously*. The stopping rule of Perkins' problem will only stop paths when they reach a new running extremum. This justifies the representation of τ_{P_0} as a hitting time of the process $(\overline{B}, \underline{B})$. Note that since we do not allow μ to hold any mass in 0, we have $\tau_{P_0} > 0$ and by properties of Brownian motion then $\overline{B}_{\tau_{P_0}} > \underline{B}_{\tau_{P_0}}$ a.s. These two facts imply that whenever we consider two paths stopped by this stopping rule at the same terminal value, either their running minima or their running maxima coincide. However, now it becomes very easy to decide on which one of those two paths to stop if we look at the other running extremum.

If we now allow for random starting we can also encounter (among other problems) the following situation: Two paths $(f, s), (g, t) \in S$ stop at the same value, that is $f(s) = g(t)$, however - lets

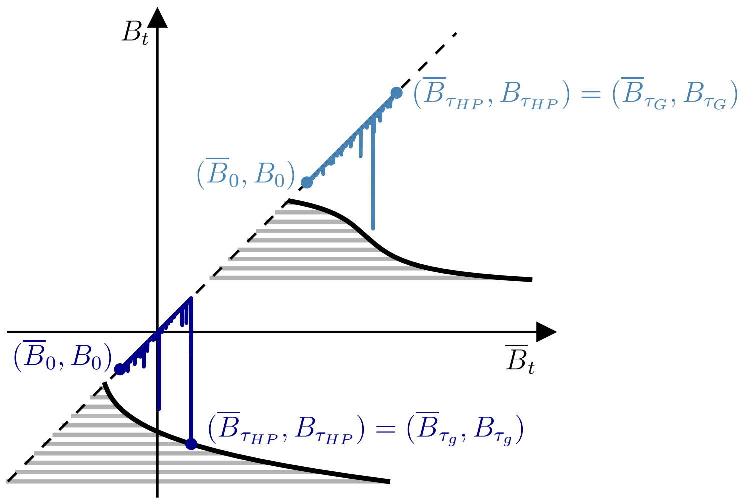


Figure 3: The Hobson-Pedersen solution to ($\mathrm{SEP}_{\lambda,\mu}$) with non-deterministic starting law λ. We see the lower path being stopped by τ_g due to hitting the barrier and the upper path being stopped by τ_G.

say f - does so by reaching a new running minimum and g by reaching a new running maximum. Then $\overline{f} > \overline{g}$ and $\underline{f} > \underline{g}$. It is therefore no longer obvious, which of those two paths should be stopped and we can see that the solution can no longer be identified as the hitting time of a set serving both optimization problems simultaneously.

3.2 Perkins embedding with random starting

The Hobson-Pedersen solution

A first take on Perkins' embedding with random starting is due to Hobson and Pedersen in [22]. We will give a brief sketch of their solution.

The authors define a function $g : \mathbb{R} \to \mathbb{R}$ and a random variable G which is independent of the randomly started Brownian motion (B_t). Both are explicitly determined by the measures λ and μ. They further define the two stopping times

$$\tau_G = \inf\left\{t > 0 : \overline{B}_t \geq G\right\} \text{ and } \tau_g = \inf\left\{t > 0 : B_t \leq g(\overline{B}_t)\right\},$$

where one should note the resemblance to the Azéma-Yor solution of the second stopping time. The solution to the Perkins problem with random starting is then given by

$$\tau_{HP} = \tau_G \wedge \tau_g,$$

see Figure 3 for an illustration. Due to this external randomization given via the random variable G this solution is not adapted to the filtration of the underlying Brownian motion. Moreover this external randomization generally remains to be needed in the deterministic case of $\lambda = \delta_0$ as seen in Figure 4. In particular, the original solution by Perkins is not a special case of the Hobson-Pedersen solution. The authors also discover that simultaneous optimization is no longer possible for general initial distributions and choose to prioritize minimizing the law of the running maximum.

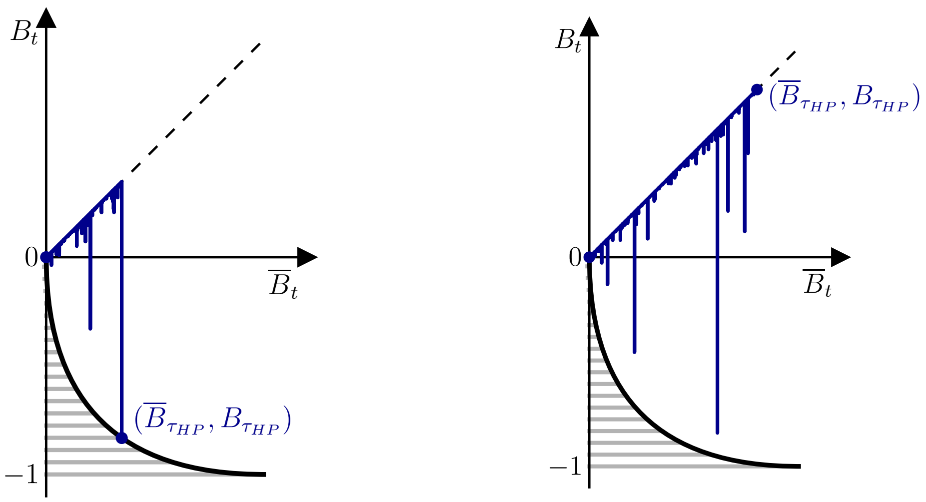


Figure 4: The Hobson-Pedersen solution to ($\mathrm{SEP}_{\lambda,\mu}$) for $\lambda = \delta_0$ and $\mu = U[-1, 1]$ (see Example A in [22]). We see that external randomization is still needed (otherwise paths could only stop in $[-1, 0]$) even though the measures λ and μ are mutually singular.

Existence of a Perkins embedding allowing for a general starting law

By interpreting Perkins' problem with general starting law as an ($\mathrm{OptSEP}_{\lambda,\mu}$) for a well chosen γ as introduced in Section 2 we can use the methods and results therein to prove existence of a solution that is given as the hitting time of the process $(\overline{B}, \underline{B})$ of a specifically structured target set.

It was explained above why we can no longer optimize simultaneously over the running minimum and the running maximum. We will therefore decide - as in the Hobson-Pederson solution presented above - that from now on the running maximum is more important to us. However, it should be obvious that all following results and calculations work analogously if our choice fell on the running minimum instead.

The following theorem provides a generalization of Perkins' theorem in the slightly weaker formulation of optimization in expectation over an auxiliary function φ. While this is needed in order to apply the results of the previous section, the proof of Theorem 1 can be concluded in the next chapter by showing that all such solutions must coincide almost surely independently of the choice of φ. Thus our solution will optimize over all such φ which will conclude optimization in first order stochastic dominance.

We no longer exclude the possibility of λ and μ sharing mass. This leads to the situation of some randomization needed at time 0.

Theorem 5. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a continuous bounded strictly increasing function. Then there exists a stopping time $\tilde{\tau}$ which minimizes $\mathbb{E}\left[\varphi\left(\overline{B}_\tau\right)\right]$ over all solutions of ($\mathrm{SEP}_{\lambda,\mu}$) and maximizes $\mathbb{E}\left[\varphi\left(\underline{B}_\tau\right)\right]$ over all stopping times satisfying the former. For $A \in \mathcal{B}(\mathbb{R})$ we have

$$P\left[\tilde{\tau} = 0, B_0 \in A\right] = (\lambda \wedge \mu)(A)$$

and there exists a set $R \in \mathbb{R}^2$ such that on $\{\tilde{\tau} > 0\}$ we have

$$\tilde{\tau} = \inf\left\{t \geq 0 : \left(\overline{B}_t, \underline{B}_t\right) \in R\right\}.$$

Proof. We will consider the function $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4) : S \to \mathbb{R}^4$ given by

$$\begin{aligned}
\gamma_1((f,s)) &:= \varphi(\overline{f}),\\
\gamma_2((f,s)) &:= -\varphi(\underline{f}),\\
\gamma_3((f,s)) &:= -\varphi(\overline{f})f(s)^2,\\
\gamma_4((f,s)) &:= -\varphi(-\underline{f})f(s)^2.
\end{aligned}$$

Then all γ_j are bounded from below in the sense of (2.1) (due to the boundedness of φ). Thus Theorem 2 guarantees the existence of a minimizer $\tilde{\tau} \in \mathrm{RST}_\lambda(\mu)$ which by Theorem 3 is supported by a γ-monotone Borel set $\Gamma \subseteq S$.
The stop-go conditions amount to the following:
Let $((f,s),(g,t)) \in S \times S$ such that $f(s) = g(t)$, then

$$\mathbb{E}\left[\varphi\left(\overline{f} \vee \left(f(s) + \overline{B}_\sigma\right)\right)\right] + \varphi(\overline{g}) \geq \varphi(\overline{f}) + \mathbb{E}\left[\varphi\left(\overline{g} \vee \left(g(s) + \overline{B}_\sigma\right)\right)\right], \qquad (\mathrm{SGC}_1)$$

$$\mathbb{E}\left[\varphi\left(\underline{f} \wedge \left(f(s) + \underline{B}_\sigma\right)\right)\right] + \varphi(\underline{g}) \leq \varphi(\underline{f}) + \mathbb{E}\left[\varphi\left(\underline{g} \wedge \left(g(s) + \underline{B}_\sigma\right)\right)\right], \qquad (\mathrm{SGC}_2)$$

$$\begin{aligned}
\mathbb{E}\Big[\varphi\left(\overline{f} \vee \left(f(s) + \overline{B}_\sigma\right)\right)\left(f(s) + B_\sigma\right)^2\Big] + \varphi(\overline{g})g(t)^2 \leq&\\
\varphi(\overline{f})f(s)^2 + \mathbb{E}\Big[\varphi\left(\overline{g} \vee \left(g(t) + \overline{B}_\sigma\right)\right)\left(g(t) + B_\sigma\right)^2\Big]&, \qquad (\mathrm{SGC}_3)\\
\mathbb{E}\Big[\varphi\left(-\underline{f} \wedge \left(f(s) + \underline{B}_\sigma\right)\right)\left(f(s) + B_\sigma\right)^2\Big] + \varphi(-\underline{g})g(t)^2 \leq&\\
\varphi(-\underline{f})f(s)^2 + \mathbb{E}\left[\varphi\left(-\underline{g} \wedge \left(g(t) + \underline{B}_\sigma\right)\right)\left(g(t) + B_\sigma\right)^2\right]&. \qquad (\mathrm{SGC}_4)
\end{aligned}$$

Let us first consider $\{\tilde{\tau} > 0\}$. By standard properties of Brownian motion we then may assume that for $(f,s) \in \Gamma$ we have $\overline{f} > \underline{f}$.

To legitimate that on $\{\tilde{\tau} > 0\}$ the stopping time $\tilde{\tau}$ is given as the hitting time of the process $(\overline{B}, \underline{B})$ we will start by showing that $\tilde{\tau}$ will only stop Brownian motion when it is reaching a new running minimum or running maximum.

Consider a path which stops somewhere between its current running extrema, that is a path $(f,s) \in S$ such that $\underline{f} < f(s) < \overline{f}$. Let r be the time where f hits its last new extremum. As f does not reach a new extremum at time s we can consider the initial segment of this path up to a time point $\tilde{s} < r$ such that $f(\tilde{s}) = f(s)$ and either $\underline{f}_{\tilde{s}} = \underline{f}$ or $\overline{f}_{\tilde{s}} = \overline{f}$, where

$$\overline{f}_{\tilde{s}} = \max_{u \leq \tilde{s}} f(u) \quad \text{and} \quad \underline{f}_{\tilde{s}} = \min_{u \leq \tilde{s}} f(u).$$

Let $\tilde{f} := f|_{[0,\tilde{s}]}$, then we claim that $((\tilde{f},\tilde{s}),(f,s)) \in \mathsf{SG}_\gamma$.

1. Case: $\underline{f}_{\tilde{s}} = \underline{f}$ and $\overline{f}_{\tilde{s}} < \overline{f}$ as depicted in Figure 5, that is the last new extremum hit was a new maximum.
Assume $\overline{f}_{\tilde{s}} < f(\tilde{s}) + \overline{B}_\sigma$, then $\overline{f}_{\tilde{s}} \vee \left(f(\tilde{s}) + \overline{B}_\sigma\right) = f(\tilde{s}) + \overline{B}_\sigma$ and (SGC_1) reads

$$\begin{aligned}
&\mathbb{E}\left[\varphi\left(f(\tilde{s}) + \overline{B}_\sigma\right)\right] + \varphi(\overline{f}) \geq \varphi(\overline{f}_{\tilde{s}}) + \varphi(\overline{f}) && \text{if} \quad \overline{f} \geq f(s) + \overline{B}_\sigma\\
&\mathbb{E}\left[\varphi\left(f(\tilde{s}) + \overline{B}_\sigma\right)\right] + \varphi(\overline{f}) \geq \varphi(\overline{f}_{\tilde{s}}) + \mathbb{E}\left[\varphi\left(f(s) + \overline{B}_\sigma\right)\right] && \text{if} \quad \overline{f} < f(s) + \overline{B}_\sigma
\end{aligned}$$

and we see that in both cases a strict inequality holds due to our assumptions.
On the other hand, if $\overline{f}_{\tilde{s}} \geq f(\tilde{s}) + \overline{B}_\sigma$, then $\overline{f}_{\tilde{s}} \vee \left(f(\tilde{s}) + \overline{B}_\sigma\right) = \overline{f}_{\tilde{s}}$ as well as $\overline{f} \vee \left(f(s) + \overline{B}_\sigma\right) = \overline{f}$ and this leads to an equality in (SGC_1). Since $\underline{f}_{\tilde{s}} = \underline{f}$ we also have an equality in (SGC_2) and therefore jump to our third condition (SGC_3):

$$\mathbb{E}\left[\varphi(\overline{f}_{\tilde{s}})(f(\tilde{s}) + B_\sigma)^2\right] + \varphi(\overline{f})f(s)^2 \leq \varphi(\overline{f}_{\tilde{s}})f(\tilde{s})^2 + \mathbb{E}\left[\varphi(\overline{f})(f(s) + B_\sigma)^2\right]$$

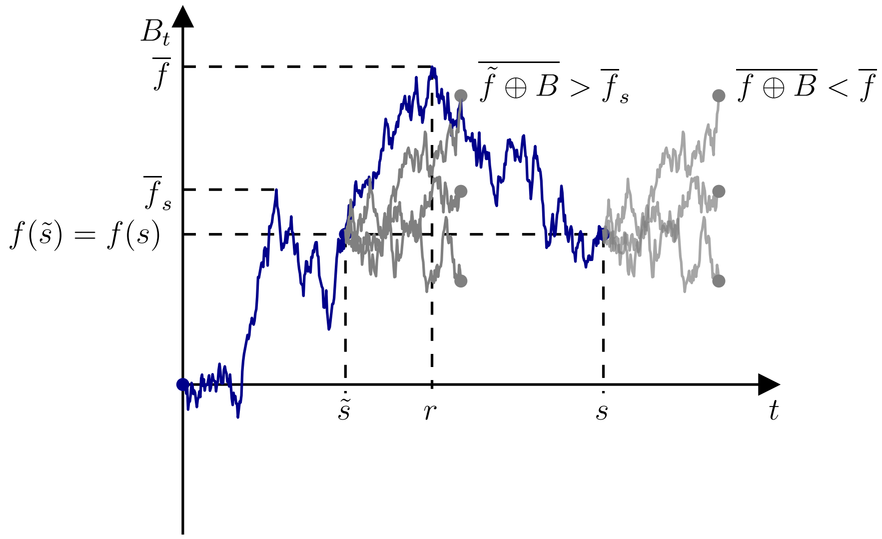


Figure 5: Stop-Go pairs: We see possible continuations of our path f. If these continuations are attached at time s the running maximum of f remains unchanged. Attached at time $\tilde{s}$ however, the running maximum of $(\tilde{f}, \tilde{s}) = (f|_{[0,\tilde{s}]}, \tilde{s})$ will be increased.

Again, due to our assumptions $f(\tilde{s}) = f(s)$, $\overline{f}_{\tilde{s}} < \overline{f}$ and as φ is strictly increasing, a strict inequality holds and thus $((\tilde{f}, \tilde{s}), (f, s)) \in \mathsf{SG}_\gamma$.

2. Case: $\overline{f}_{\tilde{s}} = \overline{f}$ and $\underline{f}_{\tilde{s}} > \underline{f}$, that is the last new extremum hit was a new minimum. Now (SGC_1) as well as (SGC_3) will always exhibit an equality and conditions (SGC_2) and (SGC_4) can be treated analogously to the previous case.

As now due to γ-monotonicity $((\tilde{f}, \tilde{s}), (f, s)) \notin \Gamma^< \times \Gamma$ it follows that $\Gamma \cap \{(f, s) \in S : \underline{f} < f(s) < \overline{f}\} = \emptyset$, that is, when stopped we will almost surely have reached a new running minimum or a new running maximum.

Summing up, we now know that $\tilde{\tau}$ stops paths of Brownian motion only when they reach a new running minimum or a new running maximum.

We will denote the set of stopped paths satisfying this condition by

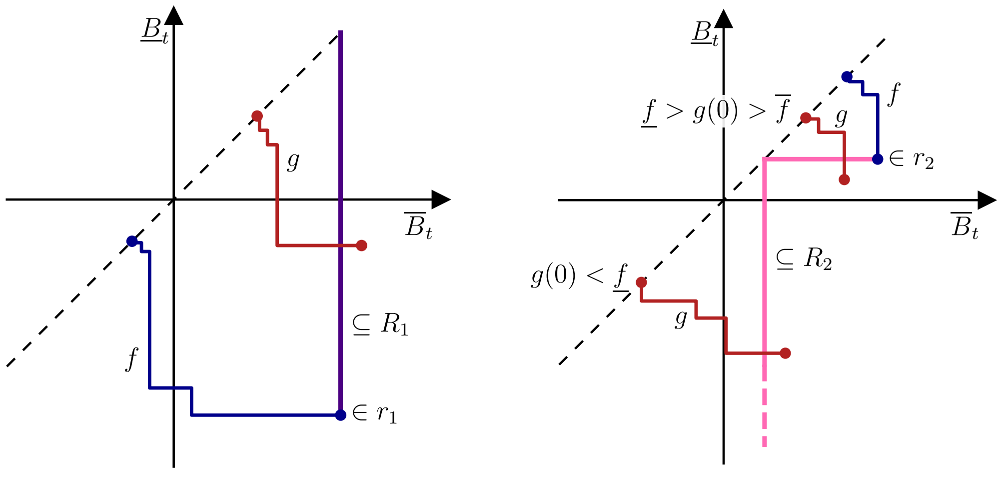
$$\tilde{S} := \{(f, s) \in S : f(s) = \underline{f} \text{ or } f(s) = \overline{f}\}.$$

This justifies to consider the phase space $(\overline{B}, \underline{B})$. All possible paths will lie below the diagonal, i.e. in the set $H_D := \{(x, y) \in \mathbb{R}^2 : x \geq y\}$.

We propose that there are two sets of points $r_1, r_2 \subseteq H_D$, such that $R = R_1 \cup R_2$, where

$$R_1 := \{\{x\} \times [y, x] : (x, y) \in r_1\} \quad \text{and}$$
$$R_2 := \{([y, x] \times \{y\}) \cup (\{y\} \times [y, \infty)) : (x, y) \in r_2\}.$$

Let us legitimate this target set structure:
Assume we know that there is a path $(f, s) \in \Gamma$ such that $s > 0$ and $f(s) = \overline{f}$, that is we stop at a new running maximum. We claim that it is impossible for a trajectory of the process $(\overline{B}, \underline{B})$ to *traverse* the $\{\overline{f}\} \times [\underline{f}, \overline{f})$ line-segment and then be stopped as seen in Figure 6 (A).



(A) The path (f, s) is stopped at a new running maximum.

(B) The path (f, s) is stopped at a new running minimum.

Figure 6

Consider a path $(g, t) \in \tilde{S}$ such that $g(0) \in (\underline{f}, \overline{f}]$, $\overline{g} \geq \overline{f}$ and $\underline{g} > \underline{f}$. Then there has to exist a timepoint $\tilde{t} \leq t$ such that $g(\tilde{t}) = \overline{g}_{\tilde{t}} = \overline{f} = f(s)$ and note that still $\underline{g}_{\tilde{t}} > \underline{f}$ has to hold. However, this situation equals the 1. case of the above discussion, hence again $((g, \tilde{t}), (f, s)) \in \mathsf{SG}_\gamma$. By γ-monotonicity it now follows that $(g, \tilde{t}) \notin \Gamma^<$, therefore $(g, t) \notin \Gamma$ and our claim is proven.
It becomes clear that by choosing $r_1 := \left\{(\overline{f}, \underline{f}) : (f, s) \in \Gamma \text{ and } f(s) = \overline{f}\right\}$ the definition of R_1 as above is reasonable. Now assume we know that there is a path $(f, s) \in \Gamma$ such that $s > 0$ and $f(s) = \underline{f}$, that is we stop at a new running minimum. We claim that it is impossible for a trajectory of the process $(\overline{B}, \underline{B})$ to *traverse* either the $[\underline{f}, \overline{f}) \times \{\underline{f}\}$ or the $\{\underline{f}\} \times [\underline{f}, \infty)$ line-segment and then be stopped as seen in Figure 6 (B).

Consider a path $(g, t) \in \tilde{S}$ such that $\overline{g} \in [\underline{f}, \overline{f})$ and $\underline{g} \leq \underline{f}$. Let us first assume $g(0) \in [\underline{f}, \overline{f})$. Then again there has to exist a time point $\tilde{t} \leq t$ such that $g(\tilde{t}) = \underline{g}_{\tilde{t}} = \underline{f} = f(s)$ while $\overline{g}_{\tilde{t}} < \overline{f}$. Again $((g, \tilde{t}), (f, s)) \in \mathsf{SG}_\gamma$ as in the 1. case above.

Now assume $g(0) < \underline{f}$. It is then possible to find a timepoint $\tilde{t} < t$ such that $g(\tilde{t}) = \overline{g}_{\tilde{t}} = \underline{f} = f(s)$ and still $\overline{g}_{\tilde{t}} < \overline{f}$. We remember that (SGC$_1$) exhibits an equality if $\overline{g}_{\tilde{t}} \geq g(\tilde{t}) + \overline{B}_\sigma$, which in our setting is equivalent to $\overline{B}_\sigma = 0$. If on the other hand $\overline{B}_\sigma > 0$, we will have a strict inequality. However, as trivial stopping times are excluded, the later will always happen with positive probability and thus taking the expectation (SGC$_1$) will always lead to a strict inequality, implying that $((g, \tilde{t}), (f, s)) \in \mathsf{SG}_\gamma$.

This concludes the proof of $(g, t) \notin \Gamma$.
Again, in analogy to the above let us take $r_2 := \left\{(\overline{f}, \underline{f}) : (f, s) \in \Gamma \text{ and } f(s) = \underline{f}\right\}$ to justify the definition of R_2.

Now define the following two target sets

$$
\begin{aligned}
R_{\mathrm{CL}} &:= R = R_1 \cup R_2, \\
R_{\mathrm{OP}} &:= \left\{\{x\} \times (y, x] : (x, y) \in r_1\right\} \cup \left\{\left([y, x) \times \{y\}\right) \cup \left(\{y\} \times [y, \infty)\right) : (x, y) \in r_2\right\}.
\end{aligned}
$$

and consider

$$
\tau_{\mathrm{CL}} := \inf\left\{t \geq 0 : (\overline{B}_t, \underline{B}_t) \in R_{\mathrm{CL}}\right\} \leq \tau_{\mathrm{OP}} := \inf\left\{t \geq 0 : (\overline{B}_t, \underline{B}_t) \in R_{\mathrm{OP}}\right\}.
$$

Note that since Γ is Borel, the sets r_1 and r_2 are analytic sets since they are continuous images of Borel sets. This implies that R_1 and R_2 are analytic sets and we see that τ_{CL} and τ_{OP} are stopping times.

We would like to show that $\tau_{\mathrm{CL}} = \tilde{\tau} = \tau_{\mathrm{OP}}$ a.s. as our claim then follows.
As $\Gamma \cap \{(f,s) \in S : \underline{f} < f(s) < \overline{f}\} = \emptyset$ it is obvious that $\tau_{\mathrm{CL}} \leq \tilde{\tau}$ a.s. by definition of τ_{CL}.

To show that $\tilde{\tau} \leq \tau_{\mathrm{OP}}$ a.s. let us assume that this is not the case. Choose $\omega \in \Omega$ such that $((B_t(\omega))_{t \leq \tilde{\tau}(\omega)}, \tilde{\tau}(\omega)) \in \Gamma$ and assume $\tau_{\mathrm{OP}}(\omega) < \tilde{\tau}(\omega)$. Then there has to exist an $s \in [\tau_{\mathrm{OP}}(\omega), \tilde{\tau}(\omega))$ such that for $f = (B_t(\omega))_{t \leq s}$ we have $(\overline{f}, \underline{f}) \in R_{\mathrm{OP}}$. As $s \leq \tilde{\tau}(\omega)$ it follows, that $(f,s) \in \Gamma^{<}$, that is (f,s) is a going path. By definition of τ_{OP} we can find find a point $(x,y) \in r_1$ such that $(\overline{f}, \underline{f}) \in \{x\} \times (y,x]$ or a point $(x,y) \in r_2$ such that $(\overline{f}, \underline{f}) \in \big([y,x) \times \{y\}\big) \cup \big(\{y\} \times [y,\infty)\big)$. However, we then find ourselves in the same situation of traversing line-segments as above. More precisely, by considering the path $(g,t) \in \Gamma$ corresponding to this r_1 respectively r_2 point we again find a SG-pair contradicting the γ-monotonicity of Γ. Hence $\tilde{\tau} \leq \tau_{OP}$ a.s.

By standard properties of Brownian motion now $\tau_{\mathrm{OP}} = \tau_{\mathrm{CL}}$ a.s. which concludes the proof in the $\{\tilde{\tau} > 0\}$ case.
Let us now consider stopping in time 0. Note that $P[\tilde{\tau} = 0, B_0 \in A] \leq (\lambda \wedge \mu)(A)$. We want to show that a strict inequality stands in conflict with the γ-monotonicity. If $P[\tilde{\tau} = 0, B_0 \in A] < (\lambda \wedge \mu)(A)$ then there has to exists some $x \in A$ such that there are paths in Γ starting in x but not immediately stopping and also paths which stop in $x \in A$ at a strictly positive time. By above discussion we see that this would constitute SG-pairs and we can conclude

$$P[\tilde{\tau} = 0, B_0 \in A] = (\lambda \wedge \mu)(A).$$

□

4 Uniqueness

In the previous section we have seen that the Perkins solution with random starting is given as a hitting time of the process $(\overline{B}, \underline{B})$ of a specifically structured target set. As mentioned in the introduction it was proved by Loynes [25] that Root's barrier type solution is essentially unique. In this chapter we will briefly discuss the extension of this argument to *barrier type solutions* of the form

$$\tau_R = \inf\{t \geq 0 : (A_t, B_t) \in R\}$$

for sufficiently regular processes A_t.

We will then propose a setting in which the Perkins solution with random starting can also be seen as a barrier type solutions and show how the Loynes argument extends to this setting. This will enable us to conclude the proof of Theorem 1.

4.1 The Loynes uniqueness result

Root initially defined barriers as topologically *closed* subsets of $\mathbb{R}$. Loynes uniqueness argument relies on this fact by using that we are actually inside the barrier when we stop. A suitable generalization for our purposes would be to ask our process A_t to be sufficiently regular such that (A_t, B_t) is jointly Markov satisfying the Blumenthal-Getoor 0-1-law. Instead of topological closures we then consider *fine closures* with respect to the process (A_t, B_t).

The *fine closure* of a set $R \subseteq \mathbb{R}^2$ with respect to a jointly Markov process (A_t, B_t) will be denoted by R^* and is defined as

$$R^* := R \cup \left\{(t,x) \in \mathbb{R}^2 : P\left[\tau_R = 0 \middle| (A_0, B_0) = (t,x)\right] = 1\right\}.$$

By this definition follows $\tau_R = \tau_{R^*}$ a.s. and $(A_{\tau_{R^*}}, B_{\tau_{R^*}}) \in R^*$ by Blumenthal-Getoor. Moreover we also have $A_{\tau_R} \sim A_{\tau_{R^*}}$ and $B_{\tau_R} \sim B_{\tau_{R^*}}$.

For details on fine closures refer to [7], see also the arguments in [26].
We see that taking fine closures does not alter the stopping properties and we will therefore assume without loss of generality that our barriers are always finely closed with respect to (A_t, B_t).

Proposition 4.1. Let R and S be two barriers such that

$$\tau_R := \inf\{t \geq 0 : (A_t, B_t) \in R\} \text{ and } \tau_S := \inf\{t \geq 0 : (A_t, B_t) \in S\}$$

are stopping times both generating the same law μ. Then $R \cup S$ also generates μ and the corresponding stopping time is given by

$$\tau_{R\cup S} = \tau_R \wedge \tau_S.$$

Proof. Consider the set

$$K := \{z \in \mathbb{R} : \inf\{y \in \mathbb{R} : (y,z) \in R\} < \inf\{y \in \mathbb{R} : (y,z) \in S\}\},$$

and define

$$R_K := \{(x,y) \in R : y \in K\}, \quad R_{K^c} := \{(x,y) \in R : y \in K^c\},$$
$$S_K := \{(x,y) \in S : y \in K\}, \quad S_{K^c} := \{(x,y) \in S : y \in K^c\}.$$

Note that $S_K \subseteq R_K$ as well as $R_{K^c} \subseteq S_{K^c}$.

Now assume that $B_{\tau_S} \in K$. By definition of τ_S and by the above discussion we have $(A_{\tau_S}, B_{\tau_S}) \in S_K \subseteq R_K$. This means, $\tau_R \leq \tau_S$ and it is now impossible to have $(A_{\tau_R}, B_{\tau_R}) \in R_{K^c}$ (otherwise B_{τ_S} would have stopped in K^c). This implies that we cannot have $B_{\tau_R} \in K^c$ and therefore $P[B_{\tau_S} \in K, B_{\tau_R} \in K^c] = 0$.
As $B_{\tau_R} \sim B_{\tau_S}$ we have:

$$\begin{aligned} P[B_{\tau_R} \in K] &= P[B_{\tau_R} \in K, B_{\tau_S} \in K] + P[B_{\tau_R} \in K, B_{\tau_S} \in K^c] \\ &= P[B_{\tau_S} \in K, B_{\tau_R} \in K] + P[B_{\tau_S} \in K, B_{\tau_R} \in K^c] = P[B_{\tau_S} \in K], \end{aligned}$$

and altogether

$$0 = P[B_{\tau_R} \in K, B_{\tau_S} \in K^c] = P[B_{\tau_S} \in K, B_{\tau_R} \in K^c].$$

It is now clear that the sets

$$\Omega_1 := \{B_{\tau_R} \in K, B_{\tau_S} \in K\} \quad \text{where} \quad \tau_R \leq \tau_S$$

and

$$\Omega_2 := \{B_{\tau_R} \in K^c, B_{\tau_S} \in K^c\} \quad \text{where} \quad \tau_S \leq \tau_R$$

are essentially disjoint and their union has full probability. Therefore $\tau_{R\cup S} = \tau_R \wedge \tau_S$ a.s.
That $\tau_{R\cup S}$ generates the same law is now obvious (due to decomposition into Ω_1 and Ω_2). □

Corollary 4.2. If τ is a barrier type solution to the $(\mathrm{SEP}_{\lambda,\mu})$, then τ is a.s. unique.

Proof. For any solution τ to $(\mathrm{SEP}_{\lambda,\mu})$ we have $\mathbb{E}[\tau] < \infty$ and thus $\mathbb{E}[B_\tau^2] = \mathbb{E}[\tau]$. Now if $\tilde{\tau}$ is another solution to $(\mathrm{SEP}_{\lambda,\mu})$ and $\tilde{\tau} \leq \tau$, then $\mathbb{E}[\tilde{\tau}] = \mathbb{E}[B_{\tilde{\tau}}^2] = \mathbb{E}[B_\tau^2] = \mathbb{E}[\tau]$, hence $\tau = \tilde{\tau}$ a.s. and we call such a solution *minimal*. Let R be the barrier inducing the solution $\tau = \tau_R$ and let τ_S be a different solution induced by the barrier S. By Proposition 4.1 we have that $\tau_R \wedge \tau_S$ is also a solution to the $(\mathrm{SEP}_{\lambda,\mu})$. Since trivially $\tau_R \wedge \tau_S \leq \tau_R$ we have that $\tau_R \wedge \tau_S = \tau_R$ due to the minimality observation before. Now analogously $\tau_R \wedge \tau_S = \tau_S$, hence $\tau_R = \tau_S$ a.s. concluding the proof of uniqueness. □

We see that this uniqueness result cannot immediately be applied to the solutions found in Theorem 5. However, an analogous uniqueness result holds in this setting. The target set constructed in the proof of Theorem 5 can be seen as an inverse barrier in the sense of Rost and the Loynes type uniqueness result can be extended to this setting.

4.2 Uniqueness of the Perkins solution

To give a Loynes type uniqueness result for the Perkins embedding we need to identify the right space and setting to be able to consider vh-barriers as barriers in the classic sense. We aim to represent vh-barriers as *inverse* barriers in the Rost sense.

Let $(\mathrm{R}', \preceq)$ be a disjoint copy of the real numbers where the order is flipped, i.e. for $x, y \in \mathrm{R}'$ we have $x \preceq y$ if and only if $x \geq y$ in the usual order of the real numbers. We keep R' distinguishable from $\mathbb{R}$. Consider $\mathrm{D} := \mathrm{R}' \cup \mathbb{R}$ and define an order $\leq_{\mathrm{D}}$ on D in the following way.

$$x, y \in \mathrm{D}, \text{ then } x \leq_{\mathrm{D}} y :\Leftrightarrow \begin{cases} x \in \mathrm{R}' \text{ and } y \in \mathbb{R} \\ x, y \in \mathrm{R}' \text{ and } x \preceq y \\ x, y \in \mathbb{R} \text{ and } x \leq y \end{cases}$$

This order enables us to depict the set D as linearly ordered from left to right. Moreover we can consider R' as the *left half* of D while considering $\mathbb{R}$ as the *right half* of D. To emphasize this we consider the map

$$\iota : \mathbb{R} \to \mathrm{R}', \quad \iota(x) = x$$

which embeds $\mathbb{R}$ into R'. Then

$$\begin{gathered} \cdots \leq_{\mathrm{D}} \iota(1) \leq_{\mathrm{D}} \cdots \leq_{\mathrm{D}} \iota(0) \leq_{\mathrm{D}} \cdots \leq_{\mathrm{D}} \iota(-1) \leq_{\mathrm{D}} \cdots \\ \cdots \leq_{\mathrm{D}} -1 \leq_{\mathrm{D}} \cdots \leq_{\mathrm{D}} 0 \leq_{\mathrm{D}} \cdots \leq_{\mathrm{D}} 1 \leq_{\mathrm{D}} \cdots, \end{gathered}$$

thus we may consider R' as $\mathbb{R}$ reflected in the origin.

We will now construct *barriers* in $\mathrm{D} \times \mathbb{R} = (\mathrm{R}' \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{R})$. Note that in the proof of Theorem 5 the vh-barrier R was established as $R = R_1 \cup R_2$ where R_1 was induced by r_1, the endpoints of paths being stopped at a new running maximum while R_2 was induced by r_2, the endpoints of paths being stopped at a new running minimum. While R is a barrier in $\mathbb{R}^2$ we can also use r_1 and r_2 to define a corresponding barrier $R_{\mathrm{D}} \subseteq \mathrm{D} \times \mathbb{R}$ in the following way. For every $(x, y) \in r_1$ we have $(x, y) \in \mathrm{R}' \times \mathbb{R} \subseteq \mathrm{D} \times \mathbb{R}$ and also $(\tilde{x}, y) \in R_{\mathrm{D}}$ for all $\tilde{x} \in \mathrm{D}$ such that $\tilde{x} \leq_{\mathrm{D}} x$. Note that as $x \in \mathrm{R}'$ of course we have $\tilde{x} \in \mathrm{R}'$ for all $\tilde{x} \leq_{\mathrm{D}} x$. Analogously for $(x, y) \in r_2$ we have $(x, y) \in \mathbb{R} \times \mathbb{R} \subseteq \mathrm{D} \times \mathbb{R}$ and also $(\tilde{x}, y) \in R_{\mathrm{D}}$ for all $\tilde{x} \in \mathrm{D}$ such that $\tilde{x} \leq_{\mathrm{D}} x$. Note that this especially implies $(\tilde{x}, y) \in R_D$ for all $\tilde{x} \in \mathrm{R}'$.

Thus every path stopped at a new running maximum creates a line in the left half of $\mathrm{D} \times \mathbb{R}$ while every path stopped at new running minimum creates a line in the right half of $\mathrm{D} \times \mathbb{R}$ which continues on into the left half creating a structure of Rost barrier type. Figure 7 gives an illustration of how the vh-barrier picture in Figure 1 (B) translates into a barrier in $\mathrm{D} \times \mathbb{R}$ which will be of inverse barrier structure.

We want to consider the path $(\overline{B}, \underline{B})$ in $\mathrm{D} \times \mathbb{R}$, more precisely we want to consider the path on the lhs $\mathrm{R}' \times \mathbb{R}$ as well as on the rhs $\mathbb{R} \times \mathbb{R}$ of $\mathrm{D} \times \mathbb{R}$ separately but simultaneously. So we define the following two maps embedding $\mathbb{R}^2$ into the respective space.

$$\begin{aligned} \overline{\psi} &: \mathbb{R}^2 \to \mathrm{R}' \times \mathbb{R}, \quad \overline{\psi}(x, y) = (\iota(y), x) \\ \underline{\psi} &: \mathbb{R}^2 \to \mathbb{R} \times \mathbb{R}, \quad \underline{\psi}(x, y) = (x, y) \end{aligned}$$

While $\underline{\psi}$ will plainly embed the path $(\overline{B}, \underline{B})$ into the right half $\mathbb{R} \times \mathbb{R}$, the map $\overline{\psi}$ will reflect the path $(\overline{B}, \underline{B})$ along the main diagonal before embedding it into the left half $\mathrm{R}' \times \mathbb{R}$. As we have 'flipped' the order on R' we will perceive the action of $\overline{\psi}$ as a counter clockwise rotation by 90°. Due to the order defined on D these two paths will now both travel from left to right. In the left half the path $\overline{\psi}\left((\overline{B}_t, \underline{B}_t)\right)$ will move vertically upwards whenever a new running maximum is reached and will move only horizontally when a new running minimum is reached. Furthermore this implies that the barrier can only be hit when a new running maximum is reached and no stopping will happen in the left hand side due to a new running minimum. In the right half the path $\underline{\psi}\left((\overline{B}_t, \underline{B}_t)\right)$ will

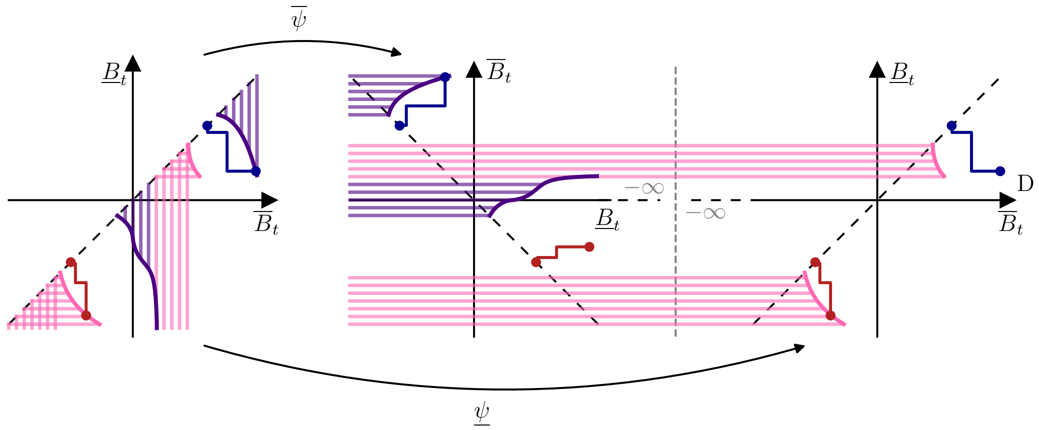


Figure 7: Translating the Perkins barrier picture in Figure 1 (B) into $\mathrm{D} \times \mathbb{R}$.

move vertically downwards whenever a new running minimum is reached but will travel perfectly horizontal at the level of the current running minimum when a new running maximum is reached. In other words, all stopping due to new running maxima will happen on the left hand side $\mathrm{R}' \times \mathbb{R}$ while all stopping due to new running minima will happen on the right hand side $\mathbb{R} \times \mathbb{R}$. We consider the respective stopping times of the embedded processes

$$\overline{\tau}_R := \inf\left\{t \geq 0 : \overline{\psi}\left(\left(\overline{B}_t, \underline{B}_t\right)\right) \in R_{\mathrm{D}}\right\},$$
$$\underline{\tau}_R := \inf\left\{t \geq 0 : \underline{\psi}\left(\left(\overline{B}_t, \underline{B}_t\right)\right) \in R_{\mathrm{D}}\right\}.$$

Now if $R \subseteq \mathbb{R}^2$ and $R_{\mathrm{D}} \subseteq \mathrm{D} \times \mathbb{R}$ are both induced by the same sets r_1 and r_2 found in Theorem 5, then

$$\tau_R = \inf\left\{t \geq 0 : \left(\overline{B}_t, \underline{B}_t\right) \in R\right\} = \overline{\tau}_R \wedge \underline{\tau}_R,$$

thus both representations of the hitting time of the barrier will be equivalent. We will assume R_{D} to be finely closed with respect to $\overline{\psi}\left(\left(\overline{B}_t, \underline{B}_t\right)\right)$ as well as $\underline{\psi}\left(\left(\overline{B}_t, \underline{B}_t\right)\right)$ and give a Loynes type uniqueness argument for the stopping time τ_R.

