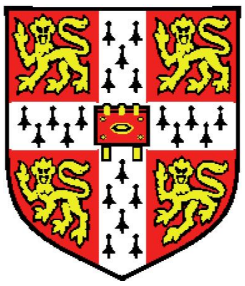
Brownian motion and multidimensional decision making



Judge Business School
University of Cambridge

Rutger-Jan Lange

King's College

A thesis submitted for the degree of

Doctor of Philosophy

30 September 2011

[This page was intentionally left blank]

Acknowledgements

I would like to thank my supervisor, Professor Danny Ralph, for his support and guidance throughout the years. Also, I would like to thank the Electricity Policy Research Group (EPRG) for supporting me financially me during my PhD, as well as for the many thought-provoking discussions. Further thanks go to my uncle Paul for his useful feedback on how this thesis fits within the wider literature. Many thanks to my Management Science buddies for making the PhD room such a worthwhile place as a starting point for discussions and coffee. Thanks to Alison for her brilliant editing skills. And lastly, I would like to thank my parents and my sister for being great; much more generally than just with regard to this PhD project.

[This page was intentionally left blank]

Declaration of Originality

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and acknowledgements. The work has not been previously submitted in part or in whole to any university for any degree or other qualification. In accordance with the regulations of the Judge Business School the thesis contains no more than 80,000 words of text.

Rutger-Jan Lange

King's College
Cambridge
30 September 2011

[This page was intentionally left blank]

Summary

This thesis consists of three self-contained parts, each with its own abstract, body, references and page numbering:

Part I **Potential theory, path integrals and the Laplacian of the indicator**
We write the transition density of absorbed or reflected Brownian motion in a d-dimensional domain as a Feynman-Kac functional involving the Laplacian of the indicator, thereby relating the hitherto unrelated fields of classical potential theory and path integrals.

Part II **The problem of alternatives**
We consider parallel investment in alternative technologies or drugs developed over time, where there can be only *one* winner. Parallel investment accelerates the search for the winner, and increases the winner's expected performance, but is also costly. To determine which candidates show sufficient performance and/or promise, we find an integral equation for the boundary of the optimal continuation region.

Part III **Optimal support for renewable deployment**
We consider the role of government subsidies for renewable technologies. Rapidly diminishing subsidies are cheaper for taxpayers, but could prematurely kill otherwise successful technologies. By contrast, high subsidies are not only expensive but can also prop up uneconomical technologies. To analyse this trade-off we present a new model for technology learning that makes capacity expansion endogenous.

There are two reasons for this standalone structure:

1. The target readership is divergent: Part I concerns mathematical physics, Part II operations research, and Part III policy. Readers interested in specific parts can thus read these in isolation. Those interested in the thesis as a whole may prefer to read the three introductions first.

2. The separate parts are only partially interconnected. Each uses some theory from the preceding part, but not all of it; e.g. Part II uses only a subset of the theory from Part I. The quickest route to Part III is therefore not through the entirety of the preceding parts. Furthermore, those instances where results from previous parts are used are clearly indicated.

[This page was intentionally left blank]

SUBMITTED (IN PARTS)

Part I — Potential theory, path integrals and the Laplacian of the indicator

Rutger-Jan Lange

University of Cambridge,
792 King's College, Cambridge, CB2 1ST, United Kingdom

E-mail: rjl63@cam.ac.uk

ABSTRACT: This paper unifies the field of potential theory — i.e. boundary value problems for the heat and Laplace equations — and that of the Schrödinger equation, by postulating the following seemingly ill-defined potential:

$$V(x) := \mp \frac{\sigma^2}{2} \nabla_x^2 \mathbb{1}_{x \in D}$$

where the volatility is the reciprocal of the mass (i.e. $m = 1/\sigma^2$) and $\hbar = 1$. The Laplacian of the indicator can be interpreted using the theory of distributions: it is the d-dimensional analogue of the Dirac δ'-function, which can formally be defined as $\partial^2/\partial x^2 \mathbb{1}_{x>0}$.
Regarding potential theory, our unified approach automatically produces the classical single and double boundary layer series. Regarding the Schrödinger equation, it automatically produces what is known as the Born series (or Born perturbation expansion). Apart from reproducing two known series solutions, our approach shows the *equality* of both solutions for a particular singular potential V, when this potential has the (scaled) Laplacian of the indicator as its limit. The sign of the potential depends whether the value (Dirichlet boundary condition) or derivative (Neumann boundary condition) is specified at the boundary.
Lastly, we demonstrate that the mode of convergence of the obtained series solutions is as follows:

mode of convergence	absorbed propagator	reflected propagator
convex domain	*alternating*	*monotone*
concave domain	*monotone*	*alternating*

As an independent contribution, we provide a new interpretation of the Feynman rules in a probabilistic setting, linking for the first time the Feynman-Kac formula to the Pascal matrix.

KEYWORDS: classical potential theory, boundary value problem, path integral, Brownian motion, Dirichlet problem, absorbed Brownian motion, Neumann problem, reflected Brownian motion, single boundary layer, double boundary layer, first passage, last passage, Feynman, Feynman-Kac, point interaction, Dirac delta, Dirac delta prime, Laplacian of the indicator, path decomposition expansion, multiple reflection expansion

Contents

1 Introduction

This paper considers the modified Dirichlet and Neumann boundary value problems for the heat and Laplace equations in d dimensions, and for a general class of domains D that allows Green's theorem — allowing a finite number of edges, corners and cusps, and where the value and normal derivative, respectively, are prescribed on the boundary. Our approach will be probabilistic in nature, interpreting the heat kernel as the absorbed or reflected transition density of a Brownian motion.

We will, first, contrast our approach with that of classical potential theory and its *ansatz* of single and double boundary layers. Second, we will consider a new approach to the Feynman-Kac functional. Finally, we will propose the synthesis of classical potential theory and path integral theory by postulating the following seemingly ill-defined potential:

$$V(x) := \mp\frac{\sigma^2}{2}\nabla_x^2 \mathbb{1}_{x\in D}.$$

This connects, as a by-product, potential theory to the study of Brownian local time. The potential can be viewed as the 'acceleration' of the time spent in D, by the Brownian particle, when the boundary points of D move outwards in the normal direction.

This introduction will motivate and sketch the arguments in the order mentioned above, thus proceeding roughly chronologically through the literature.

Considering that this paper deals with two grand fields of mathematical study, it has a relatively modest list of references. The reason for this is twofold. First, the problems in potential theory are amongst the most studied mathematical problems in the world. Although an almost endless list of references is possible, this would add little value to most readers. We have thus confined ourselves to referencing 1) some well-known standard reference works, 2) some references of historical interest, and, finally, 3) some specific references with which to contrast our results. Second, the approach is this paper is rather intuitive and care has been taken to make the paper self-contained.

As a final disclaimer it is worth noting that we examine the historical developments only superficially, without claiming to be precise or exhaustive. A much more detailed history of potential theory can be found in e.g. [1] (there was a lot of history as early as 1929!), or in [2], a reprint of the 1984 edition, which involves more recent discoveries as well.

1.1 Classical potential theory

Classical potential theory has a long history. Kellogg's (1929) widely quoted textbook [1] introduces potential theory by quoting Newton's Principia from another 242 years earlier — and we shall follow suit:

> Every particle of matter in the universe attracts every other particle, with a force whose direction is that of the line joining the two, and whose magnitude is directly as the product of their masses, and inversely as the square of their distance from each other.

19$^{\text{th}}$ century physicists, having learned of Newton's law of universal gravitation (1687) and of Coulomb's inverse-square law for electrostatic forces (1783), realised that the fundamental forces of nature could be written as the gradient of a scalar function. This scalar function was coined the 'potential'. The gravitational potential, for example, is defined as the scalar function that vanishes at infinity and such that the negative gradient equals the gravitational force — in both direction and magnitude. A simple calculation shows that we have

$$-\nabla_y \frac{-1}{|y-x|} = -\frac{y-x}{|y-x|^3},$$

where x and y are vectors, where ∇ is the gradient operator, and $|\cdot|$ indicates the norm. The gravitational potential at y as caused by a mass at x is identified as $\sim \frac{-1}{|y-x|}$. It increases with the distance between the particles, inspiring the concept of 'potential' energy. While infinite at $x = y$, at points away from masses it satisfies the Laplace equation

$$\nabla_y^2 \frac{-1}{|y-x|} = 4\pi\delta(|y-x|), \tag{1.1.1}$$

where this equation is to be read in a distributional sense. The Laplace equation also characterises the steady vibrations, the steady flow of heat and fluids, and is one of the most important equations in mathematical physics.

In moving from 19$^{\text{th}}$ century physics to 19$^{\text{th}}$ century mathematics, the emphasis shifts from the potential function, which satisfies the Laplace equation, to the Laplace equation itself and all functions that satisfy it. The concept of *harmonic* functions was introduced, for example, and a function f is defined as harmonic if it is twice continuously differentiable and satisfies the Laplace equation $\nabla^2 f = 0$ in some domain D. In one dimension the Laplace equation allows only straight lines. As a result, maxima and minima must occur at the boundary of the domain. Also, the value at any point x in the interior is equal to the average value taken over a set of equidistant points around x. It turns out that these properties persist in higher dimensions: harmonic functions have no local maxima or minima (and if they do, then they are constant) and harmonic functions satisfy the mean value property in the sense of equidistant points, i.e. spheres.

The development of harmonic functions gave rise to a multitude of questions. The Dirichlet problem, as posed by [3], asks for the existence and uniqueness of a function that is harmonic in D and takes certain prescribed boundary values on ∂D. It was long believed that there is always a solution, but Zaremba [4] gave the first counterexample. He considered a punctured open ball in $d \geq 2$, with prescribed boundary values 0 at the origin and 1 at the outer boundary. Because the boundary at the origin consists of an isolated point, there is no way to match a harmonic function to both boundary values.

Lebesgue [5] realised that not only do *isolated* boundary points cause problems, but so do other boundary points of zero measure, such as the tip of a *thorn* in $d \geq 3$. Lebesgue imagined pushing needle into a deformable sphere — thereby creating an inward pointing thorn (the terms *spine*, *cone* and *cusp* are also in use). Lebesgue shows that if the thorn is very sharp and if the prescribed boundary values are 1 on most of the sphere and 0 on most of the thorn, then the value at the tip of the thorn is not 0, as the Dirichlet problem demands, but positive.

Poincaré [6] had already used barriers, which are roughly equivalent to tangent planes, to show that the problem is solvable if every boundary point has one, and finally Wiener [7] gave necessary and sufficient conditions for the existence of the Dirichlet solution, precluding such examples as had been given by Zaremba and Lebesgue. As Kellogg [1] notes, 'no proof can ever be valid unless it places some restrictions on the region' (p. 278).

Because of the intimate relationship between potential theory and the theory of the Laplace equation, the Dirichlet problem is also referred to as the first boundary value problem of potential theory. The second boundary value problem of potential theory asks for a function that is harmonic in a region and that has a prescribed normal derivative at every boundary point, and is also known as the Neumann problem. The third boundary value problem prescribes a linear combination of the value and the normal derivative at the boundary. For all three boundary value problems similar issues of existence and uniqueness must be addressed, with similar requirements on the smoothness of the domain. The interior Neumann solution is unique up to an additive constant, for example. This paper adresses the Dirichlet and Neumann problems for the Laplace and heat equations.

To overcome the difficulties of existence as presented by Zaremba and Lebesgue, we will consider not the classical but the modified versions of these boundary value problems, as discussed in the next subsection.

1.2 The modified Dirichlet problem

To physicists, such as Green [8], it had always been 'clear' that the Dirichlet problem is solvable — because nature solves it. Suppose that a positive electrical charge is placed inside a perfect conductor, such that electrons are completely free to flow on its surface, and that the conductor is 'grounded' (i.e. it has an infinite supply of electrons). Then

the positive charge inside will induce a negative charge distribution on the conductor, such that the total configuration is charge neutral. The electrons on the conductor are attracted by the positive charge inside, but repelled by each other, leading very quickly to a static charge distribution. The induced charge on the conductor can be considered a continuous charge distribution, because electrons are phenomenally small; smaller than 10^{-15} m.

The electrical force is proportional to the gradient of the electrical potential, and thus an electrostatic equilibrium demands that the potential is constant everywhere on the conductor. If the potential is constant on the conductor, then the potential does not change in any tangential direction and therefore the gradient must point in the direction normal to the surface. The electrical force thus also points in the normal direction, but no electron can move in that direction and thus an equilibrium has been obtained. (If there were a force in any tangential direction at all, then some electrons would move.) The combined potential that results from the original charge and the induced charge solves the Laplace equation everywhere inside the conductor — except at the location of the positive charge, where it is infinite. And because it is zero (or at least constant) on the boundary, it solves the homogeneous Dirichlet problem.

From this observation, Green [8] inferred that the Dirichlet problem is always solvable. This is not technically correct since Green did not consider such geometries as proposed by Zaremba and Lebesgue. However, even for such irregular shapes as punctured disks and cones, it is clear that *some* charge distribution on the conductor must exist.

As far as Zaremba's punctured disk is concerned, we could hypothesise that no infinitesimally small bit of conductor would every carry a finite bit of charge, as the repulsive force between the electrons would become infinite. As far as the conductor is concerned, therefore, the punctured point is not really there — it would never put any finite amount of charge there. Only 1 electron could ever amass at an isolated boundary point and that is a negligible quantity.

As far as Lebesgue's thorn is concerned, it is clear that a conductor with such a shape could exist, and that it would carry *some* induced charge distribution. We can hypothesise that an equilibrium distribution is obtained when the potential on the conductor is zero everywhere, except, possibly, at the tip of the thorn. At the tip of the thorn multiple 'normal' directions exist and therefore a force pointing in any of the normal directions would be allowed, given that no electron could move in that direction. We would have to admit that the induced charge distribution can be discontinuous, if the conductor has a very irregular shape, but we would insist that *some* charge distribution exists.

Concluding, we see that as long as the potential deviates from its prescribed value at a set of points with measure zero, then an equilibrium distribution is obtained. The Dirichlet problem can thus be solved also for irregular shapes, albeit in a slightly weaker sense: the

prescribed boundary values are met almost everywhere, at all *regular* boundary points.

Not all the standard reference works discuss the modified Dirichlet problem. [9], [10], [11] and [12] are otherwise excellent references for this paper, but they discuss the classical and not the modified Dirichlet problem. The modified Dirichlet problem, sometimes called the generalised Dirichlet problem, is discussed by [13], [14], [15] and [2].

In this paper, by a 'solution' to any boundary value problem we mean a solution in the 'modified' sense, i.e. a solution that matches the prescribed boundary conditions at all boundary points, except possibly at a set of points with zero measure on the surface. We thus allow domains with corners, edges and sharp cones, but we only impose the boundary conditions at regular boundary points.

1.3 Green's identity

Green's 1828 paper [8] is important not only for historical reasons, but also because it introduces three indenties that bear his name. Here we introduce Green's second identity:

$$\int_D d\alpha\; u(\alpha) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} v(\alpha) = \oint_{\partial D} d\beta\; u(\beta) \left\{ \overleftarrow{\nabla}_\beta \cdot n_\beta - n_\beta \cdot \overrightarrow{\nabla}_\beta \right\} v(\beta). \tag{1.3.1}$$

This identify we shall find particularly useful. The notation is intended to make sense intuitively, but is also discussed in section 2: differential operators differentiate in the direction of the arrow. Green's identities are indispensable tools and we shall use this particular identity — known as Green's identity, Green's second identity and Green's theorem — repeatedly. Doob's (2001) seminal work on classical potential theory [2] opens with the claim that

> a bounded open set for which Green's [second] identity is true will be called smooth [and] a precise description of smooth sets is omitted

and, without further ado, Doob proceeds to use Green's second identity in the remaining 843 pages of his work. There can be no doubt about the validity of results that Doob obtains using Green's identity, since its application on smooth sets is allowed by definition. We will continue in a similar manner: using Green's identity throughout, ensuring that all results are true for sets that allow Green's identity, while not elaborating extensively on which sets allow Green's identity. Although the approach in [2] has the advantage of allowing use of Green's identity while evading its burden of proof, we do wish to shed *some* light on its validity. [1] discusses the divergence theorem, which is valid under the same conditions as Green's identity, at length, and he concludes (p. 118):

> The divergence theorem holds for any regular region R, with functions X, Y, Z which are continuous and piecewise continuously differentiable in R. [...] It is true that conical points, cannot, in general, occur on the boundary of a

> regular region. But by means of the second extension principle it is clear that a finite number of conical points may be admitted. More generally, if a region becomes regular by cutting out a finite number of portions by means of spheres of arbitrarily small radius, the areas of the portions of S cut out vanishing with the radius, then the theorem holds for that region.

Furthermore it is obvious, as [16] note, that when 'formulating a theorem of Green, one must certainly deal with two different kinds of assumptions, the geometrical ones and the analytical ones'. The geometrical ones have to do with the domain of integration, and the analytical ones with the functions that are integrated over. For one dimension, the divergence theorem is equivalent to the fundamental theorem of calculus, and we have that

$$F(b) - F(a) = \int_a^b f(x)\,dx,$$

where this holds if f is continuously differentiable or as long as f is locally integrable, such that $F' = f$ almost everywhere, as in [17] (p. 63, theorem 4.11). For Green's theorem, similarly, the analytical assumptions on the integrand can be relaxed to allow for functions that are merely locally integrable rather than continuously differentiable.

As far as the geometrical assumptions are concerned, we will assume the validity of Green's identity, but for the record we note that — roughly speaking — the identity holds for piecewise smooth domains, with an emphasis on piecewise: allowing edges, corners and thorns. Furthermore, we will consider both interior and exterior problems, i.e. the domain need not be bounded, as in Doob's definition of smooth.

We have emphasised in the previous subsection that we are looking for solutions to the *modified* Dirichlet and Neumann problems, allowing irregular boundary points but imposing the boundary conditions only at regular boundary points. It is thus obvious that a tool that requires a smooth boundary can never be used to solve a modified boundary value problem. The only tool that we will use, however, is Green's identity. Therefore, there is no conflict: we will try to solve modified boundary value problems, using Green's theorem along the way.

1.4 Brownian motion

In the same year in which the [8] paper appeared, the botanist Robert Brown noted the irregular movement of pollen suspended in water, which we now know is caused by random collisions with water molecules. Bachelier [18] realised that this Brownian motion could also be used to model the fluctuation of Parisian stock prices. Einstein [19] was the first to write down the transition density in 1 dimension: a normal distribution with a variance that increases linearly in time. In d dimensions, the transition density of Brownian motion

is as follows:

$$B(y,t|x,s) = \frac{1}{[2\pi\sigma^2(t-s)]^{d/2}}\; e^{-\frac{|y-x|^2}{2\sigma^2(t-s)}}. \tag{1.4.1}$$

The propagator $B(y,t|x,s)$ is equal to the (marginal) probability that a Brownian particle moves to space-time coordinate (y,t) given that it started at (x,s). Formally, Brownian motion is defined as the continuous process, with independent increments, and such that the increment during dt is normally distributed with mean zero and variance $\sigma^2 dt$. Using this, it can be proved that

$$dB_t^2 \to \sigma^2 dt \text{ almost surely.}$$

The first problem is to show that such a process actually exists, which was achieved by [20]. Many books on Brownian motion simply start with a statement on its transition density as above; see e.g. page 1 in [14]. One of the next big steps is the formulation of Itô's lemma of [21] in 1951, which utilises the above almost-sure equality in a Taylor expansion of $f(B_t)$ around $t = 0$ to show that

$$\begin{aligned} df(B_t) &= f'(B_t)dB_t + \frac{\sigma^2}{2}f''(B_t)dt, \\ f(B_t) - f(B_0) &= \int_0^t f'(B_\tau)dB_\tau + \frac{\sigma^2}{2}\int_0^t f''(B_\tau)d\tau, \end{aligned} \tag{1.4.2}$$

under the condition that f is twice differentiable. Itô's lemma is also discussed in introductory finance courses, such as [22] and [23]. In d dimensions we have that $\mathbb{E}dB = 0$ for each component of the vector B_t, and it follows for a function $f(t, B_t)$ depending on both space and time that

$$\mathbb{E}f(t+dt, x+B_{dt}) = f(t,x) + \left(\frac{\partial}{\partial t} + \frac{\sigma^2}{2}\nabla_x^2\right) f(t,x)dt,$$

such that every function, if it stays constant on average, must satisfy

$$\mathbb{E}f(t+dt, x+B_{dt}) = f(t,x) \to \left(\frac{\partial}{\partial t} + \frac{\sigma^2}{2}\nabla_x^2\right) f(t,x) = 0.$$

Of course it must hold that $B(y,t|x,s)$ itself is unbiased as the Brownian motion progresses from (x,s) to $(x+dB, s+ds)$, and therefore we must have

$$\left(\frac{\partial}{\partial s} + \frac{\sigma^2}{2}\nabla_x^2\right) B(y,t|x,s) = 0.$$

This is known as the Kolmogorov backward equation, and there is a similar forward equation. See [24], for example, for an explanation of how the forward and backward equations follow from the Kolmogorov semigroup property. We obtain both the forward and backward PDEs:

$$\begin{aligned} \text{forward PDE} \quad & \left(\frac{\partial}{\partial t} - \frac{\sigma^2}{2}\nabla_y^2\right) B(y,t|x,s) = 0, \\ \text{backward PDE} \quad & \left(\frac{\partial}{\partial s} + \frac{\sigma^2}{2}\nabla_x^2\right) B(y,t|x,s) = 0. \end{aligned} \tag{1.4.3}$$

where 'PDE' stands for partial differential equation. Whereas classical potential theory is based on the Laplace operator, the heat operator comes in two versions. [2] uses a notation where $\dot{\Delta} := \frac{\sigma^2}{2}\nabla_y^2 - \partial_t$ and $\overset{*}{\dot{\Delta}} := \frac{\sigma^2}{2}\nabla_x^2 + \partial_s$ and he writes (pp. 262 and 263):

> The potential theory based on the Laplace operator will be called *classical potential theory* below. The potential theory based on the heat operator $\dot{\Delta}$ and its adjoint $\overset{*}{\dot{\Delta}}$, called *parabolic potential theory*, will be developed [here]. [...] Parabolic potential theory is based on the pair $\dot{\Delta}$, $\overset{*}{\dot{\Delta}}$ and is similar in many respects to classical potential theory, but the fact that both $\dot{\Delta}$ and $\overset{*}{\dot{\Delta}}$ are involved means that two theories dual to each other must be considered simultaneously.

For parabolic potential theory, therefore, almost all equations of importance come in *pairs*, and, throughout this paper, we will thus present all our results in this dual manner. While it may sometimes seem superfluous to do (almost) the same calculation twice, we do in fact derive new results from this strict dual approach. Consequently we have chosen to be consistent throughout.

To complete the description of B, its explicit representation (1.4.1) shows that we have the following *pair* of short time conditions (STCs):

$$\begin{aligned} \text{forward STC } & \lim_{s \nearrow t} B(y,t|x,s) = \delta(|y-x|), \\ \text{backward STC } & \lim_{t \searrow s} B(y,t|x,s) = \delta(|y-x|). \end{aligned} \tag{1.4.4}$$

These equations say that, in a short period of time, a Brownian particle stays where it is. Because the propagator depends only on the time difference $(t-s)$, it is trivial that the STCs hold in a pair. It may therefore seem that by quoting both STCs explicitly we are being unnecessarily elaborate, but we ask for patience and promise that our persistence will pay off in the end.

Probability theory and potential theory are linked in two ways. The first is through the Green function. Note that we write 'Green's identity' but 'Green function'. The term 'Green's function' is used by [1], [10] and [25], but we write 'Green function' following [9], [15] and [2], who argues (p. 797) that 'writers who describe G as the Green's function should be condemned to differentiate the Lebesgue's measure using the Radon-Nikodym's theorem' (although we cannot help but notice that Doob uses 'Green's identity' throughout). In any case, we will use the term 'Green's identity' and 'Green function' in the sequel.

There are at least two possible probabilistic interpretations of the Green function. Suppose that a 'source' at x emitting Brownian particles (at a rate of 1 per unit of time) has been present from a time (infinitely) long ago, and we ask ourselves what the current density of particles is in space, where particles emitted in all past times contribute to the

density at y. We have the free Green function G_B as follows:

$$G_B(y,x) := \mathbb{E}_x \int_{-\infty}^{t} \delta(B_{t-s} - y)\, ds = \int_{-\infty}^{t} B(y,t|x,s)\, ds.$$

Alternatively, suppose that a source at x emits only *one* Brownian particle, and we ask what the expected amount of time is that the particle spends in the neighbourhood of any location in space, given that we observe the Brownian particle for an (infinitely) long time. We have the free Green function G_B as follows:

$$G_B(y,x) := \mathbb{E}_x \int_{s}^{\infty} \delta(B_{\tau} - y)\, d\tau = \int_{s}^{\infty} B(y,\tau|x,s)\, d\tau.$$

Since the Brownian density does not depend on either time coordinate individually, but only on the time difference, it will be clear that these two definitions (and interpretations) are identical.

However, it is also obvious, sadly, that neither of these is *guaranteed* to be finite, and it turns out that in two dimensions, we obtain indeed that G_B equals ∞. In this case Brownian motion is described as *recurrent*, since the particle returns to each area in space an infinite number of times, and spends an infinite amount of time there. We could make the motion in two dimensions *transient* by introducing an absorbing boundary. If the hitting time of the boundary is almost surely finite, and if the boundary is absorbing, then the motion can no longer be recurrent.

When the dimension is three or higher, then Brownian motion in all of space is transient, implying a finite density of paths everywhere. Although the 'number' of paths emitted by the source is ∞, we have that the 'size' of three dimensional space is also ∞, and it turns out that there is a non-trivial ratio, or density, that is finite everywhere. With this intuition, all of 2-dimensional space is simply not 'big' enough to obtain a finite ratio of particles per unit of space. In a closed and finite domain D with absorbing boundary, Brownian motion is transient in *any* dimension, since absorption happens almost surely at a finite time. For more on transient versus recurrent Brownian motion, see e.g. [12].

For a finite domain with a reflecting boundary, it is obvious that Brownian motion must be recurrent, since no particle can escape. But when the dimension is three or larger, and the domain is infinite, even reflected Brownian motion is transient, because the particle can escape to infinity.

For $d \geq 3$ the free Green function is finite and the integration can be performed to give:

$$G_B(y,x) = \frac{1}{\sigma^2} \frac{\Gamma(d/2-1)}{2\pi^{d/2}} |y-x|^{2-d}, \tag{1.4.5}$$

where Γ denotes the gamma function. To our surprise, we see that in $d = 3$ we have that the expected time spent around y when started at x equals the Newtonian gravitational potential $\frac{-1}{|y-x|}$ as in (1.1.1), up to a multiplicative constant. As [26] notes

> The relation [...] furnishes a vital link between two big things. At one end we see the Newton-Coulomb potential; at the other the normal density [...]. Can the linkage be a mere accident, or does it portend something of innate significance?

Indeed, the free Green function G_B satisfies the same differential equation that the Newtonian potential satisfies (up to a constant), namely

$$\begin{aligned}\frac{\sigma^2}{2}\nabla_y^2 G_B(y,x) &= \frac{\sigma^2}{2}\nabla_y^2 \int_s^\infty B(y,t|x,s)\,dt \\ &= \int_s^\infty \frac{\partial}{\partial t} B(y,t|x,s)\,dt \\ &= \left(\lim_{t\nearrow\infty} - \lim_{t\searrow s}\right) B(y,t|x,s) \\ &= -\delta(|y-x|).\end{aligned} \tag{1.4.6}$$

A similar calculation can be performed for the Laplacian with respect to x. Therefore we have

$$\frac{\sigma^2}{2}\nabla_y^2 G_B(y,x) = \frac{\sigma^2}{2}\nabla_x^2 G_B(y,x) = -\delta(|y-x|). \tag{1.4.7}$$

This should be compared with the equation satisfied by the Newtonian potential (1.1.1).

The second link between potential theory and Brownian motion was provided by [27], who was the first to realise that Brownian motion in $d = 2$ could be used to solve the Dirichlet problem. He noted that the first-passage distribution over the boundary of the domain, given an infinitely long observation interval, is harmonic in the starting point. In their recent book, Mörters and Peres [12] formulate in more modern language how the Dirichlet solution can be obtained as a weighted average over all first-passage times and locations (p. 70):

> [...] we can simulate the solution of the Dirichlet problem by running many independent Brownian motions, starting in $x \in U$ until they hit the boundary of U and letting $u(x)$ be the average of the values of [the given Dirichlet boundary data] on the hitting points.

Although this procedure in [12] concerns the classical and not the modified Dirichlet problem, it turns out that Brownian motion is, in fact, perfectly suited to address the latter problem. If the domain is irregular (with corners, edges and/or cusps), then a first-passage distribution still exists, even though it may be discontinuous — echoing our earlier argument that a static induced charge distribution on a conductor always exists, even if it is irregularly shaped.

In fact, the first passage will almost surely happen at a regular boundary point. Therefore the boundary data at points of zero measure are irrelevant for the macroscopic solution.

The problems of existence posed by [4] and [5] thus automatically disappear. Isolated points are *polar* for Brownian motion in $d \geq 2$; a single point will almost surely never be visited. As a result, a Brownian motion in a punctured disk will not 'feel' the isolated boundary point at the origin, since it will never hit it. The tip of an inward-pointing thorn is also irregular for a Brownian motion, because a Brownian motion started there need not leave the domain immediately. In fact, by Blumenthal's zero-one law, it will almost surely not leave the domain immediately. Thus, if the prescribed boundary values are zero on most of the thorn and one on most of the sphere, then the average first-passage value of a Brownian motion started at the tip of the thorn is indeed positive, as [5] had shown. In [26], Chung phrases it as follows

> although there may be irregular points on ∂D, no path will ever hit them. Thus they are not really there so far as the paths are concerned.

We conclude that, by defining the solution of the Dirichlet problem as in [12], the prescribed boundary values are met at all regular boundary points, just as the modified Dirichlet problem demands. We can intuitively see why this is the case: as the Brownian starting point x approaches a regular (i.e. non-singular) boundary point, the entire weight of the joint first-passage distribution (i.e. time and location) peaks at 'immediately' and 'here'. As the starting point x moves closer and closer to the boundary, therefore, the expectation over all first-passage times and locations will pick up just *one* contribution: that of the nearest boundary point.

While the 'free' Brownian process is denoted by B_t and its density by $B(y, t|x, s)$, the absorbed process is denoted by A_t and it transition density by $A(y, t|x, s)$. The absorbed density is unbiased as the particle progresses from x to $x + dB$ at $s + ds$ and thus it satisfies the same forward and backward PDEs that the free density satisfies. Since no Brownian particle can move from or to a regular boundary point without being absorbed, it satisfies the following 'forward' and 'backward' PDEs:

$$A(\beta, t|x, s) = A(y, t|\beta, s) = 0$$

for all regular boundary coordinates β. We can define the absorbed Green function as the expected time spent around y without being absorbed, i.e.

$$G_A(y, x) := \mathbb{E}_x \int_s^\infty \delta(A_t - y)\, dt = \int_s^\infty A(y, t|x, s)\, dt. \tag{1.4.8}$$

From the boundary conditions on A, it is clear that G_A equals zero for either x or y at a regular boundary location. Thus the equations satisfied by G_A are as follows:

$$\begin{gathered} \frac{\sigma^2}{2}\nabla_y^2 G_A(y, x) = \frac{\sigma^2}{2}\nabla_x^2 G_A(y, x) = -\delta(|y - x|) \\ G_A(\beta, x) = G_A(y, \beta) = 0 \end{gathered} \tag{1.4.9}$$

for all x and y in the interior and for all regular boundary points β. The electrostatic potential in a conductor satisfies almost the same differential equation (up to a factor) and exactly the same boundary condition; therefore, we conclude that the study of Green's electrostatic problem is equivalent to the study of absorbed Brownian motion.

In addition to the free and absorbed processes, the reflected process is denoted by R_t and it transition density by $R(y,t|x,s)$. The reflected density is unbiased as the particle progresses from x to $x+dB$ at $s+ds$ and thus it satisfies the same forward and backward PDEs that the free density satisfies. Since a Brownian particle is reflected in the normal direction, at each regular boundary point, it satisfies the following 'forward' and 'backward' PDEs:

$$\overrightarrow{\partial_\beta} R(\beta,t|x,s) = R(y,t|\beta,s)\overleftarrow{\partial_\beta} = 0$$

for all regular boundary coordinates β. We can define the absorbed Green function as the expected time spent around y, i.e.

$$G_R(y,x) := \mathbb{E}_x \int_s^\infty \delta(R_t - y)\, dt = \int_s^\infty R(y,t|x,s)\, dt. \tag{1.4.10}$$

From the boundary conditions on R, it is clear that G_R satisfies a set of equations as follows:

$$\begin{gathered} \frac{\sigma^2}{2}\nabla_y^2 G_R(y,x) = \frac{\sigma^2}{2}\nabla_x^2 G_R(y,x) = -\delta(|y-x|) \\ \overrightarrow{\partial_\beta} G_R(\beta,x) = G_R(y,\beta)\overleftarrow{\partial_\beta} = 0 \end{gathered} \tag{1.4.11}$$

for all x and y in the interior and for all regular boundary points β. The reflected Green function (as defined here) only exists for $d \geq 3$ and unbounded domains. If $d \leq 3$ or the domain is bounded, then the reflected particle returns to each location in space an infinite number of times, and the expected time spent in any small location is infinite. There are ways to define an 'interior' reflected Green function, see e.g. [25], but we shall not need this here.

Brosamler's (1976) [28] discovery that reflected (rather than absorbed) Brownian motion could reproduce the solution to the Neumann problem further strengthened the case for the use of stochastic processes to study the solutions of partial differential equations. For more on reflected Brownian motion and potential theory see e.g. [29], [30] and [31].

The link between probability theory and classical potential theory has inspired many articles and books with both terms in their titles, notably *Brownian Motion and Classical Potential Theory* by [14], *Green, Brown and Probability* by [26] and *Classical Potential Theory and Its Probabilistic Counterpart* by [2].

The solution to the third boundary value problem, where a linear combination of the value and derivative at the boundary is specified, can be obtained by considering 'elastic' Brownian motion. Elastic Brownian motion is reflected or absorbed with a certain

probability every time it hits the boundary. The third boundary value problem is not discussed in this paper.

1.5 Single and double boundary-layers

Green's *third* identity is not one that we will use frequently, but it shall serve as an important exposition for the classical method of obtaining solutions for boundary value problems of the Laplace equation. Take a truly harmonic function u in D, satisfying $\nabla^2 u = 0$, and take v to be the free Green function $v = G_B$ satisfying $\frac{\sigma^2}{2}\nabla^2 G_B = -\delta$, and substitute these in Green's second identity (1.3.1)

$$\frac{\sigma^2}{2}\int_D d\alpha\; u(\alpha)\left\{\overleftarrow{\nabla}^2_\alpha - \overrightarrow{\nabla}^2_\alpha\right\} G_B(\alpha, x) = \frac{\sigma^2}{2}\oint_{\partial D} d\beta\; u(\beta)\left\{\overleftarrow{\nabla}_\beta \cdot n_\beta - n_\beta \cdot \overrightarrow{\nabla}_\beta\right\} G_B(\beta, x),$$

to obtain Green's third identity

$$u(x) = \frac{\sigma^2}{2}\oint_{\partial D} d\beta\; u(\beta)\left\{\overleftarrow{\nabla}_\beta \cdot n_\beta - n_\beta \cdot \overrightarrow{\nabla}_\beta\right\} G_B(\beta, x). \tag{1.5.1}$$

Green's third identity may seem like a trivial variation on Green's second identity, but it has one profound consequence: it shows that *every* harmonic function is completely determined by its boundary behaviour. To obtain the harmonic value at x, one need only 'weigh' the boundary derivatives by G_B, and 'weigh' the boundary values by a factor proportional to the normal derivative of G_B. Closer boundary values and derivatives thus carry more weight in the determination of the value at x than do faraway ones, and so do boundary points β for which the outward normal vector points roughly in the same direction as the line joining x and β.

Unfortunately it is rarely the case that both the boundary values and the boundary derivatives are given. For the Dirichlet problem, for example, the boundary values are given but not the boundary derivatives, and the opposite holds for the Neumann problem. Instead of using the free Green function, however, we could use the absorbed Green function $v(\alpha) = G_A(\alpha, x)$ that satisfies the same differential equation to obtain:

$$u(x) = \frac{\sigma^2}{2}\oint_{\partial D} d\beta\; u(\beta)\left\{\overleftarrow{\nabla}_\beta \cdot n_\beta - n_\beta \cdot \overrightarrow{\nabla}_\beta\right\} G_A(\beta, x).$$

But because G_A is zero when evaluated on the boundary, we get

$$u(x) = \oint_{\partial D} d\beta\; u(\beta)\left\{-\frac{\sigma^2}{2} n_\beta \cdot \overrightarrow{\nabla}_\beta\right\} G_A(\beta, x) \tag{1.5.2}$$

and thus *only* the boundary values (and not the derivatives) need to be given. If we want to construct the Dirichlet solution, therefore, we need to find G_A. The absorbed Green function G_A is determined by the differential equations and boundary conditions (1.4.9) for all x and y in the interior and all regular boundary points β. Finding G_A is therefore

equivalent to solving the Dirichlet problem. In order to find G_A we need only look at (1.4.9), in which the function u does not even appear. In his classic book on electrodynamics, [25] writes

> the Green's functions satisfy simple boundary conditions, which do not depend on the detailed form of the Dirichlet (or Neumann) boundary values. Even so, it is often rather involved (if not impossible) to determine $[G.]$ because of its dependence on the shape of the surface.

It is indeed rather involved to determine $G.$, and that is one of the main aims of this paper. The absorbed Green function G_A, as defined by (1.4.8) and satisfying (1.4.9), has two physical interpretations. Either it can be interpreted as the expected time spent around a certain location by a Brownian motion that is absorbed at the boundary. Or, up to a factor, $G_A(\alpha, x)$ represents the potential at α caused by a unit charge at x in a perfect conductor. The absorbed Green function G_A satisfies the homogeneous boundary condition, where the interpretation can again be twofold. In the Brownian interpretation this is because no time can be spent by a Brownian particle at the boundary when the boundary is absorbing. In electrostatic interpretation it is because the tangential derivatives must vanish on the conductor for there to be an equilibrium. The first-passage distribution and the induced charge density are both proportional to the normal gradient at the boundary, i.e. $n_\beta \cdot \nabla_\beta G_A(\beta, x)$. This did not escape Green [8], who noted that the Dirichlet problem could be solved in some domain if one could work out the induced charge density on a perfect conductor of the same shape.

Equivalently, the Dirichlet problem can be solved if we can work out the first-passage density over the domain, and therefore the Dirichlet problem reduces to finding the transition density for an absorbed Brownian motion in a certain domain. 'All' that is needed, then, is to find the absorbed Green function. In the 'standard' approach, e.g. in Balian & Bloch [32], the following *ansatz* is made

$$G_A(y, x) = G_B(y, x) - \int_{\partial D} d\beta\, \mu_{\mathrm{DBL}}(y, \beta) \left\{ -\sigma^2\, n_\beta \cdot \overrightarrow{\nabla}_\beta \right\} G_B(\beta, x). \tag{1.5.3}$$

The German word *ansatz* is common in the physics literature, and can be taken to mean an educated guess which is later verified. Here μ is known as a 'double boundary layer', and hence we attach the subscript 'DBL'. Although the methods of single and double boundary layers appeared in Kellogg as early as 1929, it seems that Balian & Bloch [32] were the first in the physics literature to use them systematically to obtain series solutions for Green functions. For the Neumann problem the relevant quantity is the reflected Green function G_R, and [32] propose that it should look like:

$$G_R(y, x) = G_B(y, x) - \int_{\partial D} d\beta\, \mu_{\mathrm{SBL}}(y, \beta) G_B(\beta, x), \tag{1.5.4}$$

where the unknown function μ is now known as a 'single boundary layer', and hence the subscript 'SBL'. The status of single and double boundary layers has remained that of an ansatz, even in more modern handbooks on integral equations, such as [13], [33] or [34].

It is not clear, however, why the double boundary layer method should be reserved for the Dirichlet problem and why it would not be possible to find the absorbed propagator using a single boundary layer. We know from the electrostatic problem that the positive charge induces a negative charge density on the surface, and we should be able to write the potential as the sum of the direct and induced charges; i.e. as a single boundary layer. Although this intuition would lead to a different ansatz than the one that is standard, we would still be stuck with an ansatz. Therefore we propose a different method altogether. By definition of the free and absorbed Green functions, we can write:

$$\begin{aligned} G_A(y,x) &= G_B(y,x) - \frac{\sigma^2}{2}\int_D d\alpha\, G_B(y,\alpha)\left\{\overleftarrow{\nabla}^2_\alpha - \overrightarrow{\nabla}^2_\alpha\right\} G_A(\alpha,x), \\ G_A(y,x) &= G_B(y,x) + \frac{\sigma^2}{2}\int_D d\alpha\, G_A(y,\alpha)\left\{\overleftarrow{\nabla}^2_\alpha - \overrightarrow{\nabla}^2_\alpha\right\} G_B(\alpha,x). \end{aligned} \tag{1.5.5}$$

This pair should be viewed as consisting of *identities* by virtue of (1.4.7) and (1.4.9). Applying Green's second identity (1.3.1) we get

$$\begin{aligned} G_A(y,x) &= G_B(y,x) - \frac{\sigma^2}{2}\oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\overleftarrow{\nabla}_\beta\cdot n_\beta - n_\beta\cdot\overrightarrow{\nabla}_\beta\right\} G_A(\beta,x), \\ G_A(y,x) &= G_B(y,x) + \frac{\sigma^2}{2}\oint_{\partial D} d\beta\, G_A(y,\beta)\left\{\overleftarrow{\nabla}_\beta\cdot n_\beta - n_\beta\cdot\overrightarrow{\nabla}_\beta\right\} G_B(\beta,x). \end{aligned} \tag{1.5.6}$$

Because G_A disappears on the boundary, we need the arrows on the differential operators to point towards G_A:

$$\begin{aligned} G_A(y,x) &= G_B(y,x) - \oint_{\partial D} d\beta\, G_B(y,\beta)\left\{-\frac{\sigma^2}{2} n_\beta\cdot\overrightarrow{\nabla}_\beta\right\} G_A(\beta,x), \\ G_A(y,x) &= G_B(y,x) - \oint_{\partial D} d\beta\, G_A(y,\beta)\left\{-\frac{\sigma^2}{2}\overleftarrow{\nabla}_\beta\cdot n_\beta\right\} G_B(\beta,x). \end{aligned} \tag{1.5.7}$$

Let us define the following scaled inward differential operators:

$$\begin{aligned} \overrightarrow{\partial_\beta} &:= -\sigma^2 n_\beta\cdot\overrightarrow{\nabla}_\beta, \\ \overleftarrow{\partial_\beta} &:= -\sigma^2\overleftarrow{\nabla}_\beta\cdot n_\beta. \end{aligned} \tag{1.5.8}$$

Now we can write

$$\begin{aligned} G_A(y,x) &= G_B(y,x) - \oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_A(\beta,x), \\ G_A(y,x) &= G_B(y,x) - \oint_{\partial D} d\beta\, G_A(y,\beta)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} G_B(\beta,x). \end{aligned} \tag{1.5.9}$$

The signs and factorisations in the pair of equations above are carefully chosen. The absorbed Green function equals the free Green function minus all paths that have a first

passage at β, which happens with probability $\frac{1}{2}\overrightarrow{\partial_\beta} G_A(\beta, x)$, and which then propagate freely from β to y by G_B. Alternatively, the absorbed Green function equals the free Green function minus all paths that propagate to the boundary β, where they have their last passage before moving to y, with probability $G_A(y,\beta)\frac{1}{2}\overleftarrow{\partial_\beta} G_B(\beta, x)$. The equivalence of first and last passage decompositions is fully explored in section 3.

With this pair of equations we have related the absorbed Green function G_A to its boundary derivatives, and no ansatz whatsoever has been used. Instead, we have used only 1) the Laplace equation that is satisfied almost everywhere by both G_B and G_A, 2) Green's theorem and 3) the boundary conditions on G_A. Because we have only used Green's theorem this pair of equations should serve for irregular domains as well, something that is explicitly forbidden in [32] — as can be seen from the title of their that alone.

Both the results above can be used to obtain a series solution, by substituting the equation into itself. This procedure amounts to using the left-hand side of the equation as the definition for G_A appearing on the right-hand side, and the resulting infinite series is known as *Neumann's series*:

$$\begin{aligned} G_A(y,x) &= G_B(y,x) + \sum_{i=1}^{\infty}(-1)^i \left[\oint d\beta_i \dots \oint d\beta_1\right] G_B(y,\beta_i) \left[\prod_{k=2}^{i} \overrightarrow{\partial_{\beta_k}} G_B(\beta_k, \beta_{k-1})\right] \overrightarrow{\partial_{\beta_1}} G_B(\beta_1, x) \\ G_A(y,x) &= G_B(y,x) + \sum_{i=1}^{\infty}(-1)^i \left[\oint d\beta_i \dots \oint d\beta_1\right] G_B(y,\beta_i) \overleftarrow{\partial_{\beta_i}} \left[\prod_{k=1}^{i-1} G_B(\beta_{k+1}, \beta_k) \overleftarrow{\partial_{\beta_k}}\right] G_B(\beta_1, x) \end{aligned} \tag{1.5.10}$$

where the only difference between the two series is the direction of the arrows. The first series has $\overrightarrow{\partial} G_B$ as its rightmost element, in each term, and therefore it looks like a double boundary layer formulation:

$$G_A(y,x) = G_B(y,x) - \int_{\partial D} d\beta\, \mu_{\text{DBL}}(y,\beta) \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_B(\beta, x)$$

where the series definitions of μ_{DBL} can be read off. The second series has G_B as its rightmost element, in each term, and thus we see that it looks a single boundary layer formulation:

$$G_A(y,x) = G_B(y,x) - \int_{\partial D} d\beta\, \mu_{\text{SBL}}(y,\beta) G_B(\beta, x)$$

and the definition for μ_{SBL} can be read off. Thus we conclude that the absorbed Green function can be found either as a single or double boundary layer series — and in fact it turns out that both series are identical, term by term.

For the reflected Green function we again find two series, except that all terms have

positive signs in front of them:

$$G_R(y,x) = G_B(y,x) + \sum_{i=1}^{\infty} \left[\oint d\beta_i \dots \oint d\beta_1 \right] G_B(y,\beta_i) \left[\prod_{k=2}^{i} \overrightarrow{\partial_{\beta_k}} G_B(\beta_k, \beta_{k-1}) \right] \overrightarrow{\partial_{\beta_1}} G_B(\beta_1, x)$$
$$G_R(y,x) = G_B(y,x) + \sum_{i=1}^{\infty} \left[\oint d\beta_i \dots \oint d\beta_1 \right] G_B(y,\beta_i) \overleftarrow{\partial_{\beta_i}} \left[\prod_{k=1}^{i-1} G_B(\beta_{k+1}, \beta_k) \overleftarrow{\partial_{\beta_k}} \right] G_B(\beta_1, x) \tag{1.5.11}$$

We note that in the derivation of these series, we have used all the conditions that are supposed to specify the Green function. If we are optimistic, we could hope that the obtained infinite series 1) satisfies all the requirements that are used in its derivation, and 2) converges, because all the requirements, which ensure existence and uniqueness, have been used in its derivation.

Contrast this with the ansatz approach: the 'multiple reflection expansion' by [32] has been used in the physics literature by [35], [36], [37] and [38]. We will show in section 3 that the 'symmetrisation' procedure by [36] is incorrect, which is a mistake inherited by [38]. Furthermore, because the results are based on an ansatz, they must be verified after the fact, and it is thought that they are valid only for smooth domains.

In probability theory, a similar method is known as the *parametrix method*, which is also based on an ansatz, and is explored in [30] for example — which again requires a smooth boundary, as can be seen from the first sentence of the paper.

Instead we find that 1) single and double boundary layers need not be based on an ansatz, that 2) either problem may be solved with either method and that their distinction thus is arbitrary, and 3) that they may be useful for irregular as well as regular domains, by virtue of Green's theorem. In section 3 we show the following modes of convergence for the obtained series solutions:

mode of convergence	Dirichlet problem	Neumann problem
convex domain	*alternating*	*monotone*
concave domain	*monotone*	*alternating*

Our approach puts the single and double boundary layers on a more solid footing, but we also show, in subsection 3.9, how to derive some *new* integral equations:

$$G_A(y,x) = \frac{\sigma^2}{2} \int_D d\alpha \, G_A(y,\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_B(\alpha, x).$$
$$G_A(y,x) = \frac{\sigma^2}{2} \int_D d\alpha \, G_B(y,\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_A(\alpha, x). \tag{1.5.12}$$

This shows that the absorbed Green function is an eigenfunction of an integro-differential operator working on either the right or the left. Given that the problem was originally well-posed and given that in the derivation of the integro-differential equation we have again used all the conditions that are supposed to specify the solution, we could be optimistic

and expect that applying the integro-differential operator repeatedly on a trial function should give the correct answer, as a fixed point.

We note that the differentiation and integration are now over the interior of the domain rather than the boundary. While for practical purposes this might be a disadvantage because it leads to d dimensional integrals rather than $d-1$ dimensional integrals, it might be an advantage theoretically. This is because it shows that changing a single boundary location (making it irregular, for example) should have little effect if the change on the volume as a whole is negligible. We expect the integration over the volume to be somewhat more robust, in some sense, with respect to irregular boundary points.

The extension from smooth domains to piecewise smooth domains, for all these integral equations, may seem only of minor relevance. But since Kac's 1996 paper entitled 'Can one hear the shape of a drum?' [39], the topic of isospectral domains for the Dirichlet and Neumann problem has received much interest. The question can be rephrased as follows: if all the eigenvalues of the Dirichlet or Neumann solution are given, can one uniquely reconstruct the domain? It turns out that the answer is 'yes' if the domain is smooth and 'no' if sharp corners are allowed. This work provides a tool for calculating the Green function for domains of either type. For more on isospectral drums see [40], [41], [42].

In a closed domain with an absorbing boundary, a Brownian path is eventually absorbed with probability 1. Suppose, however, that while the Brownian particle is still alive there is a probability of $\lambda\, dt$, in each period of time dt, that its 'probabilistic mass' doubles. Then it proceeds as before and its probabilistic weight may double again. In some sense we could imagine that the second particle joins up with the first to create a double-decker bus. Upon two interactions, there will be 4 particles on top of each other. If an n-decker bus hits the boundary, all particles are destroyed. If λ is relatively small then some particles will be created, but eventually all particles will be absorbed by the boundary. But if λ exceeds a certain critical value then the 'probabilistic weight' of the particle that is still alive after a long time will start to dominate. While the probability that a particle is still alive after time t decreases exponentially, if its 'weight' increases exponentially at a faster rate, then the contribution of this path will start to dominate. It turns out that this critical λ is the first eigenvalue of the Dirichlet problem. The integro-differential equations above can be re-derived in the setting with particle creation at rate λ, to give

$$\begin{aligned} G_A(y,x) &= \int_D d\alpha\, G_A(y,\alpha) \left\{ \frac{\sigma^2}{2} \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha + \lambda \right\} G_B(\alpha,x), \\ G_A(y,x) &= \int_D d\alpha\, G_B(y,\alpha) \left\{ \frac{\sigma^2}{2} \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha + \lambda \right\} G_A(\alpha,x). \end{aligned} \tag{1.5.13}$$

Originally we expected a repeated application of the integro-differential operator to give rise to a convergent answer, but this is no longer true for $\lambda > \lambda_{\text{critical}} = \lambda_1$, where λ_1 is the first eigenvalue of the Dirichlet problem. Thus we have related the study of the eigenvalues

of a certain domain to the compactness of a certain operator. This is a new result and while this is an exciting field of study, this avenue is tangent to the main work of this paper and we will not pursue it further.

1.6 The Feynman-Kac formula

Seemingly unrelated to the discussion so far is the literature on Feynman path integrals and the Feynman-Kac formula. In [43], Feynman developed the path integral to describe the movement of a quantum particle, but an old probabilistic tool appears to make the path integral rather more intuitive. In probability theory, the Chapman-Kolmogorov equations describe that to propagate from starting point to ending point, a particle needs to be *somewhere* at *any* intermediate time. For any stochastic process with stationary increments, therefore, every path can be cut up into two pieces at some arbitrary intermediate time, as long as the intermediate location is integrated over its entire range.

As a physicist, Feynman visualised putting a *screen* between starting and ending point, such that every path connecting them would have to cross it, at some point and some time. The total (quantum mechanical) 'amplitude' is obtained by summing over all locations on the screen. Feynman realised that he could use as many screens as he wanted, and in the limit where the number of screens goes to infinity, the integration is no longer over one set of intermediate locations, but rather over all sets of intermediate locations, or *paths*.

This came to be known as Feynman's path integral, and from a probabilistic point of view it can be seen as a repeated application of the Chapman-Kolmogorov principle. Feynman's path integral is the solution to the complex Schrödinger equation, which differs from the heat equation only by a factor of the imaginary i in front of the time derivative. By transforming the Schrödinger equation to imaginary time, i.e. by letting $t \to -i\,t$, we recover the heat equation which governs the transition density of Brownian motion. Similarly, we can transform back to a quantum mechanical setting by letting $t \to i\,t$. The connection with probability theory is therefore clear, and Kac exploited this connection in [44] to show that the solution of the heat equation with initial conditions could be written as an expectation over all possible paths, which was dubbed the *Feynman-Kac formula*.

Historically the idea of a quantum mechanical potential was inspired by potentials in classical mechanics, but its interpretation is different. The Schrödinger equation involves a 'potential' that 'scatters' the particle. The interpretation of the potential in a probabilistic setting, however, is even further removed from the Newtonian interpretation. In a probabilistic interpretation, the potential creates or destroys Brownian particles depending on its sign, and at a rate proportional to its absolute value. For the heat equation the potential can be seen as a dissipation (heating or cooling) rate, again depending on its sign and absolute value; see for example [10], [15] or [12]. The word *potential* as used in this subsection thus has very little to do with the Newtonian interpretation of a potential.

When the potential is bounded, only a finite number of interactions with the potential can happen in any finite period of time. If the potential is positive, then each interaction kills the Brownian path. In principle there can be any number of interactions, but only paths without interactions survive. The forward and backward PDEs for this situation are as follows:

$$\begin{aligned}\left(\frac{\sigma^2}{2}\nabla_y^2 - \frac{\partial}{\partial t} - \lambda V(y)\right)\psi_V(y,t|x,s) = 0\\ \left(\frac{\sigma^2}{2}\nabla_x^2 + \frac{\partial}{\partial s} - \lambda V(x)\right)\psi_V(y,t|x,s) = 0\end{aligned} \tag{1.6.1}$$

where the symbol ψ is customary in quantum mechanics, but can be interpreted as the transition density in our setting, and where we indicate the dependence on the potential V by the subscript. The 'strength' of the potential can be tweaked through the value of the 'coupling constant' λ, and the Schrödinger equation can be re-obtained by transforming $t \to i\,t$.

The probability of an interaction with the potential (i.e. annihilation) at any given location α equals $\lambda\, V(\alpha)\, \epsilon$, where ϵ equals the amount of time spent at location α. Suppose that a path is determined by $N-1$ intermediate locations between (x,s) and (y,t), such that the time spent at each of the N intermediate locations and at the final location y equals $\epsilon = (t-s)/N$. The probability of survival is equal to the probability that no interaction with the potential occurs, and therefore the probability of survival for a given path equals

$$\prod_{i=1}^{N}(1-\lambda\, V(B_{\tau_i})\,\epsilon) \approx \prod_{i=1}^{N} e^{-\lambda\, V(B_{\tau_i})\,\epsilon} = e^{-\lambda \sum_{i=1}^{N} V(B_{\tau_i})\epsilon} \to e^{-\lambda \int_s^t V(B_\tau) d\tau}$$

where the last relationship holds in the limit for large N and where the path is no longer defined by its intermediate locations but rather by the entire, continuous, nowhere-differentiable Brownian path. If the above is the probability that a *given* path should survive (with N known intermediate locations), then the probability that *any* path should survive is obtained by taking an expectation over all possible intermediate locations, i.e. over all paths. If we want the path to end up at y then we need to take an expectation over all paths while enforcing the last position to be y. We can achieve this by plugging in a δ-function at y. We have now heuristically re-derived the Feynman-Kac formula, which postulates that the transition density from (x,s) to (y,t) in the presence of a (positive) annihilating potential V equals

$$\psi_V(y,t|x,s) = \mathbb{E}_x\left(\delta(B_t - y)\; e^{-\lambda \int_s^t V(B_\tau) d\tau}\right), \tag{1.6.2}$$

implying that in a short period of time the particle 1) stays alive and 2) stays where it is (i.e. $\lim_{t \searrow s} \psi_V = \delta$). It can be seen that the expectation is a functional: it depends on the entire Brownian path between x and y. The 'state' of being at y consists of an expectation

of all the possible ways in which the particle can move to y. Consequently, Feynman's path integral gave rise to the idea that the entire state of the universe could be expressed as a sum over all its possible 'histories'.

But there are several problems. The first is that path integrals can only be calculated exactly very occasionally, and then only for potentials for which the solution is already known through other methods. But a Taylor expansion of the Feynman-Kac exponential can be written down for almost any potential. This Taylor expansion is known as a Born expansion in the physics literature and is discussed by e.g. [45] (p. 128) or [46] (p. 161). Expanding the Feynman-Kac exponential as a Taylor series, we obtain

$$\mathbb{E}_x\left(\delta(B_t-y)\left\{1-\lambda\int_s^t V(B_\tau)d\tau+\frac{1}{2}\lambda^2\left(\int_s^t V(B_\tau)d\tau\right)^2-\frac{1}{6}\lambda^3\left(\int_s^t V(B_\tau)d\tau\right)^3+\ldots\right\}\right)$$

In section 5, using the law of iterated expectations (also known as the tower property), we obtain that this equals

$$\begin{aligned}&B(y,t|x,s)-\lambda\int_{\mathbb{R}^d}d\alpha\int_s^t d\tau B(y,t|\alpha,\tau)V(\alpha)B(\alpha,\tau|x,s)\\&+\lambda^2\int\int_{\tau_2\geq\tau_1}d\tau_2\,d\tau_1\int_{\mathbb{R}^d}d\alpha_2 B(y,t|\alpha_2,\tau_2)V(\alpha_2)\int_{\mathbb{R}^d}d\alpha_1 B(\alpha_2,\tau_2|\alpha_1,\tau_1)V(\alpha_1)B(\alpha_1,\tau_1|x,s)-\ldots,\end{aligned}$$

where B is the free Brownian propagator from above, where the integrations over the intermediate time coordinates appear in a time-ordered way, while the integrations over the intermediate spatial coordinates appear nested within the expression. The motion of the Brownian particle can be tracked by reading from right to left, with an interaction with V in between each set of propagators. While it is customary to pull all the integrations over the intermediate spatial locations towards the front of the expression, this is only allowed for nice potentials. Because we will introduce a potential for which this is not allowed, we will leave the integrations in their nested order; see also [12] (p. 214). We write the full series as

$$\begin{aligned}\psi_V(y,t|x,s)=B(y,t|x,s)+\sum_{i=1}^{\infty}(-\lambda)^i&\left[\int_{s\leq\theta_1\leq\ldots\leq\theta_i\leq t}d\theta_i\ldots\int d\theta_1\right]\\&\times\int_{\mathbb{R}^d}d\alpha_i\,B(y,t|\alpha_i,\theta_i)V(\alpha_i)\left[\prod_{k=1}^{i-1}\int_{\mathbb{R}^d}d\alpha_k\,B(\alpha_{k+1},\theta_{k+1}|\alpha_k,\theta_k)V(\alpha_k)\right]B(\alpha_1,\theta_1|x,s)\end{aligned},\tag{1.6.3}$$

where the spatial integrations automatically appear nested. We also note that for a positive potential all the integrands are positive and therefore the series should converge, if at all, in an alternating manner. The conventional 'Feynman rules' for quantum mechanics state that the zero-order term B counts paths without interactions, the first-order term in the

Taylor expansion counts all paths with exactly one interaction, and the second-order term counts all paths with exactly two interactions, and so on. In Feynman's interpretation, ψ is the sum of all these terms. [45] write (p. 123):

> With this interpretation we can describe [the propagator K_V] in the following way. K_V is, of course, a sum over alternative ways in which the particle may move from point a to point b. The alternatives are: 1. The particle may not be scattered at all, $K_0(b,a)$ 2. The particle may be scattered once, $K^{(1)}(b,a)$ 3. The particle may be scattered twice, $K^{(2)}(b,a)$. Etc. [...] The total amplitude for motion from a to b with any number of scatterings is $K_0 + K^{(1)} + K^{(2)} + ... + K^{(n)} + ...$.

and a similar interpretation can be found in [46] (p. 163). Of course Feynman was dealing with a *complex* wave-function, and the obtained series does *not* converge in an absolute sense, because the integrands are oscillating rather than vanishing at ∞. To investigate convergence, *analytic continuation* is often used: transforming the time variable $t \to -i\,t$. We have already pointed out that this transformation turns the problem into one of Brownian motion, where the positive potential V kills paths at a rate corresponding to its magnitude. The convergence of the Taylor series can be shown, and has been shown. But to date it has not been interpreted as we interpret it here.

For a positive potential, the Taylor expansion convergences in an alternating fashion — rather than in the monotone fashion that is implied by Feynman's interpretation. To explain why this is the case, we suggest a combinatorial Pascal interpretation as follows:

$$\begin{pmatrix} \text{free term} \\ \lambda^1 \text{ term} \\ \lambda^2 \text{ term} \\ \lambda^3 \text{ term} \\ \lambda^4 \text{ term} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 2 & 3 & 4 & \cdots \\ 0 & 0 & 1 & 3 & 6 & \cdots \\ 0 & 0 & 0 & 1 & 4 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix}, \tag{1.6.4}$$

where each λ^i-term is positive and defined by

$$\begin{aligned} \lambda^i\text{-term} = \lambda^i & \left[\int\limits_{s \le \theta_1 \le \ldots \le \theta_i \le t} d\theta_i \ldots \int d\theta_1 \right] \\ & \times \int\limits_{\mathbb{R}^d} d\alpha_i \, B(y,t|\alpha_i,\theta_i) V(\alpha_i) \left[\prod_{k=1}^{i-1} \int\limits_{\mathbb{R}^d} d\alpha_k \, B(\alpha_{k+1},\theta_{k+1}|\alpha_k,\theta_k) V(\alpha_k) \right] B(\alpha_1,\theta_1|x,s). \end{aligned} \tag{1.6.5}$$

Here the upper triangular Pascal matrix has appeared, and the power of λ indicates which term in the Taylor series is meant. From the first row, we can see that the free term B

counts *all* paths from (x, s) to (y, t) — regardless of the number of interactions. The first correction term, linear in λ, picks up a contribution for *every* interaction: it thus counts paths with i interactions i times. The second correction term, which goes with λ^2, counts all possible time-ordered pairs of interactions: it counts paths with i interactions 'i choose 2' times and so on. The matrices should be extended and are infinite in size. Inverting the Pascal matrix immediately gives

$$\begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & \cdots \\ 0 & 1 & -2 & 3 & -4 & \cdots \\ 0 & 0 & 1 & -3 & 6 & \cdots \\ 0 & 0 & 0 & 1 & -4 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} \text{free term} \\ \lambda^1 \text{ term} \\ \lambda^2 \text{ term} \\ \lambda^3 \text{ term} \\ \lambda^4 \text{ term} \\ \vdots \end{pmatrix} \tag{1.6.6}$$

recovering not only the expression in (1.6.3), for all paths with 0 interactions, but obtaining the probability that exactly $i > 0$ interactions occur. The full interpretation is discussed in section 5, where we show that the probability of exactly i interactions equals

$$\psi_{i \text{ interactions}}(y, t|x, s)\Big|_{\lambda=1} = \mathbb{E}_x \left(\delta(B_t - y) \frac{1}{i!} \left[\int_s^t V(B_\tau) d\tau \right]^i e^{-\int_s^t V(B_\tau) d\tau} \right), \tag{1.6.7}$$

and where the substitution $i = 0$ immediately returns the Feynman-Kac formula, providing a new interpretation of the 'Feynman rules' for a diffusion in the presence of an annihilating potential.

Furthermore, the above may be compared with the results for inhomogeneous Poisson processes. A stochastic Poisson process counts the number of events that occur within a given time interval. Events are independent and occur at each time τ with probability $\lambda(\tau)$. It is well-known that the random number N, which counts the number of events in the period $[s, t]$, is distributed as

$$\mathbb{P}\left(N = i\right) = \frac{1}{i!} \left(\int_s^t \lambda(\tau) d\tau \right)^i e^{-\int_s^t \lambda(\tau) d\tau}.$$

The resemblance with the above is clear.

1.7 A new potential

Apart from our inability to calculate path integrals exactly, a further problem is that the treatment of even the simplest boundary value problems is notoriously complicated within the path integral framework. In [47], Janke & Kleinert write

> Considering the present widespread use of path integrals [...], it is surprising how many standard text book problems of quantum mechanics have not been

> solved within this framework. [...] In this note we would like to exhibit the path integration for the particle in a box (infinite square well). While in Schrödinger theory this system has a trivial solution, a careful classification of paths is needed before Feynman's formula can be evaluated.

They then show how to evaluate Feynman's formula for the 1-dimensional particle in a box. They use a procedure that is equivalent to the 'method of images', which is well-known in both electrostatics and probability theory. See [25] for image problems in electrostatics, for example, or [12] (p. 217) for a Brownian 'iterated reflection' argument. No progress, however, has been made in evaluating Feynman's formula for higher dimensional boundary value problems.

The fundamental reason for the complexity of using path integrals for boundary value problems is that path integrals assume the possibility of movement throughout the whole of space. And thus Feynman's screens extended to infinity. Boundary value problems, on the other hand, confine the particle to a particular region of space, but the Gaussian integrals are much easier, at least analytically, if they stretch the whole real line.

It is tempting to postulate an infinite potential outside of the box, the interpretation of which can be twofold. It can be seen as an infinite cooling rate ensuring the temperature outside the box is zero, or as an infinite annihilation rate such that every Brownian path spending even a small time outside the box is annihilated. Let, for example, the potential be unity outside of the domain, i.e. $V(\alpha) = \mathbb{1}_{\alpha \notin D}$, and let $\lambda \to \infty$, so that the annihilation rate outside of the box goes to infinity. As can be seen from the Taylor series in the previous subsection, all correction terms become infinite as $\lambda \to \infty$. Even though the series (1.6.3) *formally* still converges, this obviously diminishes its practicality.

To overcome this problem, it has been suggested to use Dirac δ-function potentials, which are infinite at the edge of the box but zero beyond. In this case the correction terms are all finite and therefore the series converges in a meaningful manner, but now we are faced with a bigger problem: the potential is not strong enough to confine the particle — it can 'tunnel' through the barrier, unless we let $\lambda \to \infty$, in which case the alternating correction terms all become infinite again.

Merging the subjects of path integrals and boundary value problems has thus been difficult. Either the potential does not correspond to the desired physical situation (as it does not contain the particle), or it does, but its perturbation expansion contains terms that are all infinite. It seems impossible to reconcile the two.

But in fact, the Taylor expansion (1.6.3) of the Feynman-Kac functional looks a lot like the single and double boundary layer series. To pursue this analogy further, we define the Green function G_V as the expected time spent from x around y, in the presence of an

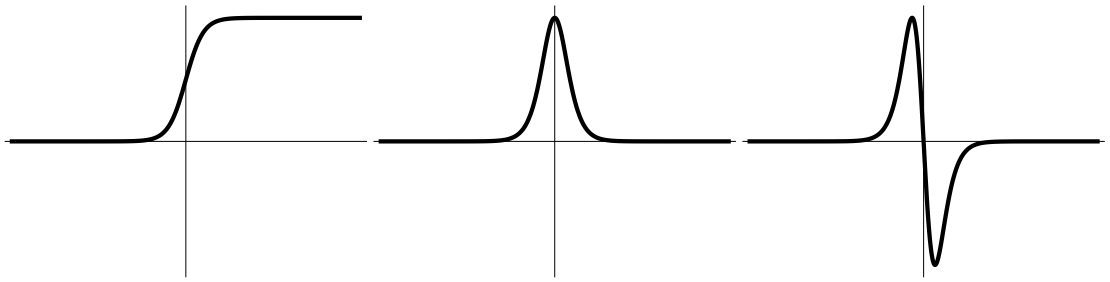


Figure 1. The function $\frac{1}{1+e^{-x/\epsilon}}$ and its first two derivatives. While for any $\epsilon > 0$ the function is continuously differentiable to all orders, for $\epsilon \to 0$ we get $\mathbb{1}_{x>0}$, $\delta(x)$ and $\delta'(x)$.

annihilating potential V:

$$G_V(y,x) := \int_s^\infty \psi_V(y,t|x,s)dt. \tag{1.7.1}$$

Integrating the series expression (1.6.3), we get

$$G_V(y,x) = G_B(y,x) + \sum_{i=1}^{\infty} (-\lambda)^i \int_{\mathbb{R}^d} d\alpha_i \, G_B(y,\alpha_i) V(\alpha_i) \left[\prod_{k=1}^{i-1} \int_{\mathbb{R}^d} d\alpha_k \, G_B(\alpha_{k+1},\alpha_k) V(\alpha_k) \right] G_B(\alpha_1,x), \tag{1.7.2}$$

which should be compared with the absorbed and reflected series (1.5.10) and (1.5.11). It can be seen that the only difference is that the integration there is over the surface of the domain, whereas here the integration is over the whole of space. It is tempting to try to take the potential as some sort of 'differential operator' on the boundary in such a way that each integration over $\mathbb{R}^d$ turns into an integration over ∂D. Would it be possible to choose the potential such that the Feynman-Kac expansion and the boundary layer expansions coincide? The answer is yes.

In section 5 of this paper we show that a Brownian motion that is absorbed or reflected at the boundary of D is consistent with a path integral formulation or Feynman-Kac functional, when the particle (or Brownian motion) is allowed in all of $\mathbb{R}^d$ but is acted upon by a potential V, where the potential is taken to be

$$V(\alpha) := \mp \frac{\sigma^2}{2} \nabla_\alpha^2 \mathbb{1}_{\alpha \in D} \tag{1.7.3}$$

The sign of the potential depends on the boundary condition (absorbing or reflecting), and $\mathbb{1}$ is the indicator function. The indicator function equals 1 if the condition in its subscript is satisfied, and 0 otherwise. The proposed potential can be seen as a generalisation of the one dimensional Dirac δ'-function, which can be defined as the double derivative of the one dimensional step function. Even though derivatives of the step function do not formally

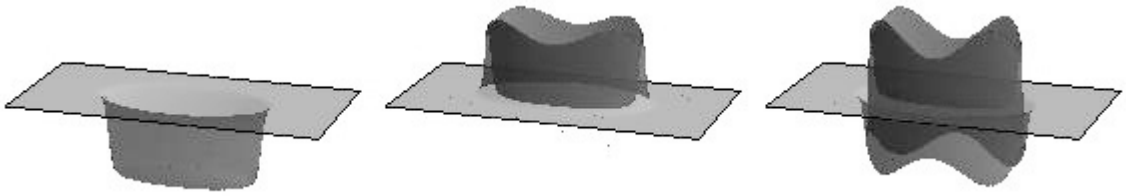


Figure 2. The mollifier $M(r,\phi) := \frac{-1}{1+e^{-(R(\phi)-r)/\epsilon}}$, $n_\phi \cdot \nabla M$ and $\nabla^2 M$. The function $R(\phi)$ is the radius of the ellipse as defined in the text. While for any $\epsilon > 0$ the function is continuously differentiable to all orders, for $\epsilon \to 0$ we get $-\mathbb{1}_{x\in D}$, $-n \cdot \nabla_x \mathbb{1}_{x\in D}$ and $-\nabla_x^2 \mathbb{1}_{x\in D}$.

exist at zero, following the usual rules of partial integration produces the correct answer. In one dimension, for example, we have

$$\int_{-\infty}^{+\infty} \frac{\partial^2 \mathbb{1}_{a<x<b}}{\partial x^2} f(x) dx = \int_{-\infty}^{+\infty} \mathbb{1}_{a<x<b} \frac{\partial^2 f(x)}{\partial x^2} dx = f'(b) - f'(a) \tag{1.7.4}$$

where the integration by parts yields no boundary terms because $\mathbb{1}_{a<x<b}$ and $\partial_x \mathbb{1}_{a<x<b}$ both vanish at infinity. In one dimension we thus obtain a 'sum' of 'outward normal derivatives' at both boundary locations a and b — and we could hypothesise that this sum becomes an integral in higher dimensions. To show that this is indeed the case, we note first that by the divergence theorem we have:

$$\int_{\mathbb{R}^d} dx\, \mathbb{1}_{x\in D} \nabla_x^2 f(x) = \int_D dx\, \nabla_x^2 f(x) = \oint_{\partial D} d\beta\, n_\beta \cdot \nabla_\beta f(\beta). \tag{1.7.5}$$

And secondly, by Green's identity, we get that

$$\int_{\mathbb{R}^d} dx\, \mathbb{1}_{x\in D} \left\{ \overleftarrow{\nabla}_x^2 - \overrightarrow{\nabla}_x^2 \right\} f(x) = \int_{\partial \mathbb{R}^d} dx\, \mathbb{1}_{x\in D} \left\{ \overleftarrow{\partial_x} - \overrightarrow{\partial_x} \right\} f(x) = 0 \tag{1.7.6}$$

where this follows from the fact that $\mathbb{1}_{x\in D}$ as well as $\nabla_x \mathbb{1}_{x\in D}$ are zero when evaluated at the 'boundary' of $\mathbb{R}^d$, which is indicated heuristically as $\partial \mathbb{R}^d$. One may object that the divergence theorem is invalid when the integrand blows up in some parts of the domain, but we may take $\mathbb{1}_{x\in D}$ to be a 'bump function'. A bump function equals 1 on D, falls off to 0 outside of D, and does so arbitrarily rapidly while still being smooth. With this 'smooth' interpretation of the indicator function, the use of the divergence theorem can be justified. Combining (1.7.5) and (1.7.6), we conclude that:

$$\int_{\mathbb{R}^d} dx\, \nabla_x^2 \mathbb{1}_{x\in D} f(x) = \oint_{\partial D} d\beta\, n_\beta \cdot \nabla_\beta f(\beta) \tag{1.7.7}$$

and thus we see that — while in one dimension the potential $\frac{\partial^2 \mathbb{1}_{a<x<b}}{\partial x^2}$ produces a sum of outward normal derivatives at a and b — in higher dimensions the potential $\nabla_x^2 \mathbb{1}_{x\in D}$ produces an integral over the outward normal derivatives over the boundary.

Although the definition is acceptable through differentiations of the step function, it is not very helpful for visualisation. For visualisation it is easier to think of a Dirac δ-function as the limit of the middle graph in Figure 1.

A smooth approximation to the step function is equally possible in higher dimensions. In Figure 2 we can see smooth approximations of $-\mathbb{1}_{x\in D}$, $-n\cdot\nabla_x\mathbb{1}_{x\in D}$ and $-\nabla_x^2\mathbb{1}_{x\in D}$, where D is taken to be a two dimensional ellipse. The ellipse is defined by providing the radius of the boundary as a function of the polar angle. In this case we have $R(\phi) = ab/\sqrt{a^2Sin(\phi)^2 + b^2Cos(\phi)^2}$ with a and b half the major and minor diameters. We see that a smooth approximation of $-\nabla_x^2\mathbb{1}_{x\in D}$ has two peaks when crossing the boundary, just as a smooth approximation of Dirac δ' has two peaks. In the one dimensional case the potential looks like a 'heartbeat', but the two dimensional landscape resembles something like a castle with a moat in front of the castle walls. In the proper limit, the castle wall and moat become infinitely high and deep — and narrow.

In terms of *why* the potential as discussed does the job, we can say the following. The second derivative of the step function is more divergent than the first and therefore the potential *is* strong enough to contain the particle. But we have also noted that positive potentials destroy paths while negative potentials create paths. Through the one dimensional analogy, we see that the Laplacian of the Heaviside step-function, $\mp\nabla_x^2\mathbb{1}_{x\in D}$, is equally positive and negative. As a result, the proposed potential conserves *particle number*. If a particle reaches the boundary of the domain, it is both copied (by the negative peak) and destroyed (by the positive peak). But these actions happen at slightly different places. If the copying happens just inside the domain, and the destroying just outside, then the boundary is reflecting from the inside: every time it hits the boundary it is destroyed just outside the domain and put back just on the inside. But if the destroying happens just inside the domain while the copying happens just outside, then the particle can get out but it can never get back in. Seen from the inside, therefore, the boundary acts as an absorbing barrier. This intuition explains why the potential for the Dirichlet and Neumann problems differ only by a sign: $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$ is reflecting from one side, and absorbing from the other. In one dimension this can easily be verified (see subsection 1.8).