Theorem 6. Let $R, S \subseteq \mathbb{R}^2$ be two vh-barriers and let τ_R resp. τ_S denote their hitting times by the process $\left(\overline{B}_t, \underline{B}_t\right)$. If τ_R and τ_S both embed the same law μ, then $\tau_R = \tau_S$ a.s.

Proof. Note that by the above discussion we must have inverse barriers $R_{\mathrm{D}}, S_{\mathrm{D}} \subseteq \mathrm{D} \times \mathbb{R}$ corresponding to the vh-barriers R resp. S such that $\tau_R = \overline{\tau}_R \wedge \underline{\tau}_R$ and $\tau_S = \overline{\tau}_S \wedge \underline{\tau}_S$. Analogously to Proposition 4.1 and Corollary 4.2 the almost sure equality $\tau_R = \tau_S$ will follow by defining a set $K \subseteq \mathbb{R}$ of those levels where barrier R is 'longer' than barrier S and showing that

$$P\left[B_{\overline{\tau}_S \wedge \underline{\tau}_S} \in K, B_{\overline{\tau}_R \wedge \underline{\tau}_R} \in K^c\right] = 0. \tag{4.1}$$

Considering the order $\leq_{\mathrm{D}}$ on D as well as the fact that R_{D} respectively S_{D} are both inverse barriers we define

$$K := \left\{z \in \mathbb{R} : \sup\left\{x \in D : (x, z) \in R_{\mathrm{D}}\right\} >_{\mathrm{D}} \inf\left\{x \in D : (x, z) \in S_{\mathrm{D}}\right\}\right\}.$$

Thus now we can consider the subset

$$R^K := \left\{(x, y) \in R_{\mathrm{D}} : y \in K\right\} \subseteq R_{\mathrm{D}}$$

and define R^{K^c}, S^K and S^{K^c} analogously. Then we have $S^K \subseteq R^K$ and $R^{K^c} \subseteq S^{K^c}$.

Consider $B_{\overline{\tau}_S \wedge \underline{\tau}_S} \in K$, then it is crucial to observe that our stopping times will only stop Brownian motion at a *new* running minimum or running maximum, hence by definition of the respective stopping times we will either have $B_{\overline{\tau}_S \wedge \underline{\tau}_S} = \overline{B}_{\overline{\tau}_S}$ or $B_{\overline{\tau}_S \wedge \underline{\tau}_S} = \underline{B}_{\underline{\tau}_S}$, thus either

$$\overline{\psi}\left((\overline{B}_t, \underline{B}_t)\right) \in S^K \subseteq R^K \quad \text{ or } \quad \underline{\psi}\left((\overline{B}_t, \underline{B}_t)\right) \in S^K \subseteq R^K.$$

Assume $B_{\overline{\tau}_S \wedge \underline{\tau}_S} = \overline{B}_{\overline{\tau}_S}$. Then $\overline{\tau}_S \le \underline{\tau}_S$ and $\overline{\tau}_R \le \overline{\tau}_S$ by the definition of K, hence $\overline{\tau}_R \wedge \underline{\tau}_R \le \overline{\tau}_S \wedge \underline{\tau}_S$ and it is impossible to have $B_{\overline{\tau}_R \wedge \underline{\tau}_R} \in K^c$ as otherwise $B_{\overline{\tau}_S \wedge \underline{\tau}_S}$ would have stopped in K^c as well. We have the analogous result when $B_{\overline{\tau}_S \wedge \underline{\tau}_S} = \underline{B}_{\underline{\tau}_S}$, hence (4.1) is clear. We can conclude the proof analogously to the proof of Proposition 4.1 and Corollary 4.2. □

We can now complete the proof of Theorem 1.

Proof. For an arbitrary choice of φ consider $\tau_P := \tilde{\tau}$ as found in Theorem 5. Let $\overline{\tau}$ be another stopping time constructed as in Theorem 5 but for a different choice of φ.

First note that the behaviour in 0 is independent of φ, hence for $A \in \mathcal{B}(\mathbb{R})$ we have

$$P\left[\tau_P = 0, B_0 \in A\right] = (\lambda \wedge \mu)(A) = P\left[\overline{\tau} = 0, B_0 \in A\right].$$

Now note that on $\{\tau_P > 0\} = \{\overline{\tau} > 0\}$ both stopping times are given as hitting times of vh-barriers and embed the same law μ, hence by Theorem 6 we have $\tau_P = \overline{\tau}$ a.s. Since the choices for φ were arbitrary we must have optimality for any such φ, thus we can conclude that τ_R minimizes resp. maximizes $\overline{B}_{\tilde{\tau}}$ resp. $\underline{B}_{\tilde{\tau}}$ in law. □

4.3 An example

We will illustrate the occurrences of the different type of lines of the Perkins barrier solution in the following simple atomic example. Consider the initial measure

$$\lambda = \frac{1}{4}\delta_{-1} + \frac{1}{2}\delta_0 + \frac{1}{4}\delta_1,$$

and for $\alpha \in [0, 1]$ the terminal measure

$$\mu = \frac{1-\alpha}{2}\delta_{-2} + \alpha\delta_0 + \frac{1-\alpha}{2}\delta_2.$$

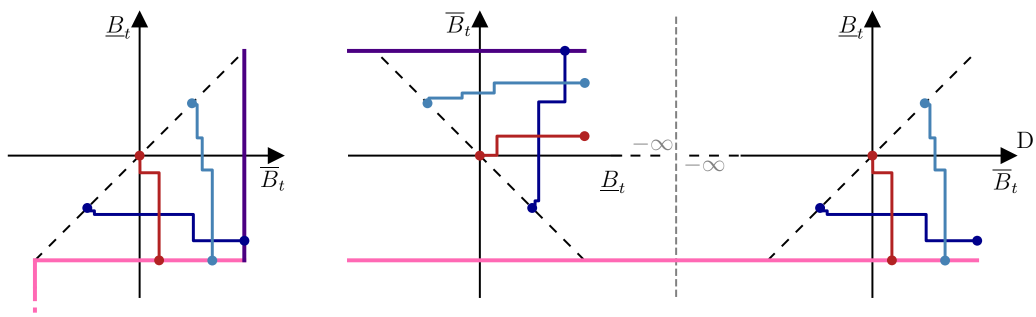


Figure 8: The Perkins embedding of μ for $\alpha \in \left[0, \frac{1}{2}\right]$.

Depending on the size α of the atom in $\{0\}$ we have the following different cases.

$\alpha \in \left[0, \frac{1}{2}\right]$: The measure λ and μ share mass $\mu(\{0\}) = \alpha \le \frac{1}{2} = \lambda(\{0\})$ in $\{0\}$, thus $(\lambda \wedge \mu)(\{0\}) = \alpha$ and all mass needed in $\{0\}$ will get stopped immediately while all other mass will be left to run to ± 2 as illustrated in Figure 8.

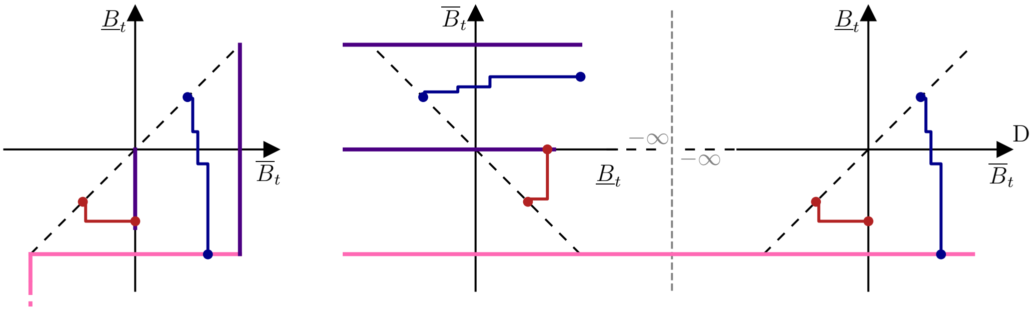


Figure 9: The Perkins embedding of μ for $\alpha \in \left(\frac{1}{2}, \frac{5}{8}\right]$.

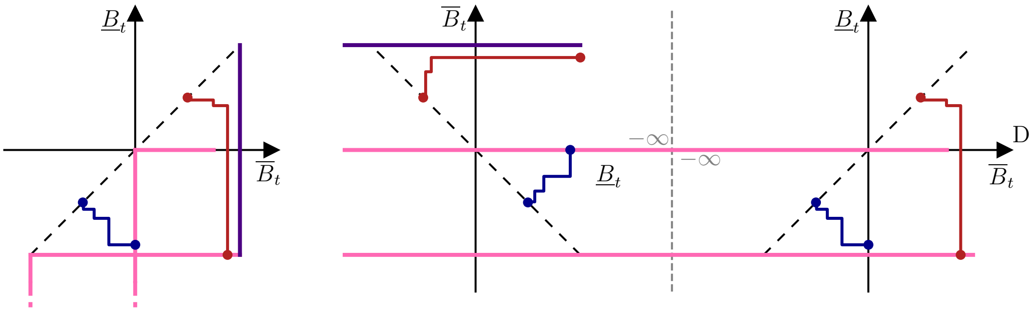


Figure 10: The Perkins embedding of μ for $\alpha \in \left(\frac{5}{8}, \frac{3}{4}\right]$.

$\alpha \in \left(\frac{1}{2}, \frac{5}{8}\right]$: Since $\mu(\{0\}) = \alpha > \frac{1}{2} = \lambda(\{0\})$ we have $(\lambda \wedge \mu)(\{0\}) = \frac{1}{2}$. Hence all mass available in $\{0\}$ will be stopped immediately but then still a little more mass is needed. An additional v-line in 0 where some of the paths starting in $B_0 = -1$ will stop upon reaching a new running maximum in 0 will provide this missing mass as illustrated in Figure 9. Note that by adding this additional v-line we can at maximum acquire an additional $\frac{1}{8}$ of mass in $\{0\}$, hence resulting in a maximal atom size of $\frac{5}{8}$ to be embedded this way.

$\alpha \in \left(\frac{5}{8}, \frac{3}{4}\right]$: As seen in the previous case we have $(\lambda \wedge \mu)(\{0\}) = \frac{1}{2}$ and all mass available in $\{0\}$ will be stopped immediately. However as discussed above merely adding a v-line will no longer suffice to acquire the $\alpha > \frac{5}{8}$ atom in $\{0\}$. Hence we also need to stop (some) paths starting in $B_0 = 1$ upon reaching a new running minimum in $\{0\}$ via an h-line in $\{0\}$ as illustrated in Figure 10. Note that by adding this horizontal line we can at most acquire an additional $\frac{1}{8}$ of mass in $\{0\}$ from the paths started in $B_0 = 1$, adding up to a total atom size of $\frac{3}{4}$ that could be embedded this way.

$\alpha \in \left(\frac{3}{4}, 1\right]$: The measures λ fails to be prior to μ in the convex order when $\alpha > \frac{3}{4}$. While technically we could provide vh-barriers such that the corresponding hitting time would embed the measure μ, the resulting stopping time would no longer be integrable hence not be a solution to $(\mathrm{SEP}_{\lambda,\mu})$.

References

[1] J. Azéma and M. Yor. Une solution simple au problème de Skorokhod. *Séminaire de Probabilités, XIII (Univ. Strasbourg, Strasbourg, 1977/78).* Vol. 721. Lecture Notes in Math. Berlin: Springer, 1979, pp. 90–115.

[2] E. Bayraktar and T. Bernhardt. On the Continuity of the Root Barrier. *Proceedings of the American Mathematical Society* 150 (July 2022).

[3] E. Bayraktar and X. Zhang. Embedding of Walsh Brownian motion. *Stochastic Processes and their Applications* 134 (2021), pp. 1–28.

[4] M. Beiglböck, A. M. G. Cox, and M. Huesmann. The geometry of multi-marginal Skorokhod Embedding. *Probab. Theory Related Fields* 176.3-4 (2020), pp. 1045–1096.

[5] M. Beiglböck, M. Nutz, and F. Stebegg. Fine properties of the optimal Skorokhod embedding problem. *J. Eur. Math. Soc. (JEMS)* 24.4 (2022), pp. 1389–1429.

[6] H. Brown, D. Hobson, and L. C. G. Rogers. Robust hedging of barrier options. *Math. Finance* 11.3 (2001), pp. 285–314.

[7] K. L. Chung and J. B. Walsh. Markov Processes, Brownian Motion, and Time Symmetry. 2nd ed. 249. Springer-Verlag New York, 2005.

[8] A. M. G. Cox and D. Hobson. A Unifying Class of Skorokhod Embeddings: Connecting the Azéma: Yor and Vallois Embeddings. *Bernoulli* 13.1 (2007), pp. 114–130.

[9] A. M. G. Cox, J. Obłój, and N. Touzi. The Root solution to the multi-marginal embedding problem: an optimal stopping and time-reversal approach. *Probab. Theory Related Fields* 173.1-2 (2019), pp. 211–259.

[10] A. M. G. Cox and J. Wang. Optimal robust bounds for variance options. *ArXiv e-prints* (Aug. 2013).

[11] A. M. G. Cox and J. Wang. Root's Barrier: Construction, Optimality and Applications to Variance Options. *Ann. Appl. Probab.* 23.3 (2013), pp. 859–894.

[12] P. Gassiat, A. Mijatović, and H. Oberhauser. An integral equation for Root's barrier and the generation of Brownian increments. *Ann. Appl. Probab.* 25.4 (2015), pp. 2039–2065.

[13] P. Gassiat, H. Oberhauser, and G. dos Reis. Root's barrier, viscosity solutions of obstacle problems and reflected FBSDEs. *Stochastic Processes and their Applications* 125.12 (2015), pp. 4601–4631.

[14] P. Gassiat, H. Oberhauser, and C. Z. Zou. A free boundary characterization of the Root barrier for Markov processes. *Probability Theory and Related Fields* 180 (2021), pp. 33–69.

[15] N. Ghoussoub, Y. Kim, and T. Lim. Optimal Brownian Stopping When the Source and Target Are Radially Symmetric Distributions. *SIAM Journal on Control and Optimization* 58.5 (2020), pp. 2765–2789.

[16] N. Ghoussoub, Y. Kim, and A. Z. Palmer. PDE methods for optimal Skorokhod embeddings. *Calc. Var. Partial Differential Equations* 58.3 (2019), Paper No. 113, 31.

[17] G. Guo, X. Tan, and N. Touzi. On the monotonicity principle of optimal Skorokhod embedding problem. *SIAM Journal on Control and Optimization* 54.5 (2016), pp. 2478–2489.

[18] P. Henry-Labordère, J. Obłój, P. Spoida, and N. Touzi. The maximum maximum of a martingale with given n marginals. *Ann. Appl. Probab.* 26.1 (2016), pp. 1–44.

[19] D. Hobson. Robust hedging of the lookback option. *Finance and Stochastics* 2 (4 1998), pp. 329–347.

[20] D. Hobson. The maximum maximum of a martingale. *Séminaire de Probabilités, XXXII.* Vol. 1686. Lecture Notes in Math. Berlin: Springer, 1998, pp. 250–263.

[21] D. Hobson. The Skorokhod embedding problem and model-independent bounds for option prices. *Paris-Princeton Lectures on Mathematical Finance 2010.* Vol. 2003. Lecture Notes in Math. Springer, Berlin, 2011, pp. 267–318.

[22] D. Hobson and J. Pedersen. The minimum maximum of a continuous martingale with given initial and terminal laws. *Ann. Probab.* 30.2 (2002), pp. 978–999.

[23] S. Jacka. Doob's inequalities revisited: A maximal H^1-embedding. *Stochastic processes and their applications* 29.2 (1988), pp. 281–290.

[24] J. Kiefer. Skorohod embedding of multivariate RV's, and the sample DF. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 24.1 (1972), pp. 1–35.

[25] R. M. Loynes. Stopping times on Brownian motion: Some properties of Root's construction. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 16 (1970), pp. 211–218.

[26] M. Beiglböck, A. Cox, and M. Huesmann. Optimal Transport and Skorokhod Embedding. *Invent. Math.* 208.2 (2017), pp. 327–400.

[27] J. Obłój. The Skorokhod embedding problem and its offspring. *Probab. Surv.* 1 (2004), pp. 321–390.

[28] J. Obłój and P. Spoida. An Iterated Azéma-Yor Type Embedding for Finitely Many Marginals. *Ann. Probab., to appear* (2016).

[29] E. Perkins. The Cereteli-Davis solution to the H^1-embedding problem and an optimal embedding in Brownian motion. *Seminar on stochastic processes, 1985.* Springer. 1986, pp. 172–223.

[30] D. H. Root. The existence of certain stopping times on Brownian motion. *Ann. Math. Statist.* 40 (1969), pp. 715–718.

[31] H. Rost. Skorokhod stopping times of minimal variance. *Séminaire de Probabilités, X (Première partie, Univ. Strasbourg, Strasbourg, année universitaire 1974/1975).* Berlin: Springer, 1976, 194–208. Lecture Notes in Math., Vol. 511.

[32] H. Rost. The stopping distributions of a Markov Process. *Invent. Math.* 14 (1971), pp. 1–16.

[33] A. V. Skorohod. Issledovaniya po teorii sluchainykh protsessov (Stokhasticheskie differentsialnye uravneniya i predelnye teoremy dlya protsessov Markova). Izdat. Kiev. Univ., Kiev, 1961, p. 216.

[34] A. V. Skorokhod. Studies in the theory of random processes. Translated from the Russian by Scripta Technica, Inc. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1965, pp. viii+199.

[35] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.* 36 (1965), pp. 423–439.

[36] P. Vallois. Le probleme de Skorokhod sur $\mathbb{R}$: une approche avec le temps local. *Séminaire de Probabilités XVII 1981/82*. Springer, 1983, pp. 227–239.

Disease Momentum: Estimating the Reproduction Number in the Presence of Superspreading

K. D. Johnson, M. Beiglböck, M. Eder, A. M. Grass, J. Hermisson, G. Pammer, J. Polechová, D. Toneian and B. Wölfl

Abstract

A primary quantity of interest in the study of infectious diseases is the average number of new infections that an infected person produces. This so-called reproduction number has significant implications for the disease progression. There has been increasing literature suggesting that superspreading, the significant variability in number of new infections caused by individuals, plays an important role in the spread of SARS-CoV-2. In this paper, we consider the effect that such superspreading has on the estimation of the reproduction number and subsequent estimates of future cases. Accordingly, we employ a simple extension to models currently used in the literature to estimate the reproduction number and present a case-study of the progression of COVID-19 in Austria. Our models demonstrate that the estimation uncertainty of the reproduction number increases with superspreading and that this improves the performance of prediction intervals. Of independent interest is the derivation of a transparent formula that connects the extent of superspreading to the width of credible intervals for the reproduction number. This serves as a valuable heuristic for understanding the uncertainty surrounding diseases with superspreading.

Keywords: COVID-19; reproduction number; overdispersion; superspreading

1 Introduction

The reproduction number, $R_t \equiv R$, gives the average number of new infections caused by a single infected person throughout the infectious period. In contrast to the basic reproduction number R_0, which describes the reproduction of the virus in a naïve, unmitigated population, R (sometimes called the *effective* reproduction number) varies through time as the epidemic develops and the opportunities for transmission change due to, for example, behavioral response, seasonality, and changes in the relative size of the susceptible population. In every population, some individuals will cause considerably more infections than others - a phenomenon known as *superspreading.* It can be quantified using a framework provided by [13]. In this paper, we extend the model of [4] to include the phenomenon of superspreading. Our goal is to better quantify the uncertainty inherent in this type of estimate of R, *not* to derive a more accurate estimate.

Ultimately we are interested in the estimation of R and specifically the question whether, given current case numbers, we can claim with statistical guarantees that $R \leq 1$ or $R > 1$. Given the growing body of evidence about the existence and importance of superspreaders [2, 12], we incorporate this feature into our models. We observe two important effects: first, it becomes increasingly difficult to accurately estimate the reproduction number R in the presence of superspreading; second, models with superspreading produce prediction intervals for new cases that have improved coverage compared to those without superpreading. Both of these are demonstrated in our Austrian case-study in Section 3. In particular, it becomes infeasible even in early May to support the claim that $R < 1$ using our methods. This is a critical period of time as it coincides with the removal of lockdown restrictions in Austria.

In particular, the width of a credible interval for R should decrease as a function of total number of cases used during estimation and increase with the extent of superspreading. Let S be the set of days used to estimate R in the nowcasting framework presented in Section 1.1 and assume that the (average) reproduction number does not change over time. One would then expect that a $(1-\alpha)\%$ credible interval to have width approximately equal to

$$\frac{2z_{1-\alpha/2}}{\sqrt{k\sum_{s\in S} I_s}}, \tag{1}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and for values of dispersion parameter k much smaller than 1, which corresponds to scenarios with high superspreading. We derive this exact functional form in a simplified model introduced in Section 2.2.

1.1 Nowcasting

The goal of nowcasting is to get accurate estimates of the current state of an epidemic. Given that our observed infections are random observations from an underlying process, our goal is to understand the parameters of that process, particularly with respect to the reproduction number. In addition, we define a time-varying parameter we call the "momentum" of an epidemic, which is a *random* realisation of population infectiousness at a time-point which accounts for superspreading. This is introduced formally in Section 2.1.

Benchmark methods for estimating the reproduction number R include those of [4] and [20]. The method of [4] provides near real-time estimation of R and is implemented in the R software package 'EpiEstim'. An improvement of this framework is given in [18] which accounts for variability in the generation interval (defined below). A substantial extension of the EpiEstim-package ('EpiNow') was developed by a group of researchers at the London School of Hygiene and Tropical Medicine [1]. The method of [20] provides an alternate estimate for historical values of R. Contrary to the methods discussed in this paper, it requires observations from both before and after the time point at which an estimate for R is desired. An important overview of other estimation methods and challenges due to COVID-19 is given in [9] and a comparative analysis of statistical methods to estimate R is given in [16]. If the epidemic is at an early stage, the reproduction number R and the rate of exponential growth are connected by the Euler-Lotka equation [15, 19].

As we follow the framework of [4], we briefly describe their basic model. Let I_0 be the number of initial infections and $I_1, I_2, \ldots$ be the number of new infections on days $1, 2, \ldots$. By $(w_n)_{n\geq 1}$ we denote the *generation interval distribution.* If J_m denotes the number of people infected by a specific person on the m-th day after this person got infected, then we have for $m \in \mathbb{N}$

$$w_m = \frac{\mathbb{E}[J_m]}{\sum_{l=1}^{\infty} \mathbb{E}[J_l]}.$$

We assume that a newly infected individual does not cause secondary cases on the same day, corresponding to $w_0 = 0$. The generation interval can be interpreted as the infectiousness profile of infected persons.

The basic model of [4] assumes that the stochastic process of total new infections on day t, $(I_t)_{t\in\mathbb{N}}$, satisfies

$$I_t \sim \text{Poisson}\left(R_t \sum_{m=1}^{t} I_{t-m} w_m\right), \tag{2}$$

for a sequence of numbers R_t, $t \in \mathbb{N}$. In practice it is often assumed that the generation interval distribution is given as a Gamma distribution that has been discretised in such a way that $w_m = 0$ for all m larger than some cut-off number ν [9]. As a result, the sum in (2) will only have $\nu \in \mathbb{N}$ summands, and to make assertions about I_t we only have to consider the case numbers $I_{t-\nu}, \ldots, I_{t-1}$. As ν is a parameter that can vary between diseases, this term is kept and used throughout our model description in Section 2.1.

When estimating the time-varying reproduction number, [4] assume that the reproduction number has stayed constant over a window of τ days. In this case, for $s \in (t - \tau + 1, \ldots, t)$, equation (2) simplifies to

$$I_s \sim \text{Poisson}\left(R \sum_{m=1}^{\nu} I_{s-m} w_m\right). \tag{3}$$

In order to treat R as fixed in the above expression, it is necessary to only explicitly model a subset of time points, lest R be assumed constant over all time points.

Note that the reproduction number in the sense of (3) does not denote the number of people that actually have been infected by a given individual, but rather describes what one would expect in an "average" evolution of the epidemic. Furthermore, while $R = R_t$ is assumed to be constant over the window of width τ, as this window moves through time the method produces *estimates* of R that slowly vary over time.

1.2 Heterogeneity in reproduction numbers

The motivation for our hierarchical Bayesian approach follows the framework of superspreading provided in [13]. Even if the reproduction number R is constant over a small window of time, it might vary between individuals. We consider the reproduction number of a specific person with index x to be drawn randomly as

$$r_x \sim \text{Gamma}(k, \text{rate} = k/R). \tag{4}$$

This distribution has mean R and variance R^2/k. Note that the above gamma distribution will also be referred to as having dispersion parameter k. The degenerate case $k = \infty$ corresponds to the deterministic case where $r_x = R$ for all individuals and leads to the model in (3). Given $r_x = r$, this person causes Poisson(r) new infections. If one integrates out the Poisson parameter r, one is left with the unconditional number of descendants which follows a negative binomial distribution with mean R and variance $R + R^2/k$. This negative binomial model is further analyzed in Section 2.2.

A basic extension of (3) that follows the concept of random individual reproduction numbers in the sense of [13] is to assign, on day t, the individual reproduction numbers $r_1^t, \dots, r_{I_t}^t$ to the I_t individuals that got infected on this day. This leads to the recursion

$$I_t \sim \text{Poisson}\left(\sum_{m=1}^{\nu} w_m \sum_{x=1}^{I_{t-m}} r_x^{t-m}\right), \tag{5}$$

where the individual reproduction numbers r_x^m are drawn i.i.d. according to (4). Note that for the degenerate case $k = \infty$, (5) recovers (3). This forms the foundation of the model explained in detail in Section 2.1.

The theme of the present paper is close to that of [5], in which heterogeneity in R between *groups* is explicitly modeled. While the high-level descriptions of these models sound nearly identical, those models are relevantly different than ours. In particular, [5] are interested in estimating group-specific or time-varying reproduction numbers for different geographical regions and age groups. On one hand, with sufficient group-specific data, this provides tools of a much broader scope than we present here; on the other hand, it is assumed that within-group variability is negligibly small. Instead, we focus on aggregate data from a *single* geographical region but do *not* assume that individual variability is negligible. Rather, this is precisely the variability we are interested in modeling. Furthermore, our critiques of the estimability of the reproduction number transfers to their setting as well: if within-group variability exists, group-specific reproduction numbers are more difficult to estimate than previously acknowledged.

2 Methods

This section introduces two methods. First, the "momentum" model formulates the estimation problem as a Bayesian Poisson regression. Second, the "generation" model is a simplification which provides a fast approximation to the momentum model as well as an explicit formula for dependence of credible interval width on k. Both are of interest beyond COVID modeling and aim to address different goals: precise estimation (momentum) and valuable speed and heuristics (generation).

2.1 The "momentum" model

As mentioned in the introduction, we identify an unobserved random variable which we term the "momentum" of the epidemic. This follows from a simple notational change in (5) according to the observation that a sum of i.i.d. Gamma random variables is also Gamma distributed with the same dispersion parameter. We rewrite (5) as

$$I_t \sim \text{Poisson}\left(\sum_{m=1}^{\nu} w_m \theta_{t-m}\right), \tag{6}$$

where

$$\theta_t = \sum_{x=1}^{I_t} r_x^t \sim \text{Gamma}(I_t k, \text{rate} = k/R). \tag{7}$$

The terms $(\theta_t)_{t \geq 0}$ are collectively referred to as the "momentum" of the disease. They will be treated as a set of nuisance parameters of the offspring distribution, as our primary interest lies in estimating the reproduction number R. In our Bayesian framework introduced below, R is a hyperparameter of the prior distribution for $(\theta_t)_{t\geq 0}$. Equation (6) describes the distribution of I_t conditioned on its whole past, i.e., $I_s, \theta_s,\ s < t$. Analogously, equation (7) describes θ_s given its history. The difference in what we understand as the relative past originates from θ_t being conceptually determined "after" I_t.

For increased clarity of the form of the model and the estimation methods required, we recast our model as a Bayesian Poisson regression using vector notation. This is made painfully explicit by using an arrow as in $\vec{I}$ for vectors. Following [4], we estimate R by explicitly modeling a set of τ days over which we assume R to be constant. We specify the regression function for each observation in this estimation window. To condense notation, we use $[l]$, for $l \in \mathbb{N}$, to be the vector $(1, 2, \ldots, l)$. Similarly, $[l, m]$ for $l, m \in \mathbb{N}$ is shorthand for the vector $(l, l+1, \ldots, m)$, i.e., $[l] = [1, l]$. This notation will primarily be used for vector indices. Furthermore, the indices of our vectors increase in time. As such, our generation interval truncated to ν days can be condensely written as $\vec{w}_{[\nu]} = (w_1, \ldots, w_\nu)$. Similarly, the τ observations we model are given by $\vec{I}_{[t-\tau+1,\, t]} = (I_{t-\tau+1}, \ldots, I_t)$.

As a regression model for $\vec{I}_{[t-\tau+1,\, t]}$, equation (6) can be written as

$$\vec{I}_{[t-\tau+1,t]} \sim \text{Poisson}(W\vec{\theta}_{[t-\nu-\tau+1,t-1]}) \quad \text{where} \tag{8}$$
$$W = \begin{pmatrix} w_\nu & w_{\nu-1} & \ldots & w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_\nu & w_{\nu-1} & \ldots & w_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ldots & \vdots \\ 0 & \cdots & 0 & w_\nu & w_{\nu-1} & \ldots & w_1 & 0 \\ 0 & \cdots & 0 & 0 & w_\nu & w_{\nu-1} & \ldots & w_1 \end{pmatrix}$$

In the above expression, we have a fixed covariate matrix W which is a function of the generation interval $w_{[\nu]}$. The momentum parameters $\vec{\theta}_{[t-\nu+\tau-1,t-1]}$ are seen to be the regression parameters to be estimated. Note that the expressions in the previous display suppress the notation for conditioning on all observations before time $t-\tau+1$. Furthermore, given $\vec{\theta}_{[t-1]}$, I_t is independent of $\vec{I}_{[t-1]}$.

We place a prior distribution on $\vec{\theta}$ which depends on R as in equation (7), as well as a hyperprior on R to account for the previously identified uncertainty in the distribution of R as reported in [1]. As we have parameterised the gamma prior on θ_t to have mean $I_t R$, the conjugate hyperprior for R is the inverse-gamma distribution. This is transparent in the posterior distribution given by equation (9) below. Hence we use an inverse-gamma hyperprior on R, where these hyperparameters are set to match the results of [1]. As such, we assume that R has mean 2.6 and standard deviation 2, yielding shape parameter 3.69 and rate parameter 6.994:

$$R \sim \text{Inv-Gamma}(3.69,\ \text{rate} = 6.994).$$

An a priori distribution for R is itself uncertain and one could theoretically place additional hyperpriors on the parameters of this inverse-gamma distribution. That being said, the change would increase computational complexity while introducing hyper-hyperparameters that would be difficult to estimate. Hence, this proposal distribution for R is treated as fixed.

This regression formulation is important as it highlights the latent variables $\vec{\theta}$ that are required to fully determine the generative model. It also focuses attention on which observations are conditioned upon and which are treated as random, i.e., the τ observations to which we fit the model are treated as random. This is relevant as more than τ nuisance parameters are present, namely $\nu + \tau - 1$. Observe that the earliest data point is $I_{t-\tau+1}$, which itself requires a history of ν momentum values of $\vec{\theta}$ to determine.

While we also think of individual reproduction numbers as changing over time due to factors such as changes in social restrictions, the assumption of constant R over a period renders this moot. Likewise, we set k to be a constant for the results presented in Section 3, as k is best estimated with contact tracing data instead of case count data. We set $k = 0.072$, in line with the results of [11], which estimated the extent of superspreading for COVID-19 from Indian data. This is also within the range of parameter values identified in [6].

Alternatively, it is possible to consider an independently estimated distribution for k. To estimate the momentum model with random k, one can merely draw k from a proposal distribution and estimate the momentum model with this fixed value. This process is repeated for many sampled values of k, and the posterior samples for R and I_t from all k are combined. This follows the same methodology as [18], where the generation interval was estimated with a separate data set before fitting model (3) without superspreading. Brief results for this case are presented in B as none of the results change significantly. The joint estimation of k and R within the momentum model appears infeasible as k is the dispersion parameter of the nuisance parameter distribution. This makes learning about k using this data highly challenging.

A full derivation of the posterior distribution of the pair $R, \vec{\theta}_{[t]}$ given $\vec{I}_{[t]}$ is given in A. We obtain as posterior

$$\begin{aligned}
&p(R, \vec{\theta}_{[t-\tau-\nu+1,\,t-1]} | \vec{I}_{[t-\tau-\nu+1,\,t]}) \\
&\propto p(\vec{I}_{[t-\tau+1,\,t]}, \vec{\theta}_{[t-\tau+1,\,t-1]} | \vec{\theta}_{[t-\tau-\nu+1,\,t-\tau]}, \vec{I}_{[t-\tau-\nu+1,\,t-\tau]}, R) p(\vec{\theta}_{[t-\tau]}, R | \vec{I}_{[t-\tau]}) \\
&\propto \left(\prod_{s=t-\tau+1}^{t} \Big(\sum_{m<s} w_{s-m}\theta_m \Big)^{I_s} e^{-\sum_{m<s} w_{s-m}\theta_m} \right) \\
&\quad \cdot \left(\prod_{s=t-\tau+1}^{t-1} \frac{k^{I_s k}}{\Gamma(I_s k) R^{I_s k}} \theta_s^{I_s k - 1} e^{-\frac{k}{R}\theta_s} \right) \left(\prod_{s=t-\nu-\tau+1}^{t-\tau} \frac{k^{I_s k}}{\Gamma(I_s k) R^{I_s k}} \theta_s^{I_s k - 1} e^{-\frac{k}{R}\theta_s} \right) \\
&\quad \cdot \left(R^{-3.69-1} e^{-6.994/R} \right), . \qquad (9)
\end{aligned}$$

The first line of (9) specifies the distribution of the observations given all other parameters, and the third line gives the inverse-gamma prior for R. The second line describes the distribution of $\vec{\theta}$, and we have explicitly partitioned the indices into two sets. The values θ_s in the first index set $[t-\tau+1, t-1]$ require no special discussion as they depend on values I_s which are being explicitly modeled. The values of θ_s in the second index set $[t-\nu-\tau+1, t-\tau]$, however, treat the corresponding I_s values as *fixed* and *constant*. This is done so that we do not need to specify further nuisance parameters before time $t-\tau-\nu+1$. Doing so would create an infinite recursion in historical observations, requiring us to treat R_t as fixed for all t. Hence we need not only a prior for R, but also for $\vec{\theta}_{[t-\tau-\nu+1,\,t-\tau]}$. More details are provided in A.

In order to condense notation for summations in exponents, let S be the index set for the second product; i.e., $S = \{t-\nu-\tau+1, t-\nu-\tau+2, \dots, t-1\}$. The additional shorthand below drops "$s \in$" from $s \in S$. With this notation, the posterior distribution of R given $\vec{\theta}$ and $\vec{I}$ is

$$p(R|\vec{\theta}_{[t-1]}, \vec{I}_{[t]}) \propto R^{-k\sum_S I_s - 3.69 - 1} e^{(-k\sum_S \theta_s - 6.994)R^{-1}},$$

which is Inv-Gamma($k\sum_S I_s + 3.69$, $k\sum_S \theta_s + 6.994$). A perhaps counter-intuitive observation is that the posterior distribution of R does not depend on the generation interval $\vec{w}_{[\nu]}$. This is the result of conditioning on $\vec{\theta}$ versus integrating it out as done in [13]. In our case, it is infeasible to integrate out $\vec{\theta}$ as the dependence is too complex. If we truly know population infectiousness, i.e., the epidemic momentum at all points in time, then $\vec{w}_{[\nu]}$ is irrelevant for estimating R, because $\vec{w}_{[\nu]}$ just determines *how we learn* about $\vec{\theta}$ via (8). More concretely, there are no terms in (9) that include all of R, $\vec{\theta}$, and $w_{[\nu]}$.

The posterior expectation and variance of R are

$$\mathbb{E}[R|\vec{\theta}, \vec{I}] = \frac{k\sum_S \theta_s + 6.994}{k\sum_S I_s + 3.69 - 1} \qquad \text{and}$$

$$\mathrm{Var}[R|\vec{\theta}, \vec{I}] = \frac{(k\sum_S \theta_s + 6.994)^2}{(k\sum_S I_s + 3.69 - 1)^2 (k\sum_S I_s + 3.69 - 2)}.$$

The denominator of the variance picks up an additional k term, making credible intervals wider when k is small. The dependence on $\vec{\theta}$ is difficult to remove in this general setting. Section 2.2 considers a simpler setting in which $\vec{\theta}$ can be integrated out in order to derive a transparent function for credible interval width.

To estimate this model, we alternate between a Gibbs-step to sample R and a Metropolis-Hastings step to sample $\vec{\theta}$. As $\mathbb{E}[\theta_s|I_s, R] = I_s R$, we can initialise reasonable starting values for $\vec{\theta}$ using various values of R such that we require little burn-in. We find total chain length to be the more important tuning parameter for valid prediction and credible intervals. In all models presented in this paper, we set $\nu = \tau = 13$ to make valid comparisons with results from the EpiEstim framework [4]. We set $\vec{w}_{[\nu]}$ to be a discretised gamma distribution with mean 4.46 and standard deviation 2.63 per the results of [17] for Austria, which are similar to values determined elsewhere [7, 10]. Inference is conducted using the 10^6 samples that remain after a burn-in of 1,000 and thinning by 5.

While the majority of the model validation and supporting graphs is relegated to B, we address here the particular concern that we have 25 nuisance parameters in $\vec{\theta}$ for modeling 13 observations. Our simulation evidence indicates that all nuisance parameters are well-estimated, even those far in the past: coverage of $\vec{\theta}$ by credible intervals in simulated data is nearly exact. Furthermore, we see approximate coverage when predicting new cases in Section 3. As such, we do not believe that we are over-fitting the data with a larger number of nuisance parameters. This is in part due to the role of the prior distribution for θ_s. For example, the first nuisance parameter $\theta_{t-\nu-\tau+1}$ only appears in a single observation term in the posterior (9): the distribution of $I_{t-\tau+1}$. Similarly, $\theta_{t-\nu-\tau+2}$ only appears in two, etc. The prior therefore plays a larger role in determining the values of these parameters.