We conclude that the $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$ potential manages to both replicate the desired physical situation (namely reflect or absorb), while also allowing for an easily interpretable and computable perturbation series. This function has never been defined before, to the author's best knowledge. We can make sense of this seemingly ill-defined function either by 1) a limiting procedure, or by 2) using partial integrations (or Green's theorem) as if

everything is well-behaved. We conclude that the following problems are equivalent:

$$\left.\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) A(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) A(y,t|x,s) &= 0 \\
A(\beta,t|x,s) &= 0 \\
A(y,t|\beta,s) &= 0 \\
\lim_{s\nearrow t} A(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} A(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2 + V(y)\right)\psi_V(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2 - V(x)\right)\psi_V(y,t|x,s) &= 0 \\
-\tfrac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D} &= V(\alpha) \\
\\
\lim_{s\nearrow t}\psi_V(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s}\psi_V(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right. \tag{1.7.8}$$

The left-hand problem is defined for x and y in the interior of D and all regular boundary points β. The right-hand problem is defined for all x and y in $\mathbb{R}^d$. But the claim is that

$$A(y,t|x,s) = \psi_V(y,t|x,s)\ \forall x,y\in D, \tag{1.7.9}$$

and in particular

$$A(y,t|x,s) = \mathbb{E}_x\left(\delta(B_t - y)\ e^{\frac{\sigma^2}{2}\int_s^t \nabla_u^2 \mathbb{1}_{u\in D}(B_\tau)d\tau}\right). \tag{1.7.10}$$

Similarly, we conclude for the reflected transition density R that

$$\left.\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) R(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) R(y,t|x,s) &= 0 \\
n_\beta\cdot\overrightarrow{\nabla}_\beta R(\beta,t|x,s) &= 0 \\
R(y,t|\beta,s)\overleftarrow{\nabla}_\beta\cdot n_\beta &= 0 \\
\lim_{s\nearrow t} R(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} R(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2 + V(y)\right)\psi_V(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2 - V(x)\right)\psi_V(y,t|x,s) &= 0 \\
\tfrac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D} &= V(\alpha) \\
\\
\lim_{s\nearrow t}\psi_V(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s}\psi_V(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right. \tag{1.7.11}$$

where again the problem on the left-hand problem is defined for x and y in the interior of D and all regular boundary points β, while the right-hand problem is defined for all x and y in $\mathbb{R}^d$. The claim is that

$$R(y,t|x,s) = \psi_V(y,t|x,s)\ \forall x,y\in D, \tag{1.7.12}$$

and in particular

$$R(y,t|x,s) = \mathbb{E}_x\left(\delta(B_t - y)\ e^{-\frac{\sigma^2}{2}\int_s^t \nabla_u^2 \mathbb{1}_{u\in D}(B_\tau)d\tau}\right). \tag{1.7.13}$$

While we admit that the path integral cannot be calculated exactly, and that the potential $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$ looks ill defined, at least we can say we have obtained a compact *short-hand* for the expansion of the Green function — just as a generating function, for example, can produce all the Legendre polynomials. In their book on random walks and path integrals, [48] write

> a clear indication of one of the advantages of the generating function is that it represents a prescription for the construction of the special function that it generates. [...] In this sense, the generating function encapsulates all information with regard to the function that it generates. Furthermore, it contains this information in an extremely compact form.

We conclude that our expression does exactly the above — for the transition density of absorbed or reflected Brownian motion. In the words of theoretical physicist and chemist Gibbs, in [49] (p. 10):

> One of the principal objects of theoretical research in my department of knowledge is to find the point of view from which the subject appears in its greatest simplicity.

In this view, the main contribution of this paper is that it provides a solution for the heat kernel with boundary conditions — and by extension for the (modified) Dirichlet problem, as pioneered by [3] — that is 1) new and 2) very compact. If one were to communicate the solution in the least possible number of bits, then this would be a good candidate.

Another attractive feature of our solution is that it unifies the treatment of the Dirichlet and Neumann problems, and that it combines the fields of potential theory and that of path integrals for the first time.

1.8 A semi-permeable boundary in one dimension

We would like to examine the potential V in more detail, and in one spatial dimension. Consider a Brownian motion in one dimension that is started at $x > 0$. The level zero is assumed to be *transparent from above* but *reflecting from below*. Above the boundary, i.e. for $y > 0$, the density ψ is equal to the absorbed density of a halfspace, i.e. $B(y,t|x,s) - B(y,t|x^*,s)$, since particles that enter the region $y < 0$ can never return. Here x^* indicates the mirror coordinate, i.e. $x^* = -x$. Below the boundary, the density ψ is equal to all those paths that cross the boundary and are then reflected from below. It turns out that this density is equal to $2B(y,t|x,s)$. The resulting density ψ, on the whole real line, is therefore given by:

$$\psi(y,t|x,s) = \begin{cases} B(y,t|x,s) - B(y,t|x^*,s) & \text{if } y > 0, \\ 2B(y,t|x,s) & \text{if } y < 0. \end{cases} \tag{1.8.1}$$

It is consistent with our interpretation of a semi-permeable barrier that we have

$$\int_{-\infty}^{\infty} \psi(y,t|x,s)\, dy = 1 \quad \forall t \geq s. \tag{1.8.2}$$

Furthermore, it is clear that the density is discontinuous across the boundary by $2B(y,t|x,s)$. The derivatives on both sides of the boundary, however, are equal. This can be physically

understood by considering the derivative as the expected 'net flow' of particles. The net flow just above the transparent boundary is downward; as there can be no upward flow coming from an absorbing boundary. The downward flow consists of all those particles that will instantaneously cross the boundary. The flow just below the boundary consists of 1) all particles crossing the boundary from above, and 2) the flow of all particles that are reflecting off the boundary from below. Reflected particles, however, do not contribute to the *net* flow just below the boundary. As a result, the net flow on either side of the boundary consists of only those particles that are crossing the boundary in the downward direction, for the first time. We have fixed $x > 0$, and thus we can only move y around, and we see that ψ satisfies

$$\begin{aligned} \Big(\frac{\sigma^2}{2}\frac{\partial^2}{\partial y^2} - \frac{\partial}{\partial t} + \frac{\sigma^2}{2}\delta'(y)\Big)\psi(y,t|x,s) &= 0 \\ \lim_{y\to\pm\infty}\psi(y,t|x,s) &= 0 \\ \lim_{t\searrow s}\psi(y,t|x,s) &= \delta(y-x) \end{aligned} \tag{1.8.3}$$

with the emergence of a Dirac δ'-potential as promised. Comparing with (5.1.1), we see that the potential can be identified as

$$V(x) = -\frac{\sigma^2}{2}\delta'(x).$$

It is crucial, for this result, that the derivatives just above and below the boundary are equal. A discontinuity in the derivative would have produced a Dirac δ-function in the PDE. In addition, it is crucial that the pre-factor of the Laplacian matches the pre-factor of the Dirac δ'-function, i.e. both are $\sigma^2/2$. Now it is *also* obvious that if we switch the sign of the potential, that the orientation of the boundary then changes: it becomes reflecting from above and absorbing from below. The resulting density ψ on the real line now equals

$$\psi(y,t|x,s) = \begin{cases} B(y,t|x,s) + B(y,t|x^*,s) & \text{if } y > 0, \\ 0 & \text{if } y < 0, \end{cases} \tag{1.8.4}$$

because there is no way that the particle can reach $y < 0$, if the boundary is reflecting from above and we have taken $x > 0$. Again ψ is discontinuous across the boundary by $2B(y,t|x,s)$. Furthermore, the derivative of ψ is zero on both sides of the boundary. It is not hard to check that ψ satisfies

$$\begin{aligned} \Big(\frac{\sigma^2}{2}\frac{\partial^2}{\partial y^2} - \frac{\partial}{\partial t} - \frac{\sigma^2}{2}\delta'(y)\Big)\psi(y,t|x,s) &= 0 \\ \lim_{y\to\pm\infty}\psi(y,t|x,s) &= 0 \\ \lim_{t\searrow s}\psi(y,t|x,s) &= \delta(y-x) \end{aligned} \tag{1.8.5}$$

and thus the potential equals

$$V(x) = \frac{\sigma^2}{2}\delta'(x).$$

Thus we find, indeed, that the orientation of the boundary changes when the potential changes sign.

With the benefit of intuition and hindsight, it was relatively simple to *propose* a solution ψ and to *verify* that it satisfies a PDE with a Dirac δ'-potential. To obtain a similar result in d dimensions, we may generalise the Dirac δ'-function from $\partial_x^2 \mathbb{1}_{x>0}$ to $\nabla_x^2 \mathbb{1}_{x\in D}$. In either case, a distributional definition can be used to make sense of a seemingly ill-defined quantity and, while this may sound complicated, it means nothing more than performing two integrations by parts under the integral sign.

1.9 The acceleration of the occupation time

Finally, we note that the proposed potential connects the study of boundary value problems to the study of Brownian *occupation times*. The occupation time is a random variable that measures the amount of time spent by a Brownian motion in a certain region. For the domain D we have the occupation time as follows:

$$\text{occupation time} = \int_s^t \mathbb{1}_{B_\tau \in D}\, d\tau. \tag{1.9.1}$$

Occupation times are discussed at length in Geman and Horowitz's (1980) review paper [50]. The occupation time is a well-behaved random variable that cannot exceed the length of the 'observation interval' from s to t, as defined by the limits of the integral.

In [51], Lévy introduced the concept of *local time*, which can be seen as the derivative of the occupation time with respect to the spatial variable. If we consider another domain, including D, but everywhere ϵ larger than D in the outward normal direction, then for every path we must have that the occupation time of the larger domain exceeds that of the smaller domain. If we calculate the difference, divide by ϵ and let ϵ go to zero, then we obtain a non-trivial limit which is defined as the local time at the boundary.

In one dimension, and perhaps surprisingly, the local time at zero during $[s, t]$ may exceed the duration of the observation interval, which is $(t - s)$. This is because Lévy's local time is the spatial derivative with respect to the occupation time. The local time is really defined as how much more time is spent above level a as opposed to above level $a-\epsilon$, divided by ϵ. Thus it has the units of time over length and it can exceed the length of the observation interval.

It is well known that the occupation time is absolutely continuous in both space and time. It seems intuitive that the random occupation time of domain D in the interval $[s, t + dt]$ cannot exceed the occupation time in $[s, t]$ by more than dt. As far as the spatial variable is concerned, continuity may be somewhat expected since the occupation

time of the domain larger by ϵ can exceed that of the smaller domain only by a little. Differentiability, however, is less obvious. In one dimension the occupation time above level a is defined as

$$\text{occupation time} = \int_s^t \mathbb{1}_{B_\tau > a}\, d\tau, \tag{1.9.2}$$

and its derivative with respect to a provides the local time at a as

$$\text{Lévy's local time} = \int_s^t \delta(B_\tau - a)\, d\tau, \tag{1.9.3}$$

where again a 'smooth' interpretation of the indicator is needed to be able to differentiate it, along with a proper limiting procedure. The occupation time cannot exceed $(t-s)$, but Lévy's local time *can* exceed $(t-s)$. The potential introduced in this paper corresponds to the *second derivative* of the occupation time with respect to the domain, or the 'acceleration' of the occupation time, as the domain grows bigger:

$$\text{acceleration of the occupation time} = \int_s^t \delta'(B_\tau - a)\, d\tau. \tag{1.9.4}$$

What we would have *liked* to find in the literature is that the local time at ∂D is almost surely differentiable, so that

$$\int_s^t \frac{\sigma^2}{2} \nabla_u^2 \mathbb{1}_{u\in D}(B_\tau)\, d\tau$$

exists, i.e. is almost surely finite. If it is almost surely finite, then so is

$$\mathbb{E}_x \exp\Big[\pm \frac{\sigma^2}{2} \int_s^t \nabla_u^2 \mathbb{1}_{u\in D}(B_\tau)\, d\tau \Big]$$

and if the latter exists then it can be found by its series expansion — and thus we have provided the proof that its Taylor series convergences. The proof of convergence has therefore been reduced to the question whether or not the acceleration of the occupation time exists.

Unfortunately, we have not been able to find in the literature a version of the occupation time that is twice differentiable in the spatial variable, so this issue must be left for now. Although we would have liked to provide our own proof of the convergence of the single and double boundary layers, this seems impossible for the time being. Instead, we will proceed as follows:

- Section 2 discusses the notation to be used.
- Section 3 introduces absorbed and reflected Brownian motion, and derives the first- and last-passage (reflection) decompositions that result in single and double boundary layer series. It shows that the single and double boundary layer series are equal, term by term, and that they are valid not only for smooth domains but for piecewise smooth domains as well. The last subsections discuss spectral theory and the Dirichlet and Neumann problems, and can be skipped without loss of continuity.

- Section 4 gives examples. It demonstrates both the absorbed and reflected series solutions and its proposed alternating/monotone convergence for convex/concave domains: for a two dimensional ellipse, cusp and three dimensional ellipsoid.

- Section 5 introduces the 'potential' in a probabilistic context: as destroying and creating paths. We find that the perturbation series converges in an alternating or monotone fashion, depending on whether the potential is positive or negative, respectively. We contrast this with the usual Feynman rules and present a new combinatorial interpretation based on the Pascal matrix. We show that if the potential is chosen to be $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$, the Feynman-Kac formula then produces the first- and last-passage series of section 3. We thus conclude, for the first time, that boundary value problems can be transformed into potential problems, and we show that the Dirichlet and Neumann boundary conditions are even more closely related than previously thought (the potential differs only by a sign).

- Section 6 concludes.

2 Notation

2.1 The domain *D*

The number of spatial coordinates is indicated by dimension d. The starting space-time coordinate, for any process, is always (x, s). The final space-time coordinate, for any process, is always (y, t). Unless otherwise specified, we assume $t \geq s$. An open, static domain is indicated by D. D is assumed to be *piecewise* smooth, or, more generally, any domain D is allowed for which the divergence theorem or Green's theorem holds; i.e. domains with a finite number of edges, corners and cusps. The domain may be bounded or unbounded. By definition, the starting point (x,s) of the process is in the interior. Unless otherwise specified, we have both spatial coordinates in the interior, i.e. $x, y \in D$. The boundary of the open domain is indicated by ∂D. The closure of D is indicated by $\bar{D}$. When $y \notin \bar{D}$, then y is in the exterior. When $y \in \partial D$, then y is on the boundary.

2.2 Stochastic processes

The process B_t is a d-dimensional (free) Brownian motion: it passes through any boundaries unaffected. We have $B_t = \{B_{1,t}, ..., B_{d,t}\}$, and similarly for all the other processes. The process A_t is a d-dimensional absorbed Brownian motion (ABM): it is absorbed at the boundary ∂D. The process R_t is a d-dimensional reflected Brownian motion (RBM): it is reflected in the direction normal to the boundary at ∂D. The free Brownian transition-density to move from (x,s) to space-time point (y,t) is indicated by $B(y, t|x, s)$. The absorbed transition-density is indicated by $A(y,t|x,s)$. The reflected transition-density is indicated by $R(y,t|x,s)$. In any transition-density, the left-most coordinate is referred to as the forward coordinate, and the right-most coordinate is referred to as the backward coordinate.

2.3 Intermediate coordinates

For intermediate space-time coordinates we will always use Greek coordinates, in particular (α, τ) or (β, τ) or (γ, τ). For the intermediate time τ it holds that $s \leq \tau \leq t$. The intermediate spatial coordinate α represents an interior coordinate, i.e. $\alpha \in D$, and the spatial coordinates β and γ represent boundary coordinates, i.e. $\beta,\gamma \in \partial D$. Often α, β and γ will be integrated over. Integrations over the interior and over the boundary are denoted as $\int\limits_D d\alpha$ and $\oint\limits_{\partial D} d\beta$ or $\oint\limits_{\partial D} d\gamma$. While we usually assume that the domain D is finite, we will write integrations over the surface of D as $\oint\limits_{\partial D}$ even when D is infinite. (The boundary can be imagined as extending to infinity.)

When there are multiple intermediate times, then we use θ_i and we will have a time-ordering such that $\theta_1 \leq \cdots \leq \theta_n$. In all integrands we will write earlier times towards the

right, so that the motion of the particle can be tracked by reading equations from right to left.

2.4 Differentiation

The gradient with respect to some internal coordinate α is denoted by ∇_α. The Laplacian with respect to the same coordinate is denoted by ∇^2_α. Both differential operators can work towards their right or left, as indicated by the arrow, i.e. $\overrightarrow{\nabla}_\alpha$ differentiates what is towards its right. The outward normal at some boundary coordinate β is denoted by n_β. The *boundary divergence* is defined by 1) taking the gradient with respect to some interior coordinate α, 2) taking the dot product with the boundary normal of the nearest boundary location β, and 3) moving the interior boundary coordinate to that location β, and 4) multiplying by $-\sigma^2$. This sequence of actions is denoted by $-\sigma^2 \lim_{\alpha\to\beta} n_\beta \cdot \nabla_\alpha f(\alpha)$. We will use the shorthand $\partial_\beta f(\beta)$ for this. Thus

$$\begin{aligned} \overrightarrow{\partial_\beta} f(\beta,\gamma) &:= -\sigma^2 \lim_{\alpha\to\beta} n_\beta \cdot \overrightarrow{\nabla}_\alpha f(\alpha,\gamma), \\ f(\gamma,\beta)\overleftarrow{\partial_\beta} &:= -\sigma^2 \lim_{\alpha\to\beta} n_\beta \cdot \overrightarrow{\nabla}_\alpha f(\gamma,\alpha), \\ \overleftrightarrow{\partial_\beta} &:= \overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}. \end{aligned} \tag{2.4.1}$$

Taking limits does not in general commute with integration. As a result, the operator ∂_β does not in general commute with integration over the boundary.

2.5 First- and last-passage times

With the convention that $\inf\{\varnothing\} = \infty$, we have for the first-passage time

$$\tau^{\mathrm{FP}}(t|x,s) = \inf_\tau \left\{ s \le \tau \le t : B_\tau \notin D | B_s = x \right\}. \tag{2.5.1}$$

With the convention that $\sup\{\varnothing\} = -\infty$, we have for the last-passage time

$$\tau^{\mathrm{LP}}(t|x,s) = \begin{cases} \sup_\tau \left\{ s \le \tau \le t : B_\tau \notin D | B_s = x \right\} & \text{if } B_t \in \bar{D}, \\ \sup_\tau \left\{ s \le \tau \le t : B_\tau \in \bar{D} | B_s = x \right\} & \text{if } B_t \notin D. \end{cases} \tag{2.5.2}$$

2.6 Expectations and probabilities

The indicator function of the event A is denoted by $\mathbb{1}_A$, and equals 1 if the event A happens, and zero when it does not. Expectations are denoted by $\mathbb{E}$ and probabilities by $\mathbb{P}$, so we have that $\mathbb{E}(\mathbb{1}_A)$=$\mathbb{P}$(A). Expectations or probabilities that are conditional on another event are denoted by a subscript or absolute bar, i.e. $\mathbb{P}_x(A) = \mathbb{P}(A|B_s = x)$. For example we have that

$$B(y,t|x,s) = \mathbb{P}(B_t \in dy | B_s = x) = \mathbb{E}_x \delta(B_t - y),$$

and

$$\begin{aligned} A(y,t|x,s) &= \mathbb{P}(B_t = y;\, \tau^{\mathrm{FP}} > t | B_s = x), \\ A(y,t|x,s) &= \mathbb{P}(B_t = y;\, \tau^{\mathrm{LP}} < s | B_s = x), \end{aligned} \tag{2.6.1}$$

where the semicolon is used to indicate a joint probability.

2.7 Green functions

The Green functions corresponding to different stochastic processes are defined as follows. For a free Brownian motion we have $G_B(y,x) := \int_0^\infty B(y,\tau|x,0)\,d\tau$. For absorbed Brownian motion, we have $G_A(y,x) := \int_0^\infty A(y,\tau|x,0)\,d\tau$. For reflected Brownian motion, we have $G_R(y,x) := \int_0^\infty R(y,\tau|x,0)\,d\tau$.

3 Absorbed and reflected Brownian motion

In this section we will discuss absorbed and reflected Brownian motion and their transition densities. In subsections 3.1 and 3.2 we will discuss absorbed Brownian motion and its first- and last-passage decompositions. In subsections 3.3 and 3.4 we will discuss reflected Brownian motion and its first- and last-reflection decompositions. Subsection 3.5 will introduce two lemmas that may not appear very insightful at first, but in fact they are crucial in subsection 3.6. Subsection 3.6 discuss the tangent plane (TP) decompositions for both processes. Subsection 3.7 derives series solutions for both processes and discusses the equivalence of the first- and last-passage (reflection) series, thereby also proving that the single and double boundary layers are equivalent. Subsection 3.8 shows the intuition for the series solution in 1 dimension. Subsection 3.9 derives new integral equations for A and R in anticipation of results in section 5. Subsection 3.10 derives a new integral equation for the absorbed and reflected Green functions, and discusses possible extensions to spectral theory. Subsection 3.11 discusses the application to the modified Dirichlet and Neumann problems. Subsections 3.8, 3.10 and 3.11 may be skipped without loss of continuity.

3.1 Absorbed Brownian Motion

The transition density of absorbed Brownian motion (ABM) is indicated by $A(y,t|x,s)$, with forward and backward space-time coordinates (y,t) and (x,s). The absorbed transition density $A(y,t|x,s)$ satisfies the following set of equations:

$$\begin{aligned}
\text{forward PDE} \quad & \left(\frac{\partial}{\partial t} - \frac{\sigma^2}{2}\nabla_y^2\right) A(y,t|x,s) = 0, \\
\text{backward PDE} \quad & \left(\frac{\partial}{\partial s} + \frac{\sigma^2}{2}\nabla_x^2\right) A(y,t|x,s) = 0, \\
\text{forward BC} \quad & A(\beta,t|x,s) = 0, \\
\text{backward BC} \quad & A(y,t|\beta,s) = 0, \\
\text{forward STC} \quad & \lim_{s\nearrow t} A(y,t|x,s) = \delta(|y-x|), \\
\text{backward STC} \quad & \lim_{t\searrow s} A(y,t|x,s) = \delta(|y-x|).
\end{aligned} \tag{3.1.1}$$

This holds for all $x, y \in D$ and all regular (i.e. non-singular) boundary points β. PDE stands for 'partial differential equation', BC stands for 'boundary condition' and STC stands for 'short-time condition'. It can be proved that the absorbed transition density 1) exists, 2) is unique and 3) is determined by the above conditions. See for example [52] or [30]. The definition of a 'regular' boundary point is such that a Brownian path started there leaves the domain immediately with probability one, as in [10], p. 245. Existence

of the absorbed transition density is also discussed in [12], p. 79. The PDEs are satisfied because the transition density is unbiased, i.e.

$$A(y,t|x,s) = \mathbb{E}A(y-dB,t-dt|x,s)$$
$$A(y,t|x,s) = \mathbb{E}A(y,t|x+dB,s+ds)$$

and using Itô's lemma (1.4.2) gives both PDEs. The BCs are satisfied because no Brownian particle can move to or from a regular boundary point without being absorbed, and the STCs are satisfied for x and y in the interior because in the short-time limit the absorbed transition density must behave like the free transition density. Furthermore we know that the Brownian particle must be somewhere at every intermediate time, and therefore we also have the Chapman-Kolmogorov equation

$$\text{Chapman-Kolmogorov } A(y,t|x,s) = \int_D d\alpha\, A(y,t|\alpha,\tau)A(\alpha,\tau|x,s) \tag{3.1.2}$$

for any $s \leq \tau \leq t$, and where the STCs ensure that the Chapman-Kolmogorov equation also holds in the limit where τ goes to s or t. See for example [14], p. 36. The Green function associated with ABM is defined by

$$G_A(y,x) := \int_s^\infty A(y,t|x,s)\,dt \tag{3.1.3}$$

and satisfies

$$\begin{gathered}\frac{\sigma^2}{2}\nabla_y^2 G_A(y,x) = \frac{\sigma^2}{2}\nabla_x^2 G_A(y,x) = -\delta(|y-x|) \\ G_A(\beta,x) = G_A(y,\beta) = 0\end{gathered} \tag{3.1.4}$$

for all x and y in the interior and for all regular boundary coordinates β. Existence of the Green function was discussed in subsection 1.4. Because paths are absorbed at the boundary ∂D, the density of all paths that are 'alive' is decreasing. The probability that the first passage occurs at time τ is equal to the 'proportion' of paths that disappear at time τ. Therefore

$$\begin{aligned}\mathbb{P}\left(\tau^{\text{FP}} \in d\tau | B_s = x\right) &= -\frac{\partial}{\partial\tau}\int_D d\alpha\, A(\alpha,\tau|x,s) \\ &= -\int_D d\alpha\, \frac{\sigma^2}{2}\nabla_\alpha^2 A(\alpha,t|x,s) \\ &= -\oint_{\partial D} d\beta\, \frac{\sigma^2}{2} n_\beta \cdot \nabla_\beta A(\beta,t|x,s) \\ &= \frac{1}{2}\oint_{\partial D} d\beta\, \overrightarrow{\partial}_\beta A(\beta,t|x,s)\end{aligned}$$

where n is the outward normal and where $\overrightarrow{\partial}_\beta$ is the scaled inward normal derivative as defined in (2.4.1). It is a positive operator when working on the absorbed density A,

because A is zero on the boundary but positive in the interior. Because probability can only disappear at the boundary, we have that the joint probability for the first-passage time and first-passage location is

$$\mathbb{P}\left(\tau^{\text{FP}} \in d\tau;\, B_{\tau^{\text{FP}}} \in d\beta | B_s = x\right) = \frac{1}{2}\overrightarrow{\partial}_\beta A(\beta, \tau | x, s) \tag{3.1.5}$$

at any regular boundary coordinate β. The simplest boundary value problems occur when the domain considered is a halfspace. By the 'André reflection principle' as in [9] (p. 42), [10] (p. 79) or [15] (p. 26), we obtain that the absorbed density for a halfspace equals

$$A^{\text{HS}}(y, t | x, s) = B(y, t | x, s) - B(y, t | x^*, s), \tag{3.1.6}$$

where x^* equals the 'mirror-coordinate' that is obtained by taking a mirror image of x in the absorbing hyperplane. It is easily checked that A^{HS} satisfies the PDEs, the BCs and the STCs, and thus by uniqueness it must be correct. As far as the STCs are concerned, two δ-functions are obtained: at both x and x^*, but the δ-function at x^* is outside of the space of interest and therefore irrelevant. For the joint distribution of the first-passage time and location, we have

$$\begin{aligned} \mathbb{P}(\tau^{\text{FP}}_{\text{over HS}} \in d\tau;\, B_{\tau^{\text{FP}}_{\text{over HS}}} \in d\beta | B_s = x) &= \tfrac{1}{2}\overrightarrow{\partial}_\beta A^{\text{HS}}(\beta, \tau | x, s) \\ &= \overrightarrow{\partial}_\beta B(\beta, \tau | x, s) \end{aligned} \tag{3.1.7}$$

using that for every boundary coordinate β of a halfspace we have:

$$\begin{aligned} n_\beta \cdot (\beta - x) &= -n_\beta \cdot (\beta - x^*), \\ |\beta - x| &= |\beta - x^*|. \end{aligned}$$

The fact that the first-passage density over a halfspace equals $\overrightarrow{\partial}_\beta B$ — i.e. without a factor $\frac{1}{2}$ — will be important later.

3.2 First- and last-passage decompositions

The original research of this section starts here. For ABM, consider once more the Kolmogorov-Chapman equation, saying that the particle must be somewhere at any intermediate time τ. We have

$$A(y, t | x, s) = \int_D d\alpha\, A(y, t | \alpha, \tau) A(\alpha, \tau | x, s).$$

This equation tells us that for a particle to survive up to time t, it must first survive up to time τ, and then it must also survive from time τ up to time t — and thus both propagators on the right-hand side are absorbed propagators A. Now consider instead the follow quantity

$$\int_D d\alpha\, A(y, t | \alpha, \tau) B(\alpha, \tau | x, s).$$

It is obvious that this does not equal the absorbed density from (x, s) to (y, t), since the absorbing constraint is not imposed during the time interval from s to τ. Instead, the propagation from (x, s) up to time τ is a free propagation, with the only condition that the location α is actually in the interior of D. Although that condition is not enforced explicitly, it is implied by the domain of integration over the coordinate α. Quickly one realises that not only are paths counted that stay in the domain D for their entire duration, but also ones that violate the absorbing boundary condition, if they violate it *before* time τ. In particular, if passages over the boundary occurred, then the last passage over the boundary must have occurred before time τ. Therefore we propose that

$$\mathbb{P}\left(B_t \in dy;\, \tau^{\text{LP}} \leq \tau | B_s = x\right) = \int_D d\alpha\, A(y, t|\alpha, \tau) B(\alpha, \tau|x, s),$$

where the semicolon indicates a joint probability, and where

$$\tau^{\text{LP}}(t|x, s) = \begin{cases} \sup\limits_{\tau} \left\{ s \leq \tau \leq t : B_\tau \notin D | B_s = x \right\} & \text{if } B_t \in \bar{D}, \\ \sup\limits_{\tau} \left\{ s \leq \tau \leq t : B_\tau \in \bar{D} | B_s = x \right\} & \text{if } B_t \notin D, \end{cases} \tag{3.2.1}$$

with conventional (but crucial!) addition that $\sup \varnothing = -\infty$. Thus the event $\tau^{\text{LP}} \leq \tau$ counts paths with a last-passage before time τ as well as paths with no passages at all. The propagator B on the right-hand side *allows* passages before time τ while not *requiring* them. It follows that

$$\mathbb{P}(B_t \in dy;\, \tau^{\text{LP}} \in d\tau | B_s = x) = \frac{\partial}{\partial \tau} \int_D d\alpha\ A(y, t|\alpha, \tau) B(\alpha, \tau|x, s).$$

The absorbed density requires that no passages occur; first nor last passages. Therefore we subtract from the free density all paths that have had a last passage, i.e.

$$\text{LP}\ A(y, t|x, s) = B(y, t|x, s) - \int_s^t d\tau\, \frac{\partial}{\partial \tau} \int_D d\alpha\, A(y, t|\alpha, \tau) B(\alpha, \tau|x, s).$$

This we call the last-passage decomposition, or LP decomposition, for the absorbed density. Upon reflection, the reader may argue that the equation above is an *identity* that holds by virtue of the fundamental theorem of calculus and the STCs satisfied by B and A. We can show very explicitly that the identity holds, by using the fundamental theorem of calculus to obtain

$$\begin{aligned} A(y, t|x, s) &= B(y, t|x, s) - \int_s^t d\tau\, \frac{\partial}{\partial \tau} \int_D d\alpha\ A(y, t|\alpha, \tau) B(\alpha, \tau|x, s) \\ &= B(y, t|x, s) - \left(\lim_{\tau \nearrow t} - \lim_{\tau \searrow s} \right) \int_D d\alpha\ A(y, t|\alpha, \tau) B(\alpha, \tau|x, s) \\ &= B(y, t|x, s) - \int_D d\alpha\ \delta(y - \alpha) B(\alpha, t|x, s) + \int_D d\alpha\ A(y, t|\alpha, s) \delta(\alpha - x) \\ &= B(y, t|x, s) - B(y, t|x, s) + A(y, t|x, s). \end{aligned} \tag{3.2.2}$$

The last line follows as long as both δ-functions pick up a contribution. The requirement that both δ-functions pick up a contribution is equivalent to requiring that x and y are in the interior (i.e. not on the boundary). We may conclude that what we call the LP decomposition is actually more of an identity, rather than a result. When we say that something is an 'identity' we will mean that it is (almost) trivially true, and follows from at most 1 property of the transition densities; in this case from the STCs. In that sense the LP decomposition is an identity.

If we take for granted that ABM exists, even when the boundary is only piecewise smooth, and that its density is continuous up to the boundary — at least at smooth boundary points — then we only need to establish that certain relationships hold in the *interior* of the domain. Any relationship valid in the interior will also hold in the limit, near the boundary. But we should be a little bit careful, however, because taking the limit where one of the coordinates goes to the boundary is *not* the same as plugging in a boundary coordinate in the equation, for either x or y.

To demonstrate this we can try plugging in a boundary coordinate β into the x-position of the LP decomposition. If we had that x equals β, then we would like the left-hand side to be zero. For the right-hand side we get

$$0 = B(y,t|\beta,s) - \int_D d\alpha\, \delta(y-\alpha) B(\alpha,t|\beta,s) + \int_D d\alpha\, A(y,t|\alpha,s)\delta(\alpha-\beta).$$

Now we may use that a smooth part of boundary, in close-up, looks like a hyperplane, and therefore a δ-function located on a smooth part of the boundary picks up only *half* of the contribution that it otherwise would. This leads to

$$0 = \frac{1}{2} A(y,t|\beta,s)$$

which enforces — or is consistent — with the assumption that A is zero on the boundary. For y on the boundary, however, things do not work out so nicely, and we get

$$0 = B(\beta,t|x,s) - \int_D d\alpha\, \delta(\beta-\alpha) B(\alpha,t|x,s) + \int_D d\alpha\, A(\beta,t|\alpha,s)\delta(\alpha-x).$$

Again, the δ-function only picks up *half* its contribution when it is located on the boundary of the integration-domain. Thus we get

$$0 = \frac{1}{2} B(\beta,t|x,s) + A(\beta,t|x,s).$$

This is *not* consistent with — or does not enforce that — the transition-density A is zero when the forward coordinate is on the boundary. We conclude that the last-passage decomposition holds when *both* of the coordinates are in the interior of D, or when x lies on the boundary, but not when y lies on the boundary.

So we have examined the validity of the LP-decomposition for locations in the interior, and on smooth parts of the boundary, where δ-functions pick up half a contribution. But since we allow for piecewise smooth boundaries, we have not been complete — i.e. what about edges and corners? We could repeat the analysis above and we could use that a δ-function picks up $1/8^{\text{th}}$ of its contribution when it is located on a corner of a 3-dimensional box, for example. But the LP decomposition holds for all points in the interior, so that we should be allowed to take limits and reach any boundary point we want.

But still the reader may object, since for edges and corners there is no unambiguous way in which an interior point should move to the boundary; if a boundary normal is not defined, in what direction should the limit be taken? The point is well made, but the answer is, in short, that we never *need* to take limits at edges and corners.

We discussed in the introduction that the classical Dirichlet problem is not well-posed for domains with irregular boundaries. For the modified Dirichlet problem, we can *define* the solution as the weighted average over all first-passage locations, because the first passage will almost surely happen at a regular boundary point. The solution will therefore match the boundary data at all regular boundary points. Thus we conclude that it is satisfactory the LP decomposition holds for all $x, y \in D$; since we only ever need to take limits at smooth parts of the boundary.

We may repeat the exact same analysis with A and B swapped in position, to find that

$$\mathbb{P}(B_t \in dy;\, \tau^{\text{FP}} \geq \tau | B_s = x) = \int_D d\alpha\; B(y,t|\alpha,\tau) A(\alpha,\tau|x,s) \tag{3.2.3}$$

where

$$\tau^{\text{FP}}(t|x,s) = \inf_{\tau}(s \leq \tau \leq t : B_\tau \notin D | B_s = x)$$

with the conventional (but crucial!) addition that $\inf\{\varnothing\} = \infty$ such that the event $\tau^{\text{FP}} \geq \tau$ counts paths with and without passages. We recognise that the motion in the time interval after time τ is free, such that *decreasing* the intermediate τ allows for more paths. Therefore we have

$$\mathbb{P}(B_t \in dy;\, d\tau^{\text{FP}} \in \tau | B_s = x) = \left(-\frac{\partial}{\partial \tau}\right) \int_D d\alpha\; B(y,t|\alpha,\tau) A(\alpha,\tau|x,s). \tag{3.2.4}$$

To obtain the absorbed transition density, we may subtract from the free density all those paths that have passed the boundary:

$$_{\text{FP}}\, A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau\; \left(-\frac{\partial}{\partial \tau}\right) \int_D d\alpha\; B(y,t|\alpha,\tau) A(\alpha,\tau|x,s).$$

Again it is obvious that the FP decomposition holds for all points x and y in the interior, and again this may be viewed as an identity, since it holds by the virtue of the fundamental

theorem of calculus and the STCs. Combining both first- and last-passage analyses, we may summarise the FP and LP decompositions as follows:

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \left(-\frac{\partial}{\partial\tau}\right) \int_D d\alpha\; B(y,t|\alpha,\tau)A(\alpha,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_D d\alpha\; A(y,t|\alpha,\tau)B(\alpha,\tau|x,s).
\end{aligned} \tag{3.2.5}$$

In the 'derivation' of these identities, we have used the STCs but not the PDEs or BCs. Accepting that the absorbed density is fully and uniquely determined by the set of STCs, PDEs and BCs, we will be looking to obtain an integral equation that encompasses all of them. Differentiation under the integral sign is allowed and we can use the PDEs of (3.1.1), to obtain

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{\sigma^2}{2}\int_s^t d\tau \int_D d\alpha\; B(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} A(\alpha,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{\sigma^2}{2}\int_s^t d\tau \int_D d\alpha\; A(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} B(\alpha,\tau|x,s).
\end{aligned} \tag{3.2.6}$$

Next, we use Green's second identity (1.3.1) — which is valid for domains with a finite number of edges, corners and cusps — to obtain

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\; A(y,t|\beta,\tau)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta\right\} B(\beta,\tau|x,s).
\end{aligned} \tag{3.2.7}$$

Here ∂_β is again the scaled inward normal derivative (2.4.1). The BCs of (3.1.1) require that A is zero on the boundary, and thus we must have that ∂_β points towards A, so we obtain

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial}_\beta\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\beta\; A(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial}_\beta\right\} B(\beta,\tau|x,s).
\end{aligned} \tag{3.2.8}$$

In both cases, a positive term is subtracted from the free density to obtain the absorbed density. Also we recognise that we have now used all 6 PDEs, STCs and BCs of (3.1.1) in the derivation of these 2 integral equations, i.e. all the conditions that are supposed to specify A uniquely have now been used — along with Green's second identity on the domain. One immediate consequence of (3.2.8) is that A is symmetric in the spatial coordinates x and y. This deserves some attention, since Chung, for example, writes in [26] (p. 90)

> By the way, there is NO probabilistic intuition for the symmetry of [the absorbed transition density].

Our set of equations, however, can easily interpreted when we realise that

$$\begin{aligned} {}_{\text{FP}}\,\mathbb{P}\big(B_t \in dy;\, \tau^{\text{FP}} \in d\tau;\, B_{\tau^{\text{FP}}} \in d\beta \big| B_s = x\big) &= B(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s) \\ {}_{\text{LP}}\,\mathbb{P}\big(B_t \in dy;\, \tau^{\text{LP}} \in d\tau;\, B_{\tau^{\text{LP}}} \in d\beta \big| B_s = x\big) &= A(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} B(\beta,\tau|x,s) \end{aligned} \tag{3.2.9}$$

and we see that the last-passage distribution of all paths from (x,s) to (y,t) is equal to the first- passage distribution of all paths on the way back. Thus the spatial symmetry follows ultimately from a time reversal.

If it still troubles the reader that the FP and LP integral equations (3.2.8) are claimed to hold for piecewise smooth domains even though they explicitly refer to the outward normal, then it should be noted that the same is true for the (undisputed) divergence theorem itself.

In their expository paper, Port & Stone [53] (p. 146) derive the FP decomposition more or less heuristically, and almost directly from the Markov property — using that the motion after the first passage is independent of the motion before the first passage. Using this intuition they almost immediately write down that

$$A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau)\mathbb{P}\big(\tau^{\text{FP}} \in d\tau;\, B_{\tau^{\text{FP}}} \in d\beta \big| B_s = x\big).$$

But it seems that they do not realise — or at least do not write — that

$$\mathbb{P}\big(\tau^{\text{FP}} \in d\tau;\, B_{\tau^{\text{FP}}} \in d\beta \big| B_s = x\big) = \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s). \tag{3.2.10}$$

Instead, Port & Stone [53] write

> It is certainly intuitively plausible that [the first-passage decomposition] should hold, and a rigorous proof is not difficult to supply. Since the proof would involve a more thorough discussion of the measure theoretic structure of the Brownian motion process than we care to go into in this paper, we will omit the proof.

Comparatively, therefore, the approach in this paper has several advantages: 1) it suggests both a first- and a last-passage decomposition, 2) it explicitly requires Green's identity on the domain, allowing a finite number of corners, edges and thorns, 3) the result follows quite naturally, i.e. without referring to a 'measure theoretic structure', and the obtained expressions do not need to be verified *after* the fact.

Concluding, we have *derived* the FP and LP integral equations from the set of PDEs, STCs and BCs. But the reverse should also be possible. It is easily seen that as t goes down to s, that the integrals in (3.2.8) disappear, and thus A must behave like B in the short-time limit. Also we see that A must satisfy the same differential equations that B satisfies. If we are happy to believe that the FP or LP distributions peak at 'here' and 'now' when the backward (forward) location moves to the boundary, then the BCs can also be read off from the integral equations. Therefore we state the following proposition:

Proposition 1. FP and LP decompositions of ABM. *For all domains D allowing Green's theorem (1.3.1), for all $x, y \in D$, and for all regular boundary coordinates β, the following formulations of ABM are equivalent:*

$$\left.\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) A(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) A(y,t|x,s) &= 0 \\
A(\beta,t|x,s) &= 0 \\
A(y,t|\beta,s) &= 0 \\
\lim_{s\nearrow t} A(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} A(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FP}\ A(y,t|x,s) = B(y,t|x,s) \\
&\qquad -\int_s^t d\tau \oint_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s) \\
&\text{LP}\ A(y,t|x,s) = B(y,t|x,s) \\
&\qquad -\int_s^t d\tau \oint_{\partial D} d\beta\ A(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} B(\beta,\tau|x,s)
\end{aligned}\right. \tag{3.2.11}$$

where ∂_β is the scaled inward normal derivative as defined in (2.4.1).

We believe that this proposition is is new. We note that the time integrals on the right-hand side of the FP and LP decompositions in Proposition 1 are like convolutions: both propagators only depend on the time difference. As a general rule, for a convolution involving two test functions f_1 and f_2, we have that

$$\begin{aligned}
\int_s^\infty dt \int_s^t d\tau\ f_2(t-\tau) f_1(\tau - s) &= \int_s^\infty d\tau \int_\tau^\infty dt\ f_2(t-\tau) f_1(\tau - s) \\
&= \int_s^\infty d\tau \left(\int_0^\infty d\theta f_2(\theta)\right) f_1(\tau - s) \\
&= \left(\int_0^\infty f_2(\theta) d\theta\right)\left(\int_0^\infty f_1(\theta) d\theta\right)
\end{aligned}$$

Applying this to the right-hand side of Proposition 1, we obtain:

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) - \oint_{\partial D} d\beta\ G_B(y,\beta)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_A(\beta,x) \\
\text{LP}\ G_A(y,x) &= G_B(y,x) - \oint_{\partial D} d\beta\ G_A(y,\beta)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} G_B(\beta,x)
\end{aligned} \tag{3.2.12}$$

where we recall that this set of equations was also obtained in a more direct manner in the introduction, in equations (1.5.5) through (1.5.9). It follows that the following problem

formulations are equivalent:

$$\left.\begin{aligned} \frac{\sigma^2}{2}\nabla_y^2 G_A(y,x) &= -\delta(|y-x|) \\ \frac{\sigma^2}{2}\nabla_x^2 G_A(y,x) &= -\delta(|y-x|) \\ G_A(\beta,x) &= 0 \\ G_A(y,\beta) &= 0 \end{aligned}\right\} = \left\{\begin{aligned} \text{FP}\quad G_A(y,x) &= G_B(y,x) \\ &\quad - \oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_A(\beta,x) \\ \text{LP}\quad G_A(y,x) &= G_B(y,x) \\ &\quad - \oint_{\partial D} d\beta\, G_A(y,\beta)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} G_B(\beta,x) \end{aligned}\right. \tag{3.2.13}$$

Since this follows *directly* from Proposition 1, we shall not make it a proposition itself. Although this *particular* reformulation appears seems to be new, we are obviously not the first to turn a differential equation with boundary conditions into an integral equation, see for example [54] (p. 214) for a similar (but different) example.

By definition, G_B is symmetric in its two arguments. From the pair of the FP and LP decompositions, we see that G_A is also symmetric in its arguments; something that may have been expected (maybe), but that was not necessarily assumed before. In [26], Chung writes

> Incredibly, this symmetry of Green's functions persists for a general domain in any dimension. Did physicists such as the redoubtable Faraday discover this phenomenon experimentally, or did they possess the electrical perception to see that it must be so? From our Brownian point of view, [...] there does not seem to be any rhyme or reason for the paths to behave so reciprocally.

Again, we find that the symmetry of the Green function follows ultimately from a reversal of time, where first becomes last and last becomes first. In [25] (p. 40 and 52) and [26] (p. 56), the symmetry of the Green function is proved using Green's second identity, but a probabilistic interpretation is absent. This is surprising, because Chung was well aware of the probabilistic importance of the last-passage time. In [55], for example, he writes

> For some reason the notion of a last exit time, which is manifestly involved in the arguments, would not be dealt with openly and directly. This may be partially due to the fact that such a time is not an "optional" (or "stopping") time, does not belong to the standard equipment, and so must be evaded at all cost. [...] A probabilistic solution to Dirichlet's problem was obtained by Doob (1954) by considering a first exit time; here a similar solution to the so-called Robin's problem will be obtained by considering a last exit time.

Chung then proceeds to discuss the 'Robin problem' or 'equilibrium problem', which asks for the probability that a transient Brownian motion ever hits a certain set D (assuming it starts off outside of this set). The transience is crucial, since that means that the

Brownian path can escape to infinity without ever hitting D. The equilibrium problem is also discussed in more modern literature, such as [12] (p. 227) who write that 'the proof of the last exit formula is taken from Chung's beautiful paper'. Chung is the first author to use the last passage systematically, but *still* it is the case that the first passage is reserved for the Dirichlet problem, while the last passage is reserved for the Robin problem. It is unclear to us why either concept should be reserved for either problem. We have established that the absorbed propagator can be found by subtracting from the free density all paths with either a first or last passage. Equally, for the Robin problem we have

$$\begin{aligned} \text{FP}\quad \mathbb{P}\big(B_t \text{ ever hits } D \big| B_s = x\big) &= \int_{\mathbb{R}^d} dy \oint_{\partial D} d\beta\; G_B(y,\beta) \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} G_A(\beta, x) \\ \text{LP}\quad \mathbb{P}\big(B_t \text{ ever hits } D \big| B_s = x\big) &= \int_{\mathbb{R}^d} dy \oint_{\partial D} d\beta\; G_A(y,\beta) \left\{ \frac{1}{2} \overleftarrow{\partial_\beta} \right\} G_B(\beta, x) \end{aligned} \tag{3.2.14}$$

because for a set D to be visited at all, it must be visited for the first time, at some point, and for the last time, at some point — provided that the entire space is transient, e.g. for $d \geq 3$. Thus we see that there is no need to reserve either the first- or last-passage decomposition for the Robin problem — either will do.

3.3 Reflected Brownian motion

The transition density of reflected Brownian motion (RBM) is indicated by $R(y,t|x,s)$, with forward and backward space-time coordinates (y,t) and (x,s). The reflected transition density $R(y,t|x,s)$ satisfies the following set of equations:

$$\begin{aligned} \text{forward PDE} \quad & \left(\frac{\partial}{\partial t} - \frac{\sigma^2}{2} \nabla_y^2 \right) R(y,t|x,s) = 0, \\ \text{backward PDE} \quad & \left(\frac{\partial}{\partial s} + \frac{\sigma^2}{2} \nabla_x^2 \right) R(y,t|x,s) = 0, \\ \text{forward BC} \quad & n_\beta \cdot \overrightarrow{\nabla}_\beta R(\beta,t|x,s) = 0, \\ \text{backward BC} \quad & R(y,t|\beta,s) \overleftarrow{\nabla}_\beta \cdot n_\beta = 0, \\ \text{forward STC} \quad & \lim_{s \nearrow t} R(y,t|x,s) = \delta(|y-x|), \\ \text{backward STC} \quad & \lim_{t \searrow s} R(y,t|x,s) = \delta(|y-x|). \end{aligned} \tag{3.3.1}$$

This holds for all $x, y \in D$ and all regular (i.e. non-singular) boundary points β. PDE stands for 'partial differential equation', BC stands for 'boundary condition' and STC stands for 'short-time condition'. It can be proved that the absorbed transition density 1) exists, 2) is unique and 3) is determined by the above conditions. See for example [30].

The PDEs are satisfied because the transition density is unbiased, i.e.

$$R(y,t|x,s) = \mathbb{E}R(y-dB,t-dt|x,s)$$
$$R(y,t|x,s) = \mathbb{E}R(y,t|x+dB,s+ds)$$

and using Itô's lemma (1.4.2) gives both PDEs. The BCs are satisfied because a Brownian particle is reflected in the normal direction, at any regular boundary point β, and therefore $R(y,t|\beta,s)$ and $R(y,t|\beta+\epsilon,s)$ are equal to first order in ϵ, if β is a regular boundary coordinate and ϵ is small displacement in the normal direction. The STCs are satisfied for x and y in the interior because in the short-time limit the reflected transition density must behave like the free transition density. Furthermore we know that the Brownian particle must be somewhere at every intermediate time, and therefore we also have the Chapman-Kolmogorov equation

$$\text{Chapman-Kolmogorov } R(y,t|x,s) = \int_D d\alpha\, R(y,t|\alpha,\tau)R(\alpha,\tau|x,s) \tag{3.3.2}$$

for any $s \leq \tau \leq t$, and where the STCs ensure that the Chapman-Kolmogorov equation also holds in the limit where τ goes to s or t. See for example [14], p. 36. If it exists (see subsection 1.4), then the reflected Green function is defined by

$$G_R(y,x) := \int_s^\infty R(y,t|x,s)\,dt \tag{3.3.3}$$

and it satisfies

$$\begin{gathered}\frac{\sigma^2}{2}\nabla_y^2 G_R(y,x) = \frac{\sigma^2}{2}\nabla_x^2 G_R(y,x) = -\delta(|y-x|) \\ \overrightarrow{\partial_\beta} G_R(\beta,x) = G_R(y,\beta)\overleftarrow{\partial_\beta} = 0\end{gathered} \tag{3.3.4}$$

where, roughly speaking, the reflected Green function only exists if $d \geq 3$ and the domain is unbounded. The interior Neumann problem also has a solution if a certain 'compatibility equation' is satisfied, see for example [34] (p. 896), [13] (p. 221).

The simplest boundary value problems occur when the domain considered is a halfspace. By the 'André reflection principle' as in [9] (p. 42), [10] (p. 79) or [15] (p. 26), we obtain that the reflected density for a halfspace equals:

$$R^{\text{HS}}(y,t|x,s) = B(y,t|x,s) + B(y,t|x^*,s) \tag{3.3.5}$$

where x^* equals the 'mirror-coordinate' that is obtained by taking a mirror image of x in the reflecting hyperplane. It is easily checked that R^{HS} satisfies the PDEs, the BCs and the STCs, and thus by uniqueness it must be correct. As far as the STCs are concerned, two δ-functions are obtained: at both x and x^*, but the δ-function at x^* is outside of the space of interest and therefore irrelevant.

3.4 First- and last-reflection decompositions

Now that we have discussed the first- and last-passage decompositions at such length, the following first-reflection (FR) and last-reflection (LR) decompositions almost immediately suggest themselves:

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= A(y,t|x,s) + \int_s^t d\tau \left(-\frac{\partial}{\partial\tau}\right) \int_D d\alpha\, R(y,t|\alpha,\tau)A(\alpha,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= A(y,t|x,s) + \int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_D d\alpha\, A(y,t|\alpha,\tau)R(\alpha,\tau|x,s).
\end{aligned} \tag{3.4.1}$$

The nomenclature follows from considering a quantity like

$$\int_D d\alpha\, R(y,t|\alpha,\tau)A(\alpha,\tau|x,s),$$

which counts paths from (x,s) to (y,t), where the first reflection (if at all) happens *after* time τ. Alternatively, both decompositions may be seen to hold by the virtue of the fundamental theorem of calculus and the STCs. Using the PDEs of (3.3.1) under the integral sign, we get

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= A(y,t|x,s) + \frac{\sigma^2}{2}\int_s^t d\tau \int_D d\alpha\; R(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} A(\alpha,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= A(y,t|x,s) - \frac{\sigma^2}{2}\int_s^t d\tau \int_D d\alpha\; A(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} R(\alpha,\tau|x,s).
\end{aligned} \tag{3.4.2}$$

Using Green's second identity (1.3.1) — which is valid for domains with a finite number of edges, corners and cusps — we obtain

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= A(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\; R(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= A(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\; A(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} R(\beta,\tau|x,s).
\end{aligned} \tag{3.4.3}$$

Here ∂_β is again the scaled inward normal derivative (2.4.1). The BCs of (3.3.1) require that $\overrightarrow{\partial_\beta}R$ and $R\overleftarrow{\partial_\beta}$ are zero, and thus we must have that ∂_β points towards A, so we obtain

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= A(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\beta\; R(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= A(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\beta\; A(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} R(\beta,\tau|x,s).
\end{aligned} \tag{3.4.4}$$

The FR decomposition tells us that the reflected path either does not hit the boundary at all, or it behaves as an ABM until it hits it for the first time, and then proceeds onwards

as an RBM. The LR decomposition tells us that the reflected path either does not hit the boundary at all, or proceeds as an RBM until it visits the boundary for the last time, after which it moves to its endpoint as an ABM. It turns out, however, that it is more useful to write the reflected density in terms of the free density, and to do this we replace the absorbed density A by the free density B:

$$\begin{aligned} \text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \left(-\frac{\partial}{\partial\tau}\right) \int_D d\alpha\, R(y,t|\alpha,\tau)B(\alpha,\tau|x,s), \\ \text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_D d\alpha\, B(y,t|\alpha,\tau)R(\alpha,\tau|x,s). \end{aligned} \tag{3.4.5}$$

In this set of equations we have kept the names FR and LR, even though that interpretation has now become a little bit problematic. But it is obvious that both identities hold, since they *only* rely on the short-time conditions in the interior. Turning the crank one more time, we get the exact same result with A replaced by B:

$$\begin{aligned} \text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\beta\; R(y,t|\beta,\tau) \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s), \\ \text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau) \left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} R(\beta,\tau|x,s). \end{aligned} \tag{3.4.6}$$

Again, the boundary operator ∂_β must point away from R. Whereas the absorbed density A is always smaller than the free density B, it is *not* the case that the reflected density is always larger than the free density B. We see that the reflected density equals the free density plus a weighted average over all boundary densities of R at ∂D at all times τ, where the weight is given by $\frac{1}{2}\overrightarrow{\partial_\beta}B(\beta,\tau|x,s)$ or $B(y,t|\beta,\tau)\frac{1}{2}\overleftarrow{\partial_\beta}$. This weight is not always positive, but we have that

$$\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s) \geq 0 \text{ if } D \text{ is convex}$$

with strict inequalities if D is strictly convex. Thus for a convex space, the reflected density is everywhere larger than the free density. Intuitively, every point in a convex domain is like a 'focal' point, where more paths are directed than in the absence of the boundary. We will discuss this further in subsection 3.6. We conclude this subsection with the following proposition:

Proposition 2. FR and LR decompositions of RBM. *For all domains D allowing Green's theorem (1.3.1), for all $x, y \in D$, and for all regular boundary coordinates β, the*

following formulations of RBM are equivalent:

$$\left.\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) R(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) R(y,t|x,s) &= 0 \\
\overrightarrow{\partial_\beta} R(\beta,t|x,s) &= 0 \\
R(y,t|\beta,s)\overleftarrow{\partial_\beta} &= 0 \\
\lim_{s\nearrow t} R(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} R(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FR } R(y,t|x,s) = B(y,t|x,s) \\
&\qquad + \int_s^t d\tau \oint_{\partial D} d\beta\; R(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s) \\
&\text{LR } R(y,t|x,s) = B(y,t|x,s) \\
&\qquad + \int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} R(\beta,\tau|x,s)
\end{aligned}\right. \tag{3.4.7}$$

where ∂_β is the scaled inward normal derivative as defined in (2.4.1).

This proposition is new. We can integrate the right-hand side problem of Proposition 2 over time, to obtain the following:

$$\begin{aligned}
\text{FR } G_R(y,x) &= G_B(y,x) + \oint_{\partial D} d\beta\; G_R(y,\beta)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_B(\beta,x), \\
\text{LR } G_R(y,x) &= G_B(y,x) + \oint_{\partial D} d\beta\; G_B(y,\beta)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} G_R(\beta,x).
\end{aligned} \tag{3.4.8}$$

This result proves the symmetry of the reflected Green function in its arguments. It follows that the following problem-formulations are equivalent:

$$\left.\begin{aligned}
\frac{\sigma^2}{2}\nabla_y^2 G_R(y,x) &= -\delta(|y-x|) \\
\frac{\sigma^2}{2}\nabla_x^2 G_R(y,x) &= -\delta(|y-x|) \\
\overrightarrow{\partial_\beta} G_R(\beta,x) &= 0 \\
G_R(y,\beta)\overleftarrow{\partial_\beta} &= 0
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FR } G_R(y,x) = G_B(y,x) \\
&\qquad + \oint_{\partial D} d\beta\; G_R(y,\beta)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_B(\beta,x) \\
&\text{LR } G_R(y,x) = G_B(y,x) \\
&\qquad + \oint_{\partial D} d\beta\; G_B(y,\beta)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} G_R(\beta,x)
\end{aligned}\right. \tag{3.4.9}$$

for all regular boundary coordinates β, and if G_R exists. Since this follows directly from Proposition 2, we shall not make it a proposition itself.

For the interior Green function, [25] (p. 39) suggests that the simplest boundary condition is not $\partial G_R = 0$, but $\partial G_R = \frac{1}{|\partial D|}$ where $|\partial D|$ indicates the total area of the surface. We shall not pursue this, but the methods presented here can easily be adapted to produce integral equations for the interior Green function G_R.

3.5 Discontinuity relations

To make good use of Propositions 1 and 2, we shall need two lemmas that at first may seem quite technical. But they are crucial to obtain the tangent plane decompositions of

the next subsection. Recall that

$$\begin{aligned}\overrightarrow{\partial_\beta} f(\beta,\gamma) &:= -\sigma^2 \lim_{\alpha\to\beta} n_\beta \cdot \overrightarrow{\nabla}_\alpha f(\alpha,\gamma)\\ f(\gamma,\beta)\overleftarrow{\partial_\beta} &:= -\sigma^2 \lim_{\alpha\to\beta} n_\beta \cdot \overrightarrow{\nabla}_\alpha f(\gamma,\alpha)\end{aligned}$$

The operator $\overrightarrow{\partial_\beta}$ does therefore not necessarily commute with integration over the boundary. Even though differentiation under the integral sign is usually allowed, pushing the limit through the integral is often not allowed. We will need the following lemmas:

Lemma 1. *For a regular boundary coordinate β and some function f, we have*

$$\begin{aligned}\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\}\left(\int_s^t d\tau \oint_{\partial D} d\gamma\; B(\beta,t|\gamma,\tau) f(\gamma,\tau)\right) &= -\frac{1}{2} f(\beta,t)\\ &\quad + \int_s^t d\tau \oint_{\partial D} d\gamma\; \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,t|\gamma,\tau) f(\gamma,\tau)\\ \left(\int_s^t d\tau \oint_{\partial D} d\gamma\; f(\gamma,\tau) B(\gamma,\tau|\beta,s)\right)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} &= -\frac{1}{2} f(\beta,s)\\ &\quad + \int_s^t d\tau \oint_{\partial D} d\gamma\; f(\gamma,\tau) B(\gamma,\tau|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\}\end{aligned} \tag{3.5.1}$$

where it is crucial if the operators $\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\}$ and $\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\}$ appear outside the integration (left-hand side) or inside the integration (right-hand side).

Proof. We prove only the first part of the lemma, and the second part proceeds analogously. First consider that the following quantity satisfies the forward PDE for all y in the interior of D, i.e.

$$\left(\frac{\partial}{\partial t} - \frac{\sigma^2}{2}\nabla_y^2\right)\int_s^t d\tau \oint_{\partial D} d\gamma\; B(y,t|\gamma,\tau) f(\gamma,\tau) = 0.$$

By the divergence theorem, we therefore have

$$\int_D d\alpha\, \frac{\partial}{\partial t}\int_s^t d\tau \oint_{\partial D} d\gamma\, B(\alpha,t|\gamma,\tau) f(\gamma,\tau) = -\int_{\partial D} d\beta\, \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\}\int_s^t d\tau \oint_{\partial D} d\gamma\, B(\beta,t|\gamma,\tau) f(\gamma,\tau).$$

But, equally, we could have obtained

$$\begin{aligned}\int_D d\alpha\, \frac{\partial}{\partial t}\int_s^t d\tau \oint_{\partial D} d\gamma\, B(\alpha,t|\gamma,\tau) f(\gamma,\tau) &= \int_D d\alpha \lim_{\tau\nearrow t} \oint_{\partial D} d\gamma\, B(\alpha,t|\gamma,\tau) f(\gamma,\tau)\\ &\quad + \int_D d\alpha \int_s^t d\tau \oint_{\partial D} d\gamma\, \left(\frac{\partial}{\partial t}\right) B(\alpha,t|\gamma,\tau) f(\gamma,\tau).\end{aligned}$$

In the short-time limit, B behaves like a δ-function, and using the forward PDE of B in the second term, we get

$$\int_D d\alpha \frac{\partial}{\partial t} \int_s^t d\tau \oint_{\partial D} d\gamma\, B(\alpha, t|\gamma, \tau) f(\gamma, \tau) = \int_D d\alpha \oint_{\partial D} d\gamma\, \delta(|\alpha - \gamma|) f(\gamma, t) + \int_D d\alpha \int_s^t d\tau \oint_{\partial D} d\gamma \left\{ \frac{\sigma^2}{2} \overrightarrow{\nabla}_\alpha^2 \right\} B(\alpha, t|\gamma, \tau) f(\gamma, \tau).$$

The δ-function on the boundary picks up half a contribution, and using the divergence theorem for the second term we obtain

$$\int_D d\alpha \frac{\partial}{\partial t} \int_s^t d\tau \oint_{\partial D} d\gamma\, B(\alpha, t|\gamma, \tau) f(\gamma, \tau) = \frac{1}{2} \oint_{\partial D} d\beta\, f(\beta, t) - \oint_{\partial D} d\beta \int_s^t d\tau \oint_{\partial D} d\gamma \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} B(\beta, t|\gamma, \tau) f(\gamma, \tau).$$

Comparing with the previous approach, we see that the following must be equal:

$$-\int_{\partial D} d\beta \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} \int_s^t d\tau \oint_{\partial D} d\gamma\, B(\beta, t|\gamma, \tau) f(\gamma, \tau) = \frac{1}{2} \oint_{\partial D} d\beta\, f(\beta, t) - \oint_{\partial D} d\beta \int_s^t d\tau \oint_{\partial D} d\gamma \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} B(\beta, t|\gamma, \tau) f(\gamma, \tau).$$

And because this holds for *every* domain D we must have for *each* boundary-location β that

$$\left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} \int_s^t d\tau \oint_{\partial D} d\gamma\, B(\beta, t|\gamma, \tau) f(\gamma, \tau) = -\frac{1}{2} f(\beta, t) + \int_s^t d\tau \oint_{\partial D} d\gamma \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} B(\beta, t|\gamma, \tau) f(\gamma, \tau).$$

This concludes the proof. The proof of part 2 proceeds in the same fashion. □

We shall also need a second lemma, which reads

Lemma 2. *For a regular boundary coordinate β and some function f, we have*

$$\lim_{x \to \beta} \int_s^t d\tau \oint_{\partial D} d\gamma\, f(\gamma, \tau) \left\{ \frac{1}{2} \overrightarrow{\partial_\gamma} \right\} B(\gamma, \tau|x, s) = \frac{1}{2} f(\beta, s) + \int_s^t d\tau \oint_{\partial D} d\gamma\, f(\gamma, \tau) \left\{ \frac{1}{2} \overrightarrow{\partial_\gamma} \right\} B(\gamma, \tau|\beta, s)$$
$$\lim_{y \to \beta} \int_s^t d\tau \oint_{\partial D} d\gamma\, B(y, t|\gamma, \tau) \left\{ \frac{1}{2} \overleftarrow{\partial_\gamma} \right\} f(\gamma, \tau) = \frac{1}{2} f(\beta, t) + \int_s^t d\tau \oint_{\partial D} d\gamma\, B(\beta, t|\gamma, \tau) \left\{ \frac{1}{2} \overleftarrow{\partial_\gamma} \right\} f(\gamma, \tau) \tag{3.5.2}$$

Proof. The proof follows in the same fashion as the proof of Lemma 1. □

Although we have not found these lemmas anywhere else, the *time-independent* versions are in fact well-known. If we let f be independent of time and we let $t \to \infty$, then Lemma 1 and Lemma 2 imply that:

$$\begin{aligned}\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\}\left(\oint_{\partial D} d\gamma\; G_B(\beta,\gamma)f(\gamma)\right) &= -\frac{1}{2}f(\beta) + \oint_{\partial D} d\gamma\; \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} G_B(\beta,\gamma)f(\gamma)\\ \left(\oint_{\partial D} d\gamma\; f(\gamma)G_B(\gamma,\beta)\right)\left\{\tfrac{1}{2}\overleftarrow{\partial_\beta}\right\} &= -\frac{1}{2}f(\beta) + \oint_{\partial D} d\gamma\; f(\gamma)G_B(\gamma,\beta)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\}\end{aligned}$$

and

$$\begin{aligned}\lim_{x\to\beta}\oint_{\partial D} d\gamma\; f(\gamma)\left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} G_B(\gamma,x) &= \frac{1}{2}f(\beta) + \oint_{\partial D} d\gamma\; f(\gamma)\left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} G_B(\gamma,\beta)\\ \lim_{y\to\beta}\oint_{\partial D} d\gamma\; G_B(y,\gamma)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} f(\gamma) &= \frac{1}{2}f(\beta) + \oint_{\partial D} d\gamma\; G_B(\beta,\gamma)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} f(\gamma)\end{aligned}$$

and these lemmas are found in a large variety of places, such as [1] (p. 309), [33] (p. 272 and 292) or [34] (p. 893), [38] (p. 4).

3.6 Tangent plane decompositions

Recall the right-hand side of Proposition 1:

$$\begin{aligned}\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\gamma\; B(y,t|\gamma,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} A(\gamma,\tau|x,s),\\ \text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\gamma\; A(y,t|\gamma,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} B(\gamma,\tau|x,s).\end{aligned}$$

On the left-hand side we see the absorbed propagator A, while on the right-hand side the forward and backward normal derivative of A appear, respectively. The idea is to make sure that the forward and backward normal derivative appear on both sides, by applying $\frac{1}{2}\overrightarrow{\partial_\beta}$ to the left of the FP decomposition, and $\frac{1}{2}\overleftarrow{\partial_\beta}$ to the right of the LP decomposition. Thus we get

$$\begin{aligned}\text{FP}\quad \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s) &= \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,t|x,s)\\ &\quad - \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\}\int_s^t d\tau \oint_{\partial D} d\beta\; B(\beta,t|\gamma,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} A(\gamma,\tau|x,s),\\ \text{LP}\quad A(y,t|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} &= B(y,t|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\}\\ &\quad - \left(\int_s^t d\tau \oint_{\partial D} d\gamma\; A(y,t|\gamma,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} B(\gamma,\tau|\beta,s)\right)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\}.\end{aligned} \tag{3.6.1}$$

Using Lemma 1 to push the differential operators through the integral signs, we get

$$\begin{aligned}
\text{FP}\ \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s) &= \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,t|x,s) \\
&\quad - \int_s^t d\tau \oint_{\partial D} d\beta\ \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,t|\gamma,\tau) \left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} A(\gamma,\tau|x,s) \\
&\quad + \frac{1}{2}\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s) \\
\text{LP}\ A(y,t|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} &= B(y,t|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} \\
&\quad - \int_s^t d\tau \oint_{\partial D} d\gamma\ A(y,t|\gamma,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} B(\gamma,\tau|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} \\
&\quad + \frac{1}{2} A(y,t|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\}
\end{aligned} \tag{3.6.2}$$

Collecting terms, we get:

$$\begin{aligned}
\text{FP}\ \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s) &= \overrightarrow{\partial_\beta} B(\beta,t|x,s) \\
&\quad - \int_s^t d\tau \oint_{\partial D} d\beta\ \overrightarrow{\partial_\beta} B(\beta,t|\gamma,\tau) \left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} A(\gamma,\tau|x,s) \\
\text{LP}\ A(y,t|\beta,s)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} &= B(y,t|\beta,s)\overleftarrow{\partial_\beta} \\
&\quad - \int_s^t d\tau \oint_{\partial D} d\gamma\ A(y,t|\gamma,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} B(\gamma,\tau|\beta,s)\overleftarrow{\partial_\beta}
\end{aligned} \tag{3.6.3}$$

where the factors of 2 are crucial and the factorisation is carefully chosen. While the derivation may have seemed quite technical, the interpretation is very intuitive. Recall the absorbed density for a halfspace:

$$A^{\text{HS}}(y,t|x,s) = B(y,t|x,s) - B(y,t|x^*,s)$$

and the corresponding first-passage distribution

$$\begin{aligned}
\mathbb{P}(\tau^{\text{FP}}_{\text{over HS}} \in d\tau;\, B_{\tau^{\text{FP}}_{\text{over HS}}} \in d\beta | B_s = x) &= \tfrac{1}{2}\overrightarrow{\partial_\beta} A^{\text{HS}}(\beta,\tau|x,s) \\
&= \overrightarrow{\partial_\beta} B(\beta,\tau|x,s).
\end{aligned}$$

Now consider a convex space. On left-hand side of (3.6.3) we see the joint probability that the first exit from domain D occurs at location β and time t. The interpretation of the right-hand side is as follows: for a convex boundary ∂D, the joint probability that a first-passage occurs at the space-time coordinate (β,t) can be estimated by the probability that the particle hits the tangent plane defined by β for the first time at (β,t) — but this is an overestimate because the halfspace defined by the tangent plane at β allows for more paths to stay alive than the actual convex domain does. Therefore we must subtract

from this initial estimate the probability that the particle leaves the domain at some other space-time location (γ, τ) and *then* hits the tangent plane defined by β at (β, t) — and we should sum over all γ and τ. We see that the right-hand side of (3.6.3) does exactly this. This interpretation is new.

The first-passage density at β is related to the first-passage density at all *other* locations γ. This was to be expected since the entire shape of the boundary is relevant for the first-passage density at any one location.

The *only* case when the first-passage density decouples from that at other locations is when the domain is halfspace: the first-passage density for a halfspace consists only of the first term in (3.6.3). This is because the second term (3.6.3) equals zero, i.e.

$$\int_s^t d\tau \oint_{\partial D} d\gamma \, \overrightarrow{\partial}_\beta B(\beta, t|\gamma, \tau) \left\{ \frac{1}{2} \overrightarrow{\partial}_\gamma \right\} A(\gamma, \tau|x, s)$$

equals zero for a halfspace, because

$$\overrightarrow{\partial}_\beta B(\beta, t|\gamma, \tau) = n_\beta \cdot \frac{\beta - \gamma}{t - \tau} B(\beta, t|\gamma, \tau).$$

For a halfspace it is clear that $n_\beta \cdot (\beta - \gamma) = 0$, because n_β and $(\beta - \gamma)$ are perpendicular. For a halfspace, therefore, the first term in the tangent plane decomposition is the only term.

We conclude that the first-passage density, at any location β, depends on the first-passage density at all other locations γ through a certain 'weight', where this weight can be positive or negative and takes the sign of $n_\beta \cdot (\beta - \gamma)$. It is not hard to check that the following variational inequalities hold for convex and concave spaces:

$$\begin{aligned} \text{Convex domain} \quad & n_\beta \cdot (\beta - \gamma) \geq 0 \\ \text{Concave domain} \quad & n_\beta \cdot (\beta - \gamma) \leq 0 \end{aligned} \tag{3.6.4}$$

for any two boundary-coordinates β and γ. As a result, the first-passage density over a convex domain at location β is smaller than the corresponding first-passage density over the hyperplane tangent to D at β. This was to be expected — and the opposite holds for a concave domain.

The interpretation of the last-passage decomposition is similar. We may approximate the probability-density that the last-passage occurs at (β, s), before going to (y, t) as the probability-density that the tangent plane at β was crossed for the last time at (β, s). However this is an overestimate for a convex domain D, since the fact that the tangent plane at β was not crossed later than s does not necessarily imply that ∂D was not crossed after time s, etcetera.

Next, we recall the first- and last-reflection decompositions of Proposition 2:

$$\text{FR}\quad R(y,t|x,s) = B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\gamma\; R(y,t|\gamma,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} B(\gamma,\tau|x,s),$$
$$\text{LR}\quad R(y,t|x,s) = B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\gamma\; B(y,t|\gamma,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} R(\gamma,\tau|x,s).$$

Apply the operators $\lim_{x\to\beta}$ and $\lim_{y\to\beta}$ and use Lemma 2 to push the limits through the integrals, collect terms and obtain:

$$\begin{aligned}\text{FR}\quad R(y,t|\beta,s) &= 2B(y,t|\beta,s) + \int_s^t d\tau \oint_{\partial D} d\gamma\; R(y,t|\gamma,\tau)\overrightarrow{\partial_\gamma} B(\gamma,\tau|\beta,s),\\ \text{LR}\quad R(\beta,t|x,s) &= 2B(\beta,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\gamma\; B(\beta,t|\gamma,\tau)\overleftarrow{\partial_\gamma} R(\gamma,\tau|x,s).\end{aligned} \tag{3.6.5}$$

Here the factors of 2 are crucial. Recall the absorbed density for a halfspace:

$$R^{\text{HS}}(y,t|x,s) = B(y,t|x,s) + B(y,t|x^*,s),$$

and the corresponding boundary quantity

$$R^{\text{HS}}(\beta,t|x,s) = 2B(\beta,t|x,s).$$

A similar interpretation as before presents itself. For a convex domain, for example, we may estimate the probability-density R to move from (x,s) to some boundary-location β as if there was (only) a reflecting tangent plane at β. This gives rise to the first term on the right-hand side, which is $2B$. But for a convex domain with a reflecting boundary, every location is like a focal point: more paths are directed there. More paths are directed to every boundary-location of a convex domain, than are approximated by putting a reflecting tangent plane at that location. Therefore we must add to the initial estimate the probability that the particle reflects off the boundary somewhere else, and only *then* reaches the tangent plane at β for the first time at (β,t). For a concave domain it can easily be checked that the first term, $2B$, is an overestimate for which the second term corrects, and so on.

3.7 Single and double boundary-layers

For absorbed Brownian motion we have derived the first- and last-passage decompositions of Proposition 1:

$$\text{FP}\quad A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s),$$
$$\text{LP}\quad A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D} d\beta\; A(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} B(\beta,\tau|x,s).$$

We have also found the resulting (by Lemma 1) TP decompositions in (3.6.3):

$$\begin{aligned}
\text{FP}\quad \left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s) &= \overrightarrow{\partial_\beta} B(\beta,t|x,s) \\
&\quad - \int_s^t d\tau \oint_{\partial D} d\beta\, \overrightarrow{\partial_\beta} B(\beta,t|\gamma,\tau) \left\{\frac{1}{2}\overrightarrow{\partial_\gamma}\right\} A(\gamma,\tau|x,s), \\
\text{LP}\quad A(y,t|\beta,s) \left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} &= B(y,t|\beta,s)\overleftarrow{\partial_\beta} \\
&\quad - \int_s^t d\tau \oint_{\partial D} d\gamma\, A(y,t|\gamma,\tau) \left\{\frac{1}{2}\overleftarrow{\partial_\gamma}\right\} B(\gamma,\tau|\beta,s)\overleftarrow{\partial_\beta}.
\end{aligned}$$

The TP decompositions are useful not only because of their interpretation, but also because they feature the *same* quantity on both sides of the equation, i.e. $\overrightarrow{\partial} A$ appears on both sides for the FP TP decomposition, whereas $A\overleftarrow{\partial}$ appears on both sides of the LP TP decomposition.

The idea for solving integral equations like this is by the 'successive approximation method' as in [34] (p. 566, 632 and 811) or equivalently the 'Neumann series' as in [56] (p. 78). The idea is simple: use the left-hand side of the equation as the definition for the unknown quantity appearing on the right-hand side, and do this repetitively to obtain a series solution. In effect the equation is repeatedly substituted into itself, giving rise to an infinite series.

Once a series solution for $\frac{1}{2}\overrightarrow{\partial} A$ or $A\overleftarrow{\partial}\frac{1}{2}$ has been obtained, then we can substitute this series back into the expression for A. Since the TP decomposition follows directly from Proposition 1 (using Lemma 1), we should expect that the series satisfies all 6 requirements of (3.1.1). We conclude that:

Proposition 3. Formal ABM series solution. *The formal solution to problem* (3.1.1) *is given by the following first- or last-passage series:*

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty}(-1)^i \left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right]\left[\oint d\beta_i\cdots\oint d\beta_1\right] \\
&\quad \times B(y,t|\beta_i,\theta_i)\left[\prod_{k=2}^{i}\overrightarrow{\partial_{\beta_k}} B(\beta_k,\theta_k|\beta_{k-1},\theta_{k-1})\right]\overrightarrow{\partial_{\beta_1}} B(\beta_1,\theta_1|x,s) \\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty}(-1)^i \left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right]\left[\oint d\beta_i\cdots\oint d\beta_1\right] \\
&\quad \times B(y,t|\beta_i,\theta_i)\overleftarrow{\partial_{\beta_i}}\left[\prod_{k=1}^{i-1} B(\beta_{k+1},\theta_{k+1}|\beta_k,\theta_k)\overleftarrow{\partial_{\beta_k}}\right] B(\beta_1,\theta_1|x,s)
\end{aligned}$$

where the FP *and* LP *series are identical, term-by-term, and where the modes convergence are as follows:*

domain	*mode of convergence*
convex domain	alternating
concave domain	monotone

but where convergence itself is taken for granted.