2.2 Generation model

In order to directly relate the dispersion parameter k to the width of the credible interval and to provide a fast approximation to the momentum model, we consider the trivial generation interval in which an infected person is only infectious for a single day. For real data, this assumption is obviously inaccurate. Therefore, we switch to modeling infections per generation instead of infections per day. While we model generations spanning multiple days, we estimate and forecast cases for conventional days.

When the generation interval w is of this form, $\vec{w}_{[1]} = (1)$, the model is purely Markovian and the data follow a Galton-Watson process. Recall that a Poisson(λ)-distributed random variable Y, where λ is distributed according to Gamma(α, β), follows a negative binomial distribution [13]:

$$Y \sim NB\left(\alpha, \frac{1}{1+\beta}\right), \quad p(Y) = \frac{\Gamma(Y+\alpha)}{Y!\Gamma(\alpha)}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{Y}. \tag{10}$$

Applying (10) and $\vec{w}_{[1]} = (1)$ to the momentum model (6) yields the following distribution for the infections I_t:

$$I_t|\vec{I}_{[t-1]}, R, k \sim NB\left(kI_{t-1}, \frac{R}{R+k}\right), \tag{11}$$

$$p(I_t|\vec{I}_{[t-1]}, R, k) = \frac{\Gamma(I_t + kI_{t-1})}{I_t!\Gamma(kI_{t-1})}\left(\frac{k}{R+k}\right)^{kI_{t-1}}\left(\frac{R}{R+k}\right)^{I_t}. \tag{12}$$

In C, we reparameterise this model in terms of $\frac{R}{R+k}$ in order to place a suitable prior which mimics that of the momentum model. After transforming the resulting posterior back to a distribution for R and using standard normal approximation techniques [8], we derive a normal approximation of the posterior of

$$p(R|\vec{I}_{[t]}, k) \approx N\left(\frac{k(\alpha-1)}{\beta+1}, \frac{k^2(\alpha+\beta)(\alpha-1)}{(\beta+1)^3}\right).$$

where

$$\alpha = 98.82 + \sum_{s=t-\tau+1}^{t} I_s \quad \text{and} \quad \beta = 3.74 + k \sum_{s=t-\tau}^{t-1} I_s.$$

We are interested in the setting in which $R \approx 1$ and $\beta \approx k \cdot \alpha$. Note that $\sum_{s=t-\tau+1}^{t} I_s$ and $k \sum_{s=t-\tau}^{t-1} I_s$ are of this approximate ratio: the terms in these two sums almost entirely overlap. Furthermore, while the hyperparameters (98.82 and 3.74) are of moderate size, they also approximately satisfy the desired ratio. This yields the following simplification of the variance of the normal approximation:

$$\frac{k^2(\alpha+\beta)(\alpha-1)}{(\beta+1)^3} \approx \frac{k^2\alpha^2(k+1)}{k^3\alpha^3} = \frac{k+1}{k\alpha} \approx \frac{1}{k\sum_{s=t-\tau+1}^{t} I_s}.$$

Hence, the approximate length of a credible interval for R behaves like

$$\frac{2z_{1-\alpha/2}}{\sqrt{k\sum_{s=t-\tau+1}^{t} I_s}}.$$

It is clear that the assumption $\nu = 1$ is highly unrealistic for COVID-19 and most other diseases. In order to bridge this gap, we estimate the model for non-overlapping generations instead of conventional days. The length of a generation is set equal to the mean of the generation interval, i.e.,

$$D_g := \sum_{t=0}^{\nu} t\omega_t.$$

Given the modeling assumptions we have made for COVID-19, a generation comprises approximately 4.87 conventional days. The first 4.87 days after infection also accounts for 64% of the assumed infectiousness given by the generation interval. This helps explain why partitioning the data into generations produces reasonable results. When a model is defined over generations, setting $\nu = 1$ is equivalent to assuming that someone is equally infectious over D_g days. The negative binomial model estimated using generations is approximately equivalent to the momentum model estimated using conventional days.

In order to account for non-integer-valued generations, consider $D_g = \lfloor D_g \rfloor + D_{frac}$, where $D_{frac} \in [0,1)$. For simplicity, we assume that new infections are uniformly distributed during the day so that we may use standard data with records of new daily cases. In order to not confuse subscripts indexing days and generations, times in the generation model will be indicated by $\tilde{t}$ instead of t. Lastly, as we are interested in using the most recent data, we care about matching the right endpoint of our time series. As such, we compute the generations *backwards* from a reference day t.

Let day t be the maximal day in our data set. We define the corresponding generation incidence, $\tilde{I}_{\tilde{t}}$, to be

$$\tilde{I}_{\tilde{t}} = \sum_{s=0}^{\lfloor D_g \rfloor - 1} I_{t-s} + D_{frac} \cdot I_{t-\lfloor D_g \rfloor}.$$

This is merely the sum over $\lfloor D_g \rfloor$ full days, and a proportion of the remaining day. Infections for previous generations then sum similarly over the historical data such that the generations form a partition of days in our data set.

As before, some mathematical details are moved to C. With simple notational changes, however, we derive a model for generations which looks functionally identical to (11), i.e.,

$$\tilde{I}_{\tilde{t}} | R, \tilde{I}_{\tilde{t}-1} \sim NB\left(\tilde{I}_{\tilde{t}-1} k, \frac{R}{k+R}\right).$$

This formula can then be used to forecast the cumulative incidence over several generations as described in C. This yields a simple, closed form approximation of the momentum model without resorting to costly Bayesian computation methods.

3 Results

This section focuses on understanding the evolution of the reproduction number in Austria between April 1 and October 31, 2020. As the momentum model effectively needs $\tau+\nu$ observations to be fit, this is approximately as early as estimates can be provided for Austria. Our goals are three-fold: to demonstrate the increase in estimated variability of R due to superspreading, to provide valid prediction intervals for new cases, and to compare to similar models without superspreading. Some results will be shown for Croatia and Czechia as well to help establish the validity of our method, but the focus in on Austrian data. Other supporting graphs for Croatia and Czechia are given in D.

An important component of estimating the reproduction number on a given date is to account for the delay distribution between date of infection and date of confirmation as discussed in [9]. If a delay of length d occurs between infection and confirmation, then an infection observed at time t actually occurred on day $t-d$. In this case, we have a "true infection history" that is distinct from the reported case numbers. In reality, the delay d is random. [1] estimate and sample true potential infection histories given observed case numbers by sampling possible delays d. As our primary goal is to understand the uncertainty in estimating R as opposed to providing best in class predictions of R for a given date, we ignore this complication. This allows us to take as model input the historical 7-day moving average of reported cases and to compare methods with simple, transparent input. As a result, however, we are not attempting to predict the number of true infections on a given date. Instead, we are predicting the number of reported or confirmed cases on this date. In order to highlight this, axes are explicitly labeled with "Reported Cases" and "Confirmation Date".

Data on the progression of COVID-19 in Austria is shown in Figure 1. This graph includes curves for the raw infection data as reported by the European Center for Disease Prevention and Control (Raw), the 7-day moving average of Raw (Raw (MA)), each sampled infection history (Sampled Inf.), and the daily median of the sampled infection histories (Sampled Inf. (M)). Observe that the boundary of the "band" created by the sampled infection histories is not smooth, as it is created from 1,000 distinct faded lines. Note that using sampled infection histories effectively shifts the time series backward in time. In order for the infection histories to approximately match the reported case numbers, we have aligned them in time.

As mentioned in Section 2.1, we sample one million total samples of R and the momentum vector $\vec{\theta}$. To forecast future cases, we use an individual sample of parameters and run the momentum model for a specified period of time. Our graphs show results for the average number of new cases over the following week. As such, they are on the scale of daily reported cases. There is no additional smoothing of the raw data or predictions. As our input is the 7-day moving average, our prediction is the 7-day-ahead forecast of this moving average.

In all of the following graphs, we plot predictions and intervals from three models: the momentum model with $k = 0.072$, the generation model of Section 2.2 with $k = 0.072$, and the EpiEstim model of [4]. As mentioned previously and visible in D, treating k as random within a relevant region does not alter our results. We label the EpiEstim model "Epi*", as the estimates are produced directly via equation (13) below instead of using the EpiEstim R package. As in [4], we fix a generation interval, as opposed to taking samples of a generation interval estimated from a separate data source as in [18]. As a result, we are not comparing to the best in class model within the EpiEstim/EpiNow framework, but with a model of corresponding complexity to the momentum model. Other improvements to the modeling framework could then be built on top of the momentum model as they have been for the model of [4].

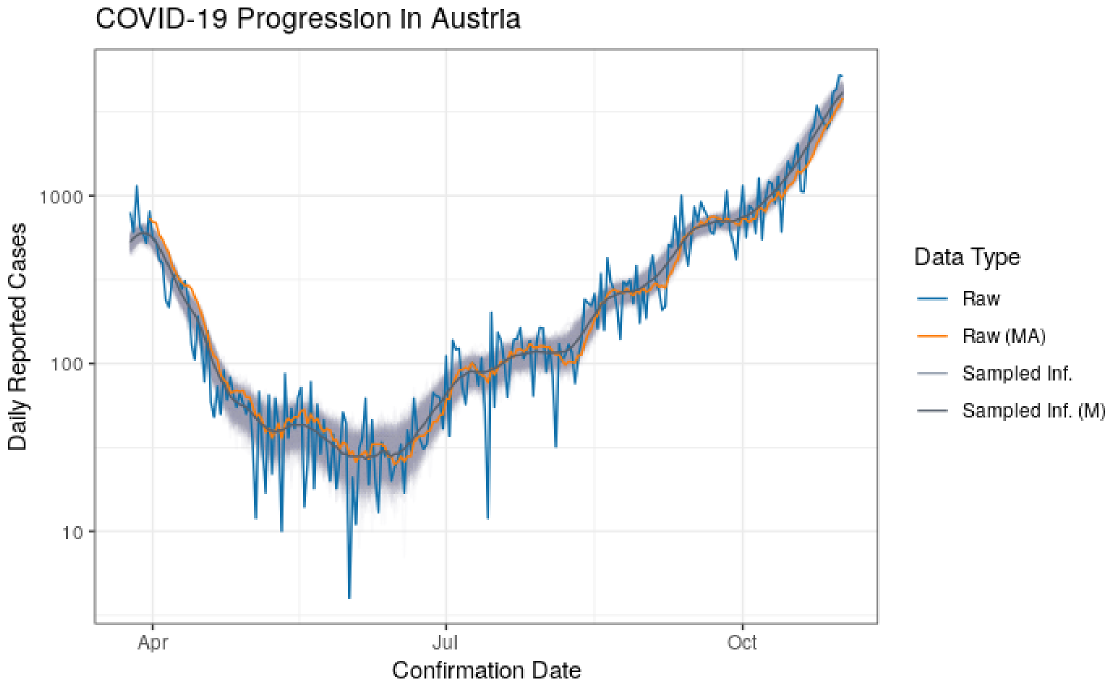


Figure 1: Summary of new cases of COVID-19 in Austria: raw infection data (Raw), the 7-day moving average of Raw (Raw (MA)), each sampled infection history (Sampled Inf.), and the daily median of the sampled infection histories (Sampled Inf. (M)).

To estimate the model of [4], we estimate the parameters of the [4] posterior distribution directly from the infection data:

$$p(R_t|I_{[t]}) = \text{Gamma}\left(a + \sum_{s=t-\tau+1}^{t} I_s, \text{ rate} = b + \sum_{s=t-\tau+1}^{t}\sum_{m=1}^{\nu} w_m I_{s-m}\right) \tag{13}$$

where a and b are the shape and rate parameter of the gamma prior distribution on R. We estimate this posterior distribution, draw one million samples for R, and run the corresponding data generating process (2) for the required number of days.

Figure 2 shows the difference between models with and without superspreading on Austrian data. In order to show a long time period, the data must be plotted on a logarithmic scale such that the low cases in the summer months are visible. As this distorts the plotting of prediction intervals in the same graph, the comparison of prediction intervals is given separately by focusing attention on the summer months between the effective end of COVID restrictions and the start of the school year.

For reference, we marked the dates of important changes in COVID-19 restrictions in Austria as vertical, dashed lines. A complete list is available at `https://regiowiki.at/wiki/Chronologie_der_Corona-Krise_in_sterreich` (in German). The events are described in Table 1. When comparing the events to both reported cases and the estimated reproduction number in Figure 3, it is necessary to keep the delay distribution in mind; i.e., the effect of an intervention will not be visible in confirmed cases and thereby the estimated reproduction number for roughly two weeks [1]. Prior to the removal of any lockdown restrictions, reported case numbers were decaying exponentially. This is visible as a linear decrease given the logarithmic scaling of the y-axis. The slope of this line changed substantially around the time that Austria began to reopen in May and June. From approximately July through the end of October, case numbers fluctuate between growing exponentially and brief periods of relative stability. These fluctuations are not modeled

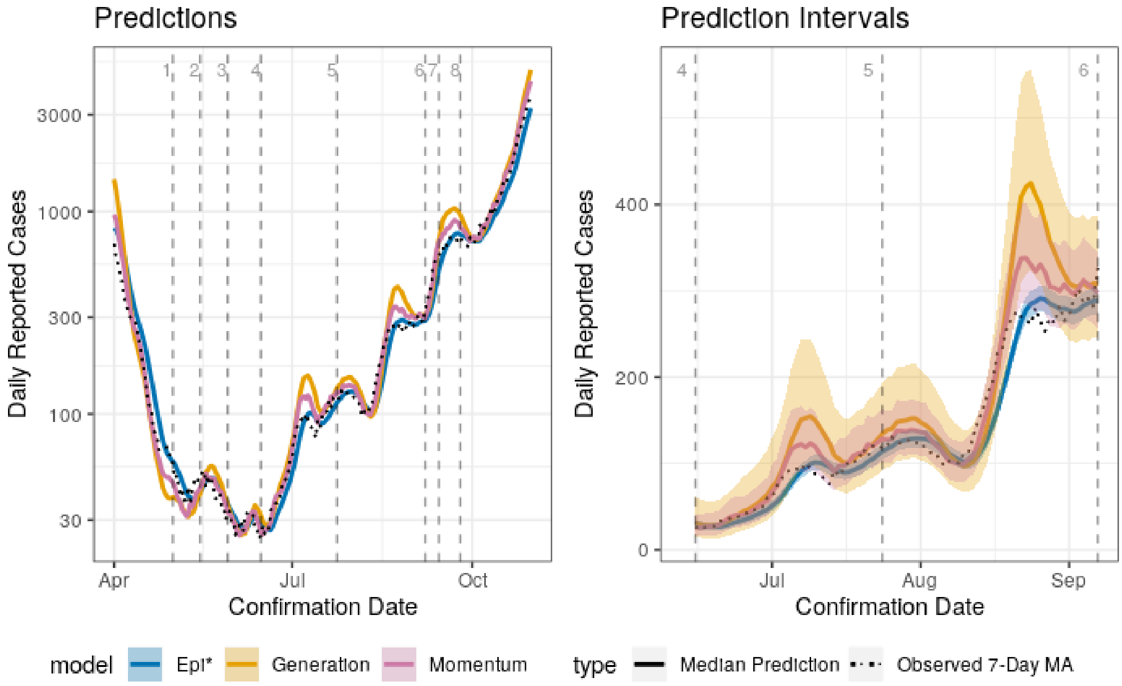


Figure 2: Predictions between April 1 and October 31, 2020, and 90% prediction intervals between two significant dates: June 15 and September 7, 2020. Predictions and intervals are for the 7-day average of new cases in the following week in Austria. Relevant event dates are given as vertical, dashed lines and are described in Table 1. The Epi* predictions consistently lag behind the observed values, whereas the other methods overshoot in the peaks due to momentum. Models with superspreading produce predictions intervals 2-3 times as wide as those without and achieve better coverage.

Table 1: Dates of important events related to COVID-19 in Austria. Changes which occur in large parts of the country but not uniformly are listed as occurring in "some regions".

Label	Date	Event
NA	2020-03-16	Start of general lock down
1	2020-05-01	Begin relaxation of movement restrictions
2	2020-05-15	Bars and restaurants can open
3	2020-05-29	Hotels and cultural sites can open
4	2020-06-15	Near complete removal of COVID restrictions
5	2020-07-24	Face masks mandatory in essential businesses
6	2020-09-07	Start of school year in some regions
7	2020-09-14	Face masks mandatory
8	2020-09-25	Bars and restaurants close early in some regions
NA	2020-11-03	Start of general soft lock down

Table 2: Coverage of the 50% and 90% prediction intervals (PI) for 7-day-ahead predictions of the 7-day moving average. Models with superspreading improve coverage significantly over that of Epi*.

Country	Model	Coverage, 50% PI	Coverage, 90% PI
Austria	Momentum, k = 0.072	0.46	0.79
	Generation, k = 0.072	0.47	0.73
	Epi*, k $\to \infty$	0.16	0.38
Croatia	Momentum, k = 0.072	0.48	0.85
	Generation, k = 0.072	0.49	0.77
	Epi*, k $\to \infty$	0.18	0.47
Czechia	Momentum, k = 0.072	0.40	0.69
	Generation, k = 0.072	0.39	0.66
	Epi*, k $\to \infty$	0.12	0.32

and reflect both noise as well as features which we do not include in our analysis, e.g., common holiday periods, changes in testing, etc. Throughout this period, some restrictions are brought back into effect without apparent substantial impact. Lockdown measures were reinstated at the end of the plotted window of time.

While all of the prediction curves track the observed cases, there are subtle but significant differences in behavior. If one looks closely, one can see that the Epi* model predictions lag behind the observed 7-day moving average: it fails to accurately estimate the rapid changes in case numbers. On the other hand, the momentum and generation model predictions "overshoot" the peaks in the time series. As the name suggests, there appears to be excess "momentum" in the process around these change points, and the model anticipates cases to continue rising as in the previous days.

The various models produce prediction intervals with drastically different widths. Most notably, the intervals for the momentum model with $k = 0.072$ are much wider than those of Epi*. The generation variant of this model produces intervals which are wider still. The momentum intervals are, on average, approximately three times as wide as those of Epi*. While the generation model provides a computationally cheap and fast estimate, it is clear that it suffers relative to the momentum model in terms of interval length. The ratio between the prediction interval lengths visible during the summer months is approximately the same throughout the entire prediction period.

To assess the validity of the prediction intervals, Table 2 shows, for each method, the proportion of true weekly new cases that fall within the prediction intervals over the prediction period. Coverage is shown for the 50% and 90% prediction intervals for the raw infection data. When cases are steadily increasing (or decreasing) prediction intervals become narrower, and when the behavior changes they become considerably wider. The prediction intervals of the momentum model cover the true values during periods of growth, while those of Epi* often fail to do so over the entire growth period. Clearly coverage is still not exact, and all models perform worse on the Czech data (see D). It is still notable that the momentum models provide approximate coverage in these cases even with the inherent messiness of the COVID-19 case data. For example, Czechia had a much higher test positivity rate than Austria and Croatia during the majority of the prediction period, which is ignored in our model.

As the reproduction number is unobserved, we are unable to compare our predictions within a supervised setting as we compared our model forecasts. Given the previous discussion though, we see that the additional variability provided by the momentum model is needed to provide prediction intervals with approximate coverage. Figure 3 shows the median predictions and 90% credible intervals for R given by the momentum, generation, and Epi* models. Intervals are, in

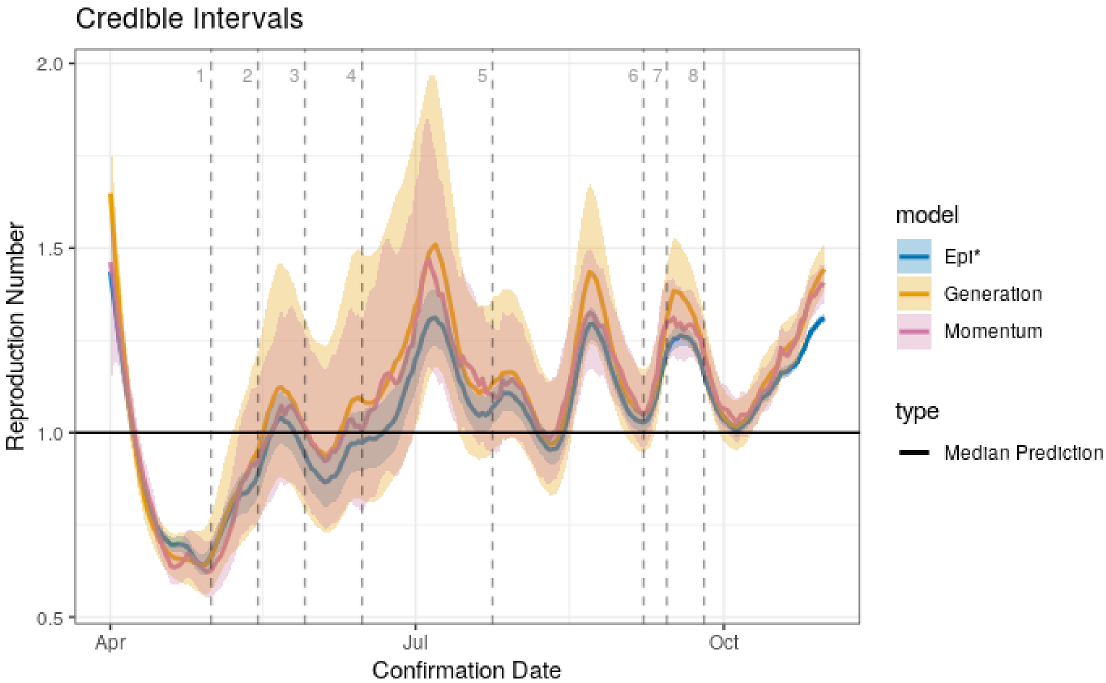


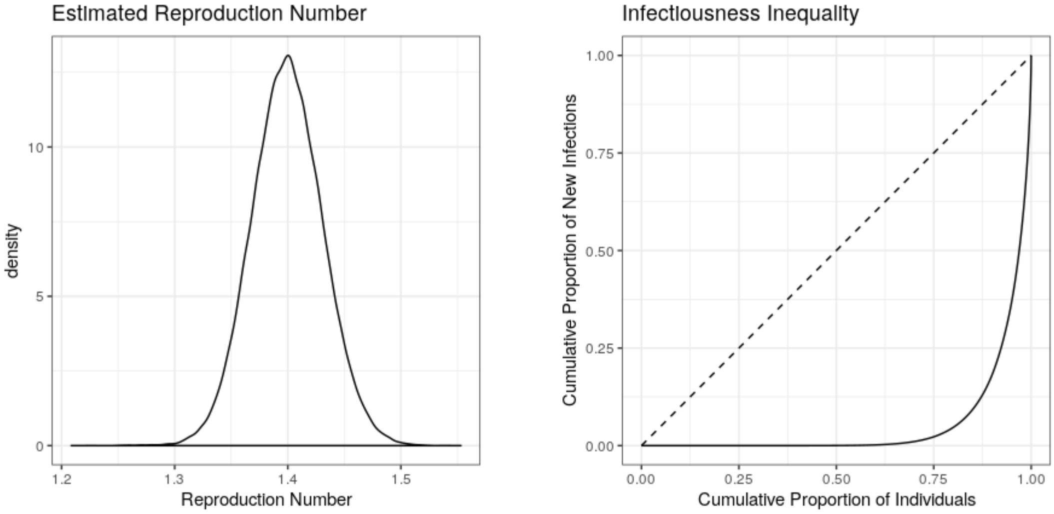
Figure 3: Credible intervals for R in Austria. The momentum and generation model predictions are consistently slightly higher than those of Epi*. They also produce credible intervals that are 2-3 times as wide. Relevant event dates are given as vertical, dashed lines and are described in Table 1. Observe that R becomes indistinguishable from 1 using our models around the time when lockdown restrictions begin to be removed.

general, asymmetric, and skewed toward higher values. The figure clearly demonstrates that the intervals for R are drastically different: with superspreading, intervals for R are roughly 2-3 *times* as wide as those without. This could have potentially large implications for policy making as we know that relatively small changes in the size of R can lead to large differences in the number of new cases if the disease is allowed to progress unchecked.

Near the beginning of our estimation period and around the time when restrictions were being relaxed in Austria, it quickly becomes infeasible to claim that the reproduction number is below 1; i.e., the credible intervals estimated during May and June include the value 1. Beginning in July and August, however, we observe long periods with reproduction numbers significantly greater than 1, even with our comparatively wide credible intervals. As before, there is a delay of approximately two weeks between when these interventions occur and any change in reproduction number could be observed. Hence any discussion of dates should be interpreted loosely.

As we see a clear improvement in coverage for switching to a model with superspreading, it is useful to have a clearer understanding of the degree of heterogeneity implied by our models. To do so, we consider the posterior samples of R from October 31, 2020. According to equation (4), each individual has a separate reproduction number, r_x, given the population reproduction number R. For each posterior sample of R, we therefore draw an individual r_x and secondary infections I_x. The Epi* models of [4] set $r_x = R$ for all individuals. Hence, it is possible to compare the degree of heterogeneity by considering a Lorenz curve of the population of values of r_x or I_x [14].

The Lorenz curve is typically used to demonstrate income inequality by showing the proportion of overall income or wealth held by the bottom x% of the people. Here we consider this to be "infectiousness inequality". The distribution of R estimated for October 31, 2020 as well as the



(A) Estimated distribution of R (B) Lorenz curve of r_x

Figure 4: Momentum model estimates of R and individual heterogeneity for October 31, 2020. 10% of individuals are expected to contribute approximately 84.6% of new infections. The dashed curve in (b) corresponds to a model without superspreading (Epi*).

implied Lorenz curve are shown in Figure 4. The Lorenz curve is a representation of the cumulative distribution function of the number of new expected infections. It allows us to visualise the degree of heterogeneity by seeing which proportion of individuals contribute to new infections. One can draw the Lorenz curve with I_x instead of r_x, which only results in a slightly rougher image with no qualitative differences.

While the population reproduction number is moderately high, this is largely driven by superspreading. The momentum model implies that the top 10% of individuals contribute 84.6% of new infections, while the top 20% contribute 98%. The usefulness of Figure 4 B is that is shows this entire distribution instead of these two common quantiles. We can clearly see that essentially no new cases are produced by nearly 75% of infected individuals. These statistics match quite closely the observed values reported in [3]. The figures can also be drawn for the estimation setting in which k is assumed to be randomly drawn from an appropriate gamma distribution. The resulting graphs look essentially identical. As such, treating k as fixed at 0.072 or fluctuating in the approximate range [.04, .2] makes little difference in the infectiousness inequality implied by the momentum model.

4 Conclusion

In this paper, we provide a simple extension of the [4] model to account for superspreading. While we explicitly use this to model the COVID-19 pandemic, the methods are easily adaptable to other diseases where superspreading is present. This "momentum" model incorporates unobserved random variables which drive the process of new infections. Even if case numbers and R are relatively small, the presence of superspreaders can increase the momentum of the disease beyond what would be expected if all individuals have the same infectiousness. We observe that this appears necessary to properly track the steep increases or decreases in reported COVID-19 cases. The momentum model produces credible intervals and posterior predictive intervals that are approximately 2-3 times as wide as those that neglect superspreading. We find that these wider intervals significantly improve the coverage of the prediction intervals. The heterogeneity in infectiousness implied by the momentum model is extremely high: 10% of individuals contribute approximately 84.6% of new infections.

As Bayesian models are time and resource intensive to estimate, we also derive a simplified model in which infected individuals are only infectious for a single day. In order to improve the fit to real data, we partition disease incidence into generations, each of which spans multiple days. The length of each generation corresponds to the generation time of the disease, and within this period an infected person is assumed to be equally infectious. This yields two main benefits. First, estimation of R and predictions of new cases are immediately available through an explicit approximation of the posterior distribution of R. Second, this model allows us to derive a simple equation to relate the width of credible intervals to the degree of superspreading. Hence, we have rigorous analysis which supports the heuristic that the approximate length of a credible interval for R behaves like

$$\frac{2z_{1-\alpha/2}}{\sqrt{k\sum_{s=t-\tau+1}^{t} I_s}},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and for values of dispersion parameter k much smaller than 1, which corresponds to scenarios with high superspreading. The model assumes that R has been constant for the preceding τ days.

A Likelihood derivation

This appendix derives the posterior distribution of R and $\vec{\theta}_{[t-\tau-\nu+1,t-1]}$ given the relevant observable past, i.e., $\vec{I}_{[t-\tau-\nu+1,t]}$. We briefly restate some basic properties and definitions of our model.

Let w_i denote the expected proportion of future infections caused by an infected person which occur on day i after infection. Let ν denote the length of infectiousness, i.e., $w_{\nu+k}=0$ for all $k>0$. Lastly, τ denotes the number of days over which we assume R is constant.

Our distributional assumptions are as follows:

$$\begin{aligned}
I_t|\vec{\theta}_{[0,t-1]},\vec{I}_{[0,t]},R &\sim \text{Poisson}\left(\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right), \text{ i.e.,}\\
p(I_t|\vec{\theta}_{[0,t-1]},\vec{I}_{[0,t]},R) &= \frac{1}{I_t!}\left(\sum_{s=1}^{\nu}\omega_s\theta_{t-s}\right)^{I_t} e^{-\sum_{s=1}^{\nu}\omega_s\theta_{t-s}}; \text{ and}\\
\theta_s|R,\vec{I}_{[0,s]},\vec{\theta}_{[0,s-1]} &\sim \text{Gamma}\left(I_s k, \text{rate}=\frac{k}{R}\right), \text{ i.e.,}\\
p(\theta_s|R,\vec{I}_{[0,s]},\vec{\theta}_{[0,s-1]}) &= \frac{\left(\frac{k}{R}\right)^{I_s k}}{\Gamma(I_s k)}\theta_s^{I_s k-1}e^{-\frac{\theta_s k}{R}}.
\end{aligned}$$

We want to calculate the joint distribution:

$$\begin{aligned}
&p(\vec{\theta}_{[t-\tau-\nu+1,t-1]},R|\vec{I}_{[t-\tau-\nu+1,t]})\\
&= p(\vec{\theta}_{[t-\tau-\nu+1,t-1]},R|\vec{I}_{[t-\tau-\nu+1,t-\tau]},\vec{I}_{[t-\tau+1,t]})\\
&\propto p(\vec{I}_{[t-\tau+1,t]}|\vec{I}_{[t-\tau-\nu+1,t-\tau]},\vec{\theta}_{[t-\tau-\nu+1,t-1]},R)p(\vec{\theta}_{[t-\tau-\nu+1,t-1]},R|\vec{I}_{[t-\tau-\nu+1,t-\tau]})\\
&= p(\vec{I}_{[t-\tau+1,t]},\vec{\theta}_{[t-\tau-\nu+1,t-1]},R|\vec{I}_{[t-\tau-\nu+1,t-\tau]})\\
&= p(I_t,\theta_{t-1}|\vec{I}_{[t-\tau-\nu+1,t-1]},\vec{\theta}_{[t-\tau-\nu+1,t-2]},R)\\
&\quad\cdot p(\vec{I}_{[t-\tau+1,t-1]},\vec{\theta}_{[t-\tau-\nu+1,t-2]},R|\vec{I}_{[t-\tau-\nu+1,t-\tau]})\\
&= p(I_t|\vec{\theta}_{[t-\nu,t-1]})p(\theta_{t-1}|I_{t-1},R)p(\vec{I}_{[t-\tau+1,t-1]},\vec{\theta}_{[t-\tau-\nu+1,t-2]},R|\vec{I}_{[t-\tau-\nu+1,t-\tau]}).
\end{aligned}$$

In the last step we used the conditional independence properties for I_t and θ_{t-1}, respectively. Repeating this process to separate $I_{[t-\tau+2,t]}$ and $\theta_{[t-\tau+1,t-1]}$ from the rest yields:

$$
\begin{aligned}
&p(\vec{\theta}_{[t-\tau-\nu+1,t-1]}, R|\vec{I}_{[t-\tau-\nu+1,t]}) \\
&\propto \prod_{s=t-\tau+2}^{t} p(I_s|\vec{\theta}_{[s-\nu,s-1]}) \\
&\quad \cdot \prod_{s=t-\tau+1}^{t-1} p(\theta_s|I_s, R) p(I_{t-\tau+1}, \vec{\theta}_{[t-\tau-\nu+1,t-\tau]}, R|\vec{I}_{[t-\tau-\nu+1,t-\tau]})
\end{aligned}
$$

Now, focusing on the last term, we have

$$
\begin{aligned}
&p(I_{t-\tau+1}, \vec{\theta}_{[t-\tau-\nu+1,t-\tau]}, R|\vec{I}_{[t-\tau-\nu+1,t-\tau]}) \\
&= p(I_{t-\tau+1}, \vec{\theta}_{[t-\tau-\nu+1,t-\tau]}|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, R) p(R|\vec{I}_{[t-\tau-\nu+1,t-\tau]}) \\
&= p(I_{t-\tau+1}|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, \vec{\theta}_{[t-\tau-\nu+1,t-\tau]}, R) p(\vec{\theta}_{[t-\tau-\nu+1,t-\tau]}|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, R) \\
&\quad \cdot p(R|\vec{I}_{[t-\tau-\nu+1,t-\tau]}) \\
&= p(I_{t-\tau+1}|\vec{\theta}_{[t-\tau-\nu+1,t-\tau]}) \prod_{s=t-\tau-\nu+1}^{t-\tau} p(\theta_s|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, R) p(R|\vec{I}_{[t-\tau-\nu+1,t-\tau]}).
\end{aligned}
$$

In the last equation, we used the fact that the individual θ_s are conditionally independent given the vector $\vec{I}$. At this point, the terms $p(\theta_s|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, R)$ become problematic. Knowledge of the terms I_m for $m > s$ certainly should shed some insight on the value of θ_s; however, it is not clear how this can be feasibly handled. It is not possible to prevent the occurrence of such terms due to the hierarchical nature of this model: the distribution of I_s requires previous θ values, which in return demand the inclusion of previous I values ad infinitum. This problem could be avoided by modeling all data from the start of the epidemic, at which point we could confidently set all values of I and θ corresponding to times prior to the onset of the epidemic to 0. This, however, would require treating the value of R as fixed for the entire epidemic, rendering our approach irrelevant as this assumption is clearly false.

As a solution, we propose putting a prior distribution on these problematic θ_s such that $p(\theta_s|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, R) \sim \Gamma(I_s k, k/R)$, essentially disregarding the additional information provided by future observations. Using a different prior, such as setting $p(\theta_s|\vec{I}_{[t-\tau-\nu+1,t-\tau]}, R) = \delta_{RI_s}$—which has the appeal of creating terms such as those in [4]—is statistically unsound, as we would draw different θ_s from different types of distributions.

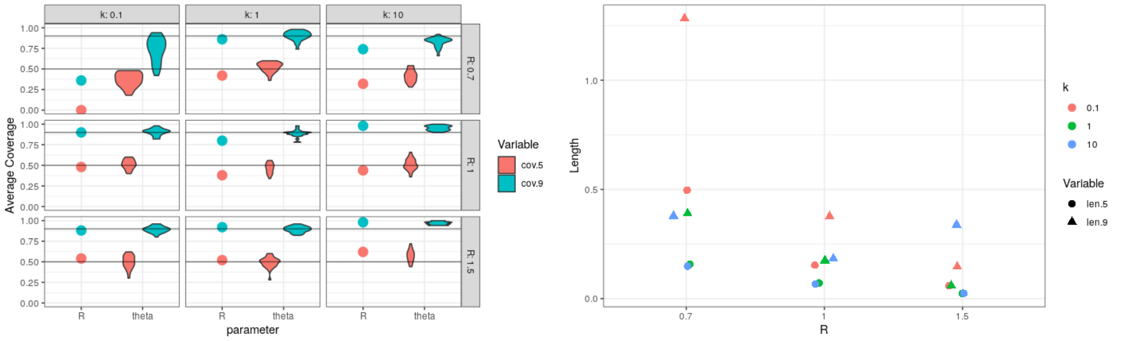
All this taken together yields:

$$
\begin{aligned}
&p(\vec{\theta}_{[t-\tau-\nu+1,t-1]}, R|\vec{I}_{[t-\tau-\nu+1,t]}) \\
&\propto \prod_{s=t-\tau+1}^{t} p(I_s|\vec{\theta}_{[s-\nu,s-1]}) \prod_{s=t-\tau+1}^{t-1} p(\theta_s|I_s, R) \prod_{s=t-\tau-\nu+1}^{t-\tau} p(\theta_s|I_s, R) \\
&\quad \cdot p(R|\vec{I}_{[t-\tau-\nu+1,t-\tau]})
\end{aligned}
$$

Using an inverse-gamma prior on R and using the densities of the other terms as discussed before evaluates to the same likelihood as in the main text.

B Model validation

Here we summarise estimation results for simulated data in order to more precisely show the effect of superspreading in a setting in which true parameters are known. The coverage and length



(A) (B)

Figure 5: Illustration of average credible interval coverage (cov.-) and length (len.-) on simulated data. As there is a single R parameter but 25 elements of $\vec{\theta}$, the coverage of the latter are summarised via a violin plot.

of intervals are shown in Figure 5. All simulations use an initial sequence of τ observations that have constant value 50. The momentum model is simulated for a further 3τ days. This complete series is then used to estimate R and $\vec{\theta}$. Simulations were repeated 50 times in order to asses coverage probabilities.

Of greatest initial import is verifying that the 90% credible intervals for R indeed cover the true value with approximately nominal probability. The case $R = 1$ is of primary importance, as it represents the bright-line between the epidemic growing or shrinking. That we have nearly exact coverage in this setting is indication that our credible intervals do not achieve coverage merely by being extremely wide. Furthermore, the intervals for $\vec{\theta}$ also cover the true values with the specified probability when $R = 1$ or $R = 1.5$. With our initial sequence of cases and $R = .7$, the epidemic sometimes dies out, which can be missed by the model. As such, coverage somewhat worse in this case.

After establishing coverage, our motivation for modeling superspreading is verified by looking at the lengths of the credible intervals: for k small, our intervals need to be extremely wide. In fact, the interval for $k = .1$ is approximately 2.5 times longer than the interval for $k = 10$ for both $R = .7$ and $R = 1$. For $R = 1.5$, the estimation problem becomes relatively easy as case numbers grow substantially. This leads to very small credible intervals.

As the explicit conditional distribution of the momentum parameters $\vec{\theta}$ is intractable, we present a summary of the samples observed through the MCMC simulation in Figure 6. This includes all 25 momentum parameters required when $\tau = \nu = 13$ as well as R. As $R = 1.5$ in this setting, one can observe that the scale increases for θ_s as s increases. It is clear that the parameters vary widely through MCMC estimation, even though the are initialised at the marginal MLE: $\hat{\theta}_s = I_s\hat{R}$. Multiple chains are run, each with a separate initial value for $\hat{R}$. When k is small, variability in $\vec{\theta}$ is large, requiring both tuning of the proposal distribution and long chains to be simulated in order to overcome high auto-correlation in the MCMC draws of $\vec{\theta}$.

Figure 7 shows how the individuals MCMC chains behave for each of the 25 momentum parameters in $\vec{\theta}$. Graphs for all parameters are shown in order to demonstrate that there is insufficient information to estimate the full set of parameters. One can also see how quickly the parameter estimates from different chains converge even when started a significantly different—and in some cases completely incorrect—starting values. Depending on the value of k, the variance of the proposal distribution for $\vec{\theta}$ must be set in order to allow θ to move slowly. If the variance is too high, then the acceptance proportion of proposed parameters is extremely low. This is due to the vector jumping to a nonsensical configuration, even if each individual θ_i is plausible in isolation.