New about this proposition is 1) that there are two series rather than one, 2) the fact the solution is based on a *derivation* rather than on an *ansatz*, i.e. it follows from the intuitive TP decomposition, and 3) its reliance on Green's theorem, allowing piecewise smooth domains, whereas the ansatz approach is thought appropriate for smooth domains only and 4) the claimed mode of convergence. The mode of convergence follows from the fact that

$$\begin{aligned}\overrightarrow{\partial_\beta} B(\beta,\theta|\gamma,\tau) &= n_\beta \cdot \frac{\beta-\gamma}{\theta-\tau} B(\gamma,\theta|\beta,\tau),\\ B(\beta,\theta|\gamma,\tau)\overleftarrow{\partial_\gamma} &= n_\gamma \cdot \frac{\gamma-\beta}{\theta-\tau} B(\beta,\theta|\gamma,\tau),\end{aligned}$$

and that for two boundary coordinates γ and β we have

$$\begin{aligned}\text{Convex domain}\ \ & n_\beta \cdot (\beta-\gamma) \geq 0,\\ \text{Concave domain}\ \ & n_\beta \cdot (\beta-\gamma) \leq 0.\end{aligned}$$

The integrands in Proposition 3 are positive for a convex domain, so that the multiplicative factor $(-1)^i$ in the sum determines the mode of convergence: alternating. For a concave domain, all terms $\overrightarrow{\partial} B$ or $B\overleftarrow{\partial}$ appearing in the *square* brackets are negative. Unfortunately, the sign of $\overrightarrow{\partial_\beta} B(\beta,\theta|x,s)$ may change as β moves along a concave boundary. And similarly for $B(y,t|\beta,s)\overleftarrow{\partial_\beta}$. On the part of the concave boundary where $\overrightarrow{\partial_\beta} B(\beta,\theta|x,s)$ has a fixed sign, the series converges in a monotone fashion. On the part of the concave boundary where $\overrightarrow{\partial_\beta} B(\beta,\theta|x,s)$ has a fixed but different sign, the series also converges in a monotone fashion — except in the other direction. Because we can split the series solution into two pieces where both converge in a monotone fashion (albeit in other directions), we say simply that the series converges in a monotone fashion.

By integrating the series solutions of Proposition 3 over time, we get two associated series for the absorbed Green function:

$$\begin{aligned}\text{FP}\ \ G_A(y,x) &= G_B(y,x)\\ &\quad + \sum_{i=1}^{\infty}(-1)^i \left[\oint d\beta_i \cdots \oint d\beta_1\right] G_B(y,\beta_i)\left[\prod_{k=2}^{i} \overrightarrow{\partial_{\beta_k}} G_B(\beta_k,\beta_{k-1})\right] \overrightarrow{\partial_{\beta_1}} G_B(\beta_1,x),\\ \text{LP}\ \ G_A(y,x) &= G_B(y,x)\\ &\quad + \sum_{i=1}^{\infty}(-1)^i \left[\oint d\beta_i \cdots \oint d\beta_1\right] G_B(y,\beta_i)\overleftarrow{\partial_{\beta_i}}\left[\prod_{k=1}^{i-1} G_B(\beta_{k+1},\beta_k)\overleftarrow{\partial_{\beta_k}}\right] G_B(\beta_1,x).\end{aligned} \tag{3.7.1}$$

The series have the same modes of convergence as those in Proposition 3. We would like to contrast this with double boundary layer ansatz in for example [32], [35], [36], [38]. We provide the following Corollary:

Corollary 1. G_A as SBL or DBL. *The absorbed Green function G_A can be found by a double or single boundary layer:*

$$\begin{aligned} \text{FP } G_A(y,x) &= G_B(y,x) - \int_{\partial D} d\beta\, \mu_{\text{DBL}}(y,\beta)\overrightarrow{\partial_\beta} G_B(\beta,x) \\ \text{LP } G_A(y,x) &= G_B(y,x) - \int_{\partial D} d\beta\, \mu_{\text{SBL}}(y,\beta) G_B(\beta,x) \end{aligned} \tag{3.7.2}$$

with the following definitions of μ_{DBL} and μ_{SBL}:

$$\begin{aligned} \text{FP } \mu_{\text{DBL}}(y,\beta) &= G_B(y,\beta) \\ &\quad + \sum_{i=1}^{\infty}(-1)^i \left[\oint d\beta_i \cdots \oint d\beta_1\right] G_B(y,\beta_i) \left[\prod_{k=2}^{i} \overrightarrow{\partial_{\beta_k}} G_B(\beta_k,\beta_{k-1})\right] \overrightarrow{\partial_{\beta_1}} G_B(\beta_1,\beta) \\ \text{LP } \mu_{\text{SBL}}(y,\beta) &= G_B(y,\beta)\overleftarrow{\partial_\beta} \\ &\quad + \sum_{i=1}^{\infty}(-1)^i \left[\oint d\beta_i \cdots \oint d\beta_1\right] G_B(y,\beta_i)\overleftarrow{\partial_{\beta_i}} \left[\prod_{k=1}^{i-1} G_B(\beta_{k+1},\beta_k)\overleftarrow{\partial_{\beta_k}}\right] G_B(\beta_1,\beta)\overleftarrow{\partial_\beta} \end{aligned} \tag{3.7.3}$$

where the DBL *naturally follows from the first-passage decomposition, and the* SBL *naturally follows from the last-passage decomposition.*

We believe that this is the first time that the absorbed Green function of the Laplace equation is written as a single boundary layer, or last-passage decomposition. What we are trying to emphasise, however, is that the difference between single and double boundary layers is arbitrary. From the starting point x, the first-passage decomposition gives rise to a double boundary layer and the last-passage decomposition gives rise to a single boundary layer. But from y, of course, the opposite holds. Therefore what looks like a first-passage or double boundary layer from the point of view of x, looks like a last passage or single boundary layer from the point of view of y. Thus, rather than celebrating the fact that we can write the absorbed Green function as a single boundary layer, we are trying to drive home the point that the symmetry in time and space ensures that last/first passages and single/double boundary layers are ultimately equivalent.

It appears that the status of single and double boundary layers has remained that of an *ansatz*, even in more modern handbooks on integral equations, such as [13], [33] or [34], where the ansatz of a double boundary layer is always connected to the absorbed Green function, and the ansatz of a single boundary layer is always connected to the reflected Green function. The double boundary layer ansatz was pioneered by Balian & Bloch in [32], who write (p. 412):

> The problem is now to determine the double layer density μ. It is a classical property that the integral equation [for the double boundary layer μ] is non-singular and has, therefore a well defined unique solution. [...] The integral equation may be solved by perturbation, and this yields for the μ the expansion [...]

They obtain only one series for G_A. Their series solution is referred to as the 'multiple reflection expansion'. [36] follow their example of a double boundary layer ansatz, and they notice that it is *not* obvious from their series solution that the absorbed Green function G_A is symmetric, or that it satisfies the boundary conditions, and they suspect that it holds only for smooth domains. To show that the series is indeed symmetric, they suggest the following symmetrisation procedure:

$$\begin{aligned} G_A(y,x) &= G_B(y,x) \\ &+ \sum_{i=1}^{\infty} (-1)^i \left[\oint d\beta_i \cdots \oint d\beta_1 \right] G_B(y,\beta_i) \left[\prod_{k=2}^{i} \frac{1}{2} \overleftrightarrow{\partial}_{\beta_k} G_B(\beta_k, \beta_{k-1}) \right] \frac{1}{2} \overleftrightarrow{\partial}_{\beta_1} G_B(\beta_1, x). \end{aligned} \tag{3.7.4}$$

Operators with arrows both ways work on both sides, i.e. $\overleftrightarrow{\partial} := \overleftarrow{\partial} + \overrightarrow{\partial}$. But this is incorrect because it produces infinite terms. The third term in their expansion, for example, looks like

$$\frac{1}{4} \left[\oint d\beta_2 \oint d\beta_1 \right] G_B(y,\beta_2) \overleftrightarrow{\partial}_{\beta_2} G_B(\beta_2,\beta_1) \overleftrightarrow{\partial}_{\beta_1} G_B(\beta_1, x).$$

This produces 4 terms, one of which is hyper-singular and does not converge:

$$\frac{1}{4} \left[\oint d\beta_2 \oint d\beta_1 \right] G_B(y,\beta_2) \left(\overrightarrow{\partial}_{\beta_2} G_B(\beta_2,\beta_1) \overleftarrow{\partial}_{\beta_1} \right) G_B(\beta_1, x) = \infty.$$

The difference seems subtle, but a quantity that does exist (and that they should have written down) is the following:

$$\frac{1}{4} \oint d\beta_2 \, G_B(y,\beta_2) \overleftrightarrow{\partial}_{\beta_2} \oint d\beta_1 \, G_B(\beta_2,\beta_1) \overleftrightarrow{\partial}_{\beta_2} G_B(\beta_1, x).$$

In this expression, the limit on $G_B(\beta_2,\beta_1)$ where β_2 goes to the boundary occurs *after* the integration over β_1 has already happened — and this makes the result finite. In [36] it is also claimed that the absorbed density can be found as an arbitrary sum of single and double boundary layers, which is not the case. The 'symmetrisation' mistake in [36] is inherited by [38].

The fact that the ansatz-based series solution must be verified after the fact is illustrated in [38], who write (p. 4)

> The validity of this [series] expression can be verified by noting that it fulfils the differential equation for [$x \notin \partial D$]. Moreover, boundary conditions can be checked using [the discontinuity equations], whereby additional contributions give rise to cancellations between successive orders of reflections.

Apart from the equivalence of single and double boundary layers, we further wish to emphasise that our result relies only on the applicability of Green's identity (1.3.1) — allowing a piecewise smooth boundary. The original [32] paper, for example, was subtitled 'Three dimensional problem with smooth boundary surface'. [33] proves that the two dimensional boundary layer problems are solvable for piecewise smooth domains (p. 306), but he concludes (p. 308) that

> The three-dimensional case, where corners, edges and even conical points of all kind may appear, cannot be treated analogously

and thus we conclude that our approach automatically includes the $d \geq 3$ piecewise smooth case — which was previously excluded.

Turning to the Neumann problem, we have found in Proposition 2 that

$$\begin{aligned} \text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\beta\; R(y,t|\beta,\tau) \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} B(\beta,\tau|x,s), \\ \text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\beta\; B(y,t|\beta,\tau) \left\{ \frac{1}{2} \overleftarrow{\partial_\beta} \right\} R(\beta,\tau|x,s). \end{aligned} \tag{3.7.5}$$

Using Lemma 2, we found the following corresponding TP decompositions:

$$\begin{aligned} \text{FR}\quad R(y,t|\beta,s) &= 2B(y,t|\beta,s) + \int_s^t d\tau \oint_{\partial D} d\gamma\; R(y,t|\gamma,\tau) \overrightarrow{\partial_\gamma} B(\gamma,\tau|\beta,s), \\ \text{LR}\quad R(\beta,t|x,s) &= 2B(\beta,t|x,s) + \int_s^t d\tau \oint_{\partial D} d\gamma\; B(\beta,t|\gamma,\tau) \overleftarrow{\partial_\gamma} R(\gamma,\tau|x,s). \end{aligned} \tag{3.7.6}$$

By the same method as for A, we find

Propostion 4. Formal RBM series solution. *The formal solution to problem* (3.3.1) *is given by the following first- and last-reflection series:*

$$\begin{aligned}
\text{FR}\ R(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty}\left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right]\left[\oint d\beta_i\cdots\oint d\beta_1\right] \\
&\quad\times B(y,t|\beta_i,\theta_i)\left[\prod_{k=2}^{i}\overrightarrow{\partial}_{\beta_k}B(\beta_k,\theta_k|\beta_{k-1},\theta_{k-1})\right]\overrightarrow{\partial}_{\beta_1}B(\beta_1,\theta_1|x,s) \\
\text{LR}\ R(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty}\left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right]\left[\oint d\beta_i\cdots\oint d\beta_1\right] \\
&\quad\times B(y,t|\beta_i,\theta_i)\overleftarrow{\partial}_{\beta_i}\left[\prod_{k=1}^{i-1}B(\beta_{k+1},\theta_{k+1}|\beta_k,\theta_k)\overleftarrow{\partial}_{\beta_k}\right]B(\beta_1,\theta_1|x,s)
\end{aligned} \tag{3.7.7}$$

where the FR *and* LR *series are identical, term-by-term, and where the mode of convergence is as follows:*

domain	*mode of convergence*
convex domain	monotone
concave domain	alternating

but where convergence itself is taken for granted.

New about this proposition is 1) the fact that we have derived two series rather than one, 2) the fact the solution is based on a *derivation* rather than on an *ansatz*, i.e. it follows from the intuitive TP decomposition, and 3) its reliance on Green's theorem, allowing piecewise smooth domains, whereas the ansatz approach is only thought appropriate for smooth domains and 4) the claimed mode of convergence. The mode of convergence is opposite to that for the absorbed density, because the factor of $(-1)^i$ is absent in the sum over i. By integrating the series in Proposition 4 over time, we get the reflected Green function, if it exists:

$$\begin{aligned}
\text{FR}\ G_R(y,x) &= G_B(y,x) + \sum_{i=1}^{\infty}\left[\oint d\beta_i\cdots\oint d\beta_1\right] G_B(y,\beta_i)\left[\prod_{k=2}^{i}\overrightarrow{\partial}_{\beta_k}G_B(\beta_k,\beta_{k-1})\right]\overrightarrow{\partial}_{\beta_1}G_B(\beta_1,x) \\
\text{LR}\ G_R(y,x) &= G_B(y,x) + \sum_{i=1}^{\infty}\left[\oint d\beta_i\cdots\oint d\beta_1\right] G_B(y,\beta_i)\overleftarrow{\partial}_{\beta_i}\left[\prod_{k=1}^{i-1}G_B(\beta_{k+1},\beta_k)\overleftarrow{\partial}_{\beta_k}\right]G_B(\beta_1,x)
\end{aligned} \tag{3.7.8}$$

with the same modes of convergence as in Proposition 3, and we have the following Corollary:

Corollary 2. G_R as SBL or DBL. *The reflected Green function G_R, if it exists, can be found by a double or single boundary layer*

$$\begin{aligned} \text{FR } G_R(y,x) &= G_B(y,x) + \int_{\partial D} d\beta\, \mu_{\text{DBL}}(y,\beta)\overrightarrow{\partial_\beta} G_B(\beta,x) \\ \text{FR } G_R(y,x) &= G_B(y,x) + \int_{\partial D} d\beta\, \mu_{\text{SBL}}(y,\beta) G_B(\beta,x) \end{aligned} \tag{3.7.9}$$

with the following definitions of μ_{DBL} and μ_{SBL}:

$$\begin{aligned} \text{FR } \mu_{\text{DBL}}(y,\beta) &= G_B(y,\beta) \\ &\quad + \sum_{i=1}^{\infty} \left[\oint d\beta_i \cdots \oint d\beta_1 \right] G_B(y,\beta_i) \left[\prod_{k=2}^{i} \overrightarrow{\partial_{\beta_k}} G_B(\beta_k,\beta_{k-1}) \right] \overrightarrow{\partial_{\beta_1}} G_B(\beta_1,\beta) \\ \text{LR } \mu_{\text{SBL}}(y,\beta) &= G_B(y,\beta)\overleftarrow{\partial_\beta} \\ &\quad + \sum_{i=1}^{\infty} \left[\oint d\beta_i \cdots \oint d\beta_1 \right] G_B(y,\beta_i)\overleftarrow{\partial_{\beta_i}} \left[\prod_{k=1}^{i-1} G_B(\beta_{k+1},\beta_k)\overleftarrow{\partial_{\beta_k}} \right] G_B(\beta_1,\beta)\overleftarrow{\partial_\beta} \end{aligned} \tag{3.7.10}$$

where the DBL *naturally follows from the first-reflection decomposition, and the* SBL *naturally follows from the last-reflection decomposition.*

Combining Propositions 3 and 4, we get

mode of convergence	absorbed series	reflected series
convex domain	*alternating*	*monotone*
concave domain	*monotone*	*alternating*

Lastly, we need to prove that the first- and last-passage series are identical, term by term. To see why this is the case, we note first that

$$\begin{aligned} \int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_D d\alpha\, B(y,t|\alpha,\tau) B(\alpha,\tau|x,s) &= \left(\lim_{\tau \nearrow t} - \lim_{\tau \searrow s} \right) \int_D d\alpha\, B(y,t|\alpha,\tau) B(\alpha,\tau|x,s), \\ &= B(y,t|x,s) - B(y,t|x,s), \\ &= 0. \end{aligned}$$

It is crucial that both x and y are in the interior, so that both δ-functions pick up a full contribution. With our usual procedure (differentiating under the integral, and using Green's second identity), we find:

$$\int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_D d\alpha\, B(y,t|\alpha,\tau) B(\alpha,\tau|x,s) = \frac{1}{2} \int_s^t d\tau \oint_{\partial D} d\beta\, B(y,t|\beta,\tau) \left\{ \overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta} \right\} B(\beta,\tau|x,s).$$

Thus, in shorthand, we have that

$$\int \oint B \overrightarrow{\partial} B = \int \oint B \overleftarrow{\partial} B \text{ for } x, y \in D.$$

This proves that the first correction-terms are equal. Is it also the case that

$$\int \oint \int \oint B \overrightarrow{\partial} B \overrightarrow{\partial} B \stackrel{?}{=} \int \oint \int \oint B \overleftarrow{\partial} B \overleftarrow{\partial} B.$$

To investigate this, we amend the identity above by

$$\int_s^t d\tau \left(\frac{\partial}{\partial \tau}\right) \int d\alpha B(y,t|\alpha,\tau) B(\alpha,t|x,s) = \begin{cases} 0 & \text{if } x \in D, \quad y \in D; \\ \frac{1}{2} B(y,t|x,s) & \text{if } x \in \partial D, \ y \in D; \\ -\frac{1}{2} B(y,t|x,s) & \text{if } x \in D, \quad y \in \partial D. \end{cases}$$

This follows directly from the fundamental theorem of calculus, the short-time conditions of B, and the fact that a Dirac δ-function on the boundary picks up half a contribution, and therefore we have

$$\int \oint \ B(y,t|\beta,\tau) \left\{ \overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta} \right\} B(\beta,\tau|x,s) = \begin{cases} 0 & \text{if } x \in D, \quad y \in D; \\ B(y,t|x,s) & \text{if } x \in \partial D, \ y \in D; \\ -B(y,t|x,s) & \text{if } x \in D, \quad y \in \partial D. \end{cases}$$

Applying this principle twice, we have

$$\begin{aligned} \int \oint \int \oint B \overrightarrow{\partial} B \overrightarrow{\partial} B &= \int \oint \int \oint B \overleftarrow{\partial} B \overrightarrow{\partial} B + \int \oint B \overrightarrow{\partial} B \\ &= \int \oint \int \oint B \overleftarrow{\partial} B \overleftarrow{\partial} B - \int \oint B \overleftarrow{\partial} B + \int \oint B \overrightarrow{\partial} B \\ &= \int \oint \int \oint B \overleftarrow{\partial} B \overleftarrow{\partial} B. \end{aligned}$$

Here we must be careful not to obtain terms that look like

$$\int \oint \int \oint B \overrightarrow{\partial} B \overleftarrow{\partial} B = \infty.$$

When both x and y are in the interior, the FP and LP series are identical, term by term, but they behave differently when one of the coordinates crosses the boundary. The FP series is continuous in y and discontinuous in x, while for the LP series the opposite holds. In either case, the coordinate connected with B is continuous, and the one connected with ∂B is discontinuous.

3.8 The one dimensional analogy

Our work remains valid for $d = 1$, which is relatively simple given that the outward normal derivatives become simple derivatives, where the sign should be fixed to get the 'outward' direction. Suppose that we have a simple one dimensional Brownian motion, started at zero, and absorbed when it reaches the positive level g, where g is a function of time, $g(0) > 0$ so that absorption does not happen immediately, and $g' > 0$ i.e. g is monotone. For the first passage over g to happen at time t, it is a necessary requirement that the

Brownian motion has not reached the constant level $g(t)$ before time t. (This follows from the fact that g is monotone increasing — and thus $g(t)$ lies above $g(\tau)$ for all $0 \leq \tau < t$. If the constant level $g(t)$ is crossed before time t, then certainly the first passage happened earlier than time t.) The constant level $g(t)$ must therefore be reached for the first time at time t. While this is a necessary requirement for the first passage to happen at t, it is not a sufficient requirement.

The analogy with convex domains should now be obvious. For a convex domain with an absorbing boundary, the Brownian particle cannot cross any of the TPs if it is to stay alive. For the first passage to happen at boundary location β and time t, it is a necessary requirement that the TP at β is crossed for the first time at (β, t). And again this requirement is necessary but not sufficient. Pursuing the one dimensional analogue, we can write

$$\begin{aligned}\mathbb{P}\Big(\tau^{\mathrm{FP}} \in t \big| B_s = x\Big) = \mathbb{P}\Big(&\text{the constant level } g(t) \text{ is crossed for the first time at time } t \big| B_x = x\Big) \\ &- \int_0^t d\tau\, \mathbb{P}\Big(\text{the constant level } g(t) \text{ is crossed for the first time at } t \big| B_\tau = g(\tau)\Big) \\ &\times \mathbb{P}\Big(\tau^{\mathrm{FP}} \in \tau \big| B_s = x\Big).\end{aligned}$$

In more mathematical language, we have

$$\begin{aligned}\left\{\frac{1}{2}\overrightarrow{\partial}_\beta\right\} A(\beta, t|x, s)\Big|_{\beta=g(t)} &= \overrightarrow{\partial}_\beta B(\beta, t|x, s)\Big|_{\beta=g(t)} \\ &\quad - \int_s^t d\tau\, \overrightarrow{\partial}_\beta B(\beta, t|g(\tau), \tau)\Big|_{\beta=g(t)} \left\{\frac{1}{2}\overrightarrow{\partial}_\gamma\right\} A(\gamma, \tau|x, s)\Big|_{\gamma=g(\tau)}\end{aligned}$$

where ∂ is now simple scaled derivative in the inward (in this case downward) direction. The TP decomposition holds for all domains allowing Green's identity — not just convex domains. The interpretation is particularly simple, however, for a convex domain. A similar story holds for the one dimensional case: the interpretation is particularly simple for a monotone increasing boundary g, but the last equation holds for *any* g! (Its derivation proceeds *exactly* like the multidimensional version.)

We have obtained an integral equation for the first-passage density at time t as a function of the first-passage density at all earlier times τ. In fact this integral equation was also obtained by [57], but without this intuition. Neither was it accompanied by its last-passage equivalent, as here.

The idea for solving integral equations like this is by 'successive approximations' as in [34] (p. 811) or equivalently the 'Neumann series' as in [56] (p. 78). The idea is simple: use the left-hand side of the equation as the definition for the unknown quantity appearing on the right-hand side, and do this repetitively to obtain a series solution. For the equation

above, we get:

$$\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s)\big|_{\beta=g(t)} = \overrightarrow{\partial_\beta} B(\beta,t|x,s)\big|_{\beta=g(t)} + \sum_{i=1}^{\infty}(-1)^i\left[\int d\theta_i \cdots \int_{s\le\theta_1\le\cdots\theta_i\le t} d\theta_1\right]\overrightarrow{\partial_\beta} B(\beta,t|g(\theta_i),\theta_i)\big|_{\beta=g(t)} \times\left[\prod_{k=2}^{i}\overrightarrow{\partial_{\beta_i}} B(\beta_k,\theta_k|g(\theta_{k-1}),\theta_{k-1})\big|_{\beta_k=g(\theta_k)}\right]\overrightarrow{\partial_{\beta_1}} B(\beta_1,\theta_1|x,s)\big|_{\gamma=g(\theta_1)} \tag{3.8.1}$$

where

$$\overrightarrow{\partial_\beta} B(\beta,t|x,s) = \frac{\beta-x}{t-s}B(\beta,t|x,s).$$

In particular,

$$\overrightarrow{\partial_\beta} B(\beta,t|\beta,s) = 0$$

and thus the 2 coordinates in each term ∂B need to be on different levels. Suppose for now that g is piecewise constant — suppose it consisted of n increasing levels where $g(t)$ is on level n. Therefore we have that

$$\underbrace{\left[\int\cdots\int\right]}_{n+1 \text{ integrations}} \underbrace{\partial B\cdots\partial B}_{n+2 \text{ terms}} = 0$$

because there is no way that $n+1$ intermediate coordinates could be put on $n+1$ different levels of g when g is piecewise constant with only n levels. And thus it can be seen that the sum in (3.8.1) terminates after n terms, i.e. we get

$$\left\{\frac{1}{2}\overrightarrow{\partial_\beta}\right\} A(\beta,t|x,s)\big|_{\beta=g(t)} = \overrightarrow{\partial_\beta} B(\beta,t|x,s)\big|_{\beta=g(t)} + \sum_{i=1}^{n}(-1)^i\left[\int d\theta_i \cdots \int_{s\le\theta_1\le\cdots\theta_i\le t} d\theta_1\right]\overrightarrow{\partial_\beta} B(\beta,t|g(\theta_i),\theta_i)\big|_{\beta=g(t)} \times\left[\prod_{k=2}^{i}\overrightarrow{\partial_{\beta_i}} B(\beta_k,\theta_k|g(\theta_{k-1}),\theta_{k-1})\big|_{\beta_k=g(\theta_k)}\right]\overrightarrow{\partial_{\beta_1}} B(\beta_1,\theta_1|x,s)\big|_{\gamma=g(\theta_1)} \tag{3.8.2}$$

and the solution is *exact*! If the Brownian particle starts at zero, but g is monotonely *decreasing* rather than increasing (i.e. $g(0) > 0$, and g piecewise constant with n pieces), then a new possibility occurs: the Brownian particle can pass between the steps. The solution is still exact if we are willing to interpret it as the *net* first-passage, where upward first-passages count as positive and downward ones as negative. If the number of steps goes to infinity, then the net first-passage and the upward first-passage will agree — because the opportunity to pass between the steps disappears. Every continuous function can be approximated by more and more steps, and while the series is exact for a finite number of steps, it converges as $n \to \infty$ for *any* continuous function g.

Unfortunately, there is no higher dimensional analogue of this result. In one dimension, the exact result can only be obtained for *one-sided* boundaries. For a one dimensional Brownian motion between 2 constant levels a and b, for example, the sum consists of an infinite number of terms.

The absorbed density can be obtained by subtracting from the free density all paths that have had first passages, i.e.

$$\mathbb{P}_x\Big(B_t = y \text{ and no passages}\big|B_s = x\Big) = B(y,t|x,s) - \int_0^t d\tau\, \mathbb{P}\Big(B_t = y|B_\tau = g(\tau)\Big)\mathbb{P}\Big(\tau^{\text{FP}} \in \tau\big|B_s = x\Big)$$

and thus, schematically

$$A(y,t|x,s) = B(y,t|x,s) + \sum_{i=1}^{\infty}(-1)^i \underbrace{\left[\int\cdots\int\right]}_{i \text{ integrations}} B(y,t|\cdot,\cdot)\left[\underbrace{\overrightarrow{\partial}B\cdots\overrightarrow{\partial}B}_{i-1 \text{ copies}}\right]\overrightarrow{\partial}B(\cdot,\cdot|x,s)$$

If we would have used the last-passage decomposition, we would have obtained

$$A(y,t|x,s) = B(y,t|x,s) + \sum_{i=1}^{\infty}(-1)^i \underbrace{\left[\int\cdots\int\right]}_{i \text{ integrations}} B(y,t|\cdot,\cdot)\overleftarrow{\partial}\left[\underbrace{B\overleftarrow{\partial}\cdots B\overleftarrow{\partial}}_{i-1 \text{ copies}}\right]B(\cdot,\cdot|x,s).$$

In one dimension and with a moving boundary, the FP and LP series are *not* identical: for a monotone boundary g one converges in a monotone fashion and one converges in an alternating fashion.

3.9 Absorbed and reflected transition densities and Feynman-Kac potentials

In this subsection we will derive a new representation and integral equation for A and R. First, recall that we found the following pair of identities:

$$\text{FP}\quad A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau\left(-\frac{\partial}{\partial\tau}\right)\int_D d\alpha\, B(y,t|\alpha,\tau)A(\alpha,\tau|x,s),$$

$$\text{LP}\quad A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau\left(\frac{\partial}{\partial\tau}\right)\int_D d\alpha\, A(y,t|\alpha,\tau)B(\alpha,\tau|x,s).$$

As usual, by using the PDEs of (3.1.1), we have seen that

$$\text{FP}\quad A(y,t|x,s) = B(y,t|x,s) - \frac{\sigma^2}{2}\int_s^t d\tau\int_D d\alpha\, B(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\}A(\alpha,\tau|x,s),$$

$$\text{LP}\quad A(y,t|x,s) = B(y,t|x,s) + \frac{\sigma^2}{2}\int_s^t d\tau\int_D d\alpha\, A(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\}B(\alpha,\tau|x,s).$$

Proceeding as before, we use Green's second identity (1.3.1) — which is valid for domains with a finite number of edges, corners and cusps — to obtain

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ A(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s).
\end{aligned}$$

Here ∂_β is again the scaled inward normal derivative (2.4.1). The BCs of (3.1.1) require that A is zero on the boundary. Now instead of discarding the boundary terms that vanish by the BCs, we may *change their sign* to obtain:

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ A(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s).
\end{aligned} \tag{3.9.1}$$

Using $\overleftrightarrow{\partial} := \overleftarrow{\partial} + \overrightarrow{\partial}$, we may write

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\overleftrightarrow{\partial_\beta}\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ A(y,t|\beta,\tau)\left\{\overleftrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s).
\end{aligned} \tag{3.9.2}$$

By the divergence theorem this equals the following:

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{\sigma^2}{2}\int_s^t d\tau \int_D d\alpha\ \nabla_\alpha^2\Big[B(y,t|\alpha,\tau)A(\alpha,\tau|x,s)\Big],\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{\sigma^2}{2}\int_s^t d\tau \int_D d\alpha\ \nabla_\alpha^2\Big[A(y,t|\alpha,\tau)B(\alpha,\tau|x,s)\Big].
\end{aligned} \tag{3.9.3}$$

Nothing stops us from extending the integration over all of $\mathbb{R}^d$ as long as we also insert an indicator function $\mathbb{1}_{\alpha\in D}$ into the integrand, i.e.

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \mathbb{1}_{\alpha\in D}\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[B(y,t|\alpha,\tau)A(\alpha,\tau|x,s)\Big],\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \mathbb{1}_{\alpha\in D}\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[A(y,t|\alpha,\tau)B(\alpha,\tau|x,s)\Big].
\end{aligned} \tag{3.9.4}$$

Suppose we took a 'smooth approximation' to the indicator function $\mathbb{1}_{\alpha\in D}$. If it were smooth, we would be able to perform an integration by parts. In one dimension, for example, we have

$$\int_{-\infty}^{+\infty} \frac{\partial^2 \mathbb{1}_{a<x<b}}{\partial x^2} f(x)dx = \int_{-\infty}^{+\infty} \mathbb{1}_{a<x<b}\frac{\partial^2 f(x)}{\partial x^2}dx = f'(b) - f'(a) \tag{3.9.5}$$

where two integrations by parts yield no boundary terms because $\mathbb{1}_{a<x<b}$ and $\partial_x \mathbb{1}_{a<x<b}$ both vanish at infinity. In higher dimensions we have by the divergence theorem

$$\int_{\mathbb{R}^d} dx\, \mathbb{1}_{x\in D} \nabla_x^2 f(x) = \int_D dx\, \nabla_x^2 f(x) = \oint_{\partial D} d\beta\, n_\beta \cdot \nabla_\beta f(\beta) \tag{3.9.6}$$

And secondly, by Green's identity, we get that

$$\int_{\mathbb{R}^d} dx\, \mathbb{1}_{x\in D} \left\{ \overleftarrow{\nabla}_x^2 - \overrightarrow{\nabla}_x^2 \right\} f(x) = \int_{\partial\mathbb{R}^d} dx\, \mathbb{1}_{x\in D} \left\{ \overleftarrow{\partial}_x - \overrightarrow{\partial}_x \right\} f(x) = 0 \tag{3.9.7}$$

where this follows from the fact that $\mathbb{1}_{x\in D}$ as well as $\nabla_x \mathbb{1}_{x\in D}$ are zero when evaluated at the 'boundary' of $\mathbb{R}^d$, which is indicated heuristically as $\partial\mathbb{R}^d$. One may object that the divergence theorem is invalid when the integrand blows up in some parts of the domain, but we may take $\mathbb{1}_{x\in D}$ to be a 'bump function'. A bump function equals 1 on D, falls off to 0 outside of D, and does so arbitrarily rapidly while still being smooth.

We conclude that with this 'smooth' interpretation of the indicator function, the integrands of (3.9.4) are smooth as α approaches the boundary of ∂D. But what about the exterior? The free Brownian density B is defined and smooth in the exterior of D, and so is the 'smooth' interpretation of $\mathbb{1}_{\alpha\in D}$. The absorbed Brownian density A, however, is *undefined* in the exterior of D. We have specified that Brownian paths get absorbed, from the inside, as soon as they hit the boundary, but we have not specified what happens afterwards.

Here comes the trick: suppose that we let each Brownian path proceed into the exterior of D after its first passage, after which it is *reflected from the outside*. In essence what we are saying is that the boundary is semi-permeable: it is permeable from the inside, and reflecting from the outside. Seen from the inside, therefore, the boundary is 'absorbing'. As a result of this definition, the density $A(y,t|x,s)$ now exists for all y in the interior as well as exterior of D. It is obvious that the density A is now discontinuous across ∂D: when y approaches ∂D from the inside A goes to zero (because the boundary is absorbing from the inside), but when it approaches ∂D from the outside it does not. The normal derivative, however, is continuous. To see why this is the case, consider a specific boundary location and time (β, t). Paths that have left the domain at some *earlier* time cannot contribute to the normal derivative at (β, t), taken from the outside, because the boundary at β is reflecting. The only paths that can contribute to the normal derivative as taken from the outside, are paths that cross the boundary from the inside for the first time at (β, t). The normal derivative taken from the in- and outside are therefore equal, at each (β, t). We conclude that although A is discontinuous for a semi-permeable boundary ∂D, the normal derivative at the boundary is continuous. This will be important later.

With this 'smooth' interpretation of the indicator function, and with the semi-permeable interpretation of the boundary (such that the normal derivative is continuous), the use of

the divergence theorem can be justified. Using Green's theorem where the boundary terms disappear, the Laplacian now operates on the indicator function as follows:

$$\begin{aligned}
\text{FP}\ A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\Big\{ -\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\Big\} A(\alpha,\tau|x,s), \\
\text{LP}\ A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ A(y,t|\alpha,\tau)\Big\{ -\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\Big\} B(\alpha,\tau|x,s).
\end{aligned} \tag{3.9.8}$$

While it seems that the Laplacian of the Heaviside step function is ill-defined, we may replace it by a *mollifier* $M_\epsilon(\alpha)$ (also known as *approximations to the identity*), which is a smooth approximation to the indicator function $\mathbb{1}_{\alpha\in D}$, where

$$\lim_{\epsilon\searrow 0} M_\epsilon(\alpha) = \mathbb{1}_{\alpha\in D}.$$

With the mollifier M_ϵ we can write the previous identities as

$$\begin{aligned}
\text{FP}\ A(y,t|x,s) &= B(y,t|x,s) - \lim_{\epsilon\searrow 0}\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\Big\{ -\frac{\sigma^2}{2}\nabla_\alpha^2 M_\epsilon(\alpha)\Big\} A(\alpha,\tau|x,s), \\
\text{LP}\ A(y,t|x,s) &= B(y,t|x,s) - \lim_{\epsilon\searrow 0}\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ A(y,t|\alpha,\tau)\Big\{ -\frac{\sigma^2}{2}\nabla_\alpha^2 M_\epsilon(\alpha)\Big\} B(\alpha,\tau|x,s).
\end{aligned} \tag{3.9.9}$$

To make sense of expressions that differentiate step functions, in one or more dimensions, we may either 1) imagine a limiting procedure as above, or 2) perform partial integrations (or Green's theorem) as if everything is well behaved.

We now also show why it is important that the normal derivative across a semi-permeable boundary is relevant. In one dimension, we would normally have that

$$\begin{aligned}
\int_{-\infty}^{\infty} dx\, \delta(x) f(x) &= f(0) \\
\int_{-\infty}^{\infty} dx\, \delta'(x) f(x) &= -f'(0)
\end{aligned}$$

if f is continuous at 0. We suppose the δ-function to be a limit of *even* functions (think e.g. of a standard normal where the variance goes to zero). In that case, if f is discontinuous, the above equalities turn into the following:

$$\begin{aligned}
\int_{-\infty}^{\infty} dx\, \delta(x) f(x) &= \frac{1}{2}f(0+) + \frac{1}{2}f(0-) \\
\int_{-\infty}^{\infty} dx\, \delta'(x) f(x) &= \frac{1}{2}f'(0+) + \frac{1}{2}f'(0-)
\end{aligned}$$

where $f(0+) := \lim_{\epsilon\searrow 0} f(\epsilon)$ and $f(0-) := \lim_{\epsilon\nearrow 0} f(\epsilon)$. If f has a jump at zero, but $f'(0+) = f'(0-)$, then

$$\int_{-\infty}^{\infty} dx\, \delta'(x) f(x) = f'(0+) = f'(0-)$$

As a result, for the determination of A in the interior of D, we may focus only on the normal derivative taken from the interior.

While all this may seem very unhelpful at the moment, we will show in section 5 that a Brownian particle that is allowed in all of $\mathbb{R}^d$ but acted upon by a potential V — which creates or destroys particles according to its sign, and at a rate corresponding to its magnitude — is given by ψ_V where ψ_V satisfies:

$$\begin{aligned}
\text{FI}\quad \psi_V(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\; B(y,t|\alpha,\tau)\, V(\alpha)\, \psi_V(\alpha,\tau|x,s),\\
\text{LI}\quad \psi_V(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\; \psi_V(y,t|\alpha,\tau)\, V(\alpha)\, B(\alpha,\tau|x,s).
\end{aligned} \tag{3.9.10}$$

Here FI and LI denote the first-interaction and last-interaction decompositions. In the physics literature, the LI decomposition is sometimes known as the Dyson equation, see for example [38]. Now we can associate the absorbing potential in either of the following ways:

$$\begin{aligned}
V(\alpha) &:= \lim_{\epsilon \searrow 0} \left(-\tfrac{\sigma^2}{2} \nabla_\alpha^2 M_\epsilon(\alpha) \right)\\
V(\alpha) &:= -\tfrac{\sigma^2}{2} \nabla_\alpha^2 \mathbb{1}_{\alpha \in D}
\end{aligned}$$

where a smooth approximation of this potential is drawn in the rightmost graph in Figure 2. In section 5 we show that positive potentials destroy paths, while negative potentials create paths. In that interpretation, $\lim_{\epsilon \searrow 0} \left(-\frac{\sigma^2}{2} M_\epsilon(\alpha) \right)$ destroys paths arriving at the boundary from the inside and creates paths at the outside. Therefore the interpretation as a semi-permeable boundary, required to ensure a continuous derivative across the boundary, is supported by the intuition!

For the reflected density R, we want to derive a similar result and we may start by recalling the following identities:

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \left(-\frac{\partial}{\partial \tau} \right) \int_D d\alpha\, R(y,t|\alpha,\tau) B(\alpha,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \left(\frac{\partial}{\partial \tau} \right) \int_D d\alpha\, B(y,t|\alpha,\tau) R(\alpha,\tau|x,s).
\end{aligned}$$

As usual we may use the PDEs of (3.3.1) under the integral sign, to get

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \frac{\sigma^2}{2} \int_s^t d\tau \int_D d\alpha\; R(y,t|\alpha,\tau) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} B(\alpha,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) - \frac{\sigma^2}{2} \int_s^t d\tau \int_D d\alpha\; B(y,t|\alpha,\tau) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} R(\alpha,\tau|x,s).
\end{aligned}$$

Using Green's second identity (1.3.1) — which is valid for domains with a finite number of edges, corners and cusps — we obtain

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ R(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} R(\beta,\tau|x,s).
\end{aligned}$$

Now instead of *discarding* the boundary terms that vanish by the BCs, we may change their sign to obtain:

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ R(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}\right\} R(\beta,\tau|x,s).
\end{aligned} \tag{3.9.11}$$

Using the divergence theorem, we have

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_D d\alpha\ \left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[R(y,t|\alpha,\tau)B(\alpha,\tau|x,s)\Big],\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_D d\alpha\ \left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[B(y,t|\alpha,\tau)R(\alpha,\tau|x,s)\Big].
\end{aligned} \tag{3.9.12}$$

Extending the integration over all of space, we have

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \mathbb{1}_{\alpha\in D}\left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[R(y,t|\alpha,\tau)B(\alpha,\tau|x,s)\Big],\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \mathbb{1}_{\alpha\in D}\left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[B(y,t|\alpha,\tau)R(\alpha,\tau|x,s)\Big].
\end{aligned} \tag{3.9.13}$$

We now define the boundary to be reflecting from the inside, and absorbing from the outside. Therefore no particle can escape, if it starts in the interior. The value across the boundary is discontinuous (i.e. R on the inside, and 0 on the outside), but the normal derivative is continuous. With a 'smooth' interpretation of the indicator function, as before, we obtain by applying Green's theorem, where the boundary terms disappear:

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ R(y,t|\alpha,\tau)\left\{\frac{\sigma^2}{2}\nabla_\alpha^2\mathbb{1}_{\alpha\in D}\right\} B(\alpha,\tau|x,s),\\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\left\{\frac{\sigma^2}{2}\nabla_\alpha^2\mathbb{1}_{\alpha\in D}\right\} R(\alpha,\tau|x,s).
\end{aligned} \tag{3.9.14}$$

Comparing with the first- and last-interaction decompositions (3.9.10), we see that can define the reflecting potential as follows:

$$V(\alpha) := \frac{\sigma^2}{2}\nabla_\alpha^2\mathbb{1}_{\alpha\in D}$$

where we realise, again, that to make sense of expressions that differentiate step functions, in one or more dimensions, to obtain the correct answer we may either 1) imagine a limiting procedure, or 2) perform partial integrations (or Green's theorem) as if everything is well behaved. This leads us to the following theorem:

Theorem 1. ABM and RBM through potentials. *For all domains D allowing Green's theorem (1.3.1), for all $x, y \in D$, and for all regular boundary coordinates β, the following formulations of ABM are equivalent:*

$$\left.\begin{aligned} \left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) A(y,t|x,s) &= 0 \\ \left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) A(y,t|x,s) &= 0 \\ A(\beta,t|x,s) &= 0 \\ A(y,t|\beta,s) &= 0 \\ \lim_{s\nearrow t} A(y,t|x,s) &= \delta(|y-x|) \\ \lim_{t\searrow s} A(y,t|x,s) &= \delta(|y-x|) \end{aligned}\right\} = \left\{\begin{aligned} &\text{FP}\ A(y,t|x,s) = B(y,t|x,s) \\ &\quad -\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} A(\alpha,\tau|x,s) \\ &\text{LP}\ A(y,t|x,s) = B(y,t|x,s) \\ &\quad -\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ A(y,t|\alpha,\tau)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} B(\alpha,\tau|x,s) \end{aligned}\right. \tag{3.9.15}$$

Similarly, for all domains D allowing Green's theorem (1.3.1), for all $x, y \in D$, and for all regular boundary coordinates β, the following formulations of RBM are equivalent:

$$\left.\begin{aligned} \left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) R(y,t|x,s) &= 0 \\ \left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) R(y,t|x,s) &= 0 \\ \overrightarrow{\partial_\beta} R(\beta,t|x,s) &= 0 \\ R(y,t|\beta,s)\overleftarrow{\partial_\beta} &= 0 \\ \lim_{s\nearrow t} R(y,t|x,s) &= \delta(|y-x|) \\ \lim_{t\searrow s} R(y,t|x,s) &= \delta(|y-x|) \end{aligned}\right\} = \left\{\begin{aligned} &\text{FR}\ R(y,t|x,s) = B(y,t|x,s) \\ &\quad -\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ R(y,t|\alpha,\tau)\left\{\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} B(\alpha,\tau|x,s) \\ &\text{LR}\ R(y,t|x,s) = B(y,t|x,s) \\ &\quad -\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\left\{\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} R(\alpha,\tau|x,s) \end{aligned}\right. \tag{3.9.16}$$

and thus we can associate the absorbing and reflecting potentials as follows:

$$V(\alpha) := \mp\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}. \tag{3.9.17}$$

This potential V gives rise to a boundary that is semi-permeable: transparent from one side, and reflecting from the other. Furthermore, the perturbation series of the 4 integral equations on the right-hand side produce exactly the first- and last-passage (reflection) series of Propositions 3 and 4, and therefore all *information about ABM or RBM is contained in the potential V.*

This theorem is new, and we believe that it is for the first time that a boundary value problem has been turned into a potential problem like this.

The 'suggestion' for this theorem is in the calculations preceding it: we used all the conditions on the left-hand side of Theorem 1 to obtain the right-hand side, and because

the left-hand side is a well-posed problem, we would *expect* that the same holds for the right-hand side. In some sense, both sides contain the same information and thus should specify the same solution.

Proof. Here we show that the perturbation series, as suggested by the right-hand side of Theorem 1, exactly matches the first- and last-passage (and reflection) series of Propositions 3 and 4. Consider

$$\text{FP } A(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha \; B(y,t|\alpha,\tau) \left\{ -\frac{\sigma^2}{2} \nabla_\alpha^2 \mathbb{1}_{\alpha \in D} \right\} A(\alpha,\tau|x,s).$$

Substitute the equation back into itself to obtain

$$\begin{aligned} \text{FP } A(y,t|x,s) = {} & B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha \; B(y,t|\alpha,\tau) \left\{ -\frac{\sigma^2}{2} \nabla_\alpha^2 \mathbb{1}_{\alpha \in D} \right\} B(\alpha,\tau|x,s) \\ & + \int_s^t d\tau_2 \int_{\mathbb{R}^d} d\alpha_2 \; B(y,t|\alpha_2,\tau_2) \left\{ -\frac{\sigma^2}{2} \nabla_{\alpha_2}^2 \mathbb{1}_{\alpha_2 \in D} \right\} \\ & \times \int_s^{\tau_2} d\tau_1 \int_{\mathbb{R}^d} d\alpha_1 \; B(\alpha_2,\tau_2|\alpha_1,\tau_1) \left\{ -\frac{\sigma^2}{2} \nabla_{\alpha_1}^2 \mathbb{1}_{\alpha_1 \in D} \right\} A(\alpha_1,\tau_1|x,s) \end{aligned}$$

Therefore the first correction term is as follows:

$$\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha \; B(y,t|\alpha,\tau) \left\{ -\frac{\sigma^2}{2} \nabla_\alpha^2 \mathbb{1}_{\alpha \in D} \right\} B(\alpha,\tau|x,s).$$

A second substitution would show that the second correction term reads:

$$\begin{aligned} & \int_s^t d\tau_2 \int_{\mathbb{R}^d} d\alpha_2 \; B(y,t|\alpha_2,\tau_2) \left\{ -\frac{\sigma^2}{2} \nabla_{\alpha_2}^2 \mathbb{1}_{\alpha_2 \in D} \right\} \\ & \times \int_s^{\tau_2} d\tau_1 \int_{\mathbb{R}^d} d\alpha_1 \; B(\alpha_2,\tau_2|\alpha_1,\tau_1) \left\{ -\frac{\sigma^2}{2} \nabla_{\alpha_1}^2 \mathbb{1}_{\alpha_1 \in D} \right\} B(\alpha_1,\tau_1|x,s). \end{aligned}$$

Consider the first correction term. Recall that

$$\int_s^t d\tau \left(\frac{\partial}{\partial \tau} \right) \int_D d\alpha \; B(y,t|\alpha,\tau) B(\alpha,t|x,s) = \begin{cases} 0 & \text{if } x \in D, \quad y \in D; \\ \frac{1}{2} B(y,t|x,s) & \text{if } x \in \partial D, \; y \in D; \\ -\frac{1}{2} B(y,t|x,s) & \text{if } x \in D, \quad y \in \partial D. \end{cases}$$

This implies that

$$\int_s^t d\tau \oint_{\partial D} d\beta \; B(y,t|\beta,\tau) \left\{ \overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta} \right\} B(\beta,\tau|x,s) = \begin{cases} 0 & \text{if } x \in D, \quad y \in D; \\ B(y,t|x,s) & \text{if } x \in \partial D, \; y \in D; \\ -B(y,t|x,s) & \text{if } x \in D, \quad y \in \partial D. \end{cases} \tag{3.9.18}$$

Using this to analyse the first correction term, we find that

$$\begin{aligned}
&\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} B(\alpha,\tau|x,s) \\
&= \int_s^t d\tau \int_D d\alpha\ \left\{-\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[B(y,t|\alpha,\tau)B(\alpha,\tau|x,s)\Big] \\
&= \int_s^t d\tau \int_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s) \\
&= \int_s^t d\tau \int_{\partial D} d\beta\ B(y,t|\beta,\tau)\left\{\overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s)
\end{aligned}$$

which indeed equals the first correction term in Proposition 3. For the second correction term, we need to consider

$$\begin{aligned}
&\int_s^t d\tau_2 \int_{\mathbb{R}^d} d\alpha_2\ B(y,t|\alpha_2,\tau_2)\left\{-\frac{\sigma^2}{2}\nabla_{\alpha_2}^2 \mathbb{1}_{\alpha_2\in D}\right\} \\
&\times \int_s^{\tau_2} d\tau_1 \int_{\mathbb{R}^d} d\alpha_1\ B(\alpha_2,\tau_2|\alpha_1,\tau_1)\left\{-\frac{\sigma^2}{2}\nabla_{\alpha_1}^2 \mathbb{1}_{\alpha_1\in D}\right\} B(\alpha_1,\tau_1|x,s)
\end{aligned}$$

where the order of differentiation and integration is very important. Our distributional definition of the Laplacian of the indicator is that it works from the inside. Therefore we may assume $\alpha_2 \in D$. By the analysis of the first term, we therefore obtain

$$\int_s^t d\tau_2 \int_{\mathbb{R}^d} d\alpha_2\ B(y,t|\alpha_2,\tau_2)\left\{-\frac{\sigma^2}{2}\nabla_{\alpha_2}^2 \mathbb{1}_{\alpha_2\in D}\right\} \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_1\ B(\alpha_2,\tau_2|\beta_1,\tau_1)\left\{\overrightarrow{\partial_{\beta_1}}\right\} B(\beta_1,\tau_1|x,s)$$

and proceeding we get that

$$\int_s^t d\tau_2 \oint_{\partial D} d\beta_2\ B(y,t|\beta_2,\tau_2)\left\{\frac{1}{2}\overleftrightarrow{\partial_{\beta_2}}\right\} \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_1\ B(\beta_2,\tau_2|\beta_1,\tau_1)\left\{\overrightarrow{\partial_{\beta_1}}\right\} B(\beta_1,\tau_1|x,s)$$

which, when written out, becomes:

$$\begin{aligned}
&\int_s^t d\tau_2 \oint_{\partial D} d\beta_2\ B(y,t|\beta_2,\tau_2)\left\{\frac{1}{2}\overleftarrow{\partial_{\beta_2}}\right\} \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_1\ B(\beta_2,\tau_2|\beta_1,\tau_1)\left\{\overrightarrow{\partial_{\beta_1}}\right\} B(\beta_1,\tau_1|x,s) \\
+ &\int_s^t d\tau_2 \oint_{\partial D} d\beta_2\ B(y,t|\beta_2,\tau_2)\left\{\frac{1}{2}\overrightarrow{\partial_{\beta_2}}\right\} \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_1\ B(\beta_2,\tau_2|\beta_1,\tau_1)\left\{\overrightarrow{\partial_{\beta_1}}\right\} B(\beta_1,\tau_1|x,s).
\end{aligned}$$

By Lemma 1 we can push the differential operator through the integral in the second term,

to obtain

$$\int_s^t d\tau_2 \oint_{\partial D} d\beta_2 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overleftarrow{\partial}_{\beta_2}\right\} \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_1 \; B(\beta_2,\tau_2|\beta_1,\tau_1) \left\{\overrightarrow{\partial}_{\beta_1}\right\} B(\beta_1,\tau_1|x,s)$$
$$+ \int_s^t d\tau_2 \oint_{\partial D} d\beta_2 \; B(y,t|\beta_2,\tau_2) \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_1 \; \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|\beta_1,\tau_1) \left\{\overrightarrow{\partial}_{\beta_1}\right\} B(\beta_1,\tau_1|x,s)$$
$$- \int_s^t d\tau_2 \oint_{\partial D} d\beta_2 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|x,s).$$

There are now no more differentiations that are pointing through integral operators, so we may finally pull all the integrals towards the left, to obtain

$$\int_s^t d\tau_2 \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_2 \oint_{\partial D} d\beta_1 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overleftarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|\beta_1,\tau_1) \left\{\overrightarrow{\partial}_{\beta_1}\right\} B(\beta_1,\tau_1|x,s)$$
$$+ \int_s^t d\tau_2 \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_2 \oint_{\partial D} d\beta_1 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|\beta_1,\tau_1) \left\{\overrightarrow{\partial}_{\beta_1}\right\} B(\beta_1,\tau_1|x,s)$$
$$- \int_s^t d\tau_2 \oint_{\partial D} d\beta_2 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|x,s).$$

Changing the direction of an arrow in the first term and taking into account (3.9.18), we obtain

$$\int_s^t d\tau_2 \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_2 \oint_{\partial D} d\beta_1 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|\beta_1,\tau_1) \left\{\overrightarrow{\partial}_{\beta_1}\right\} B(\beta_1,\tau_1|x,s)$$
$$+ \int_s^t d\tau_2 \oint_{\partial D} d\beta_2 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|x,s)$$
$$+ \int_s^t d\tau_2 \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_2 \oint_{\partial D} d\beta_1 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|\beta_1,\tau_1) \left\{\overrightarrow{\partial}_{\beta_1}\right\} B(\beta_1,\tau_1|x,s)$$
$$- \int_s^t d\tau_2 \oint_{\partial D} d\beta_2 \; B(y,t|\beta_2,\tau_2) \left\{\frac{1}{2}\overrightarrow{\partial}_{\beta_2}\right\} B(\beta_2,\tau_2|x,s).$$

Two terms cancel, and thus we finally obtain that the second correction term equals

$$\int_s^t d\tau_2 \int_s^{\tau_2} d\tau_1 \oint_{\partial D} d\beta_2 \oint_{\partial D} d\beta_1 \; B(y,t|\beta_2,\tau_2) \overrightarrow{\partial}_{\beta_2} B(\beta_2,\tau_2|\beta_1,\tau_1) \overrightarrow{\partial}_{\beta_1} B(\beta_1,\tau_1|x,s).$$

And thus we have shown that the first two terms in the expansion of the integral equation on the right of Theorem 1 are equal to those in Propositions 3 and 4. We can proceed in this spirit: 1) using Lemma 1 for pushing differentials through integrals, and 2) using (3.9.18) for changing direction of arrows – and proceeding like this it is not hard to see that *all* terms are equal and we reproduce the first-passage series. We thus conclude that the right-hand side of Theorem 1 uniquely specifies the same solution that the left-hand side specifies. □

3.10 Green functions and spectral theory

The new formulation in the previous subsection suggests a further extension to current methods. As before, we may start with the identities

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) - \frac{\sigma^2}{2}\int_D d\alpha\, G_B(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} G_A(\alpha,x),\\
\text{LP}\ G_A(y,x) &= G_B(y,x) + \frac{\sigma^2}{2}\int_D d\alpha\, G_A(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} G_B(\alpha,x).
\end{aligned} \tag{3.10.1}$$

As usual we may apply Green's second identity (1.3.1) to get

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) + \frac{1}{2}\oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta\right\} G_A(\beta,x),\\
\text{LP}\ G_A(y,x) &= G_B(y,x) - \frac{1}{2}\oint_{\partial D} d\beta\, G_A(y,\beta)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta\right\} G_B(\beta,x).
\end{aligned} \tag{3.10.2}$$

Because G_A disappears on the boundary, some terms in the above are zero. We may change their signs to obtain

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) - \frac{1}{2}\oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\overleftarrow{\partial}_\beta + \overrightarrow{\partial}_\beta\right\} G_A(\beta,x),\\
\text{LP}\ G_A(y,x) &= G_B(y,x) - \frac{1}{2}\oint_{\partial D} d\beta\, G_A(y,\beta)\left\{\overleftarrow{\partial}_\beta + \overrightarrow{\partial}_\beta\right\} G_B(\beta,x).
\end{aligned} \tag{3.10.3}$$

By the divergence theorem, this turns into

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) + \int_D d\alpha\, \left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[G_B(y,\alpha)G_A(\alpha,x)\Big],\\
\text{LP}\ G_A(y,x) &= G_B(y,x) + \int_D d\alpha\, \left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[G_A(y,\beta)G_B(\beta,x)\Big].
\end{aligned} \tag{3.10.4}$$

We may proceed as previously to obtain

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) - \int_{\mathbb{R}^d} d\alpha\, G_B(y,\alpha)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} G_A(\alpha,x),\\
\text{LP}\ G_A(y,x) &= G_B(y,x) - \int_{\mathbb{R}^d} d\alpha\, G_A(y,\beta)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} G_B(\beta,x).
\end{aligned} \tag{3.10.5}$$

Instead, we may apply the Laplacian on the integrand to obtain

$$\begin{aligned}
\text{FP}\ G_A(y,x) &= G_B(y,x) - G_B(y,x) - G_A(y,x)\\
&\qquad + \sigma^2\int_D d\alpha\, G_B(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha\cdot\overrightarrow{\nabla}_\alpha\right\} G_A(\alpha,x),\\
\text{LP}\ G_A(y,x) &= G_B(y,x) - G_B(y,x) - G_A(y,x)\\
&\qquad + \sigma^2\int_D d\alpha\, G_A(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha\cdot\overrightarrow{\nabla}_\alpha\right\} G_B(\alpha,x).
\end{aligned} \tag{3.10.6}$$

Thus we find

$$\begin{aligned} \text{FP } G_A(y,x) &= \frac{\sigma^2}{2}\int_D d\alpha\, G_B(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha\right\} G_A(\alpha,x), \\ \text{LP } G_A(y,x) &= \frac{\sigma^2}{2}\int_D d\alpha\, G_A(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha\right\} G_B(\alpha,x). \end{aligned} \tag{3.10.7}$$

This integro-differential equation for G_A is new. This means that the Dirichlet Green function is an eigenfunction of the operator

$$\begin{aligned} G_A(y,x) &= K * G_A(y,x), \\ G_A(y,x) &= G_A(y,x) * K, \end{aligned}$$

where the forward or backward operation of K is defined as

$$\begin{aligned} K * f(y,x) &= \frac{\sigma^2}{2}\int_D d\alpha\, \nabla_\alpha G_B(y,\alpha)\cdot\nabla_\alpha f(\alpha,x), \\ f(y,x) * K &= \frac{\sigma^2}{2}\int_D d\alpha\, \nabla_\alpha f(y,\alpha)\cdot\nabla_\alpha G_B(\alpha,x). \end{aligned}$$

This suggests a new repetitive method of finding G_A, where we start with a trial function (for example G_B), and apply the operator K repeatedly, on either the left or right. To be rigorous we would have to check that the operator is, for example, compact, but we will not go into these details. See for example [56], [13] or [33].

Instead we take the rather pragmatic approach by 1) taking for granted that the Green function exists and is unique, 2) that in the derivation of our integral equation we have used all the conditions that are supposed to specify it, and that, thus 3) we expect our series expansions or repetitive applications to converge. And if it does, then the answer must be correct. Further, it must be the solution to the modified problem, since we have only used Green's theorem when considering the domain.

We note that the differentiation and integration now concern the interior of the domain rather than the boundary. While for practical purposes this might be a disadvantage because it leads to d dimensional integrals rather than $d-1$ dimensional integrals, it might be an advantage theoretically. The reason is that it shows that changing a single boundary location (making it irregular, for example) should have little effect if the change on the volume as a whole is negligible. We expect the integration over the volume to be somewhat more robust, in some sense, when it comes to irregular boundary points.

A famous result of spectral theory (see for example [39], [40], [52] or [56]) states that the absorbed propagator A can be written as

$$A(y,t|x,s) = \sum_{i=1}^{\infty} e^{-\lambda_i(t-s)}\phi_i(y)\phi_i(x), \tag{3.10.8}$$

where λ_i are positive eigenvalues satisfying $\lambda_1 \leq \lambda_2 \leq \cdots \lambda_i \leq \cdots$ and ϕ_i are the eigenfunctions, satisfying

$$\frac{\sigma^2}{2}\nabla_y^2\phi_i(y) = -\lambda_i\phi_i(y), \tag{3.10.9}$$

and where the eigenfunctions disappear on the boundary. In a closed domain with an absorbing boundary, a Brownian path is eventually absorbed with probability one. The probability for the path to stay alive decays exponentially, with the 'ground-state' ϕ_1 surviving longest, because λ_1 is the smallest eigenvalue.

Now suppose, however, that while the Brownian particle is still alive there is a probability of $\lambda\,dt$, in each period of time dt, that another particle is created at the location of the first particle. In a sense, the probabilistic weight of the particle is doubled. And upon further interactions, it may double again to weight 4. If λ is relatively small then some 'weight' will be created but eventually the particle will be absorbed by the boundary. But if λ exceeds a certain critical value then the 'weight' of the particle that stays alive gets multiplied and multiplied further, and starts to dominate. It turns out that this critical λ is the first eigenvalue of the Dirichlet problem. In particular,

$$\lim_{t\nearrow\infty} e^{\lambda\,(t-s)}A(y,t|x,s) = \begin{cases} 0 & \text{if } \lambda < \lambda_1, \\ \phi_1(y)\phi_1(x) & \text{if } \lambda = \lambda_1, \\ \infty & \text{if } \lambda > \lambda_1. \end{cases} \tag{3.10.10}$$

At the location $x = y$, A is decreasing for all time. It turns out that if the derivative of $e^{\lambda\,(t-s)}A(x,t|x,s)$ with respect to t is *ever* positive, then it will explode to infinity when $t \to \infty$. Thus if

$$\frac{\partial}{\partial t}\left(e^{\lambda\,(t-s)}A(x,t|x,s)\right) = \lambda\left(e^{\lambda\,(t-s)}A(x,t|x,s)\right) + \left(e^{\lambda\,(t-s)}\frac{\partial}{\partial t}A(x,t|x,s)\right) \tag{3.10.11}$$

is ever zero or positive for some t, then $\lambda > \lambda_1$. Thus we need to find the smallest λ for which we have that

$$\frac{\partial}{\partial t}\left(e^{\lambda\,(t-s)}A(x,t|x,s)\right) = 0 \Rightarrow\ \lambda A(x,t|x,s) + \frac{\partial}{\partial t}A(x,t|x,s) = 0 \tag{3.10.12}$$

It is obvious that A is positive and that $\partial_t A(x,t|x,s)$ is negative for all t. Now the smallest λ such that $\lambda A + \partial_t A = 0$ for some t is equal to the first eigenvalue λ_1. Since we can calculate A as a series, we may be able to use this to find the first eigenvalue of a general domain D — which would be a new result.

For a reflected Brownian motion, the same decomposition (3.10.8) is possible, except the first eigenvalue is zero: i.e. after a long time R reaches an equilibrium distribution given by the ground state that does not decay away (as no particle can escape).

The situation of particle creation in D and absorption at the boundary can be specified by the potential

$$V(\alpha) := -\frac{\sigma^2}{2}\nabla_\alpha^2\mathbb{1}_{\alpha\in D} - \lambda\mathbb{1}_{\alpha\in D}$$

where a negative (constant) potential creates paths at rate λ. With this potential, we get that the Green function must satisfy that

$$\begin{aligned}
\text{FP}\quad G_A(y,x) &= G_B(y,x) - \int_{\mathbb{R}^d} d\alpha\; G_B(y,\alpha)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D} - \lambda \mathbb{1}_{\alpha\in D}\right\} G_A(\alpha,x),\\
\text{LP}\quad G_A(y,x) &= G_B(y,x) - \int_{\mathbb{R}^d} d\alpha\; G_A(y,\beta)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D} - \lambda \mathbb{1}_{\alpha\in D}\right\} G_B(\beta,x).
\end{aligned} \tag{3.10.13}$$

Proceeding as before, we find that the integro-differential equation above can be re-derived in the setting with particle creation at rate λ, to give

$$\begin{aligned}
G_A(y,x) &= \int_D d\alpha\, G_A(y,\alpha)\left\{\frac{\sigma^2}{2}\overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha + \lambda\right\} G_B(\alpha,x),\\
G_A(y,x) &= \int_D d\alpha\, G_B(y,\alpha)\left\{\frac{\sigma^2}{2}\overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha + \lambda\right\} G_A(\alpha,x).
\end{aligned} \tag{3.10.14}$$

Originally we expected a repeated application of the integro-differential operator to give rise to a convergent answer, but this is no longer true for $\lambda > \lambda_{\text{critical}} = \lambda_1$, where λ_1 is the first eigenvalue of the Dirichlet problem. The Green function can be seen as the expected time spent at a certain location, before absorption, and doubled in weight by the number of interactions that occur at rate λ in D. Therefore the integro-differential equation above has no solutions for $\lambda > \lambda_1$.

For the reflected Green function we may proceed similarly, and start with the identities

$$\begin{aligned}
\text{FR}\quad G_R(y,x) &= G_B(y,x) + \frac{\sigma^2}{2}\int_D d\alpha\, G_R(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} G_B(\alpha,x),\\
\text{LR}\quad G_R(y,x) &= G_B(y,x) - \frac{\sigma^2}{2}\int_D d\alpha\, G_B(y,\alpha)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} G_R(\alpha,x).
\end{aligned} \tag{3.10.15}$$

As usual we may apply Green's second identity (1.3.1) to get

$$\begin{aligned}
\text{FR}\quad G_R(y,x) &= G_B(y,x) - \frac{1}{2}\oint_{\partial D} d\beta\, G_R(y,\beta)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} G_B(\beta,x),\\
\text{LR}\quad G_R(y,x) &= G_B(y,x) + \frac{1}{2}\oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta}\right\} G_R(\beta,x).
\end{aligned} \tag{3.10.16}$$

Because ∂G_R disappears on the boundary, some terms in the above are zero. We may change their signs to obtain

$$\begin{aligned}
\text{FR}\quad G_R(y,x) &= G_B(y,x) + \frac{1}{2}\oint_{\partial D} d\beta\, G_R(y,\beta)\left\{\overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}\right\} G_B(\beta,x),\\
\text{LR}\quad G_R(y,x) &= G_B(y,x) + \frac{1}{2}\oint_{\partial D} d\beta\, G_B(y,\beta)\left\{\overleftarrow{\partial_\beta} + \overrightarrow{\partial_\beta}\right\} G_A(\beta,x).
\end{aligned} \tag{3.10.17}$$

By the divergence theorem, this turns into

$$\begin{aligned}
\text{FR}\quad G_R(y,x) &= G_B(y,x) - \int_D d\alpha\, \left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[G_R(y,\alpha)G_B(\alpha,x)\Big],\\
\text{LR}\quad G_R(y,x) &= G_B(y,x) - \int_D d\alpha\, \left\{\frac{\sigma^2}{2}\nabla_\alpha^2\right\}\Big[G_B(y,\beta)G_R(\beta,x)\Big].
\end{aligned} \tag{3.10.18}$$

We may proceed as previously to obtain

$$\begin{aligned}
\text{FR } G_R(y,x) &= G_B(y,x) + G_B(y,x) + G_R(y,x) \\
&\quad - \sigma^2 \int_D d\alpha\, G_R(y,\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_B(\alpha,x), \\
\text{LR } G_R(y,x) &= G_B(y,x) + G_B(y,x) + G_R(y,x) \\
&\quad - \sigma^2 \int_D d\alpha\, G_B(y,\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_R(\alpha,x).
\end{aligned} \tag{3.10.19}$$

Thus we find

$$\begin{aligned}
\text{FR } G_B(y,x) &= -\frac{\sigma^2}{2} \int_D d\alpha\, G_R(y,\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_B(\alpha,x), \\
\text{LR } G_B(y,x) &= -\frac{\sigma^2}{2} \int_D d\alpha\, G_B(y,\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_R(\alpha,x),
\end{aligned} \tag{3.10.20}$$

where this integral equation is new. Note that unlike in the absorbed case, we cannot expect to find G_R by a repeated application of a certain operator, because G_B rather than G_R appears on the left-hand side. Rather, G_B (which we already know) is an eigenfunction of the operator involving G_R (which we do not know)!

3.11 An application to the Dirichlet and Neumann boundary value problems

The modified Dirichlet solution is defined by a weighted expectation over all first passages, where the weight is given by w, i.e.

$$D(x) := \oint_{\partial D} d\beta\, w(\beta) \left\{ \frac{1}{2} \overrightarrow{\partial}_\beta \right\} G_A(\beta,x). \tag{3.11.1}$$

Because G_A is zero on the boundary, we may subtract a term that is zero, i.e.

$$D(x) = -\oint_{\partial D} d\beta\, w(\beta) \left\{ \frac{1}{2} \overleftarrow{\partial}_\beta - \frac{1}{2} \overrightarrow{\partial}_\beta \right\} G_A(\beta,x). \tag{3.11.2}$$

By Green's second identity this becomes

$$D(x) = \frac{\sigma^2}{2} \int_D d\alpha\, w(\alpha) \left\{ \overleftarrow{\nabla}^2_\alpha - \overrightarrow{\nabla}^2_\alpha \right\} G_A(\alpha,x). \tag{3.11.3}$$

And because $\frac{\sigma^2}{2}\nabla^2 G = -\delta$, we get

$$D(x) = w(x) + \frac{\sigma^2}{2} \int_D d\alpha\, \nabla^2_\alpha w(\alpha)\, G_A(\alpha,x). \tag{3.11.4}$$

This last representation shows very clearly that the Dirichlet solution satisfies

$$\begin{aligned}
\lim_{x\to\beta} D(x) &= w(\beta), \\
\nabla^2_x D(x) &= 0.
\end{aligned} \tag{3.11.5}$$

For the first representation of the Dirichlet solution, w was only required to be defined on ∂D. For the last representation to make sense, we need that w is defined in all of D. In fact it turns out that it is irrelevant how w is extended to the entirety of D, as long it is twice differentiable and matches the correct boundary data at $w(\beta)$. Alternatively, we could write

$$D(x) = \oint_{\partial D} d\beta\; w(\beta) \left\{ \frac{1}{2}\overrightarrow{\partial}_\beta \right\} G_A(\beta, x) = \oint_{\partial D} d\beta\; w(\beta) \left\{ \frac{1}{2}\overleftarrow{\partial}_\beta + \frac{1}{2}\overrightarrow{\partial}_\beta \right\} G_A(\beta, x), \tag{3.11.6}$$

which, by the divergence theorem, becomes

$$\begin{aligned} D(x) &= -\int_D d\alpha\; \left\{ \frac{\sigma^2}{2}\nabla_\alpha^2 \right\} \Big[w(\alpha) G_A(\alpha, x) \Big], \\ &= w(x) - \int_D d\alpha\; w(\alpha) \left\{ \frac{\sigma^2}{2}\overleftrightarrow{\nabla}_\alpha^2 \right\} G_A(\alpha, x) - \sigma^2 \int_D d\alpha\; w(\alpha) \left\{ \overleftrightarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_A(\alpha, x). \end{aligned} \tag{3.11.7}$$

Compare this with the previous representation, to find that

$$D(x) = w(x) - \frac{\sigma^2}{2} \int_D d\alpha\; w(\alpha) \left\{ \overleftrightarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_A(\alpha, x). \tag{3.11.8}$$

This representation is new, and taking $w(x) = D(x)$, we obtain another new observation:

$$0 = \frac{\sigma^2}{2} \int_D d\alpha\; D(\alpha) \left\{ \overleftrightarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_A(\alpha, x) \tag{3.11.9}$$

The exterior Neumann solution in $d \geq 3$ is defined by taking a weighted average over all boundary visits, i.e.

$$N(x) := -\oint_{\partial D} d\beta\; w(\beta) \left\{ \frac{1}{2}\overleftarrow{\partial}_\beta \right\} G_R(\beta, x). \tag{3.11.10}$$

We choose the 'exterior' because then G_R exists and can be defined as we have done. We choose the weight to be given by $-w(\beta)\left\{\frac{1}{2}\overleftarrow{\partial}_\beta\right\}$ for symmetry reasons which will become clear. Because ∂G_R is zero on the boundary, we add a term that is zero, i.e.

$$N(x) = -\oint_{\partial D} d\beta\; w(\beta) \left\{ \frac{1}{2}\overleftarrow{\partial}_\beta - \frac{1}{2}\overrightarrow{\partial}_\beta \right\} G_R(\beta, x). \tag{3.11.11}$$

By Green's second identity this becomes

$$N(x) = \frac{\sigma^2}{2} \int_D d\alpha\; w(\alpha) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} G_R(\alpha, x). \tag{3.11.12}$$

And because $\frac{\sigma^2}{2}\nabla^2 G = -\delta$, we get

$$N(x) = w(x) + \frac{\sigma^2}{2} \int_D d\alpha\; \nabla_\alpha^2 w(\alpha)\, G_R(\alpha, x). \tag{3.11.13}$$

The operator ∂_β commutes with the integration over the interior, and thus the Neumann solution satisfies

$$\begin{aligned} \partial_\beta N(\beta) &= \partial_\beta w(\beta), \\ \nabla_x^2 N(x) &= 0. \end{aligned} \tag{3.11.14}$$

Alternatively, we could subtract a term that is zero:

$$N(x) = -\oint\limits_{\partial D} d\beta\; w(\beta) \left\{ \frac{1}{2}\overleftarrow{\partial}_\beta \right\} G_R(\beta, x) = -\oint\limits_{\partial D} d\beta\; w(\beta) \left\{ \frac{1}{2}\overleftarrow{\partial}_\beta + \frac{1}{2}\overrightarrow{\partial}_\beta \right\} G_R(\beta, x). \tag{3.11.15}$$

By the divergence theorem this becomes

$$\begin{aligned} N(x) &= \int\limits_D d\alpha\; \left\{ \frac{\sigma^2}{2}\nabla_\alpha^2 \right\} \Big[w(\alpha) G_R(\alpha, x) \Big], \\ N(x) &= -w(x) + \int\limits_D d\alpha\; w(\alpha) \left\{ \frac{\sigma^2}{2}\overleftarrow{\nabla}_\alpha^2 \right\} G_R(\alpha, x) + \sigma^2 \int\limits_D d\alpha\; w(\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_R(\alpha, x). \end{aligned} \tag{3.11.16}$$

Compare this with the previous representation, to find

$$w(x) = \frac{\sigma^2}{2} \int\limits_D d\alpha\; w(\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_R(\alpha, x). \tag{3.11.17}$$

This comes as a surprise since the Neumann solution N has completely disappeared. The last equation says that *any* twice differentiable function $w(x)$ can be expressed by an integration over a domain of choice, over $w(\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_R(\alpha, x)$. The implications of this observation are yet unclear, and we leave this issue for the time being — but we note that D needs to be unbounded in $d \geq 3$ for G_R to be defined at all. Taking $w = N$ we find that

$$N(x) = \frac{\sigma^2}{2} \int\limits_D d\alpha\; N(\alpha) \left\{ \overleftarrow{\nabla}_\alpha \cdot \overrightarrow{\nabla}_\alpha \right\} G_R(\alpha, x). \tag{3.11.18}$$

Some of the (new) observations in this subsection may be useful, and some may not be. The main point, however, is that we have

$$\begin{aligned} D(x) &= w(x) + \frac{\sigma^2}{2} \int\limits_D d\alpha\; \nabla_\alpha^2 w(\alpha)\, G_A(\alpha, x) \\ N(x) &= w(x) + \frac{\sigma^2}{2} \int\limits_D d\alpha\; \nabla_\alpha^2 w(\alpha)\, G_R(\alpha, x) \end{aligned} \tag{3.11.19}$$

and thus our series solutions for G_A and G_R can be directly plugged into these expressions to find the Dirichlet or Neumann solution. Furthermore, the resulting series for $D(x)$ and $N(x)$ inherit the mode of convergence of G_A or G_R. And lastly, if we can find a compact way to express G_A and/or G_R in a way that contains all the information that is contained in its series expression, then we will automatically also have found a very compact representation for the Dirichlet problem.

4 Examples

This section shows the convergence of the obtained series solution in practice. We calculate the transition-density of an ABM or RBM using the program *Mathematica* [58] to calculate the successive terms in the series.