As a final model validation, we consider k being drawn from a suitable distribution instead of being fixed. By using $k \sim$ Gamma(6, rate=55) we achieve approximately the same 2.5%,

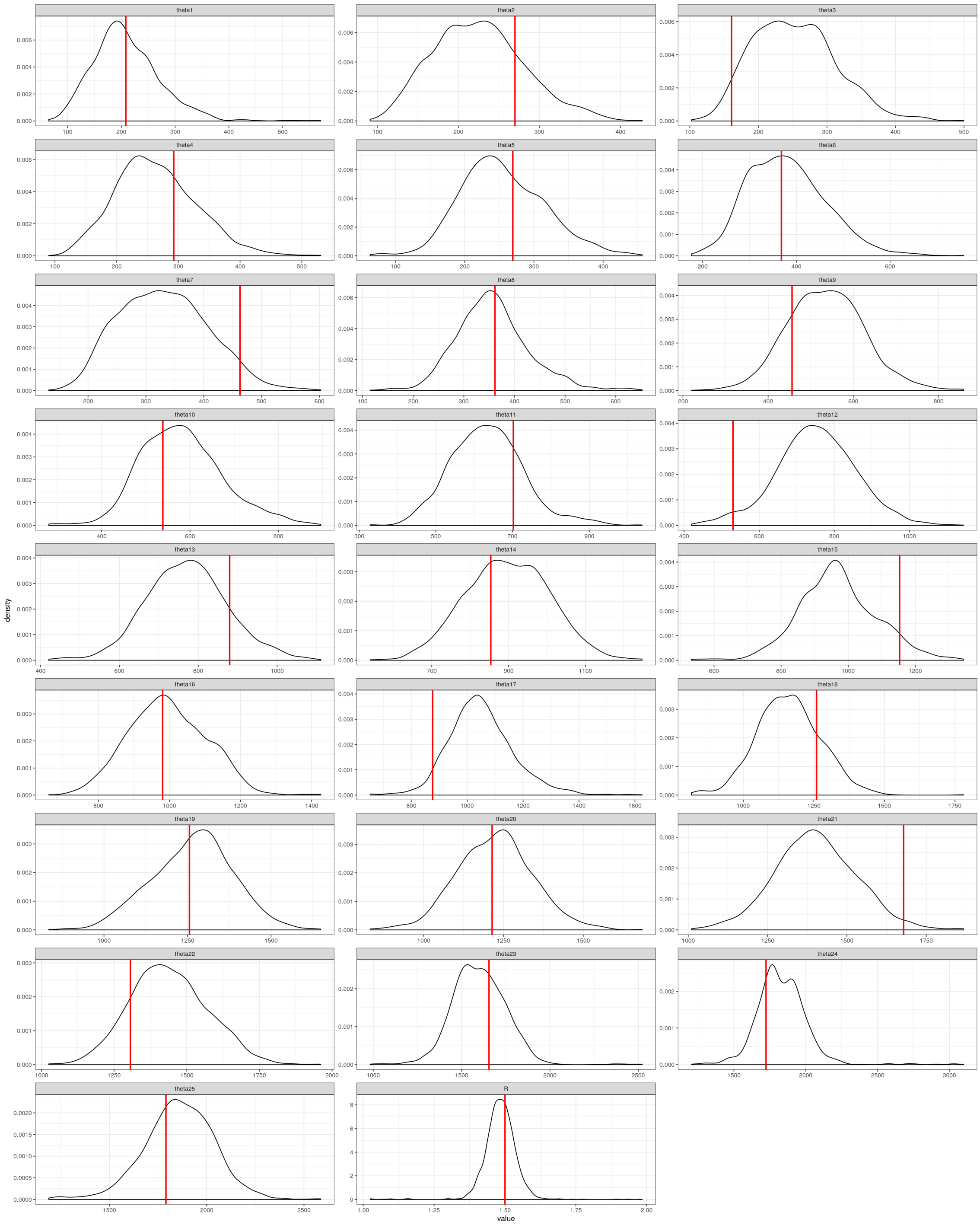


Figure 6: Samples of MCMC draws of parameters. The vertical, red lines indicate true values.

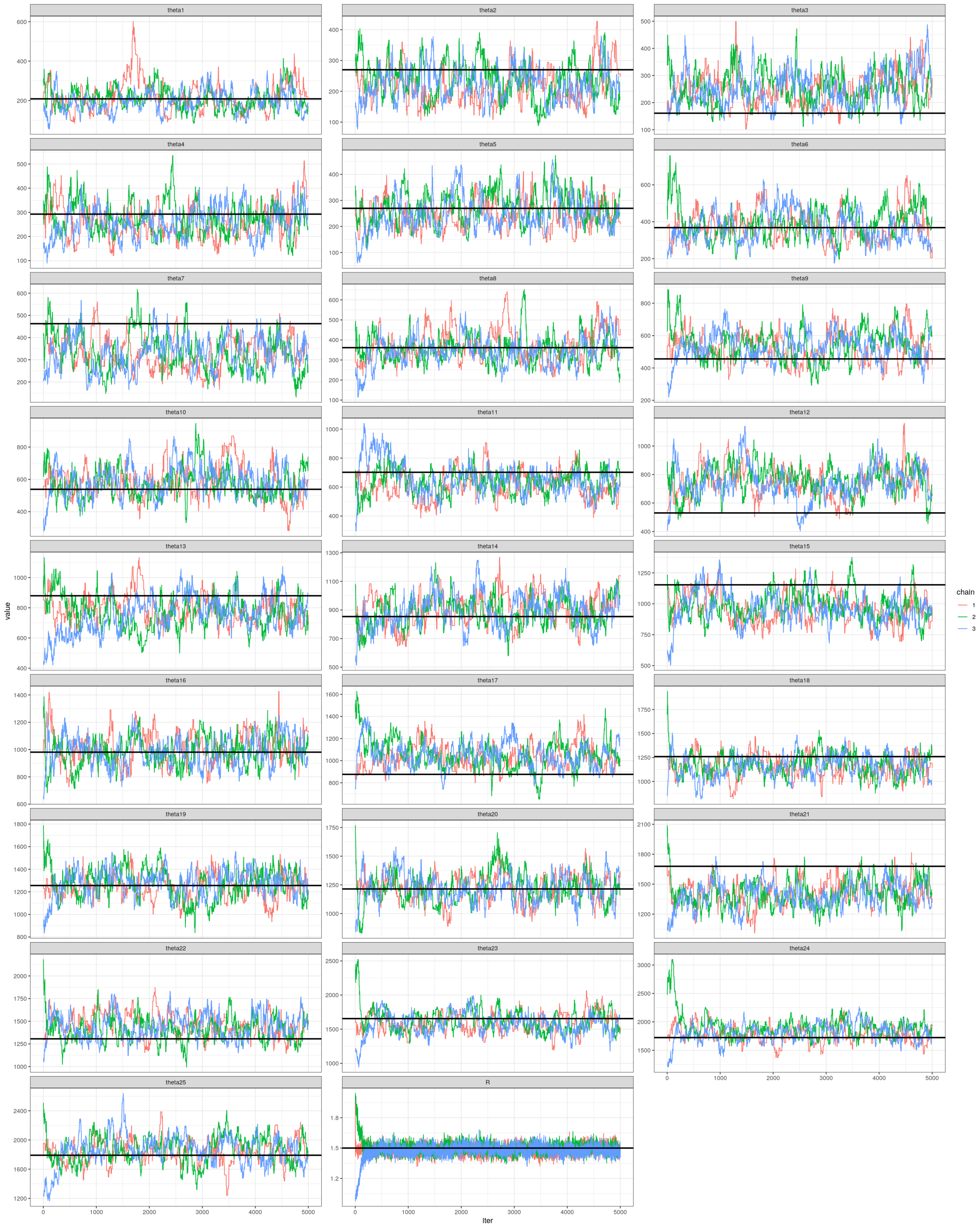


Figure 7: Each posterior distribution is composed of samples from several chains. As seen above, the chains converge quickly.

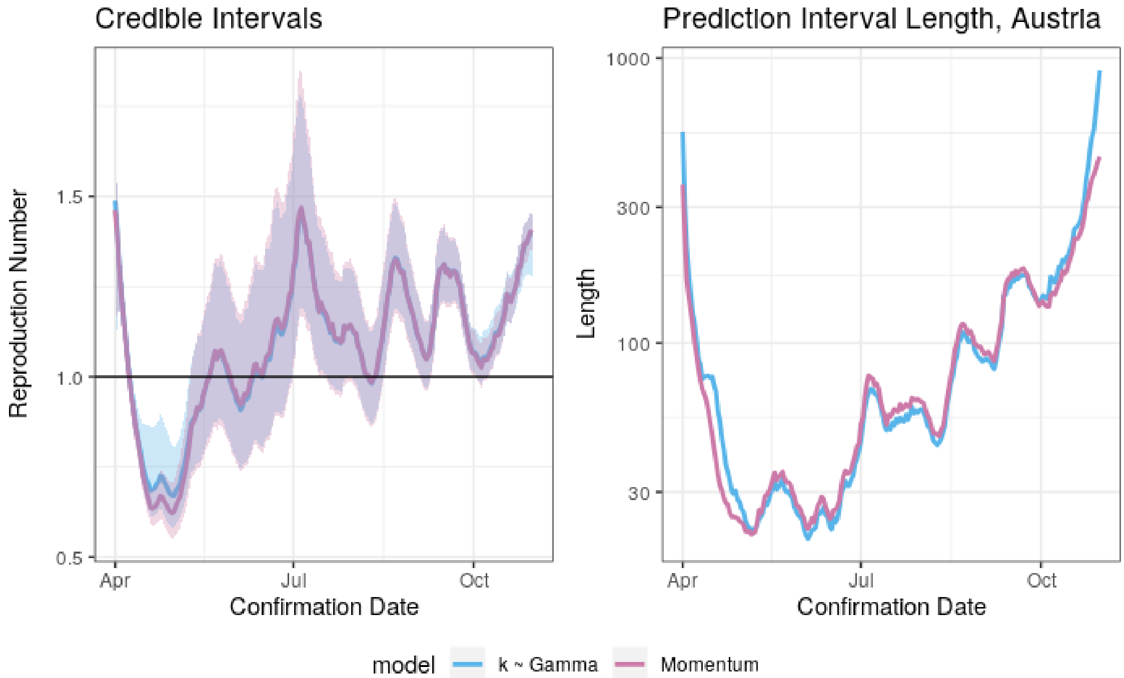


Figure 8: Comparison of selected graphs for k fixed and $k \sim Gamma()$.

50%, and 97.5% quantiles of the distribution of k given in [6]. For reference, these are .04, .1, and .2, respectively. Figure 8 shows credible intervals for R and prediction interval lengths for the momentum model with $k = 0.072$ and k random as above. Only these summary graphs are shown because no differences are visible in the missing figures. The only notable difference in the estimation of the reproduction number occurs when observed cases are very low. In this region, treating k as random yields slightly larger estimates for R as well as wider confidence intervals. Lastly, we note that the intervals for R are not as symmetric as for the k-fixed case as they are skewed slightly left. Furthermore, there is less heterogeneity in infectiousness. Our models estimate that 10% of infected individuals contribute 81% of new infections while 20% contribute 95% of new infections.

C Generation model derivations

C.1 Normal approximation

This appendix derives the normal approximation to the posterior distribution $p(R|\vec{I}_{[t]}, k)$ used in Section 2.2. As we can iteratively condition on previous values, the joint distribution of $\vec{I}_{[t-\tau+1,t]}|\vec{I}_{t-\tau}, R, k$ decomposes into a product of factors of the form (12). We have

$$
\begin{aligned}
p(\vec{I}_{[t-\tau+1,t]}|\vec{I}_{[t-\tau]}, R, k) &= \prod_{s=t-\tau+1}^{t} p(I_s|\vec{I}_{[s-1]}, R, k) \\
&= \prod_{s=t-\tau+1}^{t} \frac{\Gamma(I_s + kI_{s-1})}{I_s!\Gamma(kI_{s-1})} \left(\frac{k}{R+k}\right)^{kI_{s-1}} \left(\frac{R}{R+k}\right)^{I_s}.
\end{aligned}
$$

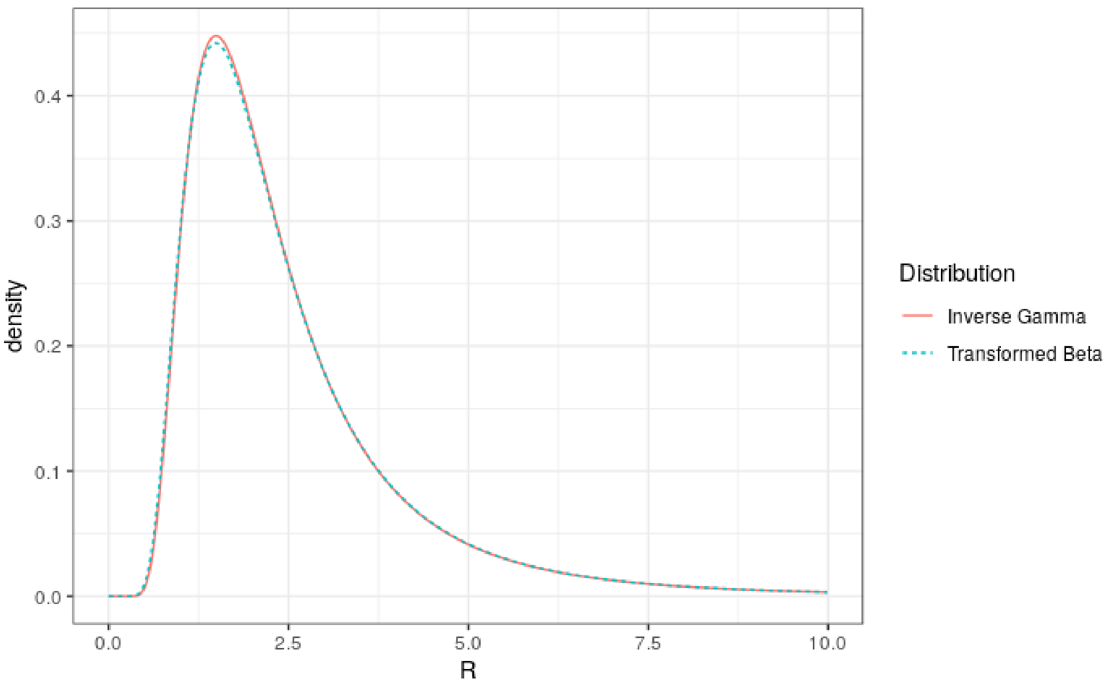


Figure 9: Comparison of priors on R and $R/(R+k)$.

The structure of this likelihood suggests estimating $R/(R+k)$ instead of R. When treating $\vec{I}_{[t-\tau]}$ and k as fixed, Bayes' theorem yields the posterior distribution of $R/(R+k)$:

$$p\left(\frac{R}{R+k}\middle|\vec{I}_{[t]},k\right) \propto \left(\frac{k}{R+k}\right)^{k\sum_{s=t-\tau}^{t-1} I_s} \left(\frac{R}{R+k}\right)^{\sum_{s=t-\tau+1}^{t} I_s} p\left(\frac{R}{R+k}\middle|I_{[t-\tau]},k\right). \tag{14}$$

Given this functional form, it is natural to put a beta prior on $R/(R+k)$ to maintain conjugacy. For small k, as shown in Figure 9, this corresponds to putting an appropriate inverse-gamma prior on R, while for larger k this would correspond to a gamma prior. Therefore, to mimic the $R \sim$ Inv-Gamma(3.69, rate $=$ 6.994) prior distribution used in Section 2.1, we can use a Beta($\tilde{\alpha} = 98.82$, $\tilde{\beta} = 3.74$) prior on $R/(R+k)$. The posterior distribution of $R/(R+k)$ then has a Beta distribution with parameters

$$\alpha = \tilde{\alpha} + \sum_{s=t-\tau+1}^{t} I_s, \quad \beta = \tilde{\beta} + k\sum_{s=t-\tau}^{t-1} I_s.$$

While the hyperparameter values are not so small as to be uninformative, they are easily outweighed by the data in most settings. By a change of variables from $R/(R+k)$ back to R, we derive the posterior distribution of R to be

$$p\left(R\middle|\vec{I}_{[t]},k\right) \propto \frac{k}{(R+k)^2}\left(\frac{R}{R+k}\right)^{\alpha-1}\left(\frac{k}{R+k}\right)^{\beta-1}. \tag{15}$$

As a final simplifying step, we compute the normal approximation of this posterior [8, Section 4.1].

To this end, the first and second derivatives of the log-posterior density are

$$\frac{d}{dR}\log p(R|\vec{I}_{[t],k}) = \frac{\alpha-1}{R} - \frac{\alpha+\beta}{R+k},$$
$$\frac{d^2}{dR^2}\log p(R|\vec{I}_{[t]},k) = -\frac{\alpha-1}{R^2} + \frac{\alpha+\beta}{(k+R)^2}.$$

Thus, the mode of the posterior is

$$\hat{R} = \frac{k(\alpha-1)}{\beta+1},$$

and the variance estimate is

$$\left(-\frac{d^2}{dR^2}p(R|\vec{I}_{[t]},k)(\hat{R})\right)^{-1} = \frac{k^2(\alpha+\beta)(\alpha-1)}{(\beta+1)^3}.$$

This yields a normal approximation of the posterior of

$$p(R|\vec{I}_{[t]},k) \approx N\left(\frac{k(\alpha-1)}{\beta+1}, \frac{k^2(\alpha+\beta)(\alpha-1)}{(\beta+1)^3}\right).$$

C.2 Generation model

It is easiest to represent the process of infections per generation if we allow the indices of the summation notation to be real numbers (hence treating the summation as integration) via

$$\sum_{s=c_1}^{c_2} I_{t-s} = (\lceil c_1\rceil - c_1)I_{t-\lceil c_1\rceil+1} + \sum_{s=\lceil c_1\rceil}^{\lfloor c_2\rfloor-1} I_{t-s} + (c_2 - \lfloor c_2\rfloor)I_{t-\lfloor c_2\rfloor}$$

for $c_1, c_2 \in \mathbb{R}$ where $c_1 < c_2$. When D_g is the length of a generation, the number of infections per generation is then given simply by

$$\tilde{I}_{\tilde{t}-i} = \sum_{s=i\cdot D_g}^{(i+1)\cdot D_g} I_{t-s},$$

for $i \in \mathbb{N}_0$.

We assume R is constant for τ days in the generation model as in the momentum model. The corresponding parameter in the generation model is $\tau_g := \tau/D_g$, where we account for non-integer values as before by summing fractional daily infections. Estimating parameters and producing forecasts requires similar modifications for non-integer values. Conceptually, however, these correspond to the same sums as before, just over generations instead of conventional days. This can be represented concisely in the notation for real-valued summation as

$$\alpha = \tilde{\alpha} + \sum_{s=0}^{\tau_g} \tilde{I}_{\tilde{t}-s}, \quad \text{and} \tag{16}$$

$$\beta = \tilde{\beta} + k\sum_{s=1}^{\tau_g+1} \tilde{I}_{\tilde{t}-s}. \tag{17}$$

This yields a negative binomial observation model as before:

$$\tilde{I}_{\tilde{t}}|R, \tilde{I}_{\tilde{t}-1} \sim NB\left(\tilde{I}_{\tilde{t}-1}k, \frac{R}{k+R}\right).$$

For prediction and comparisons used in Section 3, it is more sensible to compare the cumulative incidence of several, say $\mathbf{T}$, days. We forecast $\lceil \frac{\mathbf{T}}{D_g} \rceil$ generations $\tilde{I}_{\tilde{t}}$, and our forecast for $\mathbf{T}$ days is then

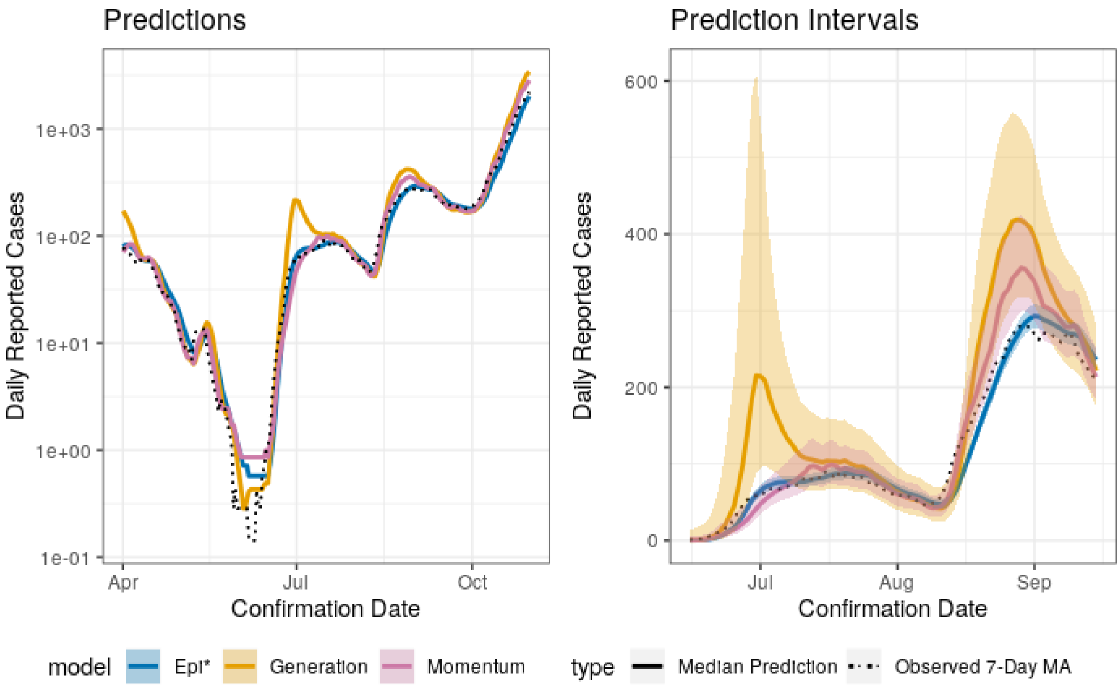
$$X_{\mathbf{T}} = \sum_{s=1}^{\lfloor \frac{\mathbf{T}}{D_g} \rfloor} \tilde{I}_{\tilde{t}+s} + \left(\frac{\mathbf{T}}{D_g} - \left\lfloor \frac{\mathbf{T}}{D_g} \right\rfloor \right) \tilde{I}_{\tilde{t}+\lceil \frac{\mathbf{T}}{D_g} \rceil}.$$

In the case study of Section 3, we forecast the total weekly cases, i.e., $\mathbf{T} = 7$. Observe that this equates to merely forecasting sufficient generations to cover the desired time period, then taking the appropriate proportion of the final forecasted generation to match the desired time window $\mathbf{T}$.

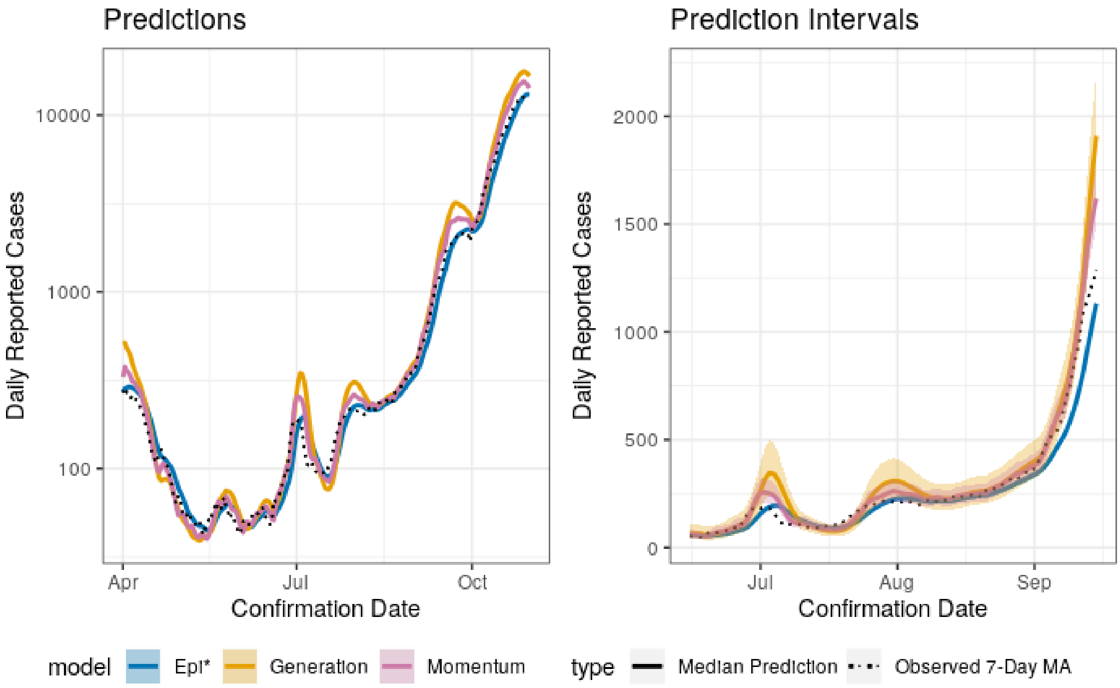
D Results for Croatia and Czechia

As further demonstration of the momentum model, Figures 10 and 11 show the same prediction and estimation results as seen in Section 3 but for Croatia and Czechia. The disease progression in Czechia is similar to that of Austria over the shown period. Croatia is a common Austrian and Czech tourist destination and the disease progression is markedly different there than in Austria. The estimated coverage probabilities of the prediction intervals are also shown in Table 2. The story remains the same as before: coverage is far better for the momentum model with superpreading than without (Epi*). Similarly, Epi* appears shifted relative to the observed cases, particularly for the Czech data. Here we see that the momentum model performs better than the generation model, particularly around peaks in the time series.

Figure 11 contains results on the estimation of R for Croatia and Czechia. The results are qualitatively the same, in that the momentum model with superspreading produces much wider credible intervals. One obvious feature of the Croatian data, however, is a steep decline and subsequent steep increase in June. This corresponds to a large increase and plateau in cases as seen in Figure 10. The Epi* model estimates that R increases to well over 2 within a short period of time before decreasing again to previous levels. Alternatively, in the same period, the momentum model provides a noticeably lower median estimate but with an incredibly wide interval. Further exploration of the feature is warranted, though it is reasonable that such a large deviation over a small window of time should produce significantly more uncertainty in the value of the underlying parameter, particularly when the model is estimated under the assumption that R is constant over $\tau = 13$ days. Within the momentum model, such short-term deviations can be captured by an increase or decrease in disease momentum instead of just an increase in R. On the other hand, this feature appears to show a flaw within the generation model, as both the estimated R and interval estimate have extreme spikes. This is likely due to the short-term nature of the case increase and the generation model only using roughly three generations for estimation.

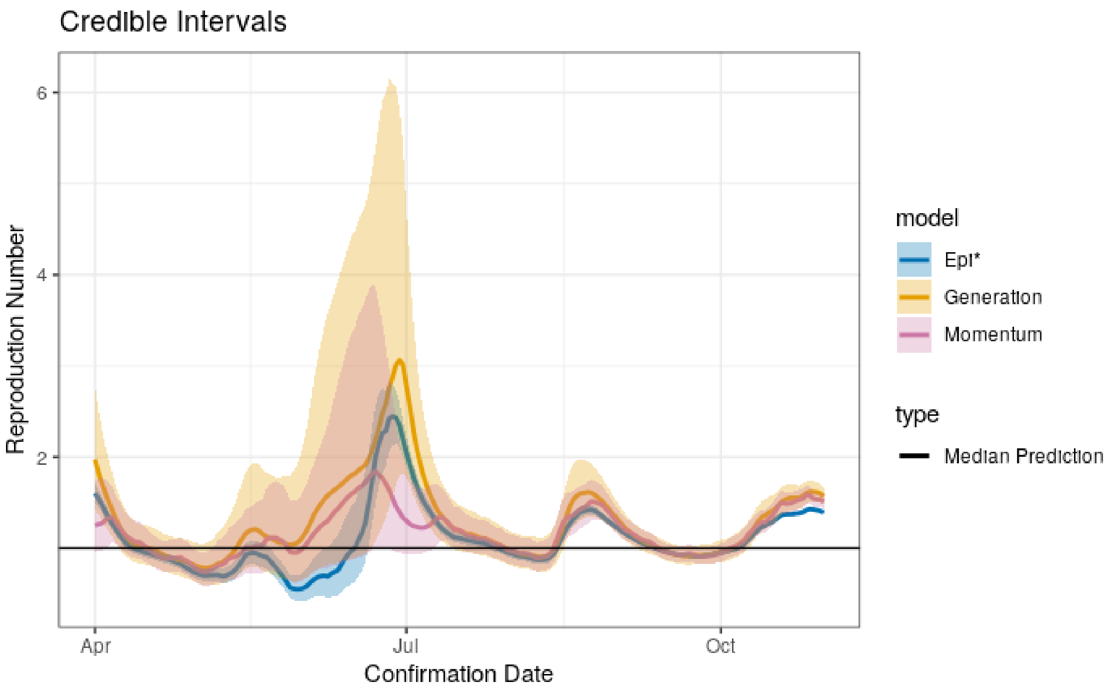


(A) Croatian Data

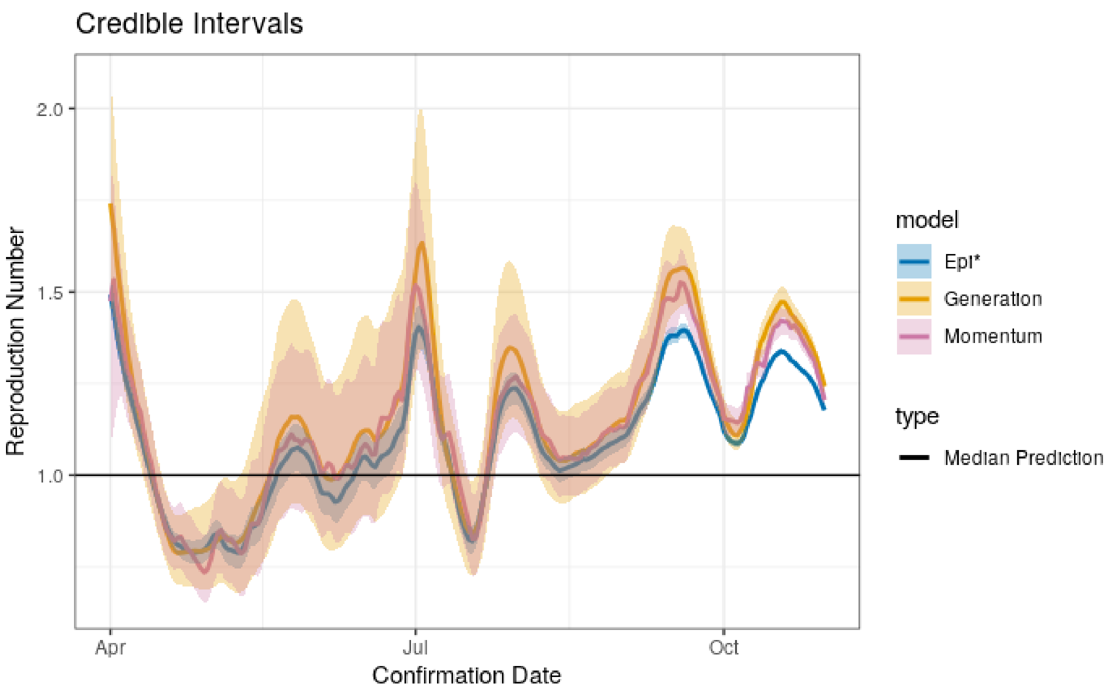


(B) Czech Data

Figure 10: Predictions between April 1 and October 31, 2020, and 90% prediction intervals between June 15 and September 7, 2020. Predictions and intervals are for the 7-day average of new cases in the following week in Croatia and Czechia.



(A) Croatian Data



(B) Czech Data

Figure 11: Credible intervals and for R in Croatia and Czechia between April 1 and October 31, 2020.

References

[1] S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research* 5.112 (2020), p. 112.

[2] D. C. Adam, P. Wu, J. Y. Wong, E. H. Lau, T. K. Tsang, S. Cauchemez, G. M. Leung, and B. J. Cowling. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine* 26.11 (2020), pp. 1714–1719.

[3] N. Arinaminpathy, J. Das, T. McCormick, P. Mukhopadhyay, and N. Sircar. Quantifying heterogeneity in SARS-CoV-2 transmission during the lockdown in India. *medRxiv* (2020).

[4] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology* 178.9 (2013), pp. 1505–1512.

[5] C. Donnat and S. Holmes. Modeling the Heterogeneity in COVID-19's Reproductive Number and its Impact on Predictive Scenarios. *arXiv preprint arXiv:2004.05272* (2020).

[6] A. Endo, S. Abbott, A. Kucharski, and S. Funk. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research* 5 (July 2020), p. 67.

[7] T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* 25.17 (2020), p. 2000257.

[8] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian Data Analysis. 2nd ed. Chapman and Hall/CRC, 2004.

[9] K. M. Gostic, L. McGough, E. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. A. Hay, P. M. De Salazar, et al. Practical considerations for measuring the effective reproductive number, Rt. *medRxiv* (2020).

[10] J. Knight and S. Mishra. Estimating effective reproduction number using generation time versus serial interval, with application to covid-19 in the Greater Toronto Area, Canada. *Infectious Disease Modelling* 5 (2020), pp. 889–896.

[11] R. Laxminarayan, B. Wahl, S. R. Dudala, K. Gopal, C. Mohan, S. Neelima, K. S. J. Reddy, J. Radhakrishnan, and J. Lewnard. Epidemiology and transmission dynamics of COVID-19 in two Indian states. *medRxiv* (2020).

[12] Y. Liu, R. M. Eggo, and A. J. Kucharski. Secondary attack rate and superspreading events for SARS-CoV-2. *The Lancet* 395.10227 (2020), e47.

[13] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature* 438.7066 (2005), pp. 355–359.

[14] M. O. Lorenz. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association* 9.70 (1905), pp. 209–219.

[15] J. Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling* 5 (2020), pp. 129–141.

[16] M. O'Driscoll, C. Harry, C. A. Donnelly, A. Cori, and I. Dorigatti. A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics with implications for the current COVID-19 pandemic. *medRxiv* (2020).

[17] L. Richter, D. Schmid, A. Chakeri, S. Maritschnik, S. Pfeiffer, and E. Stadlober. Schätzung des seriellen Intervalles von COVID19, Österreich. Tech. rep. https://www.ages.at/en/wissen-aktuell/publikationen/schaetzung-des-seriellen-intervalles-von-covid19-oesterreich/. Apr. 2020.

[18] R. Thompson, J. Stockwin, R. D. van Gaalen, J. Polonsky, Z. Kamvar, P. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics* 29 (2019), p. 100356.

[19] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences* 274.1609 (2007), pp. 599–604.

[20] J. Wallinga and P. Teunis. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology* 160.6 (Sept. 2004), pp. 509–516.

Robust Models of SARS-CoV-2 Heterogeneity and Control

K. D. Johnson, A. M. Grass, D. Toneian, M. Beiglböck and J. Polechová

Abstract

In light of the continuing emergence of new SARS-CoV-2 variants and vaccines, we create a simulation framework for exploring possible infection trajectories under various scenarios. The situations of primary interest involve the interaction between three components: vaccination campaigns, non-pharmaceutical interventions (NPIs), and the emergence of new SARS-CoV-2 variants. Additionally, immunity waning and vaccine boosters are modeled to account for their growing importance. New infections are generated according to a hierarchical model in which people have a random, individual infectiousness. The model thus includes super-spreading observed in the COVID-19 pandemic. Our simulation functions as a dynamic compartment model in which an individual's history of infection, vaccination, and possible reinfection all play a role in their resistance to further infections. We present a risk measure for each SARS-CoV-2 variant, $\rho^{\mathcal{V}}$, that accounts for the amount of resistance within a population and show how this risk changes as the vaccination rate increases. Furthermore, by considering different population compositions in terms of previous infection and type of vaccination, we can learn about variants which pose differential risk to different countries. Different control strategies are implemented which aim to both suppress COVID-19 outbreaks when they occur as well as relax restrictions when possible. We demonstrate that a controller that responds to the effective reproduction number in addition to case numbers is more efficient and effective in controlling new waves than monitoring case numbers alone. This is of interest as the majority of the public discussion and well-known statistics deal primarily with case numbers.

Keywords: COVID-19; superspreading, compartment model; SARS-CoV-2 variant risk; NPI

1 Introduction

The continued waves of the COVID-19 pandemic present unique challenges to regulatory bodies and governments. At issue is the balance between restricting behavior in order to reduce the spread of SARS-CoV-2 and the desire to return to a normal state-of-affairs. On one hand, many countries provide a deluge of statistics to measure the severity of COVID-19, even on a highly granular level. These statistics then inform complex decisions on how many restrictions to enforce. On the other hand, some countries lack sufficient testing to accurately track the spread of COVID-19. Our guiding question is, what statistics should be considered when determining if mitigation measures should be increased or decreased? Of concern is the oft-repeated scenario in which a new variant emerges which spreads more rapidly either due to increased infectiousness or vaccine escape.

To this end we create a compartment model that has compartments for each vaccinated, infected, and recovered group (for each variant), and add dynamic interactions between these groups as well as immunity waning and boosting. For example, someone could have been infected with the original SARS-CoV-2 variant, then receive a vaccine, then perhaps later become infected with a new SARS-CoV-2 variant. The resistance to further infection conferred by such a history is distinct from those who have, for example, only been vaccinated. These factors influence the effective reproduction number: the expected number of new infections caused by a currently infected individual. We can then simulate infections using this model in order to answer questions about case dynamics when a new SARS-CoV-2 variant is introduced.

We also add a controller to our simulations, which can both observe and intervene in the compartment model. The controller is thought of as a governing agent which is responsible for both keeping COVID-19 outbreaks at manageable levels and not imposing unnecessary restrictions, i.e., for keeping outbreaks under "control." In order to mimic a real-life entity such as a government, the controller must be constrained in various ways. First, the controller does not observe latent variables such as infectiousness, only raw data such as number of new cases (for each variant), which account for only a proportion of true infections equal to the detection ratio. Second, these statistics are observed with a lag, i.e., there is a delay between when infections occur and when cases are observed. Third, the controller uses non-pharmaceutical interventions (NPIs) such as mask wearing, testing and tracing, and gathering restrictions (soft lockdowns) – which mitigate the effective reproduction number of SARS-CoV-2. Lastly, these interventions cannot change continuously: there is a mandatory temporal gap after an intervention before the controller can intervene again.

Under these constraints, we are able to explore what statistics the controller needs to respond to in order to effectively suppress new outbreaks. Note that we are not advocating a particular intervention or comparing their efficiency [9, 49]. Instead, we are considering what information could best inform timely decisions on modifying NPIs. Furthermore, we note that the success of the controller we implement is often not mirrored in reality as few if any governments willingly act as rapidly as stipulated. For example, the World Health Organization issues guidelines for COVID-19 risk [17, 56], but immediate action is often not taken when a country crosses one of the thresholds they provide. This is exacerbated by the lag between infection and case observation (the "delay" parameter): governments which either fail to collect adequate data or do not make decisions using a forecast will make decisions with greater delay.

Other recent works have considered similar topics. [4] posits economic models that convert the effect of NPIs on both public health and the economy into a single cost measure. Using infection forecasts from SEIR models, they then optimise decision thresholds which depend solely on the number of cases. The infection models used do not contain any diversity in SARS-CoV-2 variants or vaccines. On a different level of analysis, [50] proposes a procedure for expert elicitation in order to synthesise the results from multiple modeling groups to improve intervention planning.

We begin by presenting a detailed description of our model. This is broken up into subsections which describe the compartment model with different vaccines and variants, the controllers we consider, and how the simulation is initialised to mimic a real outbreak. In the 3 section, we compare the modelled dynamics of the SARS-CoV-2 variants, as well as the controllers' effectiveness in containing both current and hypothesised variants. Last, we discuss broader implications of our research and further questions which could be explored within the framework.