4.1 An ellipse in 2d

First, we define an ellipse in polar coordinates:

```
a = 2; b = 1;
r[φ_] := a b / Sqrt[a^2 Sin[φ]^2 + b^2 Cos[φ]^2]
```

Then we define the volatility and space-time positions as follows:

```
σ = 1;
s = 0; x = {-1/3, 0};
t = 2; y = {1/3, 3/4};
```

We will need to go back and forth between Cartesian and spherical coordinates, and therefore we will need:

```
Radius[vector_] := Sqrt[vector[[1]]^2 + vector[[2]]^2]
Phi[vector_] := N[ArcTan[vector[[1]], vector[[2]]]]
```

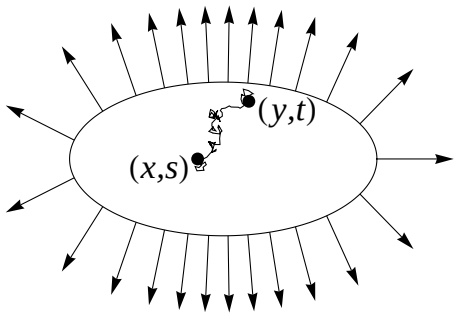
The surface- coordinate, tangent vector, outward normal, infinitesimal surface area and scaled outward normal can be defined by:

```
SurfaceCoord[φ_] := {r[φ] Cos[φ], r[φ] Sin[φ]}
Tangent[φ_] := (∂Phi SurfaceCoord[Phi] / Sqrt[∂Phi SurfaceCoord[Phi] . ∂Phi SurfaceCoord[Phi]]) //. Phi → φ
OutwardNormal[φ_] := {Tangent[φ][[2]], -Tangent[φ][[1]]}
SmallSurfaceArea[φ_] := Sqrt[r[φ]^2 + r'[φ]^2]
ScaledOutwardNormal[φ_] := SmallSurfaceArea[φ] OutwardNormal[φ]
```

We have simulated a Brownian path from x to y in a different workbook, and including it gives rise to the following sketch of the situation:

```
m = 25;
polarplot = PolarPlot[r[φ], {φ, 0, 2 π},
    PlotRange → {{-2 a, 2 a}, {-2 b, 2 b}}, PlotStyle → Black, Axes → False];
vectorplot = Graphics[{Black, Arrowheads[Small],
     Table[Arrow[{{SurfaceCoord[φ][[1]], SurfaceCoord[φ][[2]]},
        {SurfaceCoord[φ][[1]] + OutwardNormal[φ][[1]], SurfaceCoord[φ][[2]] +
          OutwardNormal[φ][[2]]}}], {φ, 0, 2 Pi - 2 Pi / m, 2 Pi / m}]}, Axes → False];
pointplotx = Graphics[{Black, PointSize[0.02], Point[x]}];
labelx = Graphics[Text[Style["(x,s)", 16, Italic], x + {-5 / 10, -1 / 10}]];
pointploty = Graphics[{Black, PointSize[0.02], Point[y]}];
labely = Graphics[Text[Style["(y,t)", 16, Italic], y + {5 / 10, -1 / 10}]];
Show[polarplot, vectorplot, pointplotx, pointploty, labelx, labely, BrownianPlot]
```

which returns the following graph:



Then we define the free propagator B its gradient, $\nabla B := -\sigma^2 n_\phi \cdot \nabla_y B$, as follows

$$\texttt{B[y\_, t\_, x\_, s\_]} := \frac{1}{\left(2\,\pi\,\sigma^2\,(\texttt{t}-\texttt{s})\right)^{\texttt{d}/2}}\, e^{-\frac{(\texttt{y}-\texttt{x})\,.\,(\texttt{y}-\texttt{x})}{2\,\sigma^2\,(\texttt{t}-\texttt{s})}}$$

$$\nabla\texttt{B[y\_, t\_, x\_, s\_]} := \texttt{ScaledOutwardNormal[Phi[y]]}\,.\,\frac{\texttt{y}-\texttt{x}}{\texttt{t}-\texttt{s}}\,\texttt{B[y, t, x, s]}$$

To be able to integrate over the surface and over a time-ordered integration range, we define the following integration operator:

```
SurfaceTimeIntegrationRange[t_, s_, n_] :=
  Flatten[Table[{{φi, 0, 2 π}, {τi, τi-1, t}}, {i, 1, n}], 1] //. τ0 → s;
SurfaceTimeIntegration[Integrand_, {MultipleRanges__}, WorkingPrec_] :=
 Re[NIntegrate[Integrand, MultipleRanges, WorkingPrecision → WorkingPrec]]
```

The n-th correction term can be defined by an integrand consisting of one B and n terms of ∇B, with n integrations over the surface as well as n time-ordered integrations, i.e.

```
CorrectionTerm{n_,WorkingPrec_}[y_, t_, x_, s_] :=
 CorrectionTerm{n,WorkingPrec}[y, t, x, s] = SurfaceTimeIntegration[
   B[y, t, SurfaceCoord[φn], τn] (∏_{i=2}^{n} ∇B[SurfaceCoord[φi], τi, SurfaceCoord[φi-1], τi-1])
    ∇B[SurfaceCoord[φ1], τ1, x, s], SurfaceTimeIntegrationRange[t, s, n], WorkingPrec]
```

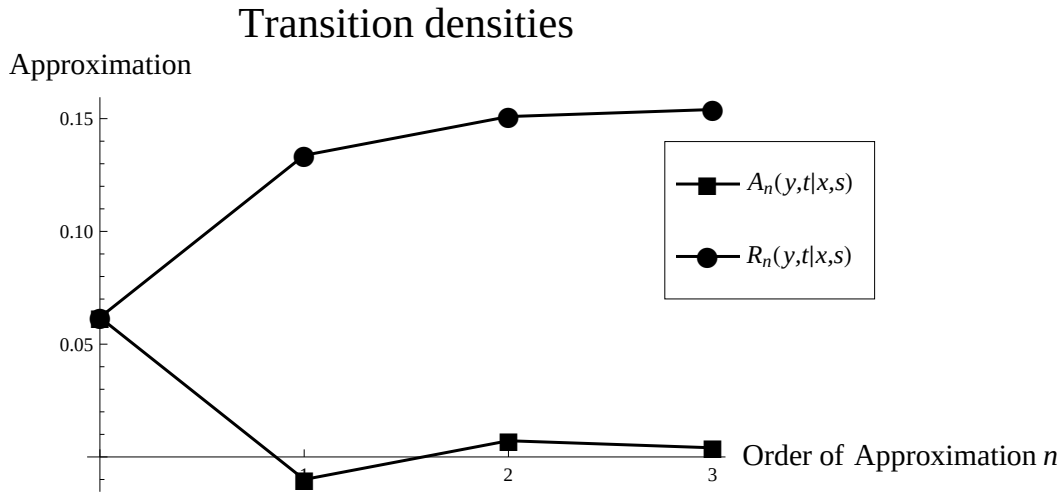
where we have chosen the 'first-passage' representation of each correction-term: with $\overrightarrow{\nabla} B$ on the right and B on the left. Now the propagators $A_{n,k}$ and $R_{n,k}$ can be defined, where n is the order of approximation and k is the working precision in *Mathematica*, i.e.

```
A{n_,WorkingPrec_}[y_, t_, x_, s_] :=
 N[B[y, t, x, s], WorkingPrec] + Sum_{i=1}^{n} (-1)^i CorrectionTerm{i,WorkingPrec}[y, t, x, s]
R{n_,WorkingPrec_}[y_, t_, x_, s_] :=
 N[B[y, t, x, s], WorkingPrec] + Sum_{i=1}^{n} CorrectionTerm{i,WorkingPrec}[y, t, x, s]
```

Having defined all we need, we can now plot the convergence of A and R as follows

```
d = 2;
WorkingPrec = 3; ApproxOrder = 3;
XFrameTicks = Table[i, {i, 0, ApproxOrder}];
Needs["PlotLegends`"]
ListPlot[{Table[{i, A{i,WorkingPrec}[y, t, x, s]}, {i, 0, ApproxOrder}],
  Table[{i, R{i,WorkingPrec}[y, t, x, s]}, {i, 0, ApproxOrder}]},
 PlotRange -> All, Joined -> True,
 InterpolationOrder -> 1, AxesOrigin -> {0, 0}, Ticks -> {XFrameTicks, Automatic},
 PlotStyle → { {Black, AbsoluteThickness[1.4]}, {Black,
  AbsoluteThickness[1.4]}},
 AxesLabel -> {Style["Order of Approximation n", 14, Plain, Black],
  Style["Approximation", 14, Plain, Black]}, PlotMarkers → {{"■", 15}, {"●", 15}},
 PlotLabel -> Style["Transition densities", 20, Plain, Black],
 PlotLegend → {Style["An(y,t|x,s)", 12, Plain, Black, Italic],
   Style["Rn(y,t|x,s)", 12, Plain, Black, Italic]}, LegendShadow → None,
 LegendSize → 0.4, LegendPosition → {+.25, -.25}, ImageSize -> 500]
```

which returns the following graph:



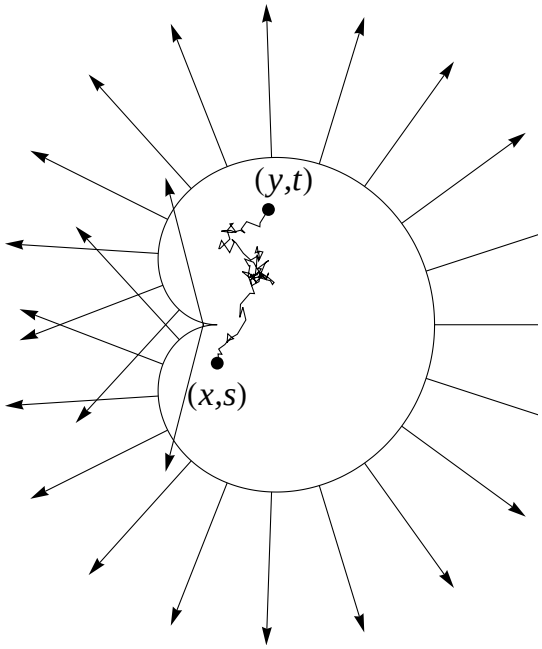
and where the modes of convergence are as promised.

4.2 A cusp in 2d

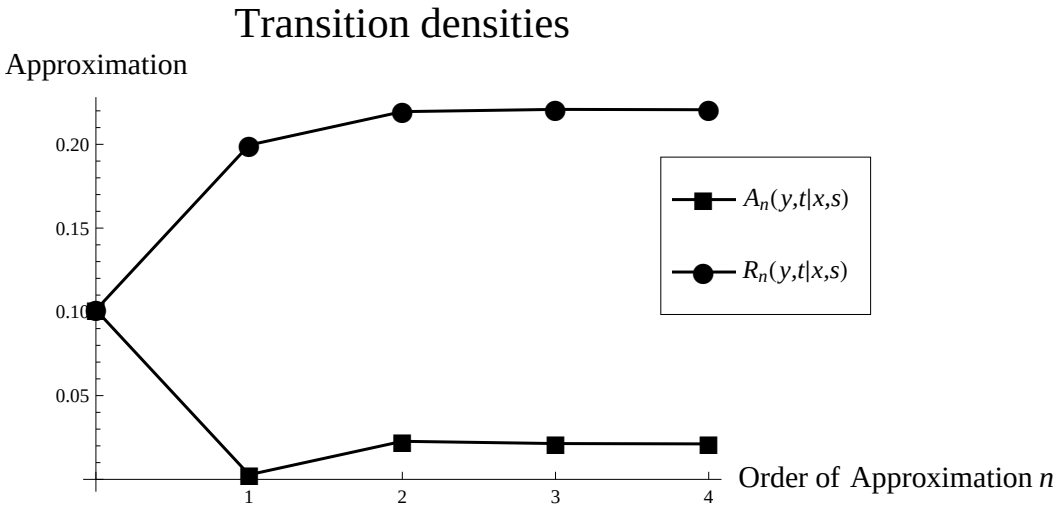
Instead of an ellipse, we could take a cusp to make the point that the method also works for piecewise smooth domains. We define the cusp as follows:

```
r[φ_] := (1 + Cos[φ])^(1/2)
σ = 1;
s = 0; x = {0, -1/4};
t = 3/4; y = {1/3, 3/4};
```

and the situation looks like



with the following propagators



which shows that the proposed method indeed works for domains that are non-smooth.

4.3 An ellipsoid in 3d

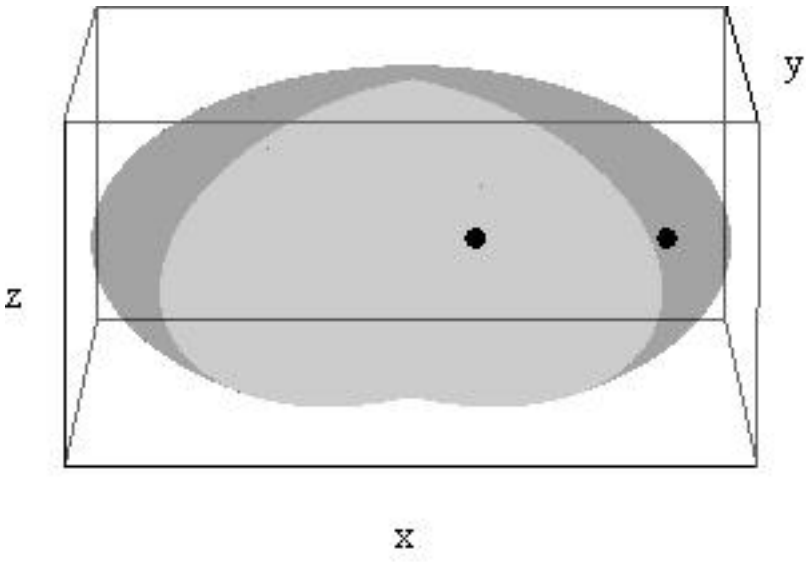
Moving to 3d, we may have an ellipsoid as follows:

```
a = 1; b = 2; c = 1 / 2;
r[θ_, φ_] := a b c / Sqrt[Cos[φ]^2 Sin[θ]^2 b^2 c^2 + Sin[φ]^2 Sin[θ]^2 a^2 c^2 + Cos[θ]^2 a^2 b^2]
```

and the following commands:

```
x = {1 / 5, 0, 0};
y = {4 / 5, 0, 0};
SphericalPlot = SphericalPlot3D[r[θ, φ], {θ, 0, π}, {φ, -2 Pi / 6, 2 π - 4 Pi / 6},
   PlotRange → All, PlotStyle → {Black, Opacity[.2]}, Mesh → None, Background → White,
   AxesStyle → Black, Boxed → True, PlotPoints → 80, BoxRatios → {a, b, c},
   AxesLabel → {Style["x", 16], Style["y", 16], Style["z", 16]}, Ticks → None];
PointPlot = Graphics3D[{{Black, PointSize[0.03], Point[y]},
    {Black, PointSize[0.03], Point[x]}}, Axes → None];
Show[SphericalPlot, PointPlot, ViewPoint → {0, -10, 1}]
```

provide the following graph of the situation:



with the following surface-coordinates and (scaled) outward normal vectors:

```
SurfaceCoord[θ_, φ_] := {r[θ, φ] Cos[φ] Sin[θ], r[θ, φ] Sin[φ] Sin[θ], r[θ, φ] Cos[θ]}
Off[N::"meprec"]
OutwardNormal[θ_, φ_] :=
 N[(Cross[∂Theta SurfaceCoord[Theta, Phi], ∂Phi SurfaceCoord[Theta, Phi]] /
       Norm[Cross[∂Theta SurfaceCoord[Theta, Phi],
         ∂Phi SurfaceCoord[Theta, Phi]]])] //. Theta → θ //. Phi → φ
ScaledOutwardNormal[θ_, φ_] := N[Cross[∂Theta SurfaceCoord[Theta, Phi],
      ∂Phi SurfaceCoord[Theta, Phi]] //. Theta → θ //. Phi → φ]
```

moving from Cartesian to spherical coordinates can be done by:

```
Radius[vector_] := Sqrt[vector[[1]]^2 + vector[[2]]^2 + vector[[3]]^2]
Theta[vector_] := N[ArcTan[vector[[3]], Sqrt[vector[[1]]^2 + vector[[2]]^2]]]
Phi[vector_] := N[ArcTan[vector[[1]], vector[[2]]]]
```

The free Brownian density is defined by:

```
d = 3; σ = 1;
B[y_, t_, x_, s_] := 1/(2 π σ^2 (t - s))^(d/2) e^(-((y-x).(y-x))/(2 σ^2 (t-s)))
∇B[y_, t_, x_, s_] :=
 Round[ScaledOutwardNormal[Theta[y], Phi[y]].(y - x)/(t - s) B[y, t, x, s], 0.01]
```

We define the integration over time and surface as follows:

```
SurfaceTimeIntegrationRange[t_, s_, n_] :=
  Flatten[Table[{{θi, 0, π}, {φi, 0, 2 π}, {τi, τi-1, t}}, {i, 1, n}], 1] //. τ0 → s;
SurfaceTimeIntegration[Integrand_, {MultipleRanges__}] :=
 Re[NIntegrate[Integrand, MultipleRanges, Method → "AdaptiveMonteCarlo"]]
```

The different correction terms are defined by:

```
CorrectionTerm_n_[y_, t_, x_, s_] :=
 CorrectionTerm_n[y, t, x, s] = SurfaceTimeIntegration[B[y, t, SurfaceCoord[θn, φn], τn]
    (∏_{i=2}^{n} ∇B[SurfaceCoord[θi, φi], τi, SurfaceCoord[θi-1, φi-1], τi-1])
   ∇B[SurfaceCoord[θ1, φ1], τ1, x, s], SurfaceTimeIntegrationRange[t, s, n]]
```

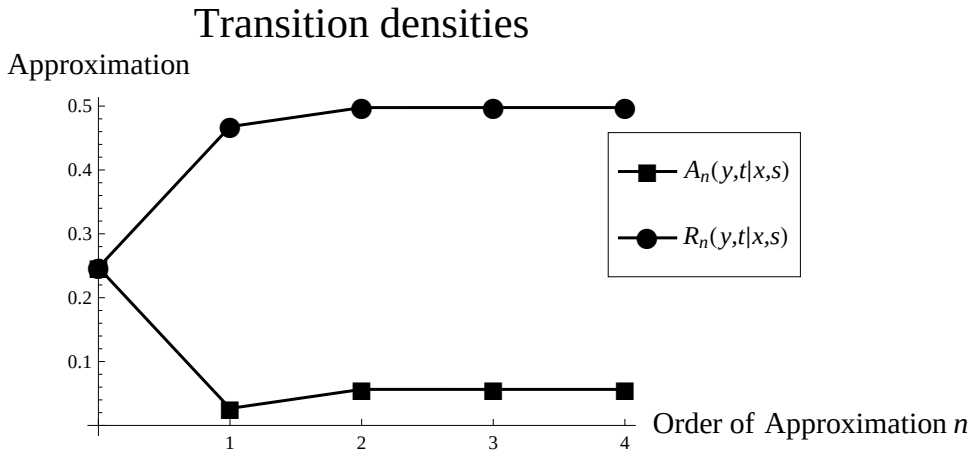
And for the absorbed and reflected transition densities this means that:

```
A_n_[y_, t_, x_, s_] := N[B[y, t, x, s]] + Σ_{i=1}^{n} (-1)^i CorrectionTerm_i[y, t, x, s]
R_n_[y_, t_, x_, s_] := N[B[y, t, x, s]] + Σ_{i=1}^{n} CorrectionTerm_i[y, t, x, s]
```

To produce a graph we enter the following commands:

```
σ = 1;
s = 0; t = 1 / 4;
n = 4;
XFrameTicks = Table[i, {i, 0, n}];
Needs["PlotLegends`"]
ListPlot[{Table[{j, Aj[y, t, x, s]}, {j, 0, n}], Table[{j, Rj[y, t, x, s]}, {j, 0, n}]},
 PlotRange -> All, Joined -> True,
 InterpolationOrder -> 1, AxesOrigin -> {0, 0}, Ticks -> {XFrameTicks, Automatic},
 PlotStyle → { {Black, AbsoluteThickness[1.4]}, {Black,
  AbsoluteThickness[1.4]}},
 AxesLabel -> {Style["Order of Approximation n", 14, Plain, Black],
  Style["Approximation", 14, Plain, Black]}, PlotMarkers → {{"■", 15}, {"●", 15}},
 PlotLabel -> Style["Transition densities", 20, Plain, Black],
 PlotLegend → {Style["An(y,t|x,s)", 12, Plain, Black, Italic],
   Style["Rn(y,t|x,s)", 12, Plain, Black, Italic]}, LegendShadow → None,
 LegendSize → 0.4, LegendPosition → {+.25, -.25}, ImageSize -> 500]
```

which returns the following graph:



5 Feynman-Kac potentials

This section has roughly the same set-up as section 3. The first subsection will discuss the problem definition (the Schrödiger equation in a probabilistic setting) and the second subsection will introduce the first- and last-interaction decompositions, resulting in two integral equations. Subsection 5.3 will derive the resulting series expansions and it will show that it matches the series expansion of the Feynman-Kac functional. Subsection 5.4 will discuss the reason for alternating/monotone convergence and will suggest a new set of 'Feynman rules' for a diffusion with a potential V. Subsection 5.5, finally, will make the link with subsection 3.9 and show how to write the solution to boundary value problems as a Feynman-Kac exponential.

5.1 The Schrödinger equation in a probabilistic setting

In quantum mechanics the motion of a physical particle is determined by the Schrödinger equation. We will transform the Schrödinger equation into a probabilistic setting by going to imaginary time ($t \to -i\,t$). Larger mass m of a particle (i.e. higher inertia) is analogous to lower variance σ^2 of a Brownian motion, suggesting we set $m = \frac{1}{\sigma^2}$. With these changes and with $\hbar = 1$, our version of the Schrödinger equation — with boundary conditions and initial conditions — reads as follows:

$$\begin{aligned}
\text{forward PDE} \quad & \left(\frac{\partial}{\partial t} - \frac{\sigma^2}{2}\nabla_y^2 + \lambda\, V(y)\right)\psi_V(y,t|x,s) \;=\; 0, \\
\text{backward PDE} \quad & \left(\frac{\partial}{\partial s} + \frac{\sigma^2}{2}\nabla_x^2 - \lambda\, V(x)\right)\psi_V(y,t|x,s) \;=\; 0, \\
\text{forward BC} \quad & \lim_{|y|\to\infty} \psi_V(y,t|x,s) \;=\; 0, \\
\text{backward BC} \quad & \lim_{|x|\to\infty} \psi_V(y,t|x,s) \;=\; 0, \\
\text{forward STC} \quad & \lim_{s\nearrow t} \psi_V(y,t|x,s) \;=\; \delta(|y-x|), \\
\text{backward STC} \quad & \lim_{t\searrow s} \psi_V(y,t|x,s) \;=\; \delta(|y-x|).
\end{aligned} \tag{5.1.1}$$

Here and elsewhere PDE stands for 'partial differential equation', BC stands for 'boundary condition' and STC stands for 'short time condition'. In the above, ψ is the usual *wave-function* of quantum mechanics, where (y,t) and (x,s) are referred to as the 'forward' and 'backward' space-time coordinates, respectively. We use the symbol ψ since this is customary in quantum mechanics, but we will think of ψ as a probability density, where the dependence on the potential is indicated through the subscript. The *coupling constant* λ measures the 'strength' of the coupling with the potential V. It can be proved under quite general conditions that the Schrödinger transition density 1) exists, 2) is unique and 3) is determined by the above conditions. See for example [59].

The BCs hold when the potential V does not grow exponentially at infinity, which we assume. The PDEs can be seen to hold through the following probabilistic interpretation of (5.1.1). Suppose that we have a Brownian motion as before, except we add the possibility that some catastrophic event happens, during time ds, annihilating the particle and reducing to zero the probability of propagation to any location, at any later time. This event we call an *interaction* with the potential. Suppose that an interaction happens with a probability that is a product of the strength of the potential at a certain location, and the time spent there. This means that during ds, and at location x, an interaction happens with probability $\lambda V(x)\,ds$. In any probabilistic interpretation we must have that the transition density ψ_V is unbiased, and the 'catastrophic event' interpretation implies that we must have

$$\psi_V(y,t|x,s) = (1-\lambda V(x)\,ds)\,\mathbb{E}\,\psi_V(y,t|x+dB,s+ds) + \lambda V(x)\,ds \times 0,$$

where with probability $(1-\lambda V(x)\,ds)$ the particle stays alive and where with probability $\lambda V(x)\,ds$ the particle gets annihilated by the potential. Using the Itô lemma (1.4.2) we obtain to first order in ds that

$$\left(\frac{\sigma^2}{2}\nabla_x^2 + \frac{\partial}{\partial s} - \lambda V(x)\right)\psi_V(y,t|x,s) = 0,$$

and similarly for the forward PDE. If the Brownian particle is not annihilated but instead its probabilistic 'weight' is doubled upon an interaction, then with probability $(1-\lambda V(x)\,ds)$ the density goes to $\mathbb{E}\,\psi(y,t|x+dB,s+ds)$, but with probability $\lambda V(x)\,ds$ it becomes twice that, i.e.

$$\begin{aligned}\psi_V(y,t|x,s) &= (1-\lambda V(x)\,ds)\,\mathbb{E}\,\psi_V(y,t|x+dB,s+ds) + (\lambda V(x)\,ds)\times 2\,\mathbb{E}\,\psi_V(y,t|x+dB,s+ds),\\ &= (1+\lambda V(x)\,ds)\,\mathbb{E}\,\psi_V(y,t|x+dB,s+ds).\end{aligned}$$

To first order in ds this leads by Itô's lemma (1.4.2) to

$$\left(\frac{\sigma^2}{2}\nabla_x^2 + \frac{\partial}{\partial s} + \lambda V(x)\right)\psi_V(y,t|x,s) = 0.$$

We realise that this equation could have been obtained immediately by switching the *sign* of V. Thus we see that a positive potential in (5.1.1) destroys a Brownian particle (upon an interaction), while a negative potential creates another Brownian particle (upon an interaction). Interactions with the potential do not change the endpoints of Brownian paths; it just gives them more or less weight. If there are one or more interactions with a positive potential, then the weight goes to zero. If there are i interactions with a negative potential, then the weight goes to 2^i. This leads to a modified transition density function.

The STCs are satisfied, finally, because the probability of an interaction is proportional to ds and thus within a very short period of time, the Brownian particle stays 1) alive and 2) where it is.

As an example, when the rate of annihilation or creation is constant, we get that the 'number' of paths arriving at (y,t) gets multiplied by

$$\lim_{n\to\infty}\left(1 \pm \frac{\lambda(t-s)}{n}\right)^n = e^{\pm\lambda(t-s)},$$

where the time interval from s to t is split up in n intervals, in each of which an interaction occurs with probability $\lambda(t-s)/n$. Therefore we get that:

$$\psi_{(V=\mp 1)}(y,t|x,s) = e^{\pm\lambda(t-s)}B(y,t|x,s),$$

for creation and annihilation of paths. A famous example of a quantum mechanical potential is the 'harmonic oscillator', where $V(x) = x \cdot x$ — see for example [60], p. 185. In the quantum mechanical interpretation, the harmonic oscillator tends to confine the particle to a region to the origin. This is because the expected 'force' equals the negative gradient of the potential, and thus the force points towards the origin and is linear in the distance away from the origin. In the probabilistic interpretation, the Brownian particle *also* tends to be found in a region close to the origin, but now because the probability of annihilation grows larger without bound, away from the origin.

We know that the Brownian particle must be somewhere at every intermediate time, and therefore we have the Chapman-Kolmogorov equation

$$\text{Chapman-Kolmogorov } \psi_V(y,t|x,s) = \int_D d\alpha\, \psi_V(y,t|\alpha,\tau)\psi_V(\alpha,\tau|x,s), \tag{5.1.2}$$

for any $s \le \tau \le t$, and where the STCs ensure that the Chapman-Kolmogorov equation also holds in the limit where τ goes to s or t. See for example [14], p. 36. The Green function exists for all positive (annihilating) potentials V and equals

$$G_V(y,x) := \int_s^\infty \psi_V(y,t|x,s)\, dt. \tag{5.1.3}$$

For potentials that are positive, the time spent by a Brownian motion at any location is finite, since the particle will almost surely be annihilated before time goes to infinity. For negative potentials, which create particles, the time spent at any location may be infinite and therefore the Green function might not exist. For positive potentials the Green function satisfies

$$\begin{gathered}\left(\frac{\sigma^2}{2}\nabla_y^2 - \lambda\, V(y)\right) G_V(y,x) = \left(\frac{\sigma^2}{2}\nabla_x^2 - \lambda\, V(x)\right) G_V(y,x) = -\delta(|y-x|), \\ \lim_{|y|\to\infty} G_V(y,x) = \lim_{|x|\to\infty} G_V(y,x) = 0.\end{gathered} \tag{5.1.4}$$

This can be verified by using (5.1.1). Because paths are annihilated by a positive potential, the density of all paths that are 'alive' is decreasing. The change in 'total density' is given

by

$$\begin{aligned}\frac{\partial}{\partial\tau}\int_{\mathbb{R}^d} d\alpha\ \psi_V(\alpha,\tau|x,s) &= \int_{\mathbb{R}^d} d\alpha\ \left(\frac{\sigma^2}{2}\nabla_\alpha^2 - \lambda V(\alpha)\right)\psi_V(\alpha,\tau|x,s),\\ &= -\lambda\int_{\mathbb{R}^d} d\alpha\ V(\alpha)\,\psi_V(\alpha,\tau|x,s),\end{aligned}$$

using the divergence theorem and the BCs of (5.1.1). This further supports our interpretation of creation and annihilation by the potential V. As we can see, the total density increases/decreases proportional to λ and to the relative likelihood of each possible position, weighted by the potential V — and decreases if V is positive, and increases if V is negative.

5.2 First- and last-interaction decompositions

In the previous subsection we have reviewed the Schrödinger equation in a probabilistic setting, without doing anything new. The original research on this topic starts here.

In this subsection we will think of V as positive, i.e. annihilating particles. While this is useful for the intuition, the calculations will never actually require V to be positive, and therefore the results will hold for all V. Using the STCs and the fundamental theorem of calculus, we can write down two *identities*, namely

$$\begin{aligned}\text{FI}\quad \psi_V(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau\ \left(\frac{\partial}{\partial\tau}\right)\int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\,\psi_V(\alpha,\tau|x,s),\\ \text{LI}\quad \psi_V(y,t|x,s) &= \psi_F(y,t|x,s) - \int_s^t d\tau\ \left(\frac{\partial}{\partial\tau}\right)\int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\,B(\alpha,\tau|x,s).\end{aligned} \tag{5.2.1}$$

The names FI and LI stand for first and last interaction. The reason that they hold is identities, is that the fundamental theorem of calculus says that

$$\begin{aligned}\text{FI}\quad \psi_V(y,t|x,s) &= B(y,t|x,s) + \left(\lim_{\tau\nearrow t} - \lim_{\tau\searrow s}\right)\int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\,\psi_V(\alpha,\tau|x,s),\\ \text{LI}\quad \psi_V(y,t|x,s) &= B(y,t|x,s) - \left(\lim_{\tau\nearrow t} - \lim_{\tau\searrow s}\right)\int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\,B(\alpha,\tau|x,s).\end{aligned} \tag{5.2.2}$$

The STCs then show that both decompositions hold by definition. The approach here should remind the reader of subsections 3.2 and 3.4. To explain the nomenclature we introduce the first- and last-interaction times. With the convention that $\inf\{\varnothing\} = \infty$ and that $\sup\{\varnothing\} = -\infty$ we have for the first- and last-interaction times

$$\begin{aligned}\text{FI}\quad \tau^{\text{FI}}(t|x,s) &= \inf_\tau\ \left\{s\le\tau\le t : \text{an interaction happens at time } \tau \middle| B_s = x\right\},\\ \text{LI}\quad \tau^{\text{LI}}(t|x,s) &= \sup_\tau\ \left\{s\le\tau\le t : \text{an interaction happens at time } \tau \middle| B_s = x\right\}.\end{aligned} \tag{5.2.3}$$

Recall Chapman-Kolmogorov, which says

$$\text{Chapman-Kolmogorov } \psi_V(y,t|x,s) = \int_D d\alpha\ \psi_V(y,t|\alpha,\tau)\,\psi_V(\alpha,\tau|x,s).$$

Both propagators are annihilated propagators ψ_V, because to survive up to time t the Brownian particle must survive first up to time τ and then from time τ up to time t. But now consider instead the following quantity:

$$\int_D d\alpha\ \psi_V(y,t|\alpha,\tau)\,B(\alpha,\tau|x,s).$$

The free propagator counts all paths from (x,s) to (α,τ), regardless of whether interactions occur or not. In effect, therefore, by making this change we *allow* interactions to happen in the time up to time τ, while not requiring them. The last interaction (if at all) must therefore have happened before time τ. Extending this reasoning, we propose that

$$\begin{aligned}
&\text{FI } \mathbb{P}\left(B_t \in dy;\, \tau^{\text{FI}} \geq \tau \middle| B_s = x\right) = \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\,\psi_V(\alpha,\tau|,s),\\
&\text{LI } \mathbb{P}\left(B_t \in dy;\, \tau^{\text{LI}} \leq \tau \middle| B_s = x\right) = \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\,B(\alpha,\tau|,s).
\end{aligned} \tag{5.2.4}$$

Recall that a positive potential V kills paths, such that the propagator ψ_V counts paths without interactions, and that the free propagator B counts all paths regardless of whether interactions happen or not. The free propagator allows interactions but does not require them, and therefore it is crucial that we specified that $\inf\{\varnothing\} = \infty$ and that $\sup\{\varnothing\} = -\infty$. Continuing, we find

$$\begin{aligned}
&\text{FI } \mathbb{P}\left(B_t \in dy;\, \tau^{\text{FI}} \in d\tau \middle| B_s = x\right) = -\left(\frac{\partial}{\partial\tau}\right) \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\,\psi_V(\alpha,\tau|,s),\\
&\text{LI } \mathbb{P}\left(B_t \in dy;\, \tau^{\text{LI}} \in d\tau \middle| B_s = x\right) = \quad \left(\frac{\partial}{\partial\tau}\right) \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\,B(\alpha,\tau|,s).
\end{aligned} \tag{5.2.5}$$

We realise that the probability that no interaction happens equals ψ_V, i.e.

$$\begin{aligned}
&\text{FI } \mathbb{P}\left(B_t \in dy;\, \tau^{\text{FI}} = \infty \middle| B_s = x\right) \quad = \psi_V(y,t|x,s),\\
&\text{LI } \mathbb{P}\left(B_t \in dy;\, \tau^{\text{LI}} = -\infty \middle| B_s = x\right) = \psi_V(y,t|x,s).
\end{aligned} \tag{5.2.6}$$

By subtracting from the free density all those paths with a first or last interaction, we get

$$\begin{aligned}
&\text{FI } \psi_V(y,t|x,s) = B(y,t|x,s) + \int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\,\psi_V(\alpha,\tau|x,s),\\
&\text{LI } \psi_V(y,t|x,s) = B(y,t|x,s) - \int_s^t d\tau \left(\frac{\partial}{\partial\tau}\right) \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\,B(\alpha,\tau|x,s).
\end{aligned} \tag{5.2.7}$$

We have re-derived the set of identities that we started with. While the interpretation of first and last interactions presents itself naturally for a positive (i.e. killing) potential, it is obvious that both identities hold for any reasonably behaved potential. Using the PDEs of (5.1.1), we get

$$\begin{aligned}
\text{FI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau) \left\{ \frac{\sigma^2}{2} \overleftarrow{\nabla}_\alpha^2 - \frac{\sigma^2}{2} \overrightarrow{\nabla}_\alpha^2 + \lambda V(\alpha) \right\} \psi_V(\alpha,\tau|x,s), \\
\text{LI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) \\
&\quad + \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau) \left\{ \frac{\sigma^2}{2} \overleftarrow{\nabla}_\alpha^2 - \frac{\sigma^2}{2} \overrightarrow{\nabla}_\alpha^2 - \lambda V(\alpha) \right\} B(\alpha,\tau|x,s).
\end{aligned} \tag{5.2.8}$$

With Green's identity (1.3.1), we can transform the integral over the 'interior' of $\mathbb{R}^d$ to one over the 'boundary' of $\mathbb{R}^d$. While it may not be obvious that Green's theorem holds for $\mathbb{R}^d$, we could approximate $\mathbb{R}^d$ by some very large domain D. Green's theorem would hold for this domain, and we would get

$$\begin{aligned}
\text{FI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) \\
&\quad + \frac{1}{2} \int_s^t d\tau \int_{\partial D} d\beta\ B(y,t|\alpha,\tau) \left\{ \overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta \right\} \psi_V(\alpha,\tau|x,s) \\
&\quad - \int_s^t d\tau \int_D d\alpha\ B(y,t|\alpha,\tau) \left\{ \lambda V(\alpha) \right\} \psi_V(\alpha,\tau|x,s), \\
\text{LI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) \\
&\quad - \frac{1}{2} \int_s^t d\tau \int_{\partial D} d\beta\ \psi_V(y,t|\alpha,\tau) \left\{ \overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta \right\} B(\alpha,\tau|x,s) \\
&\quad - \int_s^t d\tau \int_D d\alpha\ \psi_V(y,t|\alpha,\tau) \left\{ \lambda V(\alpha) \right\} B(\alpha,\tau|x,s).
\end{aligned} \tag{5.2.9}$$

When the domain D grows bigger and bigger, the BCs in (5.1.1) demand that the boundary terms disappear, and in the limit where $D \to \mathbb{R}^d$ we get

$$\begin{aligned}
\text{FI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau) \left\{ \lambda V(\alpha) \right\} \psi_V(\alpha,\tau|x,s), \\
\text{LI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau) \left\{ \lambda V(\alpha) \right\} B(\alpha,\tau|x,s).
\end{aligned} \tag{5.2.10}$$

In the physics literature these are sometimes known as Dyson's equation, or the Lippmann-Schwinger equations, but this derivation and interpretation are new.

The nomenclature that is adapted may be counter-intuitive to physicists. In the physicist's mind, the potential does not annihilate the particle, but instead it *scatters* it. There-

fore B would represent free propagation, i.e. without interactions, while ψ_V allows interactions. But we will stick with a probabilistic interpretation, where the free propagator counts *all* paths that start from (x,s) and that end at (y,t), but where only a subset of those stay alive when the potential is positive; i.e. only paths with no interactions contribute to ψ_V.

It should be noted that in the derivation of the FI and LI integral equations we have used all 6 conditions of (5.1.1). First we used the forward and backward STCs to write down two identities, then we used both PDEs under the integral sign, and, finally, we discarded the forward and backward boundary terms at spatial ∞. The two integral equations are therefore equivalent to the 6 original conditions, and if we can solve the integral equation, then, by uniqueness, we must also have obtained the solution to (5.1.1). Thus we obtain the following proposition:

Proposition 5. FI and LI decomposition of ψ_V. *For a potential V that does not grow exponentially at infinity, and for all $x,y \in \mathbb{R}^d$, the following formulations of Brownian motion in the presence of a potential are equivalent:*

$$\left.\begin{aligned}
\left(\tfrac{\partial}{\partial t} - \tfrac{\sigma^2}{2}\nabla_y^2 + \lambda V(y)\right)\psi_V(y,t|x,s) &= 0\\
\left(\tfrac{\partial}{\partial s} + \tfrac{\sigma^2}{2}\nabla_x^2 - \lambda V(x)\right)\psi_V(y,t|x,s) &= 0\\
\lim_{|y|\to\infty}\psi_V(y,t|x,s) &= 0\\
\lim_{|x|\to\infty}\psi_V(y,t|x,s) &= 0\\
\lim_{s\nearrow t}\psi_V(y,t|x,s) &= \delta(|y-x|)\\
\lim_{t\searrow s}\psi_V(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FI}\ \ \psi_V(y,t|x,s) = B(y,t|x,s)\\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\Big\{\lambda V(\alpha)\Big\}\psi_V(\alpha,\tau|x,s)\\
&\text{LI}\ \ \psi_V(y,t|x,s) = B(y,t|x,s)\\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\Big\{\lambda V(\alpha)\Big\}B(\alpha,\tau|x,s)
\end{aligned}\right. \tag{5.2.11}$$

The formal solution is given by

$$\psi_V(y,t|x,s) = B(y,t|x,s) + \sum_{i=1}^{\infty}(-\lambda)^i \left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right]$$
$$\times \int_{\mathbb{R}^d} d\alpha_i\, B(y,t|\alpha_i,\theta_i)V(\alpha_i)\left[\prod_{k=1}^{i-1}\int_{\mathbb{R}^d} d\alpha_k\, B(\alpha_{k+1},\theta_{k+1}|\alpha_k,\theta_k)V(\alpha_k)\right] B(\alpha_1,\theta_1|x,s),$$

which can be obtained by substituting the LI decomposition into itself repeatedly. The same can be done by substituting the FI decomposition into itself.