2 Methods

New infections are assumed to be generated according to the momentum model of Johnson *et al.* [26], which builds upon Cori *et al.* [16]. In the simplest version of the model, new infections I_t are the result of previous infections $I_1, \ldots, I_{t-1}$ via the following recursion:

$$I_t \sim \text{Poisson}\left(\mathcal{R}_{e,t} \sum_{m=1}^{\nu} I_{t-m} w_m\right), \tag{1}$$

where $\mathcal{R}_{e,t}$ is the time-varying effective reproduction number at time t, $\mathbf{w} = (w_1, \ldots, w_\nu)$ is the generation interval, and ν is the maximum number of days for which someone is assumed to be infectious. If J_m denotes the number of people infected by a specific person on the m-th day after this person got infected, then we have for $m \in \mathbb{N}$

$$w_m = \frac{\mathbb{E}[J_m]}{\sum_{l=1}^{\infty} \mathbb{E}[J_l]}.$$

We assume that a newly infected individual does not cause secondary infections on the same day, corresponding to $w_0 = 0$, and that infections do not occur after day ν. The generation interval can be interpreted as the infectiousness profile of infected persons. We set $\mathbf{w}$ to be a discretised gamma distribution with $\nu = 13$, mean 4.46, and standard deviation 2.63. These are values specific to Austria [45], and are similar to values determined elsewhere [21, 23, 29]. We note that the framework is general enough to allow each variant to be specified with a unique generation interval. This is potentially useful for modeling future variants, as the recently prominent Omicron (B.1.1.529) variant appears to have a shorter generation time [27].

The recursion in equation (1) assumes that all people have the same infectiousness on day t. We follow [26] and remove this assumption by explicitly drawing an infectiousness parameter for each infected person from a fixed Gamma distribution with dispersion parameter $k < 1$. This generalisation allows for superspreading: the phenomenon of extreme heterogeneity in infectiousness. We set $k = 0.1$, which corresponds to a setting in which 10% of infected individuals cause 80% of new infections [19]. This is an integral component of the difficulty of controlling COVID-19 outbreaks. Individual infectiousness can be aggregated over the infected, resulting in the following process of new infections:

$$I_t \sim \text{Poisson}\left(\sum_{m=1}^{\nu} \theta_{t-m} w_m\right) \qquad \text{where} \tag{2}$$

$$\theta_s \sim \text{Gamma}(I_s k, \text{ rate} = k/\mathcal{R}_{e,t}). \tag{3}$$

We generalise this model further in order to study the effect of combinations of variants, previous infections, vaccination strategies, and NPIs – and interactions between them – on the effective reproduction number $\mathcal{R}_{e,t}$. This is done by decomposing $\mathcal{R}_{e,t}$ into many constituent parts which depend on the compartments in our model. In order to describe this decomposition, we need notation for compartments.

Our model contains a set of compartments $\mathscr{C}$. Each compartment $\mathcal{C} \in \mathscr{C}$ is a group of people with a unique history of infection and vaccination. The history is encoded as a superscript h: $\mathcal{C}^h$. The value of h contains both digits and capital letters, where digits correspond to vaccines and letters correspond to different SARS-CoV-2 variants (when possible, the first letter of the variant name). For example, a group with label $h = A1B$ contains people that were first infected with variant A (Alpha), then vaccinated with vaccine 1, then contracted variant B (Beta). For simplicity, the digit 0 is reserved for the compartment that has neither been vaccinated nor contracted SARS-CoV-2 of any form, i.e., $\mathcal{C}^0$. As it will simplify our notation, we let $\mathscr{V}$ be the set of infectious variants: $\{WT$(wild-type), A(lpha), B(eta), D(elta), G(amma),...$\}$. For clarity, we note here that elements of $\mathscr{V}$, denoted by $\mathcal{V}$, are also valid histories h, e.g., $h = \mathcal{V}$ indicates those individuals who have only been infected with variant $\mathcal{V}$.

The only characteristics of the history that affect the model are the total set of experiences (vaccines or infections) as well as the final infection, as this determines the variant one is infectious with. Furthermore, reinfection with the same variant does not confer additional benefit. Hence we can simplify histories to those that have no repeated capital letters. Lastly, it will often be easier to write equations using the set of histories, $\mathscr{H}$, instead of the corresponding set of compartments, $\mathscr{C}$. As $h \in \mathscr{H}$ is the identifier of a compartment, we will also at times call it a compartment for ease of use.

Each group $\mathcal{C}^h$ contains the total number of people with that history, both infected (I^h) and recovered (S^h), which are subgroups with the same labels: $\mathcal{C}^h = I^h \cup S^h$. The recovered (or vaccinated) subgroup is written as S^h to emphasise that they are again susceptible to infection, though with a resistance parameter depending on h as described below. All are given subscripts t, though for consistency with the generating equations, the subscripted groups have different interpretations. $\mathcal{C}_t^h$ and S_t^h contain *all people* on day t with history h and the subset that are recovered, respectively. I_t^h gives the number of *new* infections with history h on day t. For simplicity, individuals recover $\nu + 1$ days after being infected. We assume that when one is experiencing an infection, they cannot become newly infected (or receive a vaccine). With these simplifications, we have $|\mathcal{C}_t^h| = |S_t^h| + \sum_{m=1}^{\nu} I_{t-m}^h$. Note that notation for I_t^h has been overloaded to either be the set of people with new infections with history h or the cardinality of this set. This provides consistency with the generating equation (2).

An important aspect of the simulation is that interaction groups are created dynamically. For example, someone in S_t^h can be infected with a SARS-CoV-2 variant D or become vaccinated with vaccine 4. This person then moves from S_t^h to a new group with identifier hD or $h4$, respectively. The dynamic generation of groups goes hand-in-hand with a dynamic change of population characteristics which may require different mitigation strategies. Crucially, the new group hD or $h4$ can have new resistances to infection.

There is a specific $\mathcal{R}_{e,t}^{h\mathcal{V}}$ for all compartments $h \in \mathscr{H}$ and all infectious variants $\mathcal{V} \in \mathscr{V}$, where group S_t^h is susceptible to infection with $\mathcal{V}$ on day t. This can be interpreted as the effective reproduction number of variant $\mathcal{V}$ *solely* within group $\mathcal{C}^h$. As data are often reported in terms of relative transmissibility of SARS-CoV-2 variants, each variant $\mathcal{V}$ has a basic reproduction number relative to that of the original SARS-CoV-2 variant given by $\mathcal{R}_0^{\mathcal{V}} = \lambda^{\mathcal{V}}\mathcal{R}_0 = \lambda^{\mathcal{V}}\mathcal{R}_0^{WT}$. With this group-specific notation, we can define a decomposition of $\mathcal{R}_{e,t}^{h\mathcal{V}}$ as

$$\mathcal{R}_{e,t}^{h\mathcal{V}} = \tilde{M}_t L_t \lambda^{\mathcal{V}} \mathcal{R}_0 \tilde{\gamma}^{h\mathcal{V}} \tag{4}$$

where

- $\tilde{M}_t = (1 - M_t)$, where $M_t \in [0, 1]$ is the effectiveness of NPIs at time t (mitigation of infectiousness). $\tilde{M}_t = 1$ corresponds to no mitigation (full infectiousness), whereas $\tilde{M}_t = 0$ reduces new infections to 0.
- L_t is a seasonality factor at time t.
- $\tilde{\gamma}^{h\mathcal{V}}$ is the susceptibility of group h to infection with variant $\mathcal{V}$. We consider $\tilde{\gamma}^{h\mathcal{V}} = (1 - \gamma^{h\mathcal{V}})$ where $\gamma^{h\mathcal{V}} \in [0, 1]$ is the *resistance* of group S^h to being infected by variant $\mathcal{V}$. The value of $\gamma^{h\mathcal{V}}$ depends on the unique history h.

Similar to other human coronaviruses and influenza viruses [36, 37], it is widely believed that SARS-CoV-2 follows a seasonal transmission pattern in temperate regions with transmissions peaking during the winter. Possible explanations include different viral longevity due to humidity and

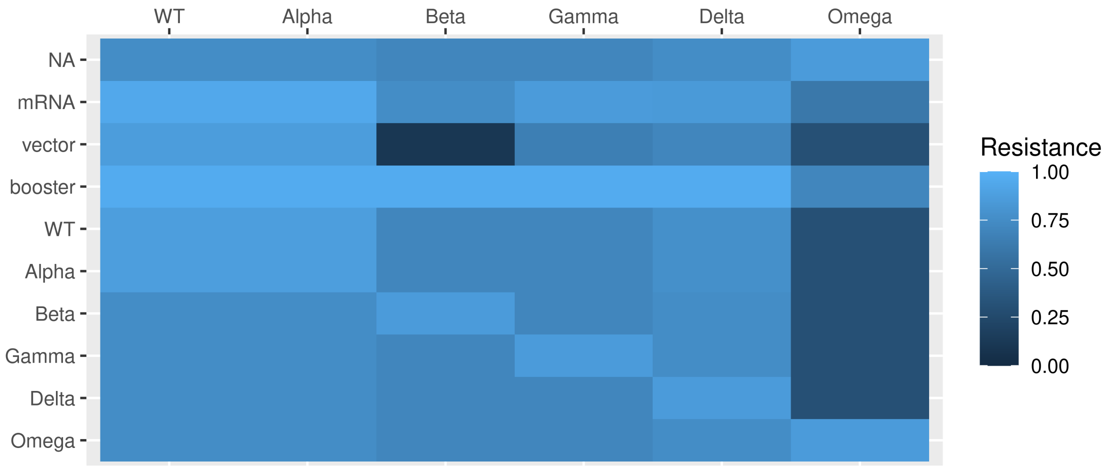


Figure 1: **Midpoint estimates of resistance parameters**. Complete statistics including confidence intervals are given in Table 1 in the 3.4.

air temperature [33, 54], reduced host airway immune response in dry winter months [32, 37], and increased indoor interactions during colder months. We model seasonality, L_t, as in [28] via a cosine transform:

$$L_t = 1 + \frac{\epsilon}{2}\left(\cos\left(2\pi\frac{t - t_{peak}}{365.25}\right) - 1\right), \tag{5}$$

where ϵ is the amplitude size and t_{peak} is the date when the transmission rate peaks. Following [22], we assume that the seasonal reduction in transmission is 40% ($\epsilon = 0.4$); the lowest transmission rate is set to July 1, while the highest transmission rate occurs on t_{peak} = January 1. This is in line with the estimates for the general seasonality of other coronaviruses in temperate climates [36].

The resistance parameters $\gamma^{h\mathcal{V}}$ evolve according to two simple rules. First, if a person with history h is given a vaccine (e.g. 1), resulting in history $g = h1$, they have variant-specific resistances equal to the maximum from those provided by h and 1. For example, history h may provide resistance .9 for infection from variant D and .98 for variant A. Vaccine 1, on the other hand, provides resistance .95 for both. The resulting resistance for history g is thus .95 for D and .98 for A. Resistance parameters are summarised in Fig 1 and are given in detail in Table 1.

Second, if a person with history h is infected by a variant $\mathcal{V}$, resulting in history $g = h\mathcal{V}$, we consider both the resistances of h and those conferred by $\mathcal{V}$. For this new history g, the resistance to an infection with SARS-CoV-2 variant $\mathcal{V}'$ is given by

$$\gamma^{g\mathcal{V}'} = 1 - (1 - \gamma^{h\mathcal{V}'})(1 - \gamma^{\mathcal{V}\mathcal{V}'}), \quad \text{equivalently} \tag{6}$$

$$\tilde{\gamma}^{g\mathcal{V}'} = \tilde{\gamma}^{h\mathcal{V}'}\tilde{\gamma}^{\mathcal{V}\mathcal{V}'}. \tag{7}$$

Thus, the susceptibility to variant $\mathcal{V}'$ declines as a product of the susceptibility to $\mathcal{V}'$ conferred by the history h, $\tilde{\gamma}^{h\mathcal{V}'}$, and the susceptibility to $\mathcal{V}'$ conferred by the infection $\mathcal{V}$, $\tilde{\gamma}^{\mathcal{V}\mathcal{V}'}$. Note that repeated infections with the same variant, i.e. $\mathcal{V} \in h$, then resistances are *not* updated.

We note here that interactions between groups are only considered in terms of resistances, not in terms of infectiousness; a vaccinated individual that nevertheless gets infected with variant $\mathcal{V}$ is considered equally infectious as an unvaccinated individual infected with variant $\mathcal{V}$. This is a simplification – in reality, the viral load in infected, vaccinated individuals appears to decline faster (and is hence lower on average) [13]. In addition, for the same nasopharyngeal viral load (C_ts), the probability of detecting an infectious virus using cell culture is also slightly lower [48]. While the model can be extended to include some estimate of lower infectiousness for an infected, vaccinated group, we chose not to do so at present; we do not have a quantitative estimate of how reduced infectiousness translates into reduction in transmission probability in real-life settings, particularly as vaccinated individuals may behave differently.

Specifying the infections created by I^h is notationally far more complex than in equations (2) and (3). The issue is that I^h is not solely responsible for creating new infections with this same history at time t: I^h_t. This problem arises even in the most simplistic multi-variant-vaccine setting. Given compartments $\mathcal{C}^0$, $\mathcal{C}^1$, $\mathcal{C}^A$, and $\mathcal{C}^{1A}$, consider the effects of infection and vaccination. New infections I^A_t are produced by both I^A and I^{1A} when they infect members of S^0 or S^A, while new infections I^{1A} are produced when either I^A or I^{1A} infect members of S^1 or S^{1A}. Vaccine 1 is administered to a random member of either S^0 or S^A, and adds members to either S^1 or S^{A1}, respectively. Given these complexities, we provide simple equations that only show the new infections of a specific history. While it is possible to provide equations for the total new infections for a variant $\mathcal{V}$, this would complicate our expressions and amounts to summing over many different compartments that are distinct in our model. Later, equation (13) provides a variant-specific equation in order to compare infectiousness of variants in a population.

For consistency of notation, our generating equations describe the number of new infections of a variant $\mathcal{V}$ *within a compartment* $\mathcal{C}^g$. To do so, consider all compartment histories which are infectious with $\mathcal{V}$: $\mathscr{H}^{\mathcal{V}} = \{h \in \mathscr{H} \; s.t.\, h = \bar{h}\mathcal{V}$, for some history $\bar{h}\}$. Each $\mathcal{C}^h$ creates new infections in S^g according to equations (2) and (3). We then sum over these groups:

$$I_t^{g\mathcal{V}} \sim \text{Poisson}\left(\tilde{M}_t L_t \tilde{\gamma}^{g\mathcal{V}} |S_t^g| N^{-1} \sum_{h \in \mathscr{H}^{\mathcal{V}}} \sum_{m=1}^{\nu} \theta_{t-m}^h w_m\right) \quad \text{where} \tag{8}$$

$$\theta_s^h \sim \text{Gamma}(I_s^h k, \text{rate} = k(\lambda^{\mathcal{V}} \mathcal{R}_0)^{-1}) \tag{9}$$

and N is the total population size, $N = \sum_{\mathcal{C} \in \mathscr{C}} |\mathcal{C}|$.

Observe that θ_s^h for $h \in \mathscr{H}^{\mathcal{V}}$, depends on $\mathcal{R}_0^{\mathcal{V}} = \lambda^{\mathcal{V}} \mathcal{R}_0$ instead of $\mathcal{R}_{e,t}^{g\mathcal{V}}$. This is because θ_s^h gives the "native infectiousness" of – and is solely a property of – I_s^h, separate from the interaction between I^h and S_t^g in the environment experienced at time t. Similarly, the second sum of the Poisson argument, $\sum_{m=1}^{\nu} \theta_{t-m}^h w_m$, does not depend on g because it represents the total infectiousness of $\mathcal{C}_t^h$. New infections, however, depend on other groups h that are infectious with $\mathcal{V}$ and the environment through the remaining parameters in equation (8). If we ignore susceptibility, mitigation, and seasonality, i.e. $\tilde{\gamma}^{g\mathcal{V}} = \tilde{M}_t = L_t = 1$, and the proportion of infected people in the population is small, then $N^{-1} \sum_{h \in \mathscr{C}} \tilde{\gamma}^{h\mathcal{V}} |S_t^h| \approx 1$. In this case, equation (8) reproduces (2) in the original setting of the momentum model for a single variant [26].

2.1 Controllers

We assume a controller is interested in constraining the process of new infections, $I_t = \sum_{h \in \mathscr{H}} I_t^h$, and can manipulate $\tilde{M}_t$. Importantly, observed cases are distinct from the underlying process of infections, I_t, as not all infections are observed. Our controllers observe only a portion of infections as determined by the detection ratio, and then only some days after infection occurred due to the delay between infection and observation. These parameter values are discussed in Section 2.4 and Section 3.

Changes in non-pharmaceutical interventions (NPIs) are concretely implemented by setting $\tilde{M}_{t+1} = \delta \tilde{M}_t$ in equation (8). Thus, increase in NPIs such as mask mandates have the effect of a multiplicative decrease in transmissibility. We have explicitly assumed a certain compound *effect* of NPIs rather than specifying them. Our goal is not to prescribe which combination of NPIs to use, but to demonstrate differences in efficiency of containment strategies that result from using different statistics to guide the decision on the timing of the NPIs.

We consider two types of controllers which react to different statistics computed from case data. Both increase and decrease mitigation, $\tilde{M}$, by some proportion $\delta \in [0, 1]$ whenever they intervene. The first controller changes the effect of NPIs when daily cases pass pre-specified boundaries and is termed a "reactive" controller. This is a controller which increases NPIs when reported daily cases are over some high threshold (e.g. 150 per 100,000 over the last 14 days) and decreases NPIs for case numbers below a low threshold (e.g. 25 per 100,000 over the last 14 days) so long as case numbers are not increasing. A second type of controller, termed "proactive", also utilises an estimate of the effective reproduction number.

Let $\bar{\mathcal{R}}_{e,t}$ be the effective reproduction number given aggregate statistics which ignore compartment history and the type of infection. This is equivalent to taking expectations of our group- and variant-specific $\mathcal{R}_{e,t}^{h\mathcal{V}}$ over individuals in the population as well as the infectiousness of strains in $\mathscr{V}$. The distribution over compartments weights by the size of the susceptible group: $|S^h|$. Similarly, the distribution over variants weights by the current total infectiousness of the variant in the population: $\sum_{h\in\mathscr{H}^{\mathcal{V}}}\sum_{m=1}^{\nu}\theta_{t-m}^{h}w_m$. We have:

$$\begin{aligned}\bar{\mathcal{R}}_{e,t} &= \mathbb{E}_{\mathcal{V}}\mathbb{E}_h[\mathcal{R}_{e,t}^{h\mathcal{V}}] \\ &= W^{-1}N^{-1}\tilde{M}_tL_t\sum_{\mathcal{V}\in\mathscr{V}}\left(\sum_{g\in\mathscr{H}}\tilde{\gamma}^{g\mathcal{V}}|S_t^g|\sum_{h\in\mathscr{H}^{\mathcal{V}}}\sum_{m=1}^{\nu}\theta_{t-m}^{h}w_m\right) \quad \text{where} \qquad (10)\\ W &= \sum_{h\in\mathscr{H}}\sum_{m=1}^{\nu}I_{t-m}^{h}w_m. \qquad (11)\end{aligned}$$

Here, we have ignored the small number of currently infected individuals by dividing by N instead of $\sum_{h\in\mathscr{H}}|S^h|$.

Equation (10) provides a useful summary as an average effect implied by our model which accounts for variant heterogeneity, population diversity, and superspreading. Unfortunately, a controller cannot compute it as it depends on the unknown parameters θ^h, which describe the "momentum" of the disease [26]. A feasible estimator does not consider θ_t^h to be known, instead using the expected total current infectiousness of $\mathcal{V}$ given $\lambda^{\mathcal{V}}$ and $\mathcal{R}_0$:

$$\hat{\mathcal{R}}_{e,t} = W^{-1}N^{-1}\tilde{M}_tL_t\sum_{\mathcal{V}\in\mathscr{V}}\left(\sum_{g\in\mathscr{H}}\tilde{\gamma}^{g\mathcal{V}}|S_t^g|\sum_{h\in\mathscr{H}^{\mathcal{V}}}\sum_{m=1}^{\nu}\lambda^{\mathcal{V}}\mathcal{R}_0 I_{t-m}^{h}w_m\right) \qquad (12)$$

We note that the statistic above is not meant to be an ideal estimate of the effective reproduction number $\mathcal{R}_{e,t}$, but instead functions as a computationally efficient way to track the spread of infections in a way that is consistent with our simulation framework. While quantities such as $|S_t^g|$ are not known in practice, they can be estimated per variant and vaccine. In fact, this is done when initialising our model and is described extensively in Section 2.4 along with estimation of (12).

A "proactive" controller changes mitigation measures based on $\hat{\mathcal{R}}_{e,t}$ and case numbers. Behavior is the same as for the reactive controller, except that there is also an upper bound specified for $\hat{\mathcal{R}}_{e,t}$: when the effective reproduction number is higher than this upper bound, restrictions are increased. Reducing restrictions requires $\hat{\mathcal{R}}_{e,t} < 1$ in addition to low case numbers.

2.2 Vaccines and variants

Our model includes two types of vaccines and six SARS-CoV-2 variants. Vaccine types are summarised in two groups: i) mRNA vaccines which include both Pfizer-BioNTech's Comirnaty (BNT162b2) and Moderna's Spikevax (mRNA-1273); and ii) vector vaccines which include AstraZeneca's Vaxzevria/Covishield (AZD1222) and Janssen's COVID-19 vaccine (JNJ-78436735).

We consider six variants: the original wild-type (WT), Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) variants, and a hypothetical variant Omega. In general, their relative advantage and effective reproduction number depend on the composition of the population. The first row of Table 1 shows the assumed relative advantage of variant $\mathcal{V}$ over the wild-type, $\lambda^{\mathcal{V}} = \mathcal{R}_0^{\mathcal{V}}/\mathcal{R}_0^{WT}$, in a naïve population. The values are computed from recent estimates of $\mathcal{R}_e^{\mathcal{V}}/\mathcal{R}_e^{\mathcal{V}'}$ using

GISAID sequences across different countries which are then aggregated to produce a summary estimate for each variant [11]. In general, this value is thus confounded with acquired advantage due to immunity escape, although the departure appears small within the time frame of the study (until June 3, 2021). The values are consistent with estimates of $\lambda^{\mathcal{V}}$ assuming substantial immune escape [20, 52] and are on the lower boundary generally accepted for the transmissibility advantage of Delta [12].

We assume that the basic reproduction number of the original strain, $\mathcal{R}_0^{WT}$, is approximately 3.5. There is a wide range of estimates of $\mathcal{R}_0^{WT}$, ranging from about 2 to 6.5 [34, 39]. As $\mathcal{R}_0^{WT}$ depends on interaction networks in a society and will therefore plausibly differ between countries and regions, we use an estimate of $\mathcal{R}_0^{WT}$ from the early Austrian case data [18, p.13], corrected for the assumed seasonality of 40%.

Table 1 summarises both the assumed effectiveness of the vaccines against infection with SARS-CoV-2 variants as well as the estimates of resistance conferred by previous infection. We use estimates from the recent analysis by [43] (United Kingdom) concerning infections (with RT-qPCR's $\mathrm{C_t} < 30$) by the Alpha and Delta variants. While [43] assesses subjects PCR-tested in weekly intervals, it is limited to people younger than 65 years old. We thus extend the lower bound of effectiveness following [38], which also gives estimates for effectiveness of both mRNA vaccines against (symptomatic) infections with Gamma/Beta variants (in Ontario, Canada). We also use estimates from Brazil [25] for the reduction of transmission of the Gamma variant following full vaccination with vector vaccines, and from Qatar [1] and South Africa [35] for the Beta variant. For computational simplicity, we use the resistances after a full vaccination (typically, two doses), and assign this 2 weeks after the first dose of a vector vaccine or 3 weeks for an mRNA vaccine. This is because the error arising from assigning full immunity 1-2 weeks after second dose would be larger than neglecting slightly lower immunity between doses [5, 42, 43, 55]. We assume the percentage of people who do not follow up with the second dose (when required) is sufficiently low that it can be ignored.

The reduction of probability of reinfection is based on the estimates by [43] for the WT, Alpha, and Delta variants. We assume that the probability of reinfection is reduced by 87% $(84-90\%)$ upon prior infection with the same variant, and it is reduced less (by 77%, $(66-85\%)$) for the Delta variant when the previous infection was by a different variant (typically WT or Alpha). There is less reliable information on the reduction of re-infection for the Beta and Gamma variants. We use the model-based estimate by [20] of 70% cross-immunity (55-80%) for the Gamma variant. Based on the relative sensitivity of the variants to convalescent sera [6, 41] – and in the absence of a direct estimate of reinfection protection for the Beta variant based on a random cross-infection survey – we employ the 70% cross-immunity for the Beta variant as well. The resistances are set to approximately 75% for cross-immunity between (typically rare) combinations where we do not have direct data, and to approximately 85% for reduction in reinfection by the same strain. See Table 1 for exact values. The last two columns show resistances after immunity has waned, which we discuss in the next subsection.

2.3 Immunity waning and boosters

The waning of immunity from both vaccines and infections plays an increasingly important role in the evolution of the COVID-19 pandemic [40, 47]. Our compartment model can be generalised to this setting. We merely need a rule to determine when and how individuals experience waning as well as an additional "intervention vaccine", i.e. booster, given to previously vaccinated individuals.

One constraint is that our simulation does not track individuals, but groups. Therefore, there is no concept of "time since fully vaccinated" that can be used for a gradual waning of immunity. As such, waning needs to be implemented as a transition of people from a susceptible category S^h to a waned category, which will in general be written as S^{h-}. In order to account for the different characteristics of vaccine and variant waning, this is represented in two different ways in our history notation. Infectious variants are still given capital letters, e.g. W and A; after waning, however, these are changed to lower case, w and a. This changes a group signature from $h = WA$ to $g = Wa$. Note that the individual was infected by A *before* their resistance from W infection waned. Vaccine waning is represented by a minus sign (overloading the general notation from before), e.g. $h = A1$ becomes $g = A1-$. We assume for simplicity that a waned group cannot wane a second time. Therefore our waned vaccine group $A1-$ cannot wane to $a1-$; waning only occurs for the most recent event experienced. We do not consider this to be a significant oversimplification as the first waning event results in the majority of the resistance reduction experienced and the use of booster vaccines returns people to effectively un-waned groups.

The transition between groups happens at a single point in time as opposed to a gradual decrease in immunity. On average, the waiting time before waning is chosen in order to mimic a gradual, population-level decrease in immunity. All individuals in the susceptible subgroup S^h have an exponential waiting time before being moved to the waned group. The waiting time is determined by the number of days, in expectation, until a lower level of resistance is observed. Each day we thus move Poisson($S_t^h/180$) people from S^h to S^{h-}, such that, in expectation, all individuals in S_t^h have waned after a fixed waning period of 180 days. Specifically, we assume that over the waning time of 180 days, the resistance to infection by Delta (and older variants) after vaccination with two doses has declined to 40% (35-45%) for vector vaccines and 50% (40-60%) for mRNA vaccines [14, 53]. For the hypothetical variant Omega, we assume both lower initial resistance after vaccination (vector: 30% (0-55%); mRNA: 60% (50-65%)) as well as stronger waning over 180 days (vector: 0% (0-0.05%); mRNA: 0.05% (0-0.1%)). We assume that the immunity post-infection declines similarly as the one after getting a vector vaccine (see Table 1).

The booster shot is implemented as a new vaccine that is only given to people who have been vaccinated, regardless of their waned status. As the real "waning event" is not uniform over the entire population and it is not known whether or not someone has truly waned when given the booster shot, this is a parsimonious representation of the effects. For simplicity, we group all booster vaccine combinations together, such that there isn't any interaction effect between the booster received and any previous infection or vaccination. As the booster is implemented as a vaccine, resistances are updated according to the same rule: the resistance after receiving the booster is given by the maximum, per variant, of current resistance and the resistance provided by the booster. For Delta and older variants, the boosted resistance starts at 96% (94-98%), and wanes only slightly over 6-months, to 85% (75-90%). For the hypothetical variant Omega, we use the tentative values for vaccine effectiveness based on resistance to symptomatic infection with the Omicron variant, and assume boosted resistance starts at 70% (60-80%) and over 6 months, wanes to 30% (20-40%) - see Table 1).

For a simple population consisting of a single group, we have an average resistance behaving as in Fig 2. In practice, many waned compartments that receive the booster will have identical resistances. This is because their resistances $\gamma^{(h-)V}$ are below those of the booster vaccine against all variants $\mathcal{V}$, and because vaccines only raise resistances to a specified level. We do not collapse the compartments, however, so that the history of a compartment can be tracked in subsequent analyses.

2.4 Initialising the simulation

Given the initial conditions, equations (8) and (9) control all new infections and vaccine schedules specify new vaccinations. To initialise the simulation, we need to specify an initial collection of compartments $\mathscr{C}$, a vaccine schedule, and a level of NPIs. These are chosen to most accurately represent the pandemic in Austria.

Epidemiological data in Austria is gathered and provided by AGES (Agentur für Gesundheit und Ernährungssicherheit GmbH), the Austrian agency for health and food safety [2]. In addition to tracking the number of cases, hospitalisations, deaths, and tests, AGES tracks the genome sequencing of SARS-CoV-2 samples gathered in Austria to monitor the prevalence of variants of concern (VOC) [3]. The first confirmed case of the Alpha variant in Austria was on January 3,

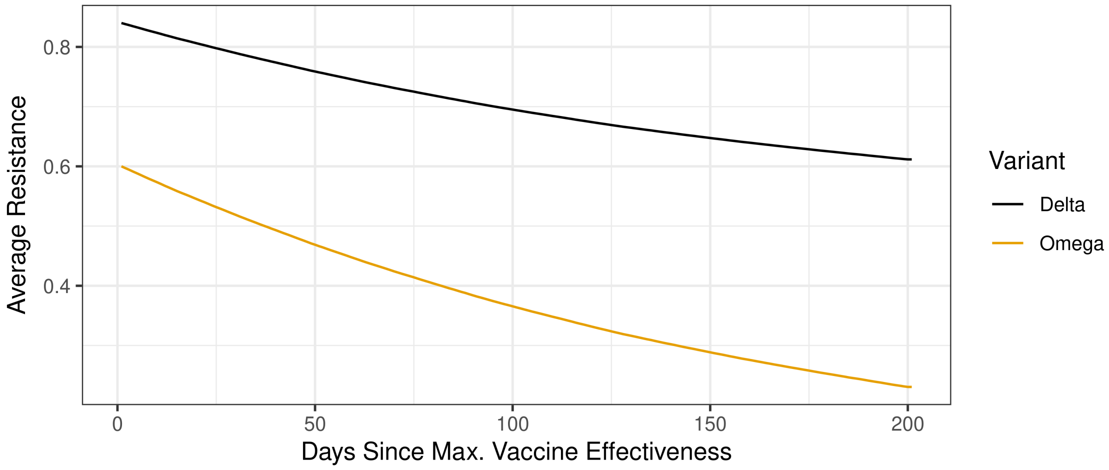


Figure 2: **Average resistance against Delta and the hypothetical variant Omega due to immunity waning in an mRNA-fully-vaccinated population**.

2021. A VOC sentinel system was subsequently established with one PCR test lab per county submitting a random sample of SARS-CoV-2 specimens for complete genome sequencing.

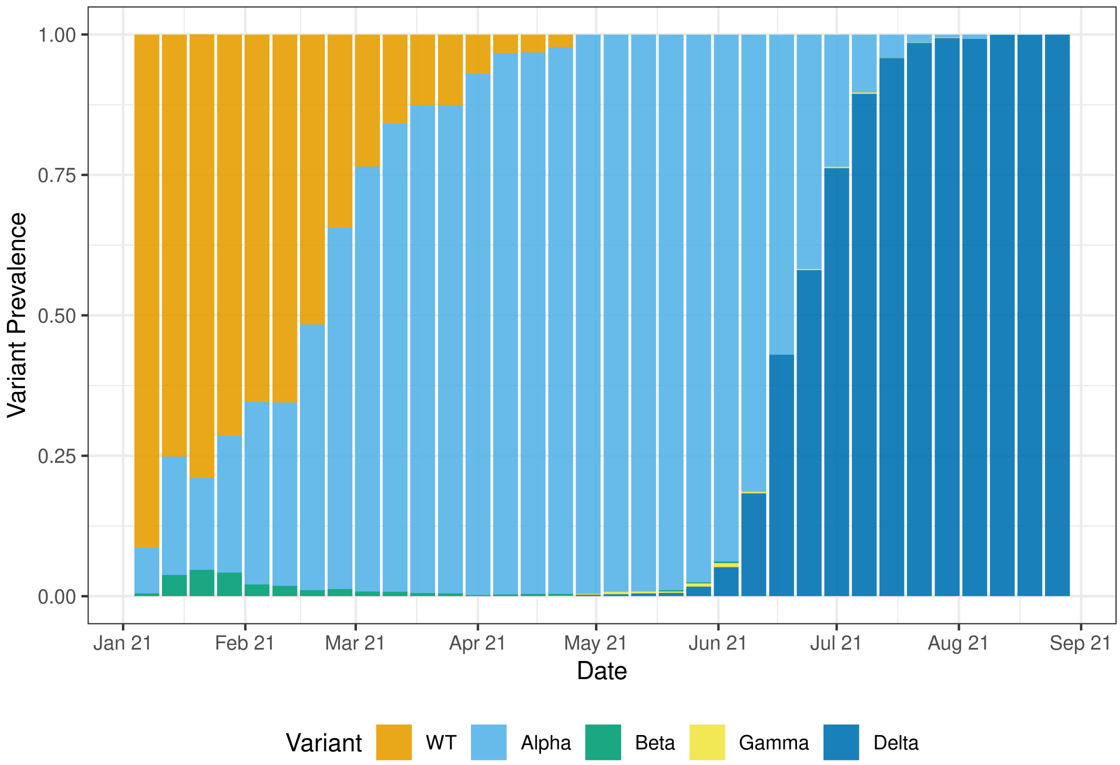


Figure 3: **Reported prevalence of different SARS-CoV-2 variants in Austria in 2021** .

VOC prevalence in Austria is published online and updated roughly every one to two weeks. Fig 3 provides the historical reported variant prevalences. 2021 has already seen two new variants emerge and quickly become dominant. In Austria, Alpha took approximately 15 weeks between emergence and dominance, whereas Delta took a mere 10 weeks. Both Beta and Gamma variants were observed in Austria, though neither variant made serious headway into the population.

When simulations are initialised for a given date, the relative size of recovered compartments are computed using data from Fig 4. To compute the number of people who were previously infected, we must account for the detection ratio throughout the entirety of the pandemic. Based on [46] and consistent with seropositivity in Austria from mid November [51], we assume that 12% of the Austrian population was infected by the wild-type before January 1, 2021. For 2021, we specify the detection ratio which measures the (age-averaged) probability that an infected individual is diagnosed and appears in the official case statistics. Due to increases in the availability and use of COVID-testing in Austria, we use the following estimates for the detection ratio in 2021: 1/2.3 for January and February, 1/2 for March and 1/1.4 for April and beyond. These are consistent with model-based estimates for Austria which use hospitalisations and deaths to learn about the proportion of unreported infections [7, 18].

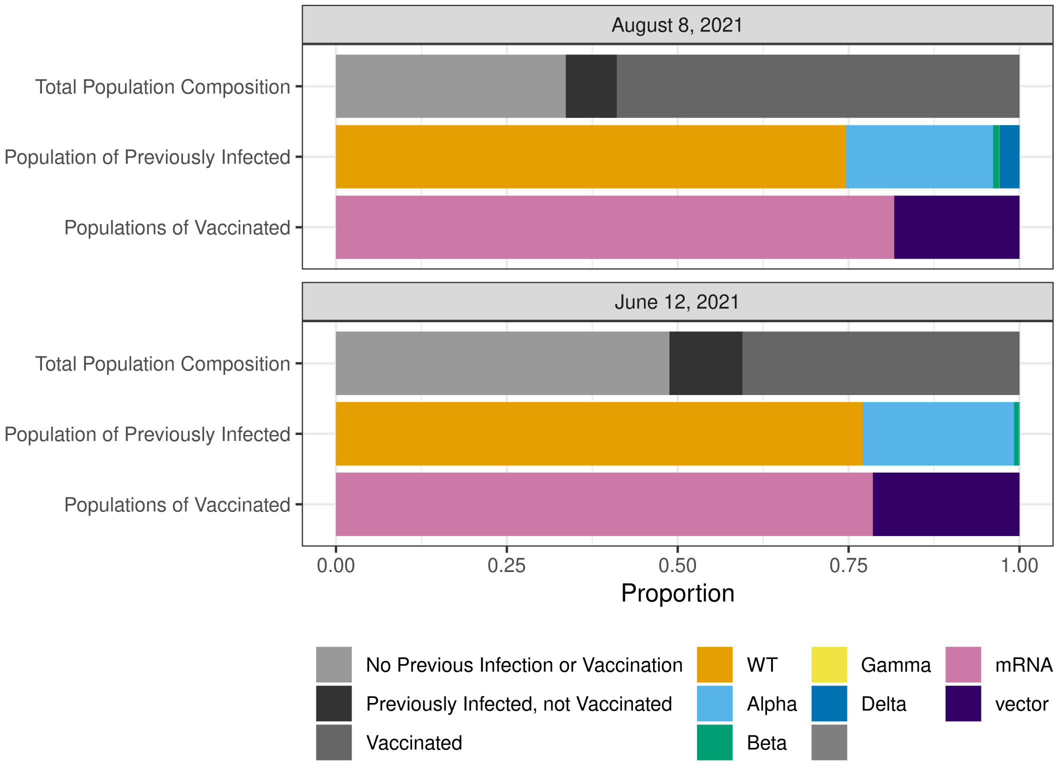


Figure 4: **Estimate of the population composition in Austria on June 12 and August 8, 2021**. Note that while Delta was dominant by August, 2021, the absolute number of infections is comparably low.

We use a vaccine schedule to match that of Austria throughout the simulation period, as the actual vaccination plan is considered to be a background setting of our model as opposed to an intervention by a controller. We use the 7-day median of administered first doses during each calendar week [10] up until August 8, 2021. When simulating beyond the window of available data, we use the latest 7-day median and administer this many doses until the expected upper bound on vaccinations is reached. Currently, this bound is 85% of the population. As of August 8, 2021, the distribution of administered vaccines was 72% Pfizer-BioNTech's Comirnaty, 10% Moderna's Spikevax, 15% AstraZeneca's Vaxzevria, and 3% Janssen's COVID-19 vaccine. Beyond August 8, the distribution of newly administered doses is 74% from Pfizer, 3% Moderna, 22%

Janssen , and 1% from AstraZeneca. The corresponding real data for booster vaccines is used until mid December 2021. Vaccine booster shots began being used in September 2021, with uptake increasing rapidly starting in October and November 2021.

A final step to calibrate our model with current case numbers is to set an initial effect of NPIs. This is done by equating the implied reproduction number from the simulation, $\hat{\mathcal{R}}_{e,t}$, to the observed $\mathcal{R}_e$ in Austria at the time the simulation starts. More concretely, $\hat{\mathcal{R}}_{e,t}$ from equation (12) is simplified at initialisation as our compartment structure features no interaction groups. We assume that all individuals that were previously infected with the wild-type or Alpha are equally likely to have been vaccinated, while people with other infections were not vaccinated as the other variants primarily appeared later. Tacitly, this assumes that those who were previously infected and then vaccinated do not receive an additional benefit due to their initial infection. This is in line with the update rule for infection followed by vaccination, though could also be considered a result of infection waning due to long-past infections. The number of recovered individuals of each type, $|S_t^h|$, is estimated by taking the cumulative cases for 2021 and scaling by the inverse detection ratio for each time period given above. A proportion of S_t^h for wild-type and Alpha are removed from these groups and placed in the vaccinated groups. This specifies $|S_t^h|$ and I_{t-m}^h for all individual variants and vaccines, allowing us to compute $\hat{\mathcal{R}}_{e,t}$ up to the missing $\tilde{M}_t$ term, which is calibrated to the observed $\mathcal{R}_e$.

3 Results

This section presents simulation results for a setting constructed to be similar to that in Austria. This serves to anchor the simulation in a realistic setting, though the high-level results are applicable beyond Austria as well. To aid comparisons to other countries, data are reported as cases/100,000 inhabitants.

3.1 Comparing variants

Ultimately, our model requires many parameters to be set which govern the resistance one variant provides to infection from others as well as resistances conferred due to vaccines. There are three relevant dimensions in which we allow variants to differ: the basic reproduction number $\mathcal{R}_0^h$, immune escape after vaccination, and immune escape after infection. It is important to view these as three separate components, and we note that increasing severity in multiple categories may over-estimate the true risks of different variants. We assume that the generation interval does not significantly differ among variants, which is consistent with known estimates for Delta and older variants [8, 24, 44]. The Supporting Information explores other settings in which variants have generation intervals with lower mean, such as is possible for Omicron.

Table 1 discusses the parameters we use and how they differ between variants. We extend that discussion here by showing how these parameters translate into dimensions of primary concern. To simplify a visual presentation, Fig 5 only shows $\mathcal{R}_0^{\mathcal{V}}$ and vaccine effectiveness, which are the two most important measures that are independent of population composition. We further distinguish vaccine effectiveness between mRNA (○) and vector (△) vaccines. Both estimates and their uncertainty are summarised in Fig 5, in which the parameters are drawn from a truncated normal distribution with mean and truncation points given in Table 1. The uncertainty represented in Fig 5 is also included in our simulations.

While Fig 5 summarises raw parameters, it is not sufficient to characterise which variants are of greater concern within a given population. We also show how these variant characteristics map to variant risk and are affected by the rise of immune resistance due to vaccinations. These are the two dimensions of primary interest: our risk measure combines both the variant profiles and the background population characteristics, while vaccinations provide the long-term solution to the pandemic. Therefore, our graphs summarise which variants are of greatest concerns to regions with

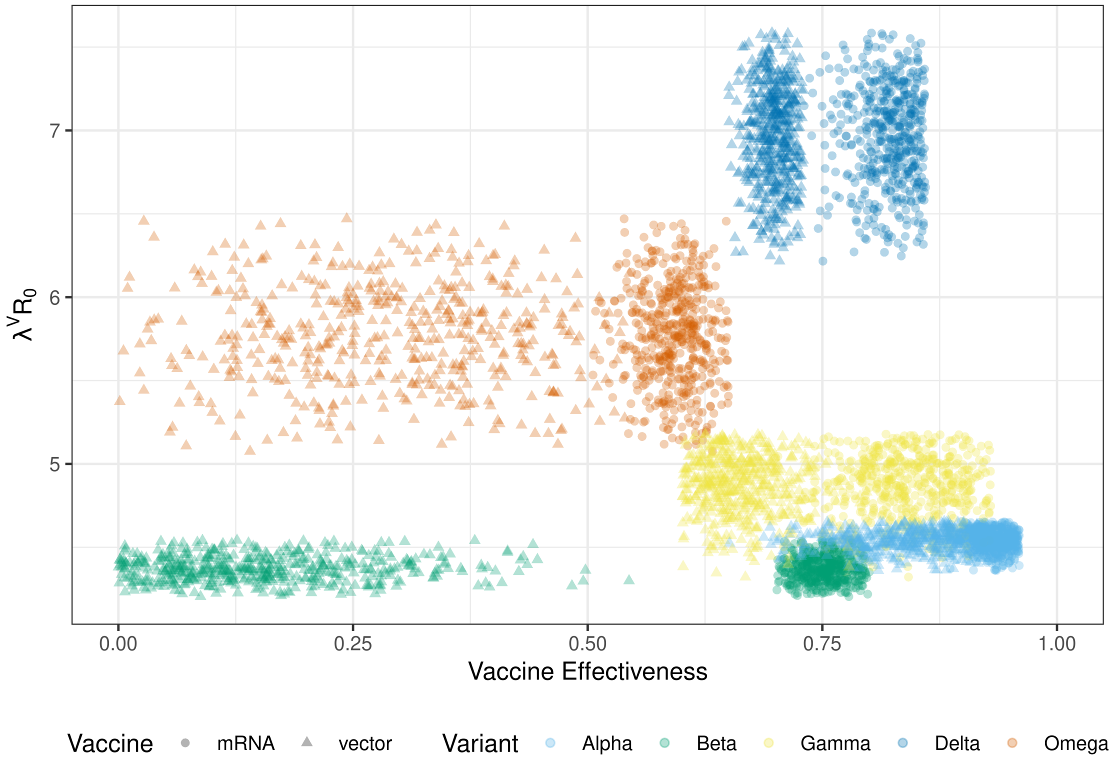


Figure 5: **Basic reproduction number vs vaccine effectiveness for all variants included in our simulations**: A(lpha), G(amma), D(elta), and O(mega). The Omega variant is constructed to analyze hypothetical scenarios. Note that WT is only present in the history of previous infections.

different vaccination rates. $\mathcal{R}_{e,t}^{h\mathcal{V}}$ from equation (4) does not easily allow one to compare variants because it specifies both the group that is being infected as well as depends on mitigation and seasonality. Therefore, we remove the time varying components ($\tilde{M}_t L_t$) and integrate $\mathcal{R}_{e,t}^{h\mathcal{V}}$ over the susceptible populations S^h:

$$\rho_t^{\mathcal{V}} = (\tilde{M}_t L_t)^{-1} \mathbb{E}_h[\mathcal{R}_{e,t}^{h\mathcal{V}}] = N^{-1} \lambda^{\mathcal{V}} \mathcal{R}_0 \sum_{h \in \mathscr{H}} \tilde{\gamma}^{h\mathcal{V}} |S_t^h|. \tag{13}$$

$\rho_t^{\mathcal{V}}$ is the conceptual driver of outbreaks in our model as it represents the reproduction number of a variant within a particular population by accounting for susceptibility due to immune evasion.

In order to plot $\rho_t^{\mathcal{V}}$ as a function of the vaccination rate $r \in [0, 1]$, we need to consider how the population composition would change and how this would be reflected in the size of our compartments. As we have created a realistic population distribution for Austria on August 8, 2021, we want to maintain this realism over the range of possible infection rates. Therefore, we split the population into two sets: vaccinated and unvaccinated. The unvaccinated cohort $\mathscr{C}_{\mathsf{uv}} = \{\mathcal{C}^h \in \mathscr{C}\, s.t.\, \mathbb{N} \cap h = \emptyset\}$ and the vaccinated cohort $\mathscr{C}_{\mathsf{va}} = \{\mathcal{C}^h \in \mathscr{C}\, s.t.\, \mathbb{N} \cap h \neq \emptyset\}$. As vaccines are assumed to be given independently of whether or not someone has been previously infected and recovered, this maintains our population distribution. Let $\mathscr{H}_{\mathsf{uv}}$ and $\mathscr{H}_{\mathsf{va}}$ contain the partitioned histories corresponding to $\mathscr{C}_{\mathsf{uv}}$ and $\mathscr{C}_{\mathsf{va}}$, respectively. Lastly, let $N_{\mathsf{va}} = \sum_{h \in \mathscr{H}_{\mathsf{va}}} |S_t^h|$ and $N_{\mathsf{uv}} = \sum_{h \in \mathscr{H}_{\mathsf{uv}}} |S_t^h|$ be the size of the vaccinated and unvaccinated susceptible populations, respectively. Note that $N \approx N_{\mathsf{va}} + N_{\mathsf{uv}}$ as we ignore the comparatively small set of people that are currently infected. We then have

$$\rho_t^{\mathcal{V}}(r) = \lambda^{\mathcal{V}} \mathcal{R}_0 \left(\frac{r}{N_{\mathsf{va}}} \sum_{h \in \mathscr{H}_{\mathsf{va}}} \tilde{\gamma}^{h\mathcal{V}} |S_t^h| + \frac{1-r}{N_{\mathsf{uv}}} \sum_{h \in \mathscr{H}_{\mathsf{uv}}} \tilde{\gamma}^{h\mathcal{V}} |S_t^h| \right). \tag{14}$$

Observe that equation (14) is merely a convex combination between $\rho_t^{\mathcal{V}}$ computed on two different populations: those who are vaccinated (first term) and those who are not (second term). For example, suppose that there is no resistance conferred by previous infection and perfect resistance conferred by vaccination ($\tilde{\gamma}^{h\mathcal{V}} = 1, \forall h \in \mathscr{H}_{\mathsf{uv}}$ and $\tilde{\gamma}^{h\mathcal{V}} = 0, \forall h \in \mathscr{H}_{\mathsf{va}}$). In this case, equation (14) simplifies to $\rho_t^{\mathcal{V}}(r) = \lambda^{\mathcal{V}} \mathcal{R}_0 (1-r)$, which is just the basic reproduction number times the proportion of unvaccinated individuals.

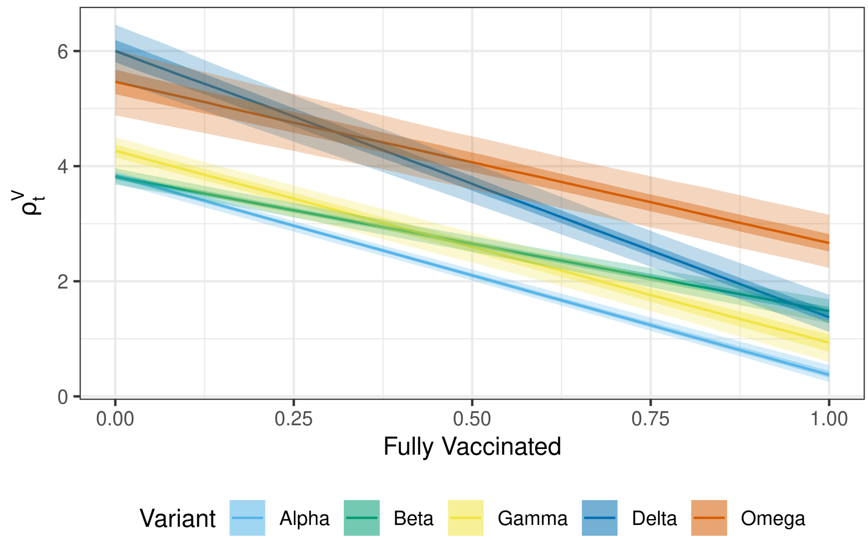


Figure 6: **The reproduction number of variants accounting for immunity within the population, before mitigation and seasonality.** $\rho_t^{\mathcal{V}}$, defined by equation (14), is shown as a function of the proportion of the population that is vaccinated. Omega is a hypothetical variant with a higher immune escape: its *relative* advantage thus increases as vaccination level does. Shaded regions correspond to 50% and 95% prediction intervals resulting from the uncertainty in viral parameters summarised in Fig 5. The assumed composition of the population is depicted in Fig 4; it reflects Austria on August 8, 2020, with about 20% of population recovered from infections, mainly by WT (75%), Alpha (22%) and Delta (3%). For simplicity, we assume that vaccination is independent of the infection history. Different population compositions and immunity waning are addressed later (Figs 7 and 16).

Fig 6 shows how $\rho^{\mathcal{V}}(r)$ changes as a function of the proportion of the population that is fully vaccinated, r. In a highly vaccinated population, the Delta and Beta variants are estimated to be similarly infectious, but they diverge significantly in populations with a lower percentage vaccinated. The bands in Fig 6 capture the uncertainty in parameter values shown in Fig 5. The ranking of risks only switches when a large proportion of the population is vaccinated (thus increasing the average resistance against variants, which were highly transmissible in a more naïve population). This area is unlikely to be reached without wide-spread and thorough vaccination campaigns. For example, only 88% of the Austrian population is in the 12+ age category, and nearly all of this category would need to be vaccinated to reach the change point.

In order to understand what a future outbreak could look like, we hypothesise a new variant, Omega, with lower basic reproduction number than the currently dominant Delta variant but also higher immune escape. This configuration was chosen in order to create a realistic problem setting that can continually affect regions even after successful vaccination campaigns or after a high number of previous infections.

As Fig 6 uses the vaccine distribution and proportion infected as observed in Austria, it is useful to further demonstrate what $\rho^{\mathcal{V}}(r)$ could look like for other populations with different vaccination campaigns as well as different histories of infections. For example, some countries such as Singapore and New Zealand have had minimal local transmissions and thus little infection-induced immunity. Others, such as the United Kingdom, have experienced higher infection numbers and thus benefit from infection-induced immunity. The behavior of outbreaks in these regions is governed by $\rho^{\mathcal{V}}(r)$ in our models, as this summarises the combined effect of population make-up and variant characteristics. Thus plotting this statistic for relevant population compositions gives insight into the variants that will be of highest risk in different regions. Fig 7 shows four extremal points for populations and vaccination strategies: 0% vs 40% previous infections and 100% mRNA vaccine usage vs. 100% vector vaccine usage. As equation (14) is linear, any point in-between these extremes can be faithfully represented as a linear interpolation between the graphs in Fig 7.

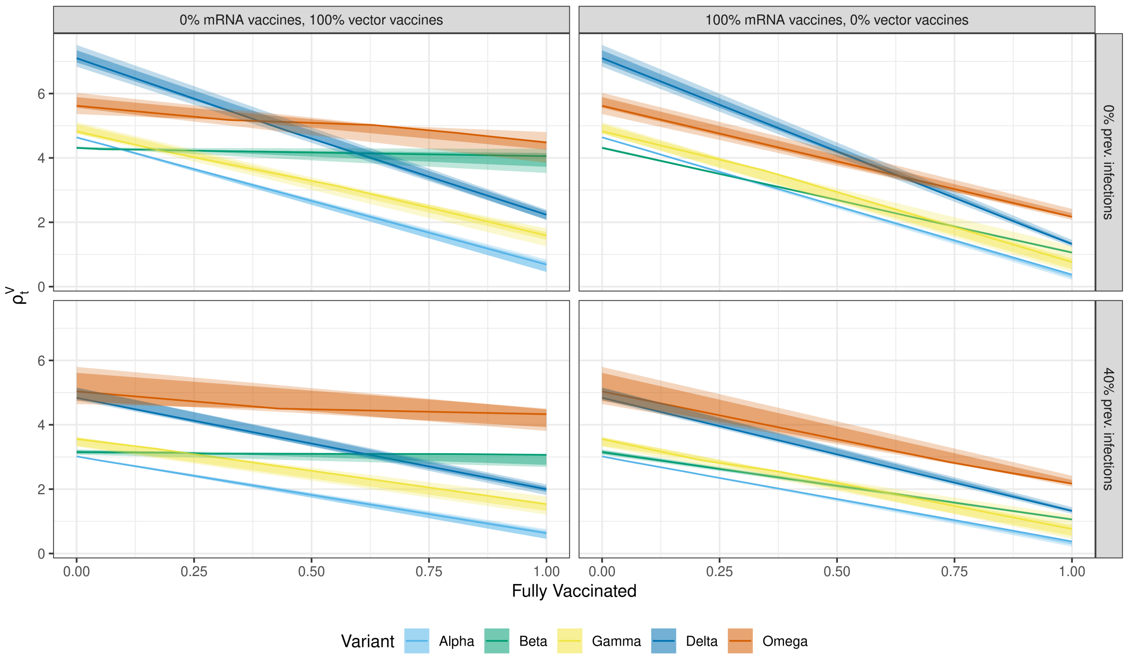


Figure 7: $\rho_t^{\mathcal{V}}$ **accounts for a population-specific reproduction number of each variant.** The charts represent the four vertices of the simplex of vaccine/infection population space. Top row assumes no previous infections; bottom row assumes 40% have acquired immunity due to previous infections. The variants of previous infections are assumed to be distributed according to the estimated proportions in Austria on August 8, 2021, given in Fig 4. We assume that vector vaccines (left) confer lower resistance against infection than mRNA vaccines (right); see Table 1. Shaded regions correspond to 50% and 95% prediction intervals resulting from the uncertainty in viral parameters summarised in Fig 5.

For example, a country with an epidemic and vaccination history similar to the UK, is represented in the bottom-left image of Fig 7: it has a high degree of previous infections and has relied heavily on vector vaccines throughout the first half of 2021. We see that Omega is the variant of highest risk essentially throughout the entire domain of vaccination proportion. Furthermore, given that we assume that vector vaccines do not protect well against infections with Omega, this risk barely decreases as vaccinations increase, although we see the ordering of risk change among the other variants as vaccination percentage changes. In the opposite corner, we see a country which has had few previous infections and has mainly used mRNA vaccines, such as Singapore. It is clear from Fig 7 that for low vaccination levels, all variants pose a higher risk than they did to the United Kingdom (as there is no prior immunity due to previous infections), with Delta being considerably riskier than Omega. When vaccination rates are higher than approximately 70%, however, the ordering switches and Omega becomes the riskiest - with the highest $\rho_t^{\mathcal{V}}$.

$\rho^{\mathcal{V}}$ is the conceptual driver of outbreaks in our model as it represents the reproduction number of a variant within a particular population, accounting for its susceptibility due to immune evasion. Furthermore, herd immunity is understood in context of current mitigation. This means that populations with greater acquired immunity (lower $\rho^{\mathcal{V}}$) can use fewer NPIs to receive the benefit of dissipating epidemics. Populations with higher $\rho^{\mathcal{V}}$ will need either need higher NPIs in order to suppress an outbreak, or require very strict border controls to prevent importing and allowing community spread of a high-risk variant.

3.2 Controller types

One long-standing question has been how best to control COVID-19 outbreaks when they arise. This subsection explores which statistics should be considered when determining whether to increase or decrease mitigation measures, particularly after the introduction of a new SARS-CoV-2 variant with higher effective reproduction number. Two controller types are considered that either respond to increases in case numbers ("reactive" control) or to increases in an estimate of the reproduction number ("proactive" control). We find that using an estimate of the effective reproduction number in addition to case numbers is a much more efficient strategy.

Responding to the effective reproduction number further helps to address a potential endogeneity effect due to increasing prevalence of COVID-19, i.e., individuals may modify their behavior when the situation either worsens or improves. Most public reporting discusses solely case numbers, so it may be reasonable to assume that individuals base decisions more on either absolute case levels or strong increases in cases. This can be problematic at the start of a wave if case numbers are incredibly low: purely measuring absolute increases or case-number thresholds may trigger a response too slowly. On the other hand, as the reproduction number is not given the same attention in the media, individuals may not change behavior dramatically when it changes. This helps connect NPIs to simulation dynamics, as individuals are not also responding to the same statistics as our NPIs.

As mentioned above, we consider two specific control settings. The first, termed "reactive", only responds to case numbers. There is an upper bound, above which stricter NPIs are used, and a lower bound, below which NPIs are relaxed. This is crafted to mimic the EU protocols for measuring the riskiness of non-essential travel, which assign color codes to regions depending on their publicly reported epidemiological data such as 14-day rate of cases, deaths, and/or tests administered [17]. Being below the lower boundary corresponds to being "green" whereas above the upper corresponds to being "red". In our results, we show a reactive controller with two different sets of thresholds. The first set, referred to as "low-positivity", uses a lower bound of 25 cases per 100,000 and an upper bound of 150 cases per 100,000 (both measured over a 14-day period). This corresponds the recommendations for a low test-positivity region. The second set, referred to as "high-positivity", uses also a lower bound of 25 cases per 100,000 but an upper bound of 50 cases per 100,000 (both measured over a 14-day period). This is far stricter and corresponds to recommendations for a high test-positivity region.

The second control, termed "proactive", responds to both case numbers and the effective reproduction number $\hat{\mathcal{R}}_{e,t}$ in equation (12). The same thresholds for case numbers are used as by the reactive control. There is also an upper threshold of 1.2 for $\hat{\mathcal{R}}_{e,t}$. The decision rule is as follows: increase restrictions if either $\hat{\mathcal{R}}_{e,t}$ or cases are above the corresponding upper threshold.

Conversely, restrictions are reduced if both $\hat{\mathcal{R}}_{e,t} < 1$ and cases are below their lower threshold. In all other instances, make no changes. For both controllers, when mitigation is modified, it is assumed that a 20% change in mitigation is made (both when strengthening and relaxing NPIs).

3.2.1 Delta

We simulate two main scenarios corresponding to both the growth of a new dominant variant and projections for infections after the variant permeates the population. For concreteness and validation in this section, the new variant is the Delta variant. We simulate infection trajectories using initial conditions when Delta accounts for 20% of current cases. This simulation provides two benefits: first, it provides valuable model validation by starting the simulation when 20% *was* accurate for Austria and comparing simulations to observed cases and statistics; second, it furnishes a sample case for what can happen when a new variant with similar transmissibility advantage reaches this threshold.

We start the simulation on June 12, 2021, as that corresponds to the AGES estimates for Delta prevalence in Austria (see Fig 7 in the Supporting Information). The same initialisation process was used as discussed previously, merely until June 12 instead of August 8. All other parameters needed to initialise our simulations are also taken from observed data on this day. This includes history of new cases, the proportion of population that is vaccinated or previously infected, etc. Fig 8 shows that our model accurately forecasts the proportion of Delta cases as measured by AGES: this holds true regardless of whether a proactive or reactive control is used, as seen in the bottom panel Fig 9. The result is independent of the controller as the controller affects all variants equally.

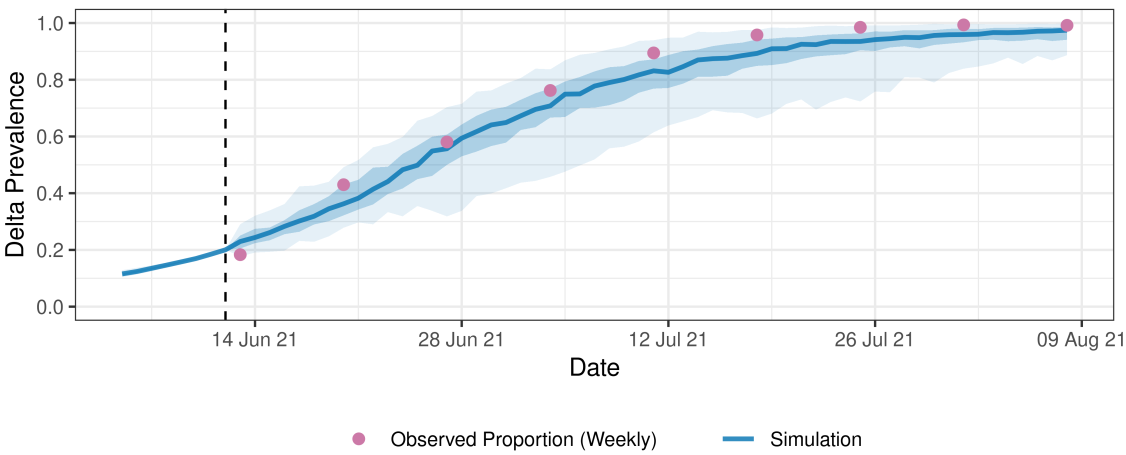


Figure 8: **The changing proportion of new Delta cases is accurately predicted**. The simulation is initialised using information available on June 12, when the observed proportion of Delta in Austria is 20%. Shaded regions correspond to 50% and 95% prediction intervals.

Simulation results for the low-positivity thresholds are shown in Fig 9, which shows daily incidences (case numbers), the effective mitigation level $\tilde{M}_t$ (reduction in $\hat{\mathcal{R}}_{e,t}$ due to NPIs), and the current estimate of the effective reproduction number $\hat{\mathcal{R}}_{e,t}$. These additional graphs can be used to more precisely monitor both the simulated COVID-19 epidemic as well as the control process. The first observation from Fig 9 is that cases under the reactive control correctly match observed cases around the first peak in mid September. Its intervention history roughly corresponds to Austria's, which loosened restrictions slightly over the summer. Notably, the reactive controller begins increasing restrictions around this peak whereas Austria did not. We note that the only real data used beyond the June 12 start date is the vaccination schedule.

Both controllers relax restrictions at roughly the same point (early July), which approximately coincides with the start of the Green Pass program for European tourism on July 1. Approximately one month later, however, the proactive control would increase mitigation measures again to prevent the start of a new outbreak. As case numbers were so low during June, merely looking at new cases yields no increase in NPIs for some time. Alternatively, increasing mitigation earlier stabilises both the effective reproduction number as well as the the number of new cases.

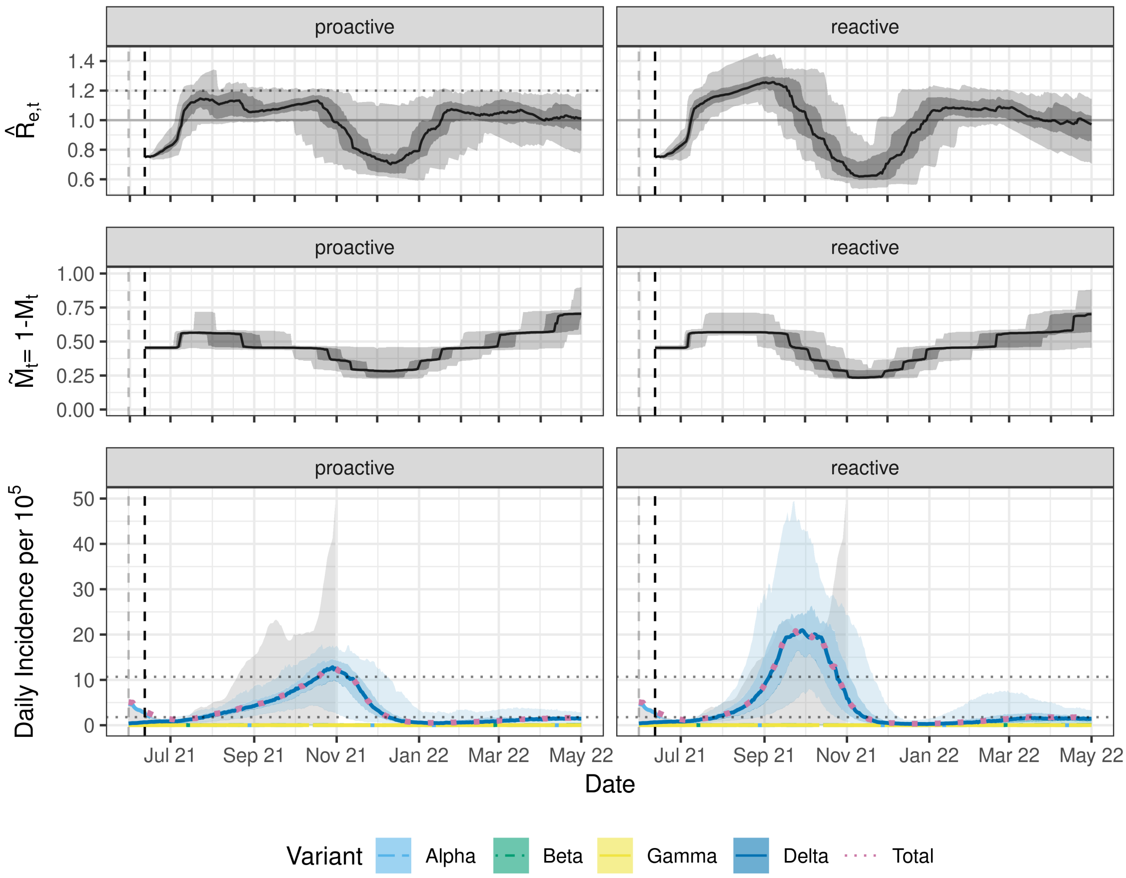


Figure 9: **Responding to the effective reproduction number is more efficient than only using case numbers.** The top row shows the effective reproduction number, middle row the effect of interventions on $\hat{\mathcal{R}}_{e,t}$ (where $\tilde{M}_t = 1$ means no NPIs), and bottom row the daily incidence per 100,000 inhabitants. The simulation starts on June 12, 2021 (black dashed vertical line) when Delta prevalence was at 20% and Alpha was the dominant variant. Initialising the model requires use case numbers from the previous 13 days (gray dashed vertical line). The case thresholds are shown as dotted horizontal lines and coincide with the WHO recommendations for change in NPIs when positivity rate is low [17, 56]); note that the thresholds, 25 resp. 150 per 100,000 within 14 days, are divided by 14 as the y-axis shows daily incidence. The shaded regions gives the 50% (dark) resp. 95% (light) prediction interval. The 7-day moving average of the actual incidence in Austria is given by the height of the gray, shaded area, and data past October is not shown as cases were allowed to dramatically increase under relaxed NPIs. The simulation under reactive control accurately predicts new cases three months after initialisation.

We can also precisely compare both the number of total infections as well as the number of active cases, i.e., those people who would be placed in quarantine. Even if infections are mild and do not require hospitalisation, economies suffer if too many people are quarantined (isolated) at any given time. We compute total number of quarantined cases as merely the sum of new cases over the preceding 10 days. In certain scenarios, more contacts of positive cases may be placed in quarantine, but this is general would just lead to a multiple of the active cases, leaving the general conclusions intact. Proactive control not only reduces the median infections and peak quarantine cases, but does so much more reliably: the observations cluster much more strongly around the median. As seen in Fig 10, reactive control has not only higher median values, but also a much more skewed distribution: it is possible to experience massive spikes before the outbreak is brought under control. Given the skewness of the distributions for the reactive controller, difference in medians is measured via a Wilcoxon Rank Sum test.

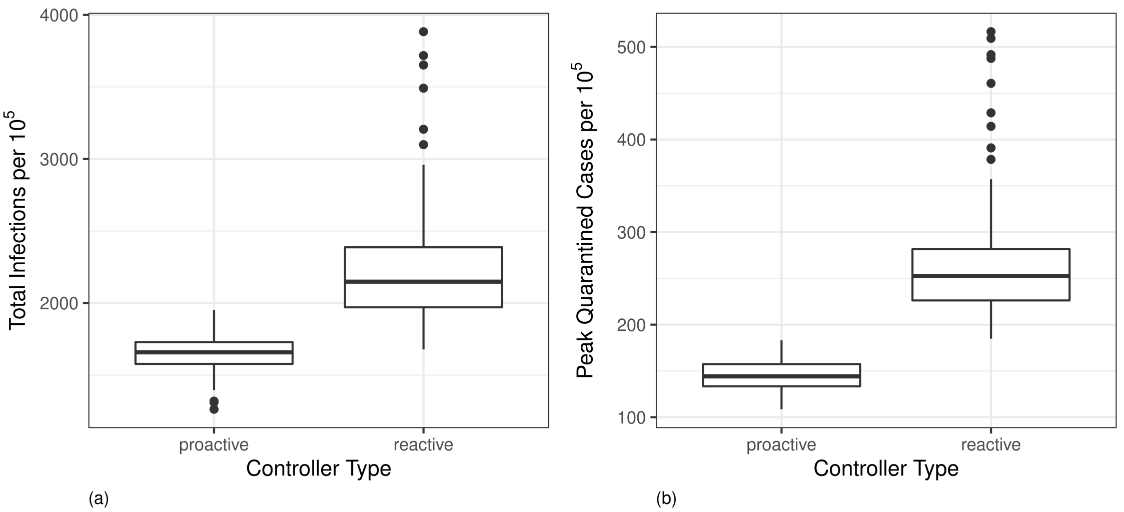


Figure 10: **Proactive control reduces total infections as well as peak numbers in quarantine.** With proactive control, the extent on NPIs reflects not only the active cases but also the estimated reproduction number $\hat{\mathcal{R}}_{e,t}$ (see Section 2.1). The difference in medians and corresponding 95% confidence intervals for total infections per 100,000 (over the simulated period) and peak quarantined cases per $100,000$ are, respectively, 509.6 (450, 580) and 108.2 (99, 118).

Comparing the distributions of the strengths of NPIs used by the two controllers provides a more rigorous understanding. Fig 11 shows the density of the difference between the observed mitigation and the "balanced" mitigation level, $\tilde{M}_t^*$, defined to be that value for which the effective reproduction number $\hat{\mathcal{R}}_{e,t} = 1$ in equation (12). Note that $\tilde{M}_t^*$ changes over time due to changes in variant prevalences, seasonality, and acquired immunity in the population. As a reminder, lower values of $\tilde{M}_t$ correspond to stronger NPIs. We see that proactive control trades more time with moderate mitigation for less time at extreme mitigation.

Next, we examine the robustness of these conclusions by using stricter case number thresholds for the controllers as recommended for a higher positivity rate. Fig 12 shows that with stricter case thresholds, simulated cases are much lower because NPIs are triggered more quickly for both controllers. Yet, under a reactive controller, case numbers and the effective reproduction number exhibit a "yoyo" effect, in which they relatively rapidly cycle through periods of increase and decrease. This effect can also be seen for the low-positivity thresholds of Fig 9, but the timescale is much longer. In general, large peaks lead to sufficiently strict NPIs that subsequent peaks appear much later in our simulations. The proactive controller is much more efficient and eliminates this

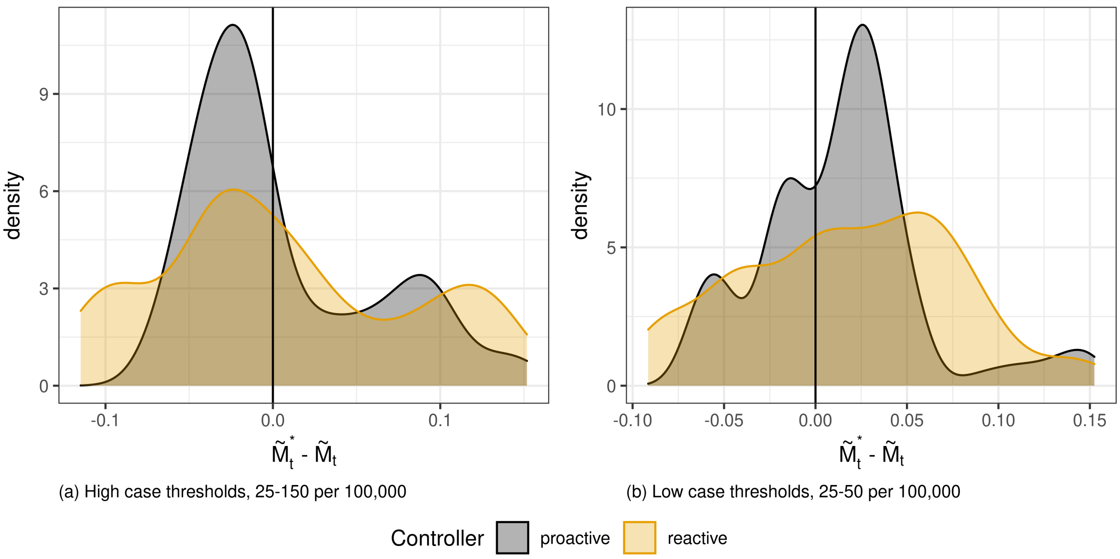


Figure 11: **Proactive control uses mitigation close to the "balanced" level.** We define the "balanced" mitigation level, $\tilde{M}^*$, to be that value for which the effective reproduction number $\hat{\mathcal{R}}_{e,t} = 1$. Note that lower values of $\tilde{M}_t$ correspond to stronger NPIs. We can see that with higher case thresholds (a), the proactive controller spends more time around this value as well as avoids extreme deviations. With lower case thresholds (b), the proactive controller fares even better, as holding mitigation $\tilde{M}_t$ slightly below $\tilde{M}^*$ is sufficient to alleviate the epidemic without unnecessarily strict NPIs.

behavior almost entirely. In particular, the left column of Fig 9 shows on average *less strict* NPIs than Fig 12.

The box plots of total infections and peak quarantined cases (see Supporting Information, Fig 1) show that with stricter thresholds, median values are closer together. Reactive control fails to provide reliable control, however, in that some simulated trajectories still produce many infections and quarantined cases. The difference in median total infections per 100,000 is 53.2 (95% CI: 29, 79) and the difference in median peak quarantined cases is 16.8 (95% CI: 14, 20). That being said, the 97.5% quantile of peak quarantined cases of the reactive controller is twice as high as that of the the proactive one. The mitigation graph for this setting is shown in Fig 11. Proactive control continues to use a more measured response, whereas reactive control prefers more extreme mitigation levels. In fact, the proactive controller primarily holds $\tilde{M}_t$ slightly below $\tilde{M}_t^*$, which is sufficient to alleviate the epidemic without unnecessarily strict NPIs.

From the point of view of feasibility, the proactive controller makes far fewer interventions than the reactive controller. While a strict reactive controller does keep case numbers low, this results in interventions which occur almost every two weeks. This is the minimum period that we specify in our model as a gap between interventions. It is unlikely that a government would be able to so regularly change policy.

3.2.2 New variant: Omega

This section introduces a new hypothetical SARS-CoV-2 variant that occupies both a reasonable and empty region of the infectiousness graph in Fig 6. The hypothetical variant, termed Omega, has a lower basic reproduction number than Delta but evades immunity after vaccination or infection by older variants more easily. This provides a scenario in which even a relatively highly vaccinated community will still experience an outbreak and the possibility to explore policies used during

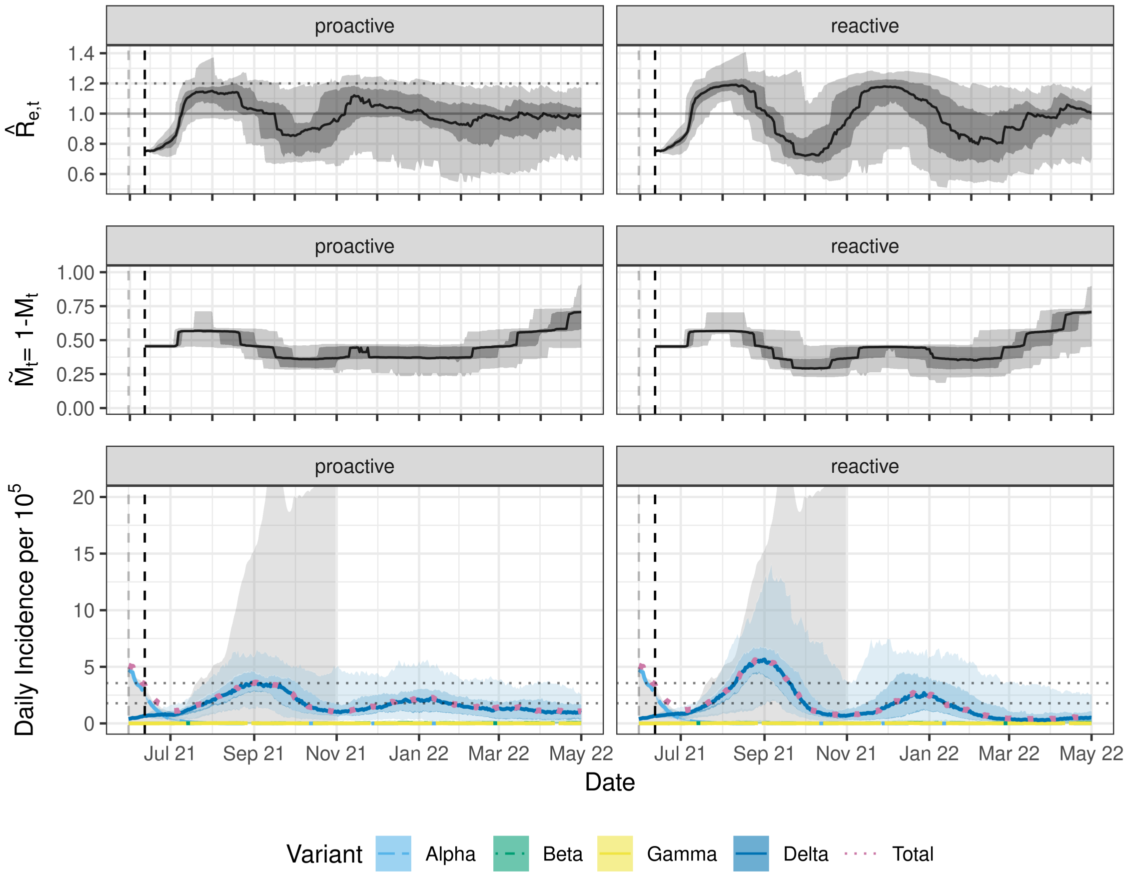


Figure 12: **The proactive control is more efficient than reactive control even when incidence thresholds are stricter**. Stricter thresholds are used such that NPIs are increased when the incidence is 50 per 100,000 inhabitants over a 14 day period. This creates a "yoyo" effect under reactive control that is effectively prevented with proactive control. All other parameters are as in Fig 9.

the winter of 2021 and spring 2022. Omega is introduced as a weekly import, and for simplicity, one case is imported per day. The distribution of imported infections has no effect on the results, regardless if infections are imported daily or staggered throughout the week. All other parameter and control settings are the same as in the previous subsection on Delta.

While this paper was under revision, variant Omicron has emerged and spread. As it is conceptually similar to our Omega variant, we updated the resistance parameters for Omega to mirror tentative estimates of these parameters for Omicron. Yet the lack of concrete values for other parameters such as mean generation interval and waning of vaccine effectiveness against infection prevent us from making concrete claims about Omicron. As such, we have maintained the name Omega in order to emphasise the inability to properly mimic Omicron at this point in time.

Fig 13 shows how the controllers manage the new Omega variant using the stricter, high-positivity thresholds (25 and 50 cases/100,000 over 14 days). The first few months of each image look qualitatively the same as those in Fig 12: the increased mitigation in the fall delays Omega from being established. The second simulated wave, however, is driven by Omega due to its immune escape. Given the population vaccination levels and compartment structure in Austria, Omega out-competes Delta when the proportion of vaccinated individuals exceeds approximately 30% (Fig 6), which happened in Austria in mid-April, 2021.

The largest difference between scenarios with and without Omega is the width of the infection prediction intervals for the reactive control. Not only does the reactive control allow larger outbreaks, but it is unable to guarantee that all simulation paths are controlled. In approximately 2.5% of simulations, infections peaked at 40 cases/100,000. The proactive control was able to provide more efficient and reliable control on all simulation instances, effectively preventing an outbreak. This can also be seen by its ability to keep the effective reproduction number stable and near 1. The difference in median total infections per 100,000 is 297.1 (95% CI: 236, 361) and the difference in median peak quarantined cases is 20.9 (95% CI: 13, 29). The difference in the 97.5% quantiles of predicted infections is over 2,000 cases/100,000, while for predicted peak cases in quarantine the difference is over 400 cases/100,000 (see Fig 2).

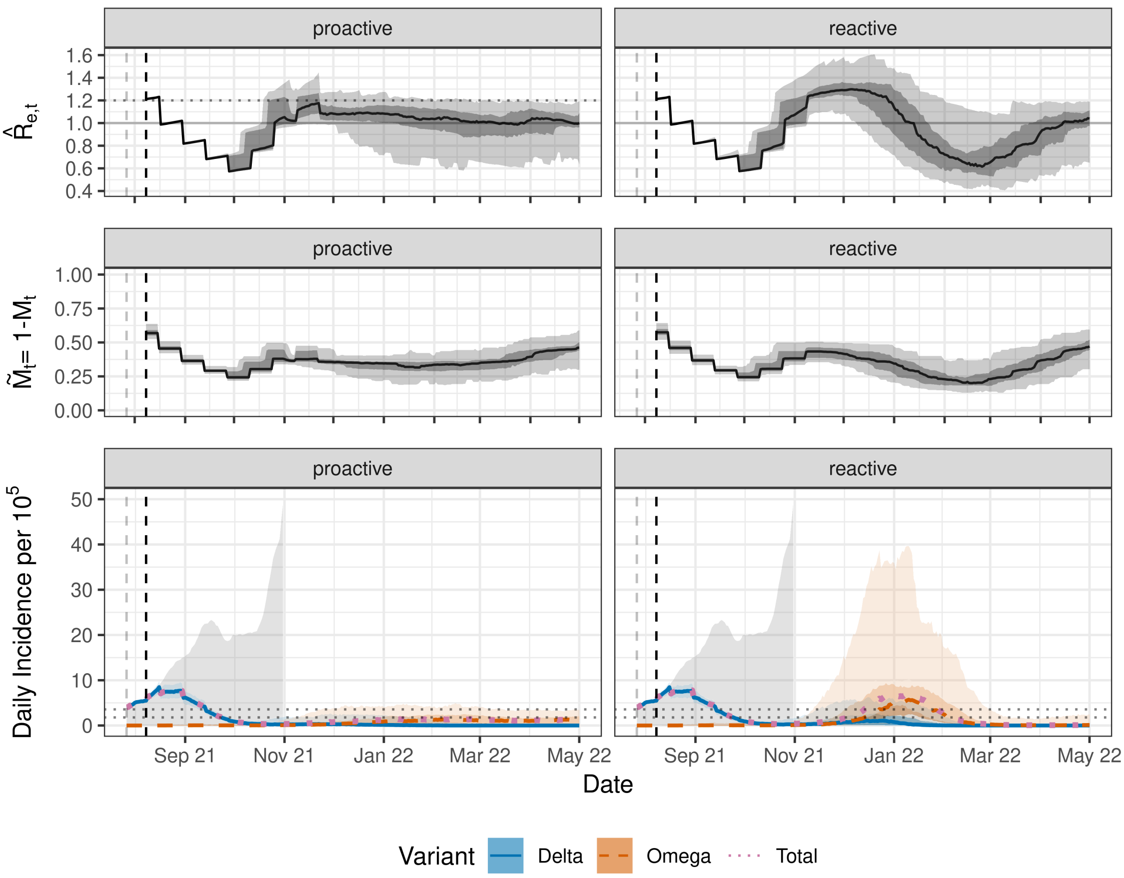


Figure 13: **The reactive control fails to contain new variants that are competitive in highly-vaccinated populations.** The figure shows a scenario with a hypothetical variant Omega that has a smaller basic reproduction number than Delta but has a greater ability to escape immunity post vaccination or infection by other variants (*cf.* Fig 6). Proactive control prevents an outbreak and uses fewer NPIs overall. Other parameters are the same as in Fig 12; note the longer gaps between peaks, which are the result of larger preceding peaks and extended time under more NPIs.

As vaccinations increase, some governments may react to hospitalised (or ICU-hospitalised) incidence instead of case numbers. The rationale is that the controller decides based on hospital capacity, rather than managing the cases. This is particularly enticing as the vaccines reduce severe illnesses or hospitalisations even more than mere infections. Yet this results in a larger delay between infections and actionable information, as there is a larger delay between infection and hospitalisation. While we do not model hospitalisations as this would require adding an age structure to all of our compartments, we can increase the delay that the controller uses for observing cases. This isolates the effect of the delayed information. If hospitalisations are a constant multiple of infections, then our results translate directly to that domain as well.

Fig 14 shows simulation results with the hypothetical Omega variant, again using the strict, high-positivity thresholds, but with a delay of 21 days (instead of 7 days). In this case, the controller is using the same decision rules to increase or decrease NPIs, but the case data informing this decision is 21 days old. This corresponds to the approximate 2-3 week delay between infection and hospitalisation at ICU [57]. We see that the initial outbreaks are significantly more pronounced (*cf.* Fig 13): both controllers begin the simulation by merely increasing mitigation as rapidly as possible. The subsequent outbreak, however, is entirely prevented by the proactive controller, even with delayed information. In fact, the proactive controller using a 21 day delay prevents outbreaks significantly better than a reactive controller with only 7 day delay (*cf.* Fig 2 and Fig 3) .

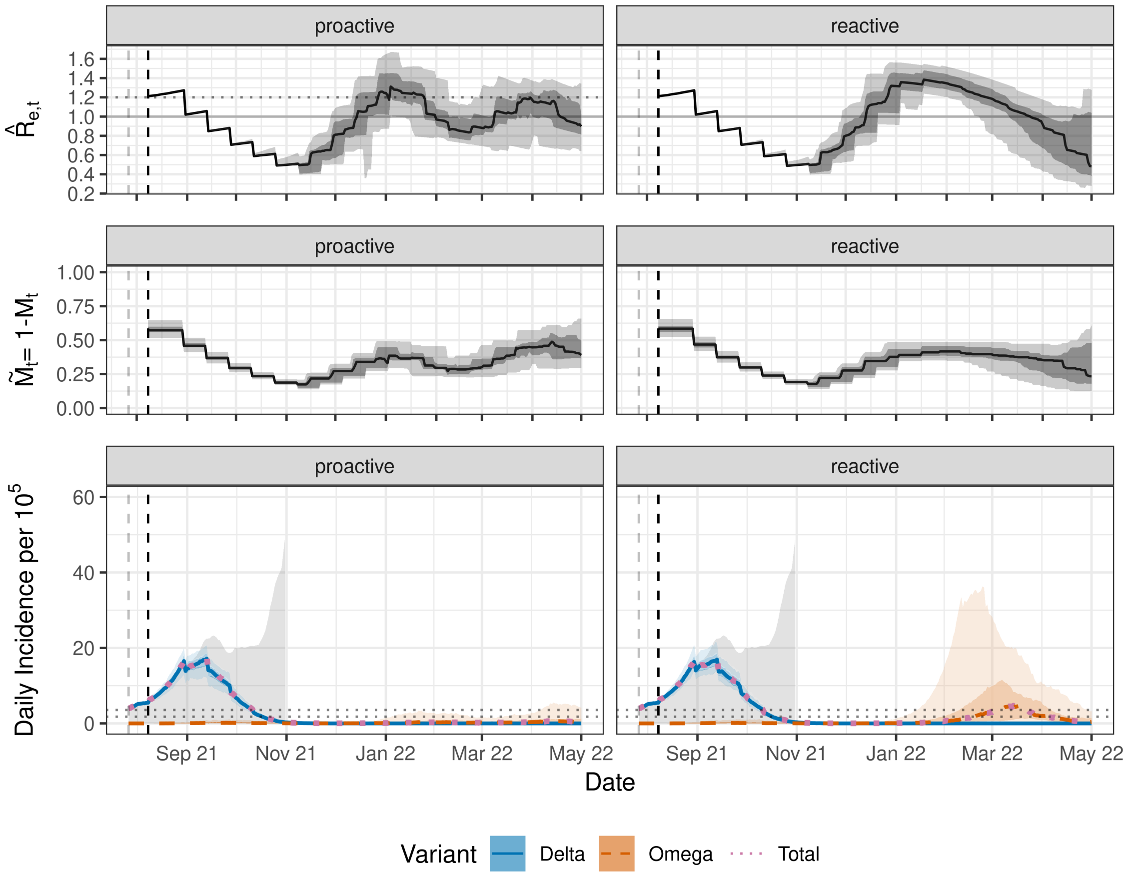


Figure 14: **Controller with a larger delay leads to significantly more pronounced outbreaks, especially with reactive control**. The figure shows a controller with a 3-week delay. A similar delay would be expected when decisions are based on hospitalisations due to COVID-19 rather than on cases. We (conservatively) assume the case thresholds would stay the same. Compare with Fig 13, where the controllers respond with a 1-week delay.

One additional effect of delayed information is that restrictions are not lifted in a timely manner. Both controllers maintain the restrictions longer than necessary. This gives rise to the more diffuse mitigation density used by the proactive control as seen in Fig 15. For the reactive controller, this has the effect of delaying the Omega outbreak into the spring. In reality, we expect a government to relax sooner, but at the cost of future waves also arriving sooner.

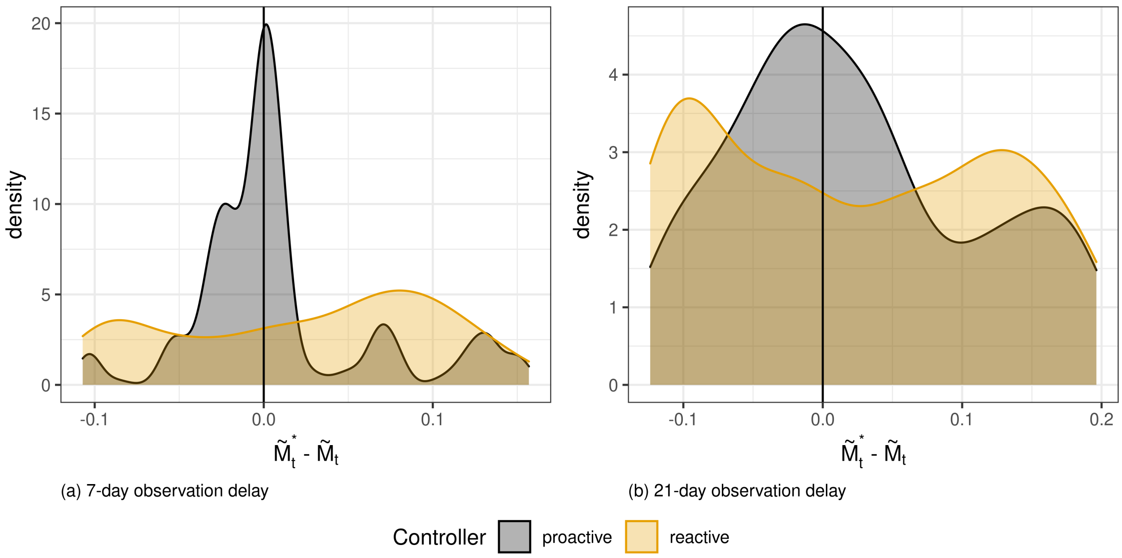


Figure 15: **Proactive control performs nearly as well under greater observation delay.** In both settings, proactive control uses mitigation much closer to the balanced value, while the reactive controller compensates for overly lax NPIs by using overly strict NPIs. The degradation in information by using a larger delay is clearly seen in the additional dispersion of the proactive mitigation density. The modal value, however, still achieves balance between strict NPIs and preventing outbreaks.

Comparing the difference in peak quarantined cases carries additional meaning in this setting as it is highly correlated with peak hospitalisations. While the medians are again similar, the maximum peak quarantined cases over simulations are nearly twice as high for the reactive controller as for the proactive controller (Fig 3). Proactive control suffers much less from looking at more delayed data than reactive control. As such, this indicates that looking at rates of change in hospital can be a good indicator, while the delay from reacting to actual hospital utilisation is costly.

3.3 Waning and boosting

In the simulation results presented above, populations always increase their immunity to new infections over time, either through vaccination or through infections. In reality, the induced immunity also wanes over time. We focus on simulations with the Omega variant, as it is assumed to be most susceptible to waning. With waning immunity and boosting, the variant risk as measured by $\rho^{\mathcal{V}}$ changes non-monotonically over time (see Fig 16). While boosting rapidly increased in October and November, its strongest effect on $\rho^{\mathcal{V}}$ appears delayed. This is because all vaccinated individuals are eligible for the booster, not just those with waned immunity. The sheer size of the vaccinated groups means many doses need to be administered before the waned groups dramatically decrease in size. The effects of this changing risk profile can also be seen in the infection outbreaks that occur in the simulations and the subsequent controller behavior.

Fig 17 shows the simulations with waning but without boosting. The corresponding graph with boosting is in the Supporting Information (Fig 4), as adding boosting is sufficient to suppress the Omega outbreak in the winter. For simplicity, we focus only on the strict case thresholds of 25 and 50/100,000/14 days. In the summer, the relative advantage of Omega is too low to see

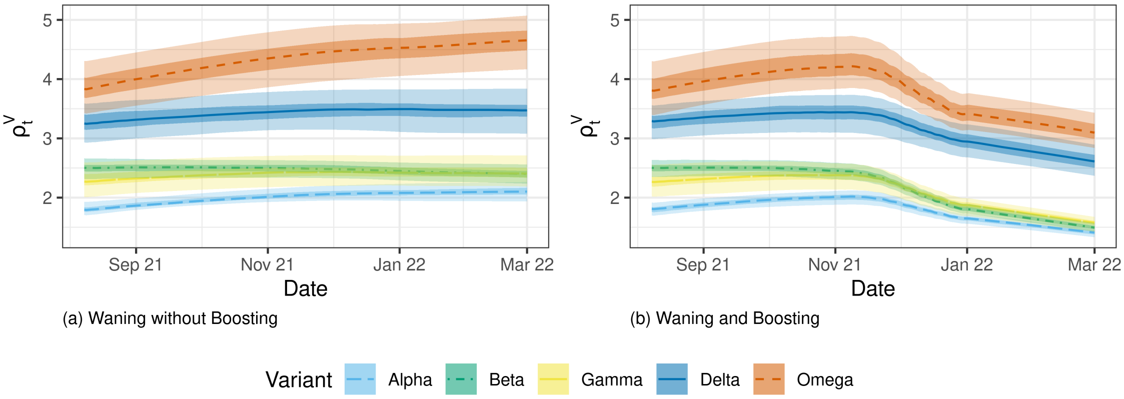


Figure 16: **Waning and boosting change the population-specific reproduction number $\rho_t^\mathcal{V}$.** Without waning, $\rho_t^\mathcal{V}$ steadily decreases throughout the simulation due to infections and vaccination. (a) Waning of immunity to Omega is strong enough to reverse this effect. (b) Booster shots compensate for immunity waning, and one can clearly see the effect of the wide-scale boosting that rapidly increased in October and November.

an outbreak in the presence of relatively strong mitigation. Waning immunity then leads to a larger outbreak of the Omega variant than observed without waning in Fig 13, though only under reactive control. The winter outbreak is even suppressed when the proactive controller uses looser, 25-150/100,000/14 days case thresholds (Fig 5).

In terms of comparison of the controllers, the results of the previous section (without waning and boosting) still hold. There are significant differences in the medians for both total infections and peak quarantined cases between reactive and proactive control. The differences in the upper quantiles of these predicted distributions is even more extreme. We can see in Fig 18 that for both low and high thresholds, the reactive controller tends to use extremely high or low mitigation levels, consistently failing to find a balance between suppressing outbreaks and NPI use. In fact, switching to stricter case thresholds exacerbates this problem.

Lastly, we briefly explored simulation dynamics when Omega is modified to spread even more rapidly. In addition to the high immune escape of our baseline Omega variant, we increased its base reproduction number to 2–matching Delta–and reduced the mean of the generation interval from 4.6 days to 3 days (CI: 1.5, 4.5). Such a variant has a significant potential to cause extremely large outbreaks if not controlled efficiently. We confirmed that while the proactive controller suppresses the potential outbreak, the reactive controller does not, even with strict case control thresholds (see Fig 6).

3.4 Discussion

The model and simulation that we develop provides significant insight into efficient control strategies of COVID-19 outbreaks. The key behavior which we wanted to capture in our simulation was a complex interaction between diverse groups in a compartment model. Our simulation creates compartments for various vaccines, multiple SARS-CoV-2 variants, and specifies well-supported parameters for them all. The interaction rules between compartments are transparent. This allows us to simulate complex scenarios in a realistic and dynamic setting: COVID-19 epidemic spread in Austria. Having such a realistic model is an integral component to controlling outbreaks, as having a good model a system is a prerequisite for regulating it efficiently [15].

It is well known that timely restrictions are more efficient in controlling an epidemic [30, 31]; yet timely decisions can only be taken with a good controller. One can consider addressing this

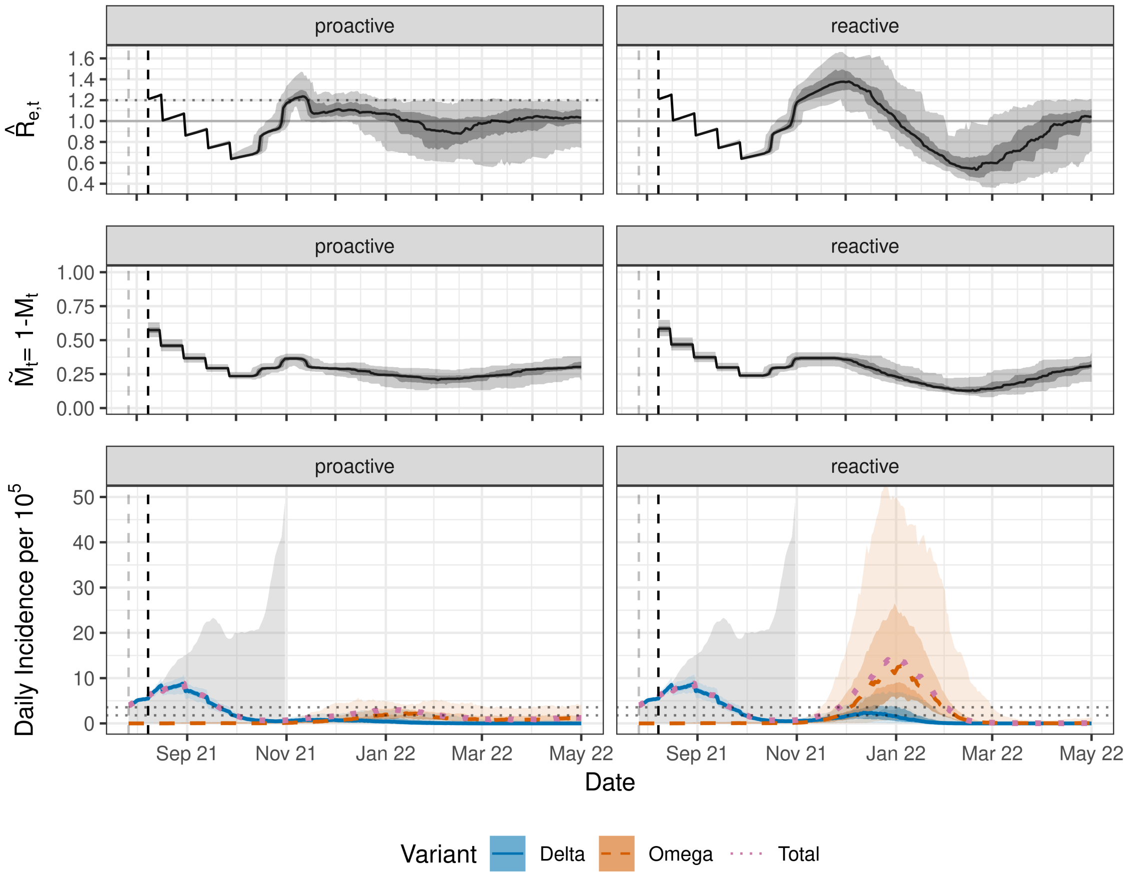


Figure 17: **Waning leads to a larger outbreak of the Omega variant under reactive but not proactive control**. Parameters as in Fig 13, with waning immunity but no boosting.

in two distinct ways: 1, use stricter thresholds of a given statistic to motivate change; or 2, use a better statistic, preferably informed by a model of the epidemic [15]. Our study demonstrates that the second type of solution, using the effective reproduction number $\mathcal{R}_{e,t}$ to guide intervention decisions in addition to case numbers, is more efficient and successful at curbing the epidemic. We provide a quantitative comparison of a continuous controller of two types - one which only reacts to cases (reactive), and one which also reacts to the effective reproduction number (proactive). We show that the proactive controller is more efficient and effective at controlling infection outbreaks than the reactive controller, even when it uses data that has a larger delay (such as using $\mathcal{R}_{e,t}$ estimated from hospitalisations). In contrast, the NPIs imposed by the reactive controller are further away from the more efficient “balanced” minimum intervention policy which keeps $\mathcal{R}_{e,t}$ close to 1. By oscillating between over- and under-regulating, the reactive controller fails to provide reliable control of new outbreaks: some simulation paths exhibit large spikes in infections and peak quarantined cases, even when the median values are controlled.

A potential limitation of our simulation is our lack of observation level model which would allow us to simulate a fluctuating detection ratio. While some view positivity-rate to be informative about the detection ratio, we - in the absence of a well-informed model for testing strategies and its saturation - instead consider a range of case-based thresholds which span the gamut of positivity rate scenarios. In all of these settings, we find consistent support for using a proactive control strategy that responds to changes in the effective reproduction number.

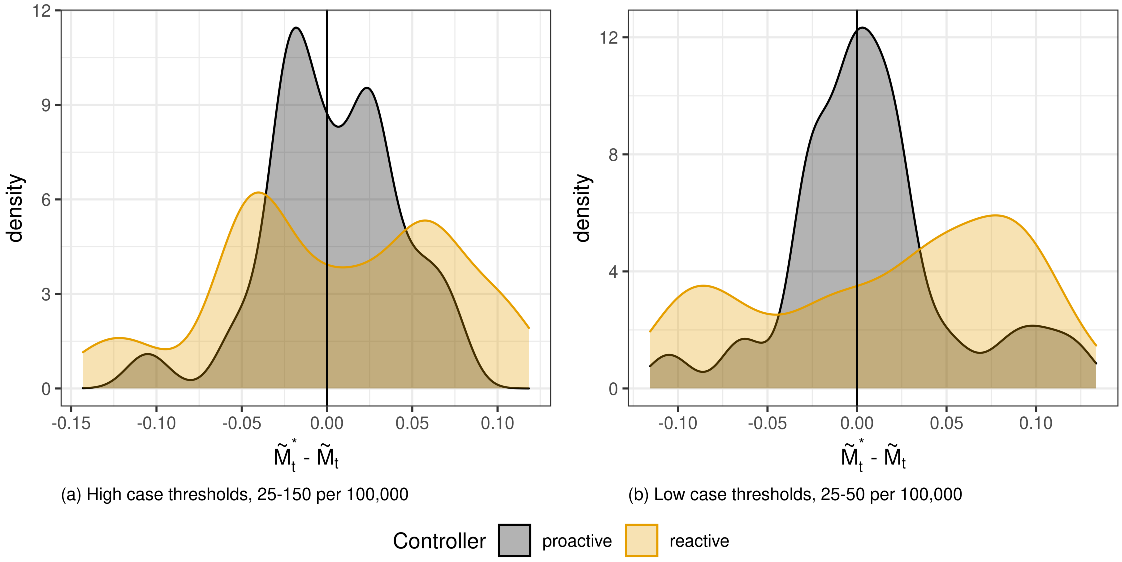


Figure 18: **With waning, which accelerates and strengthens the Omega outbreak, reactive control fails to balance NPI use.** For both high (a) and low (b) control case thresholds, proactive controller is more efficient. The rising $\rho^{\mathcal{V}}$ values in simulations with waning apply the greatest pressure toward an epidemic outbreak, which the proactive controller contains efficiently, without resorting to extreme NPIs.

Our model does not explicitly incorporate any network structure for individuals' interactions; each individual in the simulation interacts independently with all other members. In reality, infections take place in the household, work, and social environments. The different cross-contamination rates in these environments lead to clusters of observed infections, not only in terms of infections occurring, but also in terms of identifying them via contact tracing. This is partially alleviated by our model accounting for super-spreading. Large super-spreading events are often caused by high infectiousness coupled with a particular network structure. By incorporating super-spreading natively in our model, we are able to produce cluster-like effects due to people that are significantly more contagious than others. Furthermore, other changes in network structure are captured by seasonality or mitigation; seasonality can be caused by increased indoor interaction during winter months, and social distancing rules are common NPIs.

Our model contains many parameters governing diverse characteristics such as vaccine effectiveness, resistance to reinfection (including cross-infection by other variants), and basic reproduction numbers. A natural question is which of these has a stronger effect on the simulation. The difficulty in answering lies in the distribution of the population across the compartments we describe. Populations with lower vaccination rates but higher rates of previous infection will naturally be more sensitive to cross-infection rates and vice versa. That is why we defined the $\rho^{\mathcal{V}}$ parameter, which characterises a decisive component of the effective reproduction number. It depends on the basic reproduction number as well as the reduction in transmissibility arising from the (partial) immunity acquired by previous infections and vaccinations within a particular population. We believe $\rho^{\mathcal{V}}$ is an important summary parameter of great relevance.

The paper is not intended to forecast what the future of SARS-CoV-2 will bring: potentially vaccine resistant variants, or variants with even higher base reproduction number, etc. The next important VOC may well have different characteristics to the hypothesised Omega. It is important to keep in mind though, that same VOCs which appear to be currently out-competed by (say) Delta may have a competitive advantage later. The possibility of change in the relative advantage between variants is especially relevant when the emerging variant leads to more severe symptoms. The main claims of this paper, however, hold true for all of these possibilities. Regardless of the process leading to future waves, one certainty is that they will occur. In this eventuality, governments must design methods to identify and react to changes in the pandemic. Our results focus on this common denominator.

Acknowledgements

We would like to thank Astrid Christine Erber, Joachim Hermisson, Michal Hledík , Pavel Payne and Claudia Zimmermann for insightful discussions and comments on earlier drafts.

Appendix

Table 1: Model parameters for different variants. The first row gives the assumed increase of $\mathcal{R}_0$ relative to wild-type (WT) based on [11]. The rest give the immunity against new infection following vaccination or previous infection. The upper row per label gives the median estimate; the lower row gives a confidence interval. The references are given in the superscript of the median value. See the text for more details.

Label	WT	Alpha	Beta	Gamma	Delta	Omega	WT,A,B,G,D	Omega
$\lambda^{\mathcal{V}}$	1	1.3	1.25	1.40	2	1.55	6 months	6 months
	-	(1.24,1.33)	(1.2,1.3)	(1.22,1.48)	(1.76,2.17)	(1.35,1.75)	waned	waned
mRNA	0.94[38, 43]	0.94[38, 43]	0.75[1]	0.85[38]	0.84[38, 43]	0.6[53]	0.5[14, 53]	0.05[53]
	(0.85,0.96)	(0.85,0.96)	(0.7,0.8)	(0.7 ,0.93)	(0.7,0.86)	(0.5,0.65)	(0.4,0.6)	(0,0.1)
vector	0.86 [43]	0.86[43]	0.1[35]	0.65[25]	0.7[43]	0.3[53]	0.4[14, 53]	0[53]
	(0.65,0.93)	(0.65,0.93)	(0,0.55)	(0.6,0.8)	(0.65,0.73)	(0,0.55)	(0.35,0.45)	(0,0.05)
booster	0.96[53]	0.96[53]	0.96[53]	0.96[53]	0.96[53]	0.7[53]	0.85[53]	0.3[53]
	0.94,0.98	0.94,0.98	0.94,0.98	0.94,0.98	0.94,0.98	0.6,0.8	0.75,0.9	0.2,0.4
WT	0.87[43]	0.87[43]	0.7[20]	0.7[6, 41]	0.77[43]	0.3	0.4	0
	(0.84,0.9)	(0.84,0.9)	(0.55,0.8)	(0.55,0.8)	(0.66,0.85)	(0,0.55)	(0.35,0.45)	(0,0.05)
Alpha	0.87[43]	0.87[43]	0.7[20]	0.7[6, 41]	0.77[43]	0.3	0.4	0
	(0.84,0.9)	(0.84,0.9)	(0.55,0.8)	(0.55,0.8)	(0.66,0.85)	(0,0.55)	(0.35,0.45)	(0,0.05)
Beta	0.75	0.75	0.85	0.7[6, 41]	0.75	0.3	0.4	0
	(0.65,0.85)	(0.65,0.85)	(0.8,0.9)	(0.55,0.8)	(0.65,0.85)	(0,0.55)	(0.35,0.45)	(0,0.05)
Gamma	0.75	0.75	0.7[20]	0.85	0.75	0.3	0.4	0
	(0.65,0.85)	(0.65,0.85)	(0.55,0.8)	(0.8 ,0.9)	(0.65,0.85)	(0,0.55)	(0.35,0.45)	(0,0.05)
Delta	0.75	0.75	0.7[20]	0.7[6, 41]	0.85	0.3	0.5	0
	(0.65,0.85)	(0.65,0.85)	(0.55,0.8)	(0.55,0.8)	(0.8,0.9)	(0,0.55)	(0.35,0.6)	(0,0.05)
Omega	0.75	0.75	0.7	0.7	0.75	0.85	0	0.5
	(0.65,0.85)	(0.65,0.85)	(0.55,0.8)	(0.55,0.8)	(0.65,0.85)	(0.8,0.9)	(0,0.05)	(0.35,0.6)

References

[1] L. J. Abu-Raddad, H. Chemaitelly, and A. A. Butt. Effectiveness of the BNT162b2 Covid-19 vaccine against the B.1.1.7 and B.1.351 variants. *New England Journal of Medicine* 385 (2 2021).

[2] Agentur für Gesundheit und Ernährungssicherheit GmbH. AGES Dashboard COVID19. `https://covid19-dashboard.ages.at/`. 2020.

[3] Agentur für Gesundheit und Ernährungssicherheit GmbH. SARS-CoV-2-Varianten in Österreich. `https://www.ages.at/themen/krankheitserreger/coronavirus/sars-cov-2-varianten-in-oesterreich/`. Accessed: 2021-09-13. 2021.

[4] H.-S. Ahn, J. Silberholz, X. Song, and X. Wu. Optimal COVID-19 containment strategies: Evidence across multiple mathematical models. *Available at SSRN* (2021).

[5] L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine* 384.5 (2021), pp. 403–416.

[6] T. A. Bates, H. C. Leier, Z. L. Lyski, S. K. McBride, F. J. Coulter, J. B. Weinstein, J. R. Goodman, Z. Lu, S. A. Siegel, P. Sullivan, et al. Neutralization of SARS-CoV-2 variants by convalescent and BNT162b2 vaccinated serum. *Nature Communications* 12.1 (2021), pp. 1–7.

[7] M. R. Bicher, C. Rippinger, G. R. Schneckenreither, N. Weibrecht, C. Urach, M. Zechmeister, D. Brunmeir, W. Huf, and N. Popper. Model Based Estimation of the SARS-CoV-2 Immunization Level in Austria and Consequences for Herd Immunity Effects. *medRxiv* (2021).

[8] F. Blanquart, C. Abad, J. Ambroise, M. Bernard, G. Cosentino, J.-M. Giannoli, and F. Débarre. Spread of the Delta variant, vaccine effectiveness against PCR-detected infections and within-host viral load dynamics in the community in France. Tech. rep. 2021.

[9] J. M. Brauner, S. Mindermann, M. Sharma, D. Johnston, J. Salvatier, T. Gavenčiak, A. B. Stephenson, G. Leech, G. Altman, V. Mikulik, et al. Inferring the effectiveness of government interventions against COVID-19. *Science* 371.6531 (2021).

[10] Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz. COVID-19 vaccination data for Austria. Tech. rep. 2021.

[11] F. Campbell, B. Archer, H. Laurenson-Schafer, Y. Jinnai, F. Konings, N. Batra, B. Pavlin, K. Vandemaele, M. D. Van Kerkhove, T. Jombart, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance* 26.24 (2021), p. 2100509.

[12] CDC. Delta Variant:Infections and Spread. Tech. rep. 2021.

[13] P. Y. Chia, S. W. X. Ong, C. J. Chiew, L. W. Ang, J.-M. Chavatte, T.-M. Mak, L. Cui, S. Kalimuddin, W. N. Chia, C. W. Tan, L. Y. A. Chai, S. Y. Tan, S. Zheng, R. T. P. Lin, L. Wang, Y.-S. Leo, V. J. L. Lee, D. C. Lye, and B. E. Young. Virological and serological kinetics of SARS-CoV-2 Delta variant vaccine-breakthrough infections: a multi-center cohort study. *MedRXiV* (2021).

[14] B. A. Cohn, P. M. Cirillo, C. C. Murphy, N. Y. Krigbaum, and A. W. Wallace. SARS-CoV-2 vaccine protection and deaths among US veterans during 2021. *Science* (2021), eabm0620.

[15] R. C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science* 1.2 (1970), pp. 89–97.

[16] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology* 178.9 (2013), pp. 1505–1512.

[17] Council of the European Union. Draft Council Recommendation on a coordinated approach to the restriction of free movement in response to the COVID-19 pandemic. `https://data.consilium.europa.eu/doc/document/ST-11689-2020-REV-1/en/pdf`. Accessed: 2021-09-14. 2020.

[18] M. Eder, J. Hermisson, M. Hledik, C. Hütter, E. Iofinova, R. Pisupati, J. Polechova, G. Puixeu, S. Sarikas, B. Wölfl, and C. Zimmermann. A COVID-19 compartment model for Austria. en. Tech. rep. 2020.

[19] A. Endo, S. Abbott, A. Kucharski, and S. Funk. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research* 5 (July 2020), p. 67.

[20] N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. Crispim, F. C. Sales, I. Hawryluk, J. T. McCrone, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* 372.6544 (2021), pp. 815–821.

[21] T. Ganyani, C. Kremer, D. Chen, A. Torneri, C. Faes, J. Wallinga, and N. Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* 25.17 (2020), p. 2000257.

[22] T. Gavenčiak, J. T. Monrad, G. Leech, M. Sharma, S. Mindermann, J. M. Brauner, S. Bhatt, and J. Kulveit. Seasonal variation in SARS-CoV-2 transmission in temperate climates. *medRxiv* (2021).

[23] W. S. Hart, E. Miller, N. J. Andrews, P. Waight, P. K. Maini, S. Funk, and R. N. Thompson. Generation time of the Alpha and Delta SARS-CoV-2 variants. *medRxiv* (2021).

[24] W. Hart, S. Abbott, A. Endo, J. Hellewell, E. Miller, N. Andrews, P. Maini, S. Funk, and R. Thompson. Inference of SARS-CoV-2 generation times using UK household data. *medRxiv* (2021).

[25] M. Hitchings, O. T. Ranzani, M. Dorion, T. L. D'Agostini, R. C. de Paula, O. F. P. de Paula, E. F. de Moura Villela, M. S. S. Torres, S. B. de Oliveira, W. Schulz, et al. Effectiveness of the ChAdOx1 vaccine in the elderly during SARS-CoV-2 Gamma variant transmission in Brazil. *medRxiv* (2021).

[26] K. D. Johnson, M. Beiglböck, M. Eder, A. Grass, J. Hermisson, G. Pammer, J. Polechová, D. Toneian, and B. Wölfl. Disease momentum: Estimating the reproduction number in the presence of superspreading. *Infectious Disease Modelling* 6 (2021), pp. 706–728.

[27] D. Kim, J. Jo, J.-S. Lim, and S. Ryu. Serial interval and basic reproduction number of SARS-CoV-2 Omicron variant in South Korea. *medRxiv* (2021).

[28] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 368.6493 (2020), pp. 860–868.

[29] J. Knight and S. Mishra. Estimating effective reproduction number using generation time versus serial interval, with application to covid-19 in the Greater Toronto Area, Canada. *Infectious Disease Modelling* 5 (2020), pp. 889–896.

[30] E. S. Knock, L. K. Whittles, J. A. Lees, P. N. Perez-Guzman, R. Verity, R. G. FitzJohn, K. A. Gaythorpe, N. Imai, W. Hinsley, L. C. Okell, et al. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Science Translational Medicine* (2021).

[31] J. R. Koo, A. R. Cook, M. Park, Y. Sun, H. Sun, J. T. Lim, C. Tam, and B. L. Dickens. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. *The Lancet Infectious Diseases* 20.6 (2020), pp. 678–688.

[32] N. Kronfeld-Schor, T. Stevenson, S. Nickbakhsh, E. Schernhammer, X. Dopico, T. Dayan, M. Martinez, and B. Helm. Drivers of infectious disease seasonality: potential implications for COVID-19. *Journal of Biological Rhythms* 36.1 (2021), pp. 35–54.

[33] T. Kwon, N. N. Gaudreault, and J. A. Richt. Seasonal stability of SARS-CoV-2 in biological fluids. *Pathogens* 10.5 (2021), p. 540.

[34] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. en. *Journal of Travel Medicine* 27.2 (Mar. 2020), taaa021.

[35] S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, A. L. Koen, L. Fairlie, S. D. Padayachee, K. Dheda, S. L. Barnabas, Q. E. Bhorat, et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 vaccine against the B.1.351 variant. *New England Journal of Medicine* 384.20 (2021), pp. 1885–1898.

[36] A. S. Monto, P. DeJonge, A. P. Callear, L. A. Bazzi, S. Capriola, R. E. Malosh, E. T. Martin, and J. G. Petrie. Coronavirus occurrence and transmission over 8 years in the HIVE cohort of households in Michigan. *The Journal of infectious diseases* (2020).

[37] M. Moriyama, W. J. Hugentobler, and A. Iwasaki. Seasonality of respiratory viral infections. *Annual review of virology* 7 (2020).

[38] S. Nasreen, S. He, H. Chung, K. A. Brown, J. B. Gubbay, S. A. Buchan, S. E. Wilson, M. E. Sundaram, D. B. Fell, B. Chen, M. Tadrous, K. Wilson, S. E. Wilson, and J. C. Kwong. Effectiveness of COVID-19 vaccines against variants of concern in Ontario, Canada. *Medrxiv* (2021).

[39] M. Park, A. R. Cook, J. T. Lim, Y. Sun, and B. L. Dickens. A systematic review of COVID-19 epidemiology based on current evidence. *Journal of clinical medicine* 9.4 (2020), p. 967.

[40] A. Pegu, S. O'Connell, S. D. Schmidt, S. O'Dell, C. A. Talana, L. Lai, J. Albert, E. Anderson, H. Bennett, K. S. Corbett, B. Flach, L. Jackson, B. Leav, J. E. Ledgerwood, C. E. Luke, M. Makowski, P. C. Roberts, M. Roederer, P. E. Rebolledo, C. A. Rostad, N. G. Rouphael, W. Shiw, W. Lingshu, A. T. Widge, E. S. Yang, the mRNA-1273 Study Group, J. H. Beigel, J. R. Graham Barney S Mascola, M. S. Suthar, A. McDermott, and N. A. Doria-Rose. Durability of mRNA-1273 vaccine–induced antibodies against SARS-CoV-2 variants. *Science* 373.6561 (2021), pp. 1372–1377.

[41] D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M. M. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* (2021), pp. 1–7.

[42] F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. P. Marc, E. D. Moreira, C. Zerbini, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine* (2020).

[43] K. B. Pouwels, E. Pritchard, P. C. Matthews, N. Stoesser, D. W. Eyre, K.-D. Vihta1, T. House, J. Hay, J. I. Bell, J. N. Newton, J. Farrar, D. Crook, D. Cook, E. Rourke, R. Studley, T. Peto, I. Diamond, S. A. Walker, and the COVID-19 Infection Survey Team. Impact of Delta on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. en. *medRxiv* (2021).

[44] R. Pung, T. M. Mak, A. J. Kucharski, and V. J. Lee. Serial intervals in SARS-CoV-2 B.1.617.2 variant cases. *The Lancet* 398 (10303 2021), pp. 837–838.

[45] L. Richter, D. Schmid, A. Chakeri, S. Maritschnik, S. Pfeiffer, and E. Stadlober. Schätzung des seriellen Intervalles von COVID19, Österreich. Tech. rep. `https://www.ages.at/en/wissen-aktuell/publikationen/schaetzung-des-seriellen-intervalles-von-covid19-oesterreich/`. Apr. 2020.

[46] C. Rippinger, M. Bicher, C. Urach, D. Brunmeir, N. Weibrecht, G. Zauner, G. Sroczynski, B. Jahn, N. Mühlberger, U. Siebert, et al. Evaluation of undetected cases during the COVID-19 epidemic in Austria. *BMC Infectious Diseases* 21.1 (2021), pp. 1–11.

[47] C. M. Saad-Roy, S. E. Morris, C. J. E. Metcalf, M. J. Mina, R. E. Baker, J. Farrar, E. C. Holmes, O. G. Pybus, A. L. Graham, S. A. Levin, et al. Epidemiological and evolutionary considerations of SARS-CoV-2 vaccine dosing regimes. *Science* 372.6540 (2021), pp. 363–370.

[48] M. C. Shamier, A. Tostmann, S. Bogers, J. De Wilde, J. IJpelaar, W. A. van der Kleij, H. De Jager, B. L. Haagmans, R. Molenkamp, B. B. O. Munnink, C. van Rossum, J. Rahamat-Langendoen, N. van der Geest, C. P. Bleeker-Rovers, H. Wertheim, M. P. G. Koopmans, and C. H. GeurtsvanKessel. Virological characteristics of SARS-CoV-2 vaccine breakthrough infections in health care workers. *medRxiv* (2021).

[49] M. Sharma, S. Mindermann, C. Rogers-Smith, G. Leech, B. Snodin, J. Ahuja, J. B. Sandbrink, J. T. Monrad, G. Altman, G. Dhaliwal, et al. Understanding the effectiveness of government interventions in Europe's second wave of COVID-19. *medRxiv* (2021).

[50] K. Shea, M. C. Runge, D. Pannell, W. J. M. Probert, S.-L. Li, M. Tildesley, and M. Ferrari. Harnessing multiple models for outbreak management. *Science* 368.6491 (2020), pp. 577–579.

[51] Statistik Austria. Press-conference 12.396-236/20. Tech. rep. 2020.

[52] P. Stefanelli, F. Trentini, G. Guzzetta, V. Marziano, A. Mammone, P. Poletti, C. M. Grane, M. Manica, M. del Manso, X. Andrianou, et al. Co-circulation of SARS-CoV-2 variants B.1.1.7 and P.1. *medRxiv* (2021).

[53] UK Health Security Agency. SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 33. Tech. rep. 2021.

[54] N. Van Doremalen, T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, S. I. Gerber, et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England journal of medicine* 382.16 (2020), pp. 1564–1567.

[55] M. Voysey, S. A. C. Clemens, S. A. Madhi, L. Y. Weckx, P. M. Folegatti, P. K. Aley, B. Angus, V. L. Baillie, S. L. Barnabas, Q. E. Bhorat, et al. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet* 397.10269 (2021), pp. 99–111.

[56] World Health Organization. Critical preparedness, readiness and response actions for COVID-19: interim guidance, 27 May 2021. Tech. rep. World Health Organization, 2021.

[57] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Gua, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen, and B. Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* 395.10229 (2020), pp. 1054–1062.

Robust Models of SARS-CoV-2 Heterogeneity and Control: Supporting Information

K. D. Johnson, A. M. Grass, D. Toneian, M. Beiglböck and J. Polechová

1 Supplemental figures

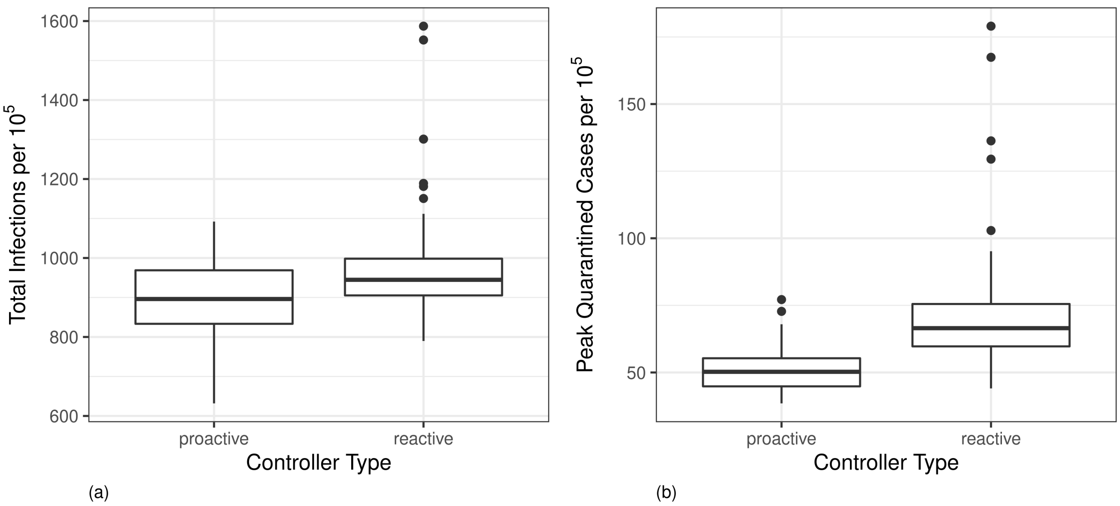
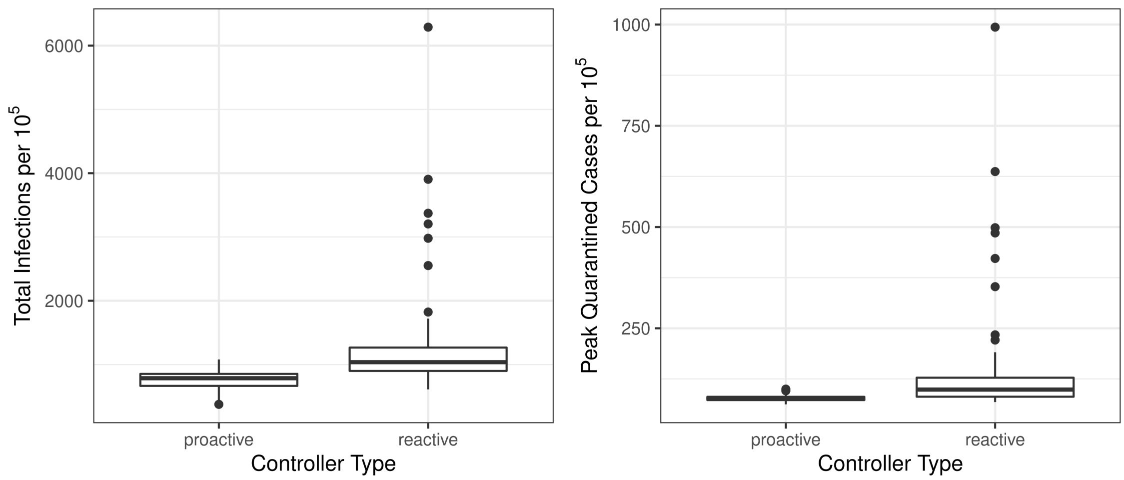


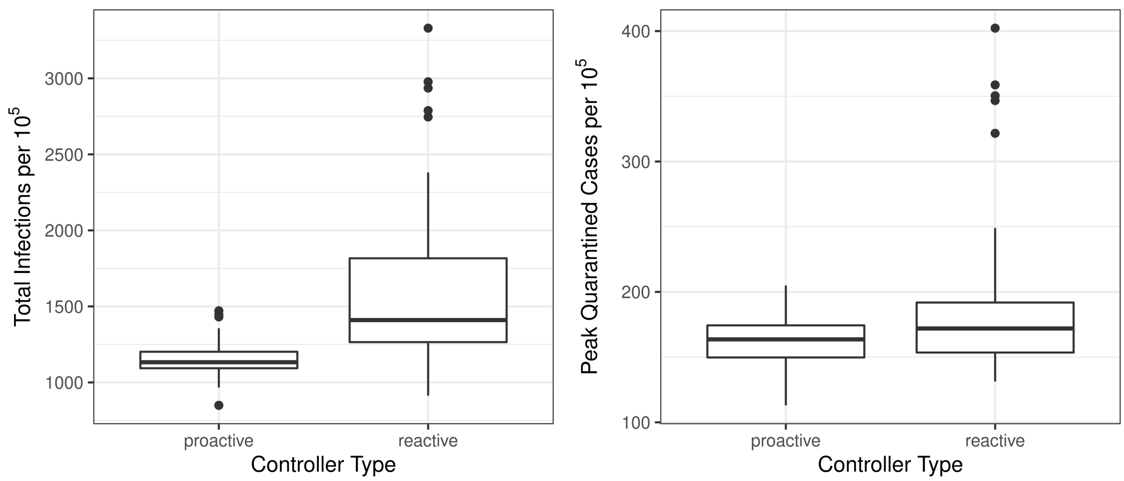
Figure 1: **Proactive control reduces total infections as well as peak numbers in quarantine.** Simulation with Delta variant and strict case thresholds at 25-50/100,000/14 days.

Below, we modified the characteristics of Omega to have a higher base reproduction number as well as a shorter generation time in order to explore the dynamics of more contagious variants. This setting includes both waning and boosting. The reproduction number matches that of Delta, with 2 and interval (1.76, 2.17). The mean of the generation interval has been reduced from original value of 4.6 to 3, with confidence interval (1.5, 4.5). Fig 6 shows that winter outbreaks are observed with both a reactive and proactive controller. While the median values look quite similar, there is a significant difference in their prediction intervals.



(a) (b)

Figure 2: **Proactive control provides better control of statistics in all simulation runs.** Simulation with Omega variant and strict case thresholds at 25-50/100,000/14 days.



(a) (b)

Figure 3: **Proactive control does not suffer due to delayed data although reactive does.** Simulation with delayed controller, Omega variant, and strict case thresholds at 25-50/100,000/14 days.

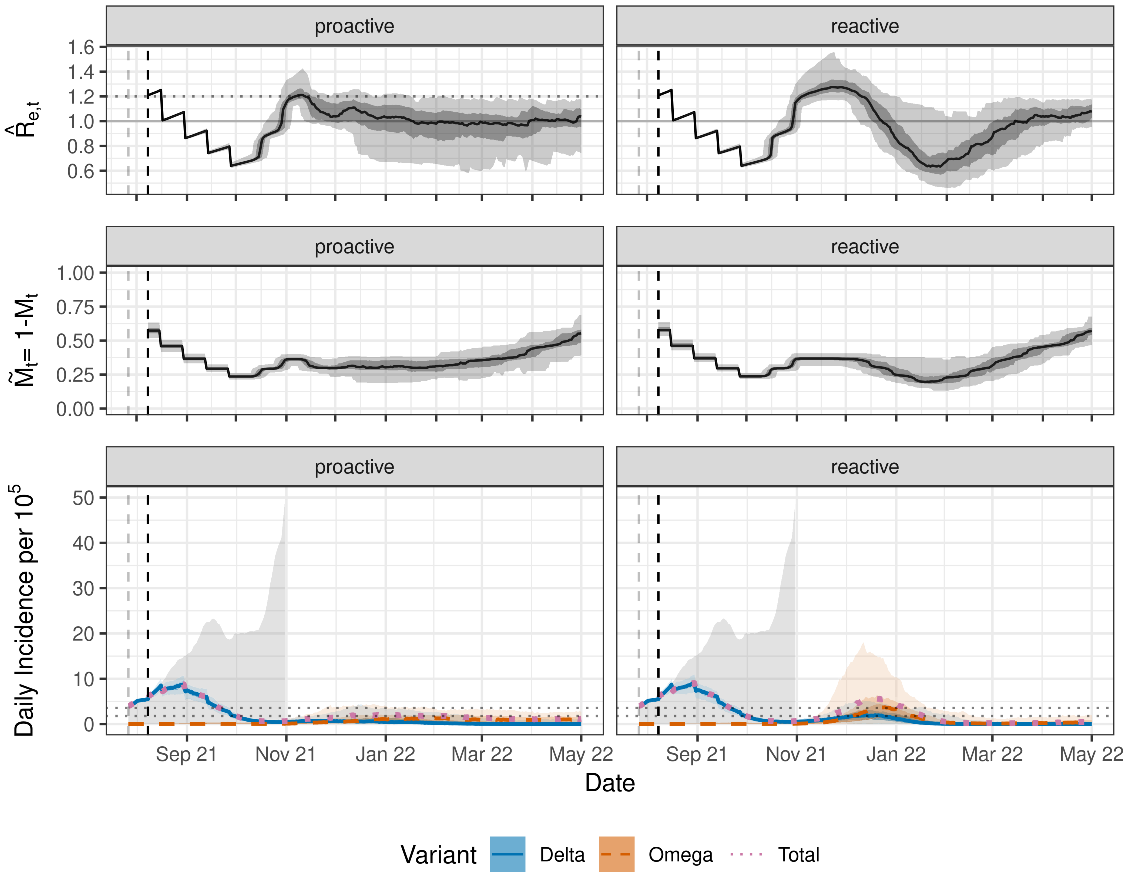


Figure 4: **Boosters primarily prevent a winter outbreak due to waning** With vaccine boosting, ρ_t^O declines fast enough so that large winter outbreaks are prevented – both under reactive and proactive control. Case thresholds are 25-50/100,000/14 days.

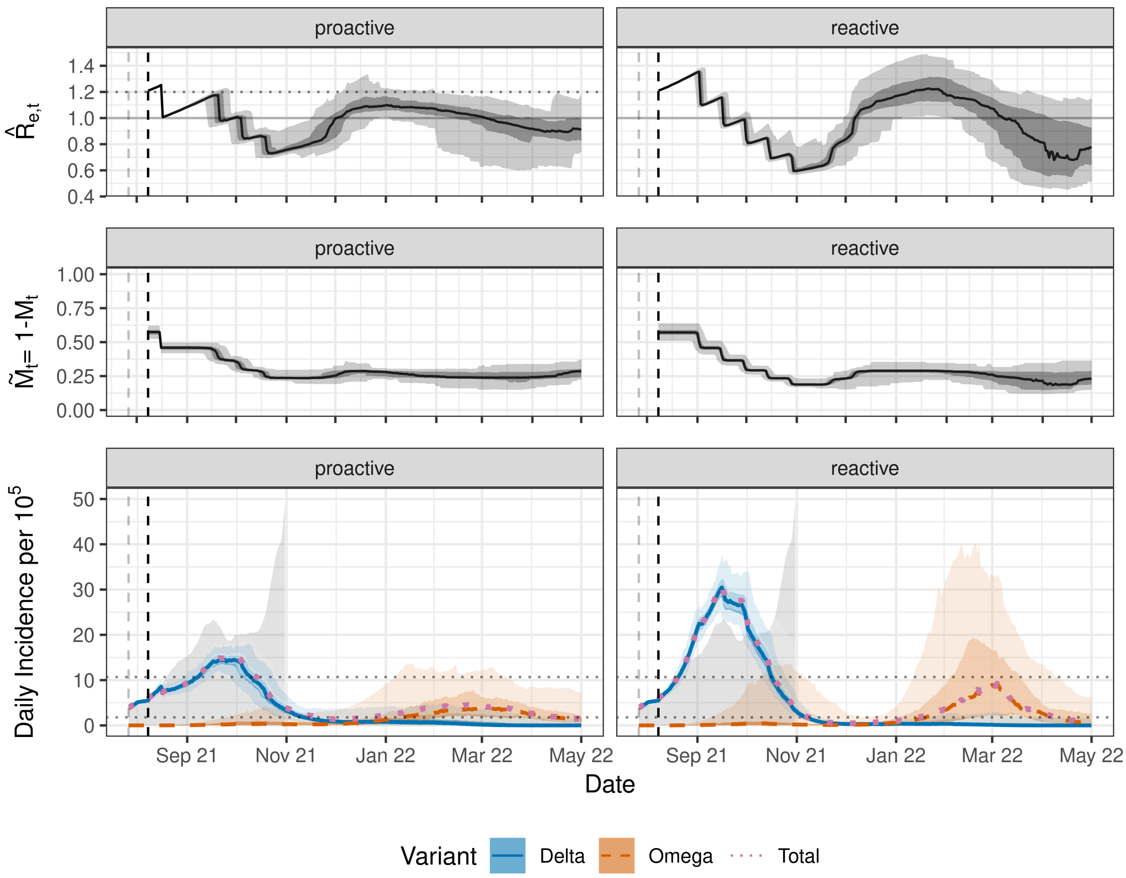


Figure 5: **Proactive control effectively prevents an Omega outbreak even with waning and high case thresholds.** Case thresholds are 25-150/100,000/14 days, other parameters as is Fig 17.

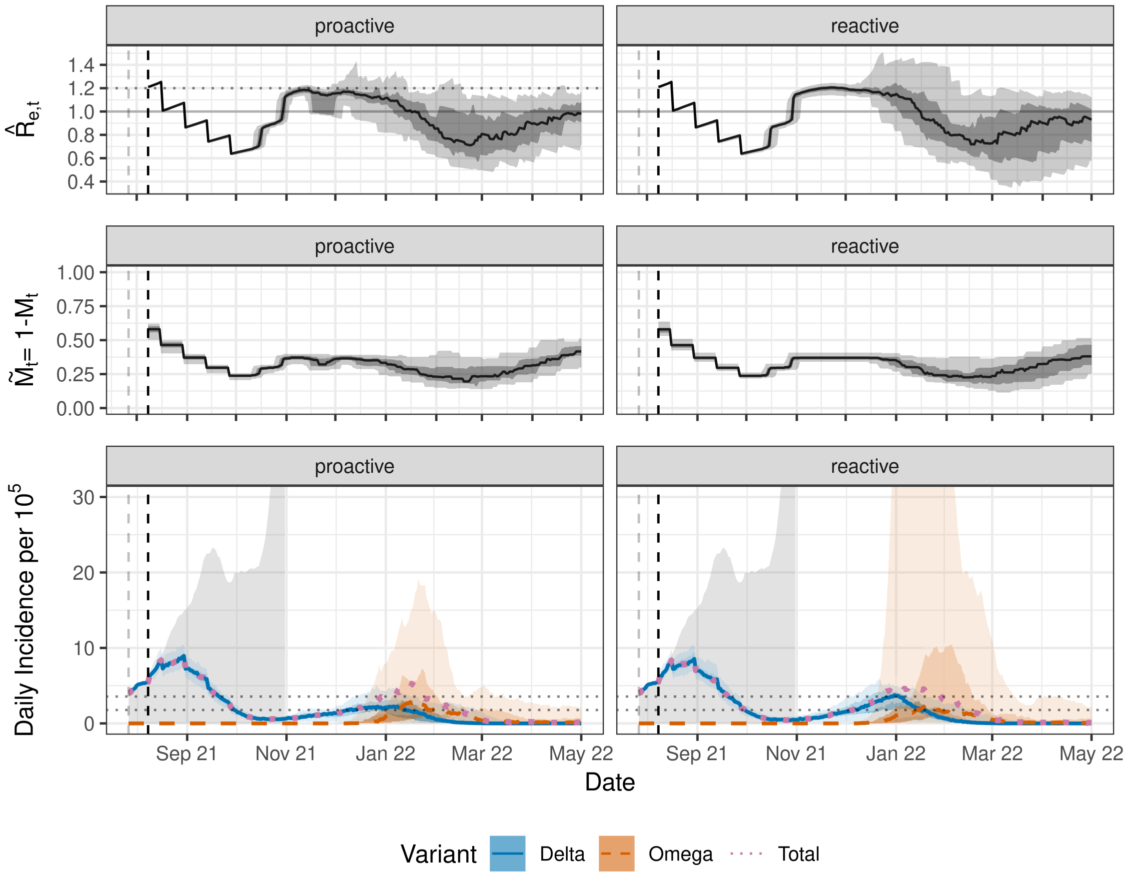


Figure 6: **Proactive control suppresses efficiently even a variant with a large $\rho_t^{\mathcal{V}}$.** In contrast, very large outbreaks are possible under reactive control. We modelled a modified Omega, which in addition to high immune escape (see Table 1), also has high basic reproduction number (same as Delta), while the mean of its generation interval is smaller at 3 days (CI: 1.5, 4.5). Case thresholds are 25-50/100,000/14 days. Both immunity waning and vaccine boosting are included. In contrast to all other simulations, here the import of Omega at the rate of 1/day starts later on 1.12.2021.

2 Sampling in practice

We use the following data and parameters specific to Austria

- As generation interval (w_t) we will choose a discretisation of a Gamma-distribution with mean 4.46 and standard deviation 2.63 cut off after day 13 as found by AGES in [5]. We use this generation interval for *all* variants. source:
- The Austrian population is given by $N = 8.932.664$ (as of 2021-01-01) [7].
- We use the data on the progress of the Austrian vaccination program provided by the Austrian Ministry of Health [3]. Individuals vaccinated with an mRNA respective vector vaccine will be added to the appropriate group 14 respective 21 days after the first dose. We take the median doses of the last 7 days of available data for projecting daily administered doses into the future.
- We use reported incidence provided by the WHO dashboard [8].

  We assume all infections that happened in 2020 were with the wild type and we allocate the incidence of 2021 according to data provided by AGES [2] (see also Fig 3). Furthermore we scale this reported incidence to account for undetected infections based on values found in [6]. that actually get reported) of 0.42 which lies within the range of values investigated for Austria in [6].
- We use an estimate of the observed reproduction number using the R package EpiNow developed by [1]. These estimates are provided *via* [4].

The sampling procedure is as follows:

(1) Import data and choose the parameters as above.

(2) Simulate the overlap between vaccinated and previously infected individuals to determine the group sizes S_t^h.

(3) Choose initial incidence values, separated into the different variants (e.g. taking constant percentages or fitting a simple model like an exponential growth model, see Section 3) as well as imported cases.

(4) Let t_0 the the the last day of observed data and let $\mathcal{R}_e$ denote the effective reproduction number observed on this day. For initialising the model we compute the initial mitigation $\tilde{M}_{t_0}$ using Equation 12 in the following way

$$\tilde{M}_{t_0} = \mathcal{R}_e W N \left(L_{t_0} \sum_{\mathcal{V} \in \mathscr{V}} \left(\sum_{g \in \mathscr{H}} \tilde{\gamma}^{g\mathcal{V}} |S_t^g| \sum_{h \in \mathscr{H}^{\mathcal{V}}} \sum_{m=1}^{\nu} \lambda^{\mathcal{V}} \mathcal{R}_0 I_{t-m}^h w_m \right) \right)^{-1}. \tag{1}$$

(5) Sample the model as described in the 2 section.

3 Estimating initial variant prevalence

In order to initialise the model with different variant prevalences as in the Delta simulations in 3.2.1, we use the following tentative estimation given the variant specific effective reproduction number $\mathcal{R}_{e,t}^{\mathcal{V}}$. For simplicity, we assume that initially all infections occur in non-interaction groups. and the only group infected with variant $\mathcal{V}$ will be group $g = 0$. Let us consider a simple exponential growth model

$$I_t^{\mathcal{V}} = c \left(\beta^{\mathcal{V}}\right)^t \tag{2}$$

for a growth rate $\beta^{\mathcal{V}} > 0$. We equate this to the expected incidence given in (1), that is

$$I_t^{\mathcal{V}} = \mathcal{R}_{e,t}^{\mathcal{V}} \sum_{m=1}^{\nu} I_{t-m}^{\mathcal{V}} w_m = \mathcal{R}_{e,t}^{\mathcal{V}} \sum_{m=1}^{\infty} I_{t-m}^{\mathcal{V}} w_m$$

(where for the last equation we set the necessary w_m to 0) to get the following relation

$$c\left(\beta^{\mathcal{V}}\right)^t = \mathcal{R}_{e,t}^{\mathcal{V}} \sum_{m=1}^{\infty} w_t c\left(\beta^{\mathcal{V}}\right)^{t-m} \Leftrightarrow \frac{1}{\mathcal{R}_{e,t}^{\mathcal{V}}} = \sum_{m=1}^{\infty} w_t \left(\beta^{\mathcal{V}}\right)^{-m} = G_\omega\left(\frac{1}{\beta^{\mathcal{V}}}\right),$$

where $G_w(x) = \sum_{t=1}^{\infty} w_t x^t$ denotes the *probability generating function* of w.

Thus we can recover the growth rate $\beta^{\mathcal{V}}$ from the group reproduction number $\mathcal{R}_{e,t}^{\mathcal{V}}$ in the following way

$$\beta^{\mathcal{V}} = \left(G_w^{-1}\left(\frac{1}{\mathcal{R}_{e,t}^{\mathcal{V}}}\right)\right)^{-1}. \tag{3}$$

Taking day t as a reference point we furthermore estimate the total growth rate $\hat{\beta}$ of the total incidence I_t using the observed reproduction number $\mathcal{R}_e$, precisely $\hat{\beta} = \left(G_w^{-1}\left(\frac{1}{\mathcal{R}_e}\right)\right)^{-1}$. If $p^{\mathcal{V}} \geq 0$ denotes the prevalence of variant $\mathcal{V}$ on day t, that is $I_t^{\mathcal{V}} = p^{\mathcal{V}} I_t$, then the estimated incidence of the previous days can be computed via

$$I_{t-s}^{\mathcal{V}} = p^{\mathcal{V}} \left(\frac{\hat{\beta}}{\beta^{\mathcal{V}}}\right)^s I_{t-s}.$$

Since the group specific reproduction number $\mathcal{R}_{e,t}^{\mathcal{V}}$ is usually not known it also needs to be tenatively estimated. For this we start with assuming constant variant prevalence first. Let $p^{\mathcal{V}_i} \geq 0$, $i = 1, \ldots, n$ such that $\sum_{i=1}^{n} p^{\mathcal{V}_i} = 1$ denote the prevalence of variant $\mathcal{V}_i$ within the total incidence I_t. Precisely $I_{t-s}^{\mathcal{V}_i} = p^{\mathcal{V}_i} I_{t-s}$ and $I_{t-s} = \sum_{i=1}^{n} p^{\mathcal{V}_i} I_{t-s}$ for $s = 0, \ldots, \nu - 1$. Using this constant initial incidence for each variant and using the observed reproduction number $\mathcal{R}_e$ we compute a tentative initial mitigation $\hat{M}_t$ using equation (1). The variant specific reproduction number $\mathcal{R}_{e,t}^{\mathcal{V}}$ can then be estimated in the following way

$$\hat{\mathcal{R}}_{e,t}^{\mathcal{V}} = \hat{M}_t L_t N^{-1} \sum_{\substack{h \in \mathscr{C} \\ s.t.\, h = \overline{h}V}} \tilde{\gamma}^{gh} |S_t^g| \lambda^{\mathcal{V}} \mathcal{R}_0 \tag{4}$$

and furthermore be used to compute the variant specific growth rate $\beta^{\mathcal{V}}$ using (3).

4 Estimating daily Delta prevalence in Austria

For Austria the VOC prevalence data is provided only with weekly granularity [2]. In order to translate this data to daily granularity when initialising the Delta simulations in 3.2.1, we again use the simple exponential growth model (2) and fit it to the weekly data. We take a reference time point $\bar{t}$ and use the following model

$$I_t^{\mathcal{V}} = \left(\beta^{\mathcal{V}}\right)^{t-\bar{t}} I_{\bar{t}}^{\mathcal{V}},\ t \in \mathbb{N},\ \beta^{\mathcal{V}} > 0.$$

If t is the end of a week, then the sum of all cases in this week is given via

$$\sum_{s=0}^{6} I_{t-s}^{\mathcal{V}} = \sum_{s=0}^{6} \left(\beta^{\mathcal{V}}\right)^{t-s-\bar{t}} I_{\bar{t}}^{\mathcal{V}} = \left(\beta^{\mathcal{V}}\right)^{t-\bar{t}} I_{\bar{t}}^{\mathcal{V}} \sum_{s=0}^{6} \left(\beta^{\mathcal{V}}\right)^{-s} = \left(\beta^{\mathcal{V}}\right)^{t-\bar{t}} I_{\bar{t}}^{\mathcal{V}} \frac{1 - \frac{1}{(\beta^{\mathcal{V}})^7}}{1 - \frac{1}{\beta^{\mathcal{V}}}}.$$

Let $\bar{J}_t^{\mathcal{V}}$ denote the cumulative weekly cases of variant $\mathcal{V}$ within the week that ends on day t. We then fit an exponential growth model via the following least squares method.

$$\inf_{(\beta^{\mathcal{V}}, I_{\bar{t}}^{\mathcal{V}}) \in \mathbb{R}^+ \times \mathbb{R}^+} \sum_t \left| \left(\beta^{\mathcal{V}}\right)^{t-\bar{t}} I_{\bar{t}}^{\mathcal{V}} \frac{1 - \frac{1}{(\beta^{\mathcal{V}})^7}}{1 - \frac{1}{\beta^{\mathcal{V}}}} - \bar{J}_t^{\mathcal{V}} \right|^2,$$

where we sum only over t corresponding to last days of the week.

To estimate the daily prevalence in Austria we fit this model to the weekly cumulative Alpha and Delta cases separately (assuming the prevalence of other variants to be negligible), thus daily Delta prevalence p_t^D can be recovered via

$$p_t^D = \frac{I_t^D}{I_t^A + I_t^D}.$$

We use the data provided by [2] between May 31, 2021 and July 21, 2021 to find that 20% Delta prevalence was most likely given on June 16, 2021 as seen in Fig 7.

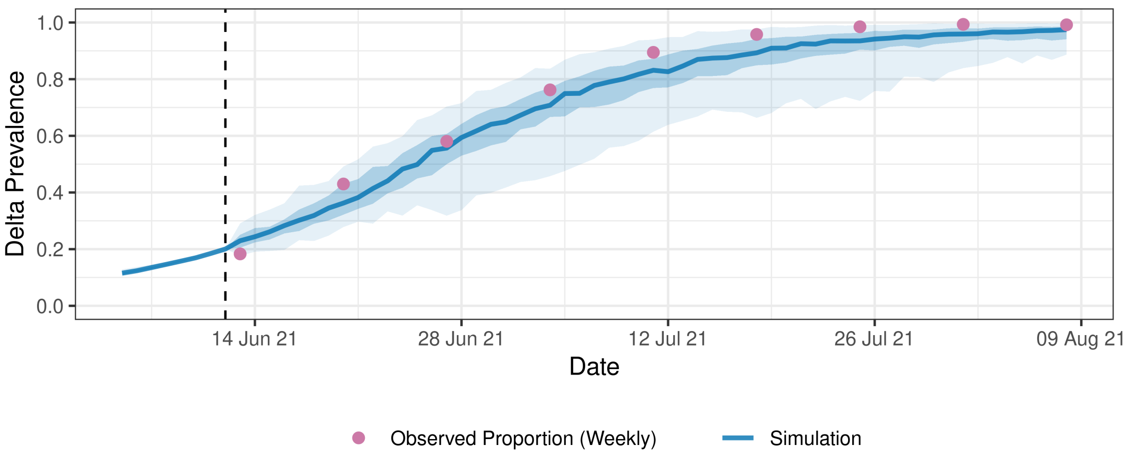


Figure 7: Reported and estimated Delta prevalence in Austria between May and August 2021.

References

[1] S. Abbott, J. Hellewell, R. N. Thompson, K. Sherratt, H. P. Gibbs, N. I. Bosse, J. D. Munday, S. Meakin, E. L. Doughty, J. Y. Chun, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research* 5.112 (2020), p. 112.

[2] Agentur für Gesundheit und Ernährungssicherheit GmbH. SARS-CoV-2-Varianten in Österreich. `https://www.ages.at/themen/krankheitserreger/coronavirus/sars-cov-2-varianten-in-oesterreich/`. Accessed: 2021-09-13. 2021.

[3] Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz. COVID-19 vaccination data for Austria. Tech. rep. 2021.

[4] M. Eder, J. Hermisson, M. Hledik, C. Hütter, E. Iofinova, R. Pisupati, J. Polechova, G. Puixeu, S. Sarikas, B. Wölfl, and C. Zimmermann. EpiMath: R-Nowcasting for Austria. Tech. rep. 2021.

[5] L. Richter, D. Schmid, A. Chakeri, S. Maritschnik, S. Pfeiffer, and E. Stadlober. Schätzung des seriellen Intervalles von COVID19, Österreich. Tech. rep. https://www.ages.at/en/wissen-aktuell/publikationen/schaetzung-des-seriellen-intervalles-von-covid19-oesterreich/. Apr. 2020.

[6] C. Rippinger, M. Bicher, C. Urach, D. Brunmeir, N. Weibrecht, G. Zauner, G. Sroczynski, B. Jahn, N. Mühlberger, U. Siebert, and N. Popper. Evaluation of undetected cases during the COVID-19 epidemic in Austria. *BMC Infectious Diseases* (2021).

[7] Statistik Austria. Population of Austria at the beginning of the year 2021 (regional status of 2021). Tech. rep. 2021.

[8] World Health Organization. WHO COVID-19 Dashboard. https://covid19.who.int/WHO-COVID-19-global-data.csv. Accessed: 2021-12-27. 2021.

CV - Annemarie Grass

Work

Since Apr 2018 **Project Assistant**, *University of Vienna*.
Associated to FWF-START Price Project Y-782, FWF-Projects P28661 and P35197 and WWTF Projekt MA16-021 under the supervision of Prof. Mathias Beiglböck

July 2017-Mar 2018 **Project Assistant**, *Vienna University of Technology*.
Associated to FWF-START Price Project Y-782 under the supervision of Prof. Mathias Beiglböck

Education

Jul 2017-Jan 2024 **Doktoratsstudium (equiv. to Ph.D) in Mathematics**, *Vienna University of Technology and University of Vienna*.

Since Mar 2017 **Magisterstudium (equiv. to M.Sc) in Statistics**, *University of Vienna*.

Mar 2016-Mar 2017 **M.Sc in Mathematics**, *University of Vienna*, *with distinction (1.0)*.
Area of Specialisation: Stochastics, Dynamical Systems and Financial Mathematics.
Master Thesis: *Uniqueness Properties of Barrier-Type Solutions to the Skorokhod Embedding Problem and Perkins Embedding with General Starting Law*, Supervisor: Prof. Mathias Beiglböck (at Vienna University of Technology).

Oct 2011-Mar 2016 **B.Sc in Mathematics**, *University of Vienna*, *with distinction (1.5)*.
- 1.Bachelor Thesis: *Das Spiel von Banach und Mazur*, Supervisor: Jakob Kellner
- 2.Bachelor Thesis: *Skorokhods Einbettungssatz*, Supervisor: Prof. Mathias Beiglböck

Sep 2005-June 2010 **Matura**, *Fashion Institute Hetzendorf*, Vienna, *with distinction (1.0)*.

Visits and Conferences

Sep 2023 **ÖMG Tagung 2023 - Meeting of the Austrian Mathematical Society**, *Karl-Franzens-University and Graz University of Technology*, (Contributed Talk: Optimal Stopping Characterizations of Solutions to the Skorokhod Embedding Problem and their Applications to Robust Mathematical Finance).

Jun 2023 **Conference on Stochastic optimal control in Economics, Finance, and Learning theory**, *ETH Zürich*, (Participation only).

Feb 2023 **Third Austrian Day of Women in Mathematics**, *Online*, (Contributed Talk: Optimal Stopping Characterizations of Solutions to the Skorokhod Embedding Problem).

Jan-Feb 2023 **Research Visit**, *Nanyang Technological University*, (Invited Talk: Optimal Stopping Characterizations of Solutions to the Skorokhod Embedding Problem and their Applications to Robust Mathematical Finance).

Jun 2022 **Conference on Advances in Mathematical Finance and Optimal Transport**, *Scuola Normale Superiore in Pisa*, (Participation only).

Mar 2021 **Conference on Stochastic Mass Transports**, *BIRS in Banff*, (Participation only).

Mar 2018-Oct 2022 **Research Visit (Reccuring)**, *University of Bath*.
For joint work with Alexander Cox.

July 2019 **Eight or Nine Talks on Contemporary Optimal Transport Problems**, *IRMA in Strasbourg*, (Participation only).

Sep 2018 **Conference on Robust Techniques in Quantitative Finance**, *University of Oxford*, (Participation only).

Teaching Experience

Summerterm 2019 - Winterterm 2020/21 **Linear Algebra for Statisticians**, *University of Vienna*, (Exercise Class).

Winterterm 2018/19 **Mathematics 1 for (international) Business Administration**, *University of Vienna*, (Exercise Class).

Winterterm 2018/19 and 2019/20 **Repetition Course on Mathematical Statistics**, *University of Vienna*.

Winterterm 2015/16 and 2016/17 **Mathematics 1 for (international) Business Administration**, *University of Vienna*, (As Teaching Assistant).

Summerterm 2015 and 2016 **Mathematics for Molecular Biologist**, *University of Vienna*, (Tutorials).

Summerterm 2013-Summerterm 2015 **Introduction to Computer Infrastructure**, *University of Vienna*, (Tutorials).

Voluntary Work and Participation

Oct 2012-June 2015 **Student Representative**, *Österreichische Hochschülerinnen und Hochschülerschaft (Austrian Students Union)*, Vienna.
- Case Worker for Ecological Affairs in the Department for Human Rights and Social Policy of the Austrian Students Union (Oct 2012-June 2013).
- Elected head of the Student Representatives of Mathematics (July 2013-June 2015).

Oct 2012-Jun 2020 **Various Committee Work**, *University of Vienna*.

Scholarships and Awards

Jan 2017-June 2017 Chosen to participate in the 'NaturTalente' programme organized by University of Vienna.

Apr 2016-Mar 2017 Research Scholarship granted via FWF-START Price Project Y-782.

2013/14-2016/17 Annual Performance Scholarship granted by the University of Vienna.

Computer-Skills

LaTex	advanced	Python	advanced
Matlab	advanced	R	advanced
Mathematica	intermediate	Java	basic

Language Skills

German (mother tongue), English (fluent), Spanish (basic), French (basic)

References

Mathias Beiglböck
Professor of Financial Mathematics
Department of Mathematics
University of Vienna
Oskar-Morgenstern-Platz 1,
Vienna, 1090, Austria
✉ mathias.beiglboeck@univie.ac.at

Irene Klein
Department of Statistics and Operation Research
University of Vienna
Oskar-Morgenstern-Platz 1,
Vienna, 1090, Austria
✉ irene.klein@univie.ac.at

Alex Cox
Professor of Probability
Department of Mathematical Sciences
University of Bath
Claverton Down
Bath BA2 7AY, United Kingdom
✉ a.m.g.cox@bath.ac.uk