It should be noted that the derivation of this proposition almost completely mirrors the derivation in subsections 3.2 and 3.4, where we derived the first- and last-passage (reflection) decompositions, and stated them in Propositions 1 and 2. This close connection has not been revealed before, and it is not coincidental that we present the results on

the Schrödinger equation in such a similar manner. The fact that the same approach is fruitful is remarkable already, because boundary value problems and potential problems are considered as separate objects of study. To pursue the analogy further, we will derive here the counterpart to the series solution of Propositions 3 and 4. A simple method is to substitute the equations on the right-hand side of Proposition 5 back into themselves repeatedly, leading to a series solution. But with an eye on a future application, we will be a little bit more formal. We define the operator K as operating on some test-function f, either on the left or right side, as follows

$$\begin{aligned} K * f(y,t|x,s) &:= \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha \; B(y,t|\alpha,\tau)\, V(\alpha)\, f(\alpha,\tau|x,s), \\ f(y,t|x,s) * K &:= \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha \; f(y,t|\alpha,\tau)\, V(\alpha)\, B(\alpha,\tau|x,s). \end{aligned} \tag{5.2.12}$$

We say that the operator K is positive (negative) if V is positive (negative). With this notation we can write the FI and LI decompositions as follows

$$\begin{aligned} \text{FI}\quad & \psi_V(y,t|x,s) = B(y,t|x,s) - \lambda\; K * \psi_V(y,t|x,s) \\ \text{LI}\quad & \psi_V(y,t|x,s) = B(y,t|x,s) - \lambda\; \psi_V(y,t|x,s) * K \end{aligned} \tag{5.2.13}$$

Inviting the 'solution' to be written as follows:

$$\begin{aligned} \text{FI}\quad & \psi_V(y,t|x,s) = \frac{1}{1+\lambda K*}\; B(y,t|x,s), \\ \text{LI}\quad & \psi_V(y,t|x,s) = B(y,t|x,s)\; \frac{1}{1+*K\lambda}. \end{aligned} \tag{5.2.14}$$

Here and elsewhere, the action of an operator in the denominator is defined by its series, i.e.

$$\begin{aligned} \text{FI}\quad & \frac{1}{1+\lambda K*}\; B(y,t|x,s) = \big(1 - \lambda K * + \lambda^2 K * K * - \lambda^3 K * K * K * + \dots\big) B(y,t|x,s), \\ \text{LI}\quad & B(y,t|x,s)\; \frac{1}{1+*K\lambda} = B(y,t|x,s)\big(1 - *K\lambda + *K * K\lambda^2 - *K * K * K\lambda^3 + \dots\big). \end{aligned} \tag{5.2.15}$$

It can easily be checked that both series are equal, term by term. For a positive potential we get convergence in an alternating fashion, just like the expansion for $1/(1+x)$ converges in an alternating fashion, provided that $0 < x < 1$. For a negative potential we get convergence in a monotone fashion, just like the expansion for $1/(1-|x|)$ converges in a monotone fashion, provided that $|x| < 1$. This indicates that the propagator ψ_V may also exist for negative potentials as long as the potential V is finite.

5.3 The Feynman-Kac formula

The Feynman-Kac formula that appeared in [44] suggests itself through the interpretation of V as a rate of killing. It is useful because it allows us to write the series solution of

Proposition 5 in a very compact manner. We will first derive it heuristically, and afterwards we will show that its Taylor expansion agrees with the series suggested by Proposition 5.

We slice the time from s to t up such that there are N intermediate locations. Therefore we have that the length of each time interval is $\epsilon = (t-s)/(N+1)$. Using τ_i as the intermediate times, such that $\tau_i = s + i\,\epsilon$, then we have for i running from 0 to $N+1$ that $\tau_0 = s$ and $\tau_{N+1} = t$ and all the other τ_i for $1 \leq i \leq N$ refer to N intermediate times. If a path from (x,s) to (y,t) is defined by its N intermediate locations $\{B_{\tau_1}, \cdots, B_{\tau_N}\}$ then we have that the probability of survival of this path is a product of $N+1$ probabilities: one for each intermediate location, and one for the end-point (it is assumed that the particle is not annihilated at the starting point). Therefore the probability of survival becomes

$$\prod_{i=1}^{N+1} (1 - \lambda\,\epsilon\,V(B_{\tau_i})) \approx \prod_{i=1}^{N+1} e^{-\lambda\,\epsilon\,V(B_{\tau_i})} = e^{-\lambda \sum_{i=1}^{N+1} V(B_{\tau_i})\,\epsilon} \to e^{-\lambda \int_s^t V(B_\tau) d\tau} \tag{5.3.1}$$

and where the last relationship holds in the limit for large N and where the path is no longer defined by its intermediate locations but rather by the entire, continuous, nowhere-differentiable Brownian path. If the above is the probability that a *given* path should survive (with N known intermediate locations), then the probability that *any* path should survive is obtained by taking an expectation over all possible intermediate locations, i.e. over all paths. If we want the path to end up at y then we need to take an expectation over all paths while enforcing the last position to be y. We can achieve this by plugging in a δ-function at y. Thus we are lead to propose that

$$\psi_V(y,t|x,s) = \mathbb{E}\left(\delta(B_t - y)\, e^{-\lambda \int_s^t V(B_\tau)\, d\tau} \big| B_s = x\right) \tag{5.3.2}$$

where we note that if the potential (or coupling constant) is zero, then the free propagator B results as follows

$$B(y,t|x,s) = \mathbb{E}\left(\delta(B_t - y)|B_s = x\right) = \int d\alpha\, \delta(\alpha - y)\, B(\alpha,t|x,s) = B(y,t|x,s).$$

Again we note that a positive potential, which kills paths, leads to a propagator ψ_V which is smaller than the free propagator B, while a negative potential ensures that ψ_V is larger than B. Like the Feynman path integral, the Feynman-Kac exponential can usually not be calculated exactly. But it is easy enough to write down its perturbation series:

$$\mathbb{E}\left(\delta\left(B_t - y\right) e^{-\lambda \int_s^t V(B_\tau)\, d\tau} \big| B_s = x\right) = \sum_{i=0}^{\infty} \frac{1}{i!}\, \mathbb{E}\left(\delta\left(B_t - y\right) \left(-\lambda \int_s^t V\left(B_\tau\right) d\tau\right)^i \big| B_s = x\right). \tag{5.3.3}$$

The resulting series is called the Born series by physicists, such as [46] (p. 163), but it does not have a special name to probabilists, who simply 'expand the exponential as a power series', such as in [12] (p. 214). The expansion should behave properly (i.e. converge)

when the time-integral of $V(.)$ is finite. Let us define the coefficients of the expansion as follows:

$$\mathbb{E}\left(\delta(B_t - y)\, e^{-\lambda \int_s^t V(B_\tau)\, d\tau} | B_s = x\right) = B(y,t|x,s) + \sum_{i=1}^{\infty} a_i(y,t|x,s).$$

Further note that we have

$$\frac{1}{i!}\left(-\lambda \int_s^t V\left(B_\tau\right) d\tau\right)^i = \frac{(-\lambda)^i}{i!} \int_s^t d\theta_i \cdots \int_s^t d\theta_1 \prod_{j=1}^{i} V\left(B_{\theta_j}\right).$$

We note that the integrand, being the product of i copies of V, is symmetric in the exchange of any of its arguments. With i terms there are $i!$ ways of ordering them, but all possible orderings of the integrand give the same contribution to the overall integral. Therefore we may enforce that $\theta_i \geq \theta_{i-1} \geq ... \geq \theta_1$, rather than allowing all θ_i to have independent integration ranges spanning $[s,t]$. This time-ordered integration range is $i!$ times smaller than the original region of integration, matching exactly the combinatorial pre-factor. Therefore we have that

$$\frac{1}{i!}\left(-\lambda \int_s^t V\left(B_\tau\right) d\tau\right)^i = (-\lambda)^i \left[\int d\theta_i \cdots \int d\theta_1\right]_{s \leq \theta_1 \leq \cdots \leq \theta_i \leq t} \prod_{j=1}^{i} V\left(B_{\theta_j}\right)$$

where the integrals on the right-hand side are time-ordered, and thus we obtain

$$a_i(y,t|x,s) = (-\lambda)^i \left[\int d\theta_i \cdots \int d\theta_1\right]_{s \leq \theta_1 \leq \cdots \leq \theta_i \leq t} \mathbb{E}\left(\delta(B_t - y) \prod_{j=1}^{i} V(B_{\theta_j}) \Big| B_s = x\right).$$

The reason that time-ordered integrands are useful under expectation signs, is that we may now make use of the Markov property. Recall that the law of iterated expectations says that for any time τ smaller than t we have

$$\mathbb{E}f(B_t) = \mathbb{E}(\mathbb{E}(f(B_t)|\mathcal{F}_\tau)) = \mathbb{E}(\mathbb{E}(f(B_t)|B_\tau))$$

which is also known as the tower-property. We want to use the tower property on a product involving i terms, where $s \leq \theta_1 \leq \cdots \theta_i \leq t$. For the first application of the tower property, there are two possible actions

- Condition on $\theta_2, \cdots, t$, and take an expectation at time θ_1.
- Condition on $\theta_1, \cdots, \theta_i$, and take an expectation at time t.

In essence, we can work our way through the product in the forward or backward time direction. The first choice gives that

$$a_i(y,t|x,s) = (-\lambda)^i \left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right] \times \int_{\mathbb{R}^d} d\alpha_i\, B(y,t|\alpha_i,\theta_i)V(\alpha_i) \left[\prod_{k=1}^{i-1}\int_{\mathbb{R}^d} d\alpha_k\, B(\alpha_{k+1},\theta_{k+1}|\alpha_k,\theta_k)V(\alpha_k)\right] B(\alpha_1,\theta_1|x,s) \tag{5.3.4}$$

where the inner-most integration is over α_1 and at time θ_1. The second choice gives

$$a_i(y,t|x,s) = (-\lambda)^i \left[\int_{s\le\theta_1\le\cdots\le\theta_i\le t} d\theta_i\cdots d\theta_1\right] \times \int_{\mathbb{R}^d} d\alpha_1\, B(\alpha_1,\theta_1|x,s)V(\alpha_1) \left[\prod_{k=1}^{i}\int_{\mathbb{R}^d} d\alpha_k\, V(\alpha_k)B(\alpha_k,\theta_k|\alpha_{k-1},\theta_{k-1})\right] B(y,t|\alpha_i,\theta_i) \tag{5.3.5}$$

where the innermost integration is over α_i at time θ_i. The integrations over α_j span all of $\mathbb{R}^d$ and it is tempting to pull them all to the front of the expression. It would be nice if all the integrations were written at the front, as opposed to in a concatenated fashion within the expression. When the potential V is a nice function, then it is useful (and customary) to pull all the space-integrations to the left of the formula, as in [12] (p. 214) and [46] (p. 163). But we will shortly introduce a potential V that is *not* nice, and that does not commute with integration, invalidating any such procedure of 'pulling integrations to the left'. Therefore we will leave the expansion as it is: with the integrals over the intermediate coordinates nested within the expression. We notice further that we have that

$$\begin{aligned} a_i(y,t|x,s) &= (-\lambda)\, K * a_{i-1}(y,t|x,s) \\ a_i(y,t|x,s) &= (-\lambda)\, a_{i-1}(y,t|x,s) * K \end{aligned} \tag{5.3.6}$$

where the operator K is defined as before in (5.2.12). And because

$$\begin{aligned} a_1(y,t|x,s) &= (-\lambda)\, K * B(y,t|x,s) \\ a_1(y,t|x,s) &= (-\lambda)\, B(y,t|x,s) * K \end{aligned} \tag{5.3.7}$$

we find that the entire series solution can be written in two ways, i.e.

$$\begin{aligned} \psi_V(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty} (-\lambda)^i\, (K*)^i\, B(y,t|x,s) \\ \psi_V(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty} (-\lambda)^i\, B(y,t|x,s)\, (*K)^i \end{aligned} \tag{5.3.8}$$

and we realise that the Feynman-Kac exponential produces exactly the same series solution that the integral equations of Proposition 5 produce. Of course both series solutions are equal, and this can easily be seen when the potential V is nice. It may therefore seem a little *artificial* to treat the two series so separately. However, as can be seen from the integral equations derived in the last subsection, i.e.

$$\begin{aligned} \text{FI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\Big\{\lambda\, V(\alpha)\Big\}\psi_V(\alpha,\tau|x,s) \\ \text{LI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\Big\{\lambda\, V(\alpha)\Big\}B(\alpha,\tau|x,s) \end{aligned} \tag{5.3.9}$$

we *naturally* get two series — by a repeated substitution of the equation into itself, where 1) the integrations over time are automatically time-ordered, and 2) the spatial integration will automatically appear nested within the expression. Furthermore we have seen in section 3 that for boundary value problems, one of them leads to a 'single boundary layer' whereas the other leads to a 'double boundary layer'. We argued that both arise through the first/last passage decompositions, and so their equivalence essentially derives from the symmetry in time. The same is true for Feynman-Kac potentials, and we realise that the *one* exponential can be associated with two integral equations, and thus we conclude that

Proposition 6. Feynman-Kac and 2 integral equations. *For all $x, y \in \mathbb{R}^d$, and for a potential V that does not grow exponentially at infinity, the following problem formulations are equivalent:*

$$\left.\begin{array}{c} \psi_V(y,t|x,s) = \\ \mathbb{E}_{(x,s)}\Big(\delta(B_t - y)\mathrm{Exp}\Big[-\lambda\int_s^t V(B_\tau)d\tau\Big]\Big) \end{array}\right\} = \left\{\begin{array}{l} \text{FI}\ \psi_V(y,t|x,s) = B(y,t|x,s) \\ \quad -\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\Big\{\lambda\, V(\alpha)\Big\}\psi_V(\alpha,\tau|x,s) \\ \text{LI}\ \psi_V(y,t|x,s) = B(y,t|x,s) \\ \quad -\int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\Big\{\lambda\, V(\alpha)\Big\}B(\alpha,\tau|x,s) \end{array}\right. \tag{5.3.10}$$

where both a 'forward' and a 'backward' integral equation can be associated with one *Feynman-Kac exponential, with solution given by*

$$\begin{aligned} \text{FI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty} (-\lambda)^i\, (K*)^i\, B(y,t|x,s) \\ \text{LI}\ \psi_V(y,t|x,s) &= B(y,t|x,s) + \sum_{i=1}^{\infty} (-\lambda)^i\, B(y,t|x,s)\, (*K)^i \end{aligned} \tag{5.3.11}$$

where the operation of the integral operator K is defined in (5.2.12).

Again it is obvious — from our explicit symmetric treatment of first and last interactions — that ψ_V is symmetric in the spatial coordinates x and y. We are now almost ready to state our third and last theorem, when we take V to be $V(\alpha) = \mp\frac{\sigma^2}{2}\nabla^2\mathbb{1}_{\alpha\in D}$ for absorbed and reflected Brownian motion. But first we will discuss the *interpretation* of the series expansion solution, and the reason for the different modes of convergence.

5.4 The Feynman rules for a diffusion

In this section we started with the Schrödinger equation (5.1.1) (subsection 5.1), second we derived the FI and LI integral equations using *all* the specified conditions (subsection 5.2), and thirdly we showed that the expansion of that set of integral equations leads to the same series solution that the Taylor expansion of the Feynman-Kac exponential leads to (subsection 5.3).

In the physics literature, the order of proceedings is usually very different. First, the Feynman path integral is *postulated* and *then* it is expanded in its series. Thirdly, it is realised that this series expansion could also have been obtained by expanding a *single* integral equation. [45], for example, postulate first the Feynman path integral (p. 120), then they expand it as a power series (p. 121), and lastly (p. 126) it is realised that the same series expression could have been obtained by the expansion of the 'last scattering integral equation'. A 'first scattering integral equation', however, is strikingly absent. The same procedure is followed in [46] (p. 164) where again the 'last interaction integral equation' is noticed to produce the same series that the path integral produces, but the 'first interaction integral equation' is strikingly absent.

Furthermore, the interpretation of the series expansion is very different. In Feynman's interpretation, the free term B counts paths *without* interactions, the first order term in the Taylor expansion counts all with *exactly* one interaction, the second order term counts all paths with *exactly* two interactions, and so on. In the physical picture, any number of interactions are allowed and therefore ψ_V is the sum of all these terms. [45] write (p. 123):

> The [last interaction integral equation] equation is very important and very useful, so we shall develop a special interpretation to help think about it physically. We call the interaction between the potential and the particle a *scattering*: thus we say that the potential scatters the particle and that the *amplitude to be scattered by a potential is* [proportional to V] *per unit volume and per unit time.* With this interpretation we can describe [the propagator] K_V in the following way. [The propagator] K_V is, of course, a sum over alternatives ways in which the particle may move from point a to point b. The alternatives are: 1. The particle may not be scattered at all, $K_0(b,a)$ 2. The particle may be scattered once, $K^{(1)}(b,a)$ 3. The particle may be scattered twice, $K^{(2)}(b,a)$. Etc. [...]

> The total amplitude for motion from a to b with any number of scatterings is $K_0 + K^{(1)} + K^{(2)} + ... + K^{(n)} + \ldots$.

and a similar interpretation can be found in [46] (p. 163). Of course Feynman was dealing with an *complex* wave-function, and the obtained series does *not* converge in an absolute sense, because the integrands are oscillating rather than vanishing at ∞. To investigate convergence, one often makes use of what is called 'analytic continuation': transforming the time-variable $t \to -i\,t$. We have already pointed out that this transformation turns the problem into one of Brownian motion, where the positive potential V kills paths at a rate corresponding to its magnitude. The convergence of the Taylor series can be shown, and has been shown. But it has not been interpreted as we interpret it here.

Feynman's interpretation suggests that the convergence should be monotone: each term adds more paths to the propagator. But instead we have established that the series solution converges in an alternating fashion when the potential is positive. Furthermore, in the probabilistic interpretation the free propagator counts *all* paths, regardless of the number of interactions, while ψ_V counts only paths with zero interactions, because each interaction kills the particle. Therefore we propose a combinatorial Pascal interpretation as follows:

Theorem 2. Combinatorial Feynman rules. *The Feynman rules for a Brownian particle, in the presence of a positive (i.e. killing) potential V, are as follows:*

$$\begin{pmatrix} \text{free term} \\ \lambda^1 \text{ term} \\ \lambda^2 \text{ term} \\ \lambda^3 \text{ term} \\ \lambda^4 \text{ term} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 2 & 3 & 4 & \cdots \\ 0 & 0 & 1 & 3 & 6 & \cdots \\ 0 & 0 & 0 & 1 & 4 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix} \tag{5.4.1}$$

*where we have the infinite upper-triangular Pascal matrix, and where each λ^i-*term *is positive and defined by*

$$\begin{aligned} \lambda^i\text{-term} = \lambda^i &\left[\int_{s \le \theta_1 \le \cdots \le \theta_i \le t} d\theta_i \cdots \int d\theta_1 \right] \\ &\times \int_{\mathbb{R}^d} d\alpha_i\, B(y,t|\alpha_i,\theta_i) V(\alpha_i) \left[\prod_{k=1}^{i-1} \int_{\mathbb{R}^d} d\alpha_k\, B(\alpha_{k+1},\theta_{k+1}|\alpha_k,\theta_k) V(\alpha_k) \right] B(\alpha_1,\theta_1|x,s) \end{aligned} \tag{5.4.2}$$

This interpretation implies that

$$\mathbb{P}\Big(B_t = y;\text{ exactly } i \text{ interactions}\Big|B_s = x\Big) = \mathbb{E}_x\left(\delta(B_t - y)\,\frac{1}{i!}\left[\int_s^t V(B_\tau)d\tau\right]^i e^{-\int_s^t V(B_\tau)d\tau}\right) \tag{5.4.3}$$

which returns the classical Feynman-Kac formula (with a δ-function as the initial condition) by substituting $i = 0$.

The claim of Theorem 2 is that the free term B counts all paths from (x,s) to (y,t) — regardless of the number of interactions. The first correction term, linear in λ picks up a contribution for *every* interaction: it thus counts paths with i interactions i times. The second correction term, that goes with λ^2, counts all possible time-ordered pairs of interactions: it counts paths with i interactions 'i choose 2' times and so on. The matrices should be extended and are infinite in size. Inverting the Pascal matrix gives immediately

$$\begin{pmatrix}\text{paths with 0 interactions}\\ \text{paths with 1 interaction}\\ \text{paths with 2 interactions}\\ \text{paths with 3 interactions}\\ \text{paths with 4 interactions}\\ \vdots\end{pmatrix} = \begin{pmatrix}1 & -1 & 1 & -1 & 1 & \cdots\\ 0 & 1 & -2 & 3 & -4 & \cdots\\ 0 & 0 & 1 & -3 & 6 & \cdots\\ 0 & 0 & 0 & 1 & -4 & \cdots\\ 0 & 0 & 0 & 0 & 1 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots\end{pmatrix}\cdot\begin{pmatrix}\text{free term}\\ \lambda^1\text{ term}\\ \lambda^2\text{ term}\\ \lambda^3\text{ term}\\ \lambda^4\text{ term}\\ \vdots\end{pmatrix} \tag{5.4.4}$$

recovering not only the expression in (1.6.3), for all paths with 0 interactions, but obtaining the probability that exactly $i > 0$ interactions occur!

Proof. The interpretation follows from the fact that

$$\begin{aligned}&\mathbb{E}\Big(\text{number interactions from } (x,s) \text{ to } (y,t)\Big)\\ &= \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\; B(y,t|\alpha,\tau)\,\mathbb{P}(\text{an interaction happens at } \tau\big|B_\tau = \alpha)\,B(\alpha,\tau|x,s)\\ &= \lambda\int_s^t d\tau\int_{\mathbb{R}^d} d\alpha\; B(y,t|\alpha,\tau)\,V(\alpha)\,B(\alpha,\tau|x,s)\end{aligned} \tag{5.4.5}$$

and similarly

$$\begin{aligned}&\mathbb{E}\Big(\text{number of time-ordered interaction-pairs from } (x,s) \text{ to } (y,t)\Big)\\ &= \lambda^2\int_s^t d\tau_2\int_s^{\tau_2} d\tau_1\int_{\mathbb{R}^d} d\alpha_2\int_{\mathbb{R}^d} d\alpha_1\; B(y,t|\alpha_2,\tau_2)\,V(\alpha_2)\,B(\alpha_2,\tau_2|\alpha_1,\tau_1)\,V(\alpha_1)\,B(\alpha_1,\tau_1|x,s)\end{aligned} \tag{5.4.6}$$

More formally we may say that there is a set I consisting of n interaction-times as follows:

$$I = \{\tau_1^I, \cdots, \tau_n^I\}$$

where each interaction-time τ_i^I is defined as the first interaction after the previous one, i.e.

$$\tau_j^{\mathrm{I}} = \inf_{s \le \tau_j \le t} (\tau_j : \tau_j > \tau_{j-1} \text{ and an interaction happens at } \tau_j)$$

and where we have that $\inf \varnothing = \infty$ and $\tau_0^I = s$, such that τ_1^I is the time of the first interaction, and there are only a finite number of finite τ_i^I. For some n onwards, we have that all τ_j^I are ∞, i.e. they have simply not happened. Now consider the random variable N, the *number* of interactions, defined as follows

$$N := \sum_{j=1}^{\infty} \int_s^t d\tau \; \delta(\tau_j^I - \tau)$$

where only the τ_j^I that are finite fall inside the interval $[s, t]$ and contribute to the sum. We calculate a weighted probability to propagate to y, where the weight is determined by the number of interactions: i.e. paths with one interactions are counted once, paths with two interactions are counted twice, etcetera. We see that this equals

$$\begin{aligned}
\mathbb{E}\Big(\delta(B_t - y)\, N | B_s = x\Big) &= \mathbb{E}\left(\delta(B_t - y) \sum_{j=1}^{\infty} \int_s^t d\tau \; \delta(\tau_j^I - \tau) \Big| B_s = x\right) \\
&= \int_s^t d\tau \; \mathbb{E}\left(\delta(B_t - y) \sum_{j=1}^{\infty} \delta(\tau_j^I - \tau) \Big| B_s = x\right)
\end{aligned}$$

And by the tower property, we have

$$\mathbb{E} \sum_{j=1}^{\infty} \delta(\tau_j^I - \tau) = \mathbb{E}\left[\mathbb{E}\Big(\sum_{j=1}^{\infty} \delta(\tau_j^I - \tau) \Big| B_\tau\Big)\right] = \lambda\, \mathbb{E}\, V(B_\tau) = \lambda \int_{\mathbb{R}^d} d\alpha V(\alpha) B(\alpha, \tau | x, s)$$

and thus we have

$$\mathbb{E}\Big(\delta(B_t - y)\, N | B_s = x\Big) = \lambda \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha \; B(y, t | \alpha, \tau)\, V(\alpha)\, B(\alpha, \tau | x, s)$$

Confirming that the first correction term picks up a contribution *every* time an interaction happens — showing that the second row of the matrix-equation in Theorem 2 is correct. Next, consider

$$\begin{aligned}
\mathbb{E}\Big(\delta(B_t - y)\, N^2 | B_s = x\Big) &= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \mathbb{E}\Big(\delta(B_t - y) \int_s^t d\tau \, \delta(\tau_i^I - \tau) \int_s^t d\theta \, \delta(\tau_j^I - \theta) \Big| B_s = x\Big) \\
&= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \mathbb{E}\left(\int_s^t d\tau \int_s^t d\theta \, \delta(B_t - y)\, \delta(\tau_i^I - \tau)\, \delta(\tau_j^I - \theta) | B_s = x\right).
\end{aligned}$$

We see that the integrand is symmetric in i and j, and therefore we can separate out the sum over 'diagonal' and 'off-diagonal' terms as follows:

$$\mathbb{E}\Big(\delta(B_t-y)\,N^2|B_s=x\Big) = 2\sum_{j=1}^{\infty}\sum_{i=j+1}^{\infty}\mathbb{E}\left(\int_s^t d\tau\int_s^t d\theta\,\delta(B_t-y)\,\delta(\tau_i^I-\tau)\,\delta(\tau_j^I-\theta)|B_s=x\right),$$
$$+\sum_{i=1}^{\infty}\mathbb{E}\left(\int_s^t d\tau\int_s^t d\theta\,\delta(B_t-y)\,\delta(\tau_i^I-\tau)\,\delta(\tau_i^I-\theta)|B_s=x\right).$$

As for the 'off-diagonal' terms, we see that we are left with an integration over the square $[s,t]^2$. However, because i is always larger than j, the peaks contributions of $\delta(\tau_i^I-\tau)$ appear at later times than those of $\delta(\tau_i^I-\theta)$. Therefore we only need to integrate over the triangle (rather than the square) defined by $\tau>\theta$. Thus we have

$$\mathbb{E}\Big(\delta(B_t-y)\,N^2|B_s=x\Big) = 2\sum_{j=1}^{\infty}\sum_{i=j+1}^{\infty}\mathbb{E}\left(\int_s^t d\tau\int_s^\tau d\theta\,\delta(B_t-y)\,\delta(\tau_i^I-\tau)\,\delta(\tau_j^I-\theta)|B_s=x\right),$$
$$+\sum_{i=1}^{\infty}\mathbb{E}\left(\int_s^t d\tau\int_s^t d\theta\,\delta(B_t-y)\,\delta(\tau_i^I-\tau)\,\delta(\tau_i^I-\theta)|B_s=x\right).$$

where only the upper limit in the integral over θ has changed. The integration is now time-ordered, and we may use the tower-property to obtain

$$\mathbb{E}\Big(\delta(B_t-y)\,N^2|B_s=x\Big) = 2\lambda^2\int_s^t d\tau\int_s^\tau d\theta\,B(y,t|\alpha_2,\tau)\,V(\alpha_2)\,B(\alpha_2,\tau|\alpha_1,\theta)\,V(\alpha_1)\,B(\alpha_1,\theta|x,s)$$
$$+\lambda\int_s^t d\tau\int_{\mathbb{R}^d}d\alpha\;B(y,t|\alpha,\tau)\,V(\alpha)\,B(\alpha,\tau|x,s).$$

For example if $N=3$ interactions happens, then $N^2=9$. There are 3 possible time-ordered pairs, and the off-diagonal terms contribute twice that (i.e. 6), whereas the diagonal terms contribute 3, adding to 9 as it should. Focusing on the off-diagonal terms only, we see that the second row in the matrix-equation of Theorem 2 consists of *half* the off-diagonal terms.

It follows that the number of *time-ordered* pairs of interactions is given by

$$\mathbb{E}\Big(\delta(B_t-y)\,1/2\,N(N-1)|B_s=x\Big) =$$
$$\lambda^2\int_s^t d\tau\int_s^\tau d\theta\,B(y,t|\alpha_2,\tau)\,V(\alpha_2)\,B(\alpha_2,\tau|\alpha_1,\theta)\,V(\alpha_1)\,B(\alpha_1,\theta|x,s).$$

This term counts paths with 2 hits once (as there is one way to pick a time ordered pair out of 2 interactions), paths with 3 hits 3 times (as there are 3 ways to pick a time-ordered pair from 3 interactions), paths with 4 hits 6 times (as there are $4*3/2$ ways to pick a pair out of 4 interactions). And similarly for higher order terms.

However, we have also shown how to calculate exactly not only the probability for zero interactions, but for *any* number of interactions! Writing the factors of λ explicitly in the

Pascal matrix, we get

$$\begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & -\lambda & +\lambda^2 & -\lambda^3 & +\lambda^4 & \cdots \\ 0 & +\lambda & -2\lambda^2 & +3\lambda^3 & -4\lambda^4 & \cdots \\ 0 & 0 & +\lambda^2 & -3\lambda^3 & +6\lambda^4 & \cdots \\ 0 & 0 & 0 & +\lambda^3 & -4\lambda^4 & \cdots \\ 0 & 0 & 0 & 0 & +\lambda^4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} B \\ BVB \\ BVBVB \\ BVBVBVB \\ BVBVBVBVB \\ \vdots \end{pmatrix}$$

where the right-hand side is schematic: all integrations and arguments have been left implicit. This equation makes very clear how we can obtain the probability of *any* number of interactions, from the probability of 0 interactions. For example, to obtain the probability of 1 interaction, we may apply the operator $-\partial_\lambda$ to the first row of the equation, and then set λ equal to 1. The strength of the coupling constant can always be normalised such that it equals 1; we only want differentiate with respect to it, and then set it equal to 1. We can obtain the coefficients of other rows similarly, and we conclude that

$$\psi_{i \text{ interactions}}(y,t|x,s)\Big|_{\lambda=1} = (-1)^i \frac{1}{i!} \left(\partial_\lambda\right)^i \psi_{0 \text{ interactions}}(y,t|x,s)\Big|_{\lambda=1}$$

All these series converge in alternating fashions. Recall that we could write

$$\psi_V(y,t|x,s) = \psi_{0 \text{ interactions}}(y,t|x,s) = \frac{1}{1+\lambda K*} B(y,t|x,s)$$

With the transition to $\lambda = 1$, we obtain

$$\psi_{i \text{ interactions}}(y,t|x,s)\Big|_{\lambda=1} = \frac{(K*)^i}{(1+K*)^{i+1}} B(y,t|x,s)$$

The expansion of the right-hand side as a series gives exactly the coefficients that can also be read off from the inverted Pascal matrix above. Rewriting the last expression, we get

$$(K*)^i B = (1+K*)^{i+1} \psi_{i\,\text{interactions}} \qquad \text{where } \lambda = 1,$$

where the integral-equation for $i = 0$ brings us back to the setting where only paths with zero interactions survive, and the expansion of which (by substitution into itself) gives the same combinatorial factors as can be read off from the Pascal matrix. We have now connected a whole set of integral equations with the Pascal matrix, of which our original integral equation is only a special case, namely $i = 0$.

By differentiating the Feynman-Kac formula with respect to λ we thus get that the probability of exactly i interactions equals $\psi_{i \text{ interactions}}$ as follows:

$$\psi_{i \text{ interactions}}(y,t|x,s) = \mathbb{E}_x \left(\delta(B_t - y) \frac{1}{i!} \left[\int_s^t V(B_\tau) d\tau \right]^i e^{-\int_s^t V(B_\tau) d\tau} \right) \quad \text{for } \lambda = 1, \tag{5.4.7}$$

where we have set $\lambda = 1$ (after the differentiation). This formula returns the Feynman-Kac formula (with a Dirac δ-function as the initial condition) simply by substituting $i = 0$. Concluding, this formula provides a new interpretation for the 'Feynman-rules' of a diffusion with an annihilating potential. It turns out that we have recovered an expression that reminds us strongly of an inhomogeneous Poisson process. An inhomogenious Poisson process counts the number of events that occur within a given time interval. Arrivals are independent and occur at each time τ with probability $\lambda(\tau)$. It is well-known that the random number N, which counts the number of events in the period $[s, t]$, is distributed like this:

$$\mathbb{P}\left(N = i\right) = \frac{1}{i!}\left(\int_s^t \lambda(\tau)d\tau\right)^i e^{-\int_s^t \lambda(\tau)d\tau},$$

The resemblance with the above is clear. □

The last expression can also be understood intuitively. Consider for example the propagation from (x, s) to (y, t) with exactly 1 interaction, on either of N intermediate locations or the final location y. To obtain the probability that this happens, we must multiply $N+1$ probabilities: N for 'no interaction' and 1 for an interaction. Also, we must sum over all locations where the 1 interaction could have happened. With the probability of an interaction being proportional to V and the probability of no interaction being proportional to $1 - V$, we obtain the following heuristic expression:

$$\sum_{j=1}^{N+1} V(B_{\tau_j})\,\epsilon \prod_{i\neq j}^{N+1} (1 - \epsilon\, V(B_{\tau_i})) \approx \sum_{j=1}^{N+1} V(B_{\tau_j})\,\epsilon \prod_{i=1}^{N+1} e^{-\lambda\,\epsilon\, V(B_{\tau_i})} \to \left(\int_s^t V(B_\tau)d\tau\right) e^{-\lambda\int_s^t V(B_\tau)d\tau}$$

and for multiple interactions a similar argument can be made, where the symmetry factor $\frac{1}{i!}$ is needed not to over count.

If we can calculate the propagator corresponding to the potential V exactly, (i.e. not as a series), then we can also calculate the propagator with exactly i interactions exactly, by differentiating i times with respect to the coupling constant λ, and multiplying by $\frac{(-1)^i}{i!}$, and setting $\lambda = 1$.

When the potential is positive, we have obtained an alternating series. But when the potential is negative, then the operator K is negative, and the series converges in a monotone fashion. In terms of counting paths, what is going on? In the annihilated case, the propagator had to count *only* paths with *no* interactions with the potential, because any interaction killed the particle. For particle creation, the propagator should similarly count paths with no interactions once, because they contribute to the density at the target point once. But now paths with one interaction, however, should be counted twice: corresponding to *one* doubling of the weight. Paths with 2 interactions should be counted 4 times, and paths with 3 interactions 8 times. Each path contributes to the final

density corresponding to its weight. Again we write the λ^1 term

$$\int_{\mathbb{R}^d} d\alpha\, B(y,t|\alpha,\tau)\,\lambda|V(\alpha)|B(\alpha,\tau|x,s),$$

where the only difference with the previous case is that we need absolute bars around V to interpret it as the probability of an interaction (V itself is negative). Just as before we find that the first perturbation term counts more than just paths with one interaction. Paths with two interactions are counted twice, and paths with three interactions are counted three times. Therefore, we have again:

$$\begin{pmatrix} \text{free term} \\ |\lambda^1 \text{ term}| \\ |\lambda^2 \text{ term}| \\ |\lambda^3 \text{ term}| \\ |\lambda^4 \text{ term}| \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 2 & 3 & 4 & \cdots \\ 0 & 0 & 1 & 3 & 6 & \cdots \\ 0 & 0 & 0 & 1 & 4 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix}.$$

Previously we counted only paths with zero interactions. In this case what we want to do is count paths with i interactions 2^i times. Thus we want to obtain

$$\text{total density} = \begin{pmatrix} 1 & 2 & 4 & 8 & 16 & \cdots \end{pmatrix} \cdot \begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix}.$$

By adding the rows of the Pascal matrix above, we see that this can be obtained by

$$\text{total density} = \text{free term} + \sum_{i=1}^{\infty} \left|\lambda^i \text{ term}\right|,$$

where it is obvious that the total density should converge in a monotone fashion.

5.5 Boundary value problems as Feynman-Kac potentials

In subsection 3.9 we proved Theorem 1: for all $x, y \in D$, for all regular boundary coordinates β, and for all domains allowing Green's theorem (1.3.1), the following problem

formulations are equivalent:

$$\left.\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) A(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) A(y,t|x,s) &= 0 \\
A(\beta,t|x,s) &= 0 \\
A(y,t|\beta,s) &= 0 \\
\lim_{s\nearrow t} A(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} A(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FP}\ A(y,t|x,s) = B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} A(\alpha,\tau|x,s) \\
&\text{LP}\ A(y,t|x,s) = B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ A(y,t|\alpha,\tau)\left\{-\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} B(\alpha,\tau|x,s)
\end{aligned}\right.$$

and

$$\left.\begin{aligned}
\left(\partial_t - \tfrac{\sigma^2}{2}\nabla_y^2\right) R(y,t|x,s) &= 0 \\
\left(\partial_s + \tfrac{\sigma^2}{2}\nabla_x^2\right) R(y,t|x,s) &= 0 \\
\overrightarrow{\partial_\beta} R(\beta,t|x,s) &= 0 \\
R(y,t|\beta,s)\overleftarrow{\partial_\beta} &= 0 \\
\lim_{s\nearrow t} R(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} R(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FR}\ R(y,t|x,s) = B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ R(y,t|\alpha,\tau)\left\{\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} B(\alpha,\tau|x,s) \\
&\text{LR}\ R(y,t|x,s) = B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\left\{\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}\right\} R(\alpha,\tau|x,s)
\end{aligned}\right.$$

and in particular that the perturbation expansion of the 4 integral equations on the right-hand side produces exactly the first- and last-passage (reflection) series of Propositions 3 and 4. We may compare this with Proposition 5, which concluded that for all $x, y \in \mathbb{R}^d$, and for a potential V that does not grow exponentially at infinity, the following problem formulations are equivalent:

$$\left.\begin{aligned}
\left(\tfrac{\partial}{\partial t} - \tfrac{\sigma^2}{2}\nabla_y^2 + \lambda V(y)\right) \psi_V(y,t|x,s) &= 0 \\
\left(\tfrac{\partial}{\partial s} + \tfrac{\sigma^2}{2}\nabla_x^2 - \lambda V(x)\right) \psi_V(y,t|x,s) &= 0 \\
\lim_{y\to|\infty|} \psi_V(y,t|x,s) &= 0 \\
\lim_{x\to|\infty|} \psi_V(y,t|x,s) &= 0 \\
\lim_{s\nearrow t} \psi_V(y,t|x,s) &= \delta(|y-x|) \\
\lim_{t\searrow s} \psi_V(y,t|x,s) &= \delta(|y-x|)
\end{aligned}\right\} = \left\{\begin{aligned}
&\text{FI}\ \psi_V(y,t|x,s) = B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ B(y,t|\alpha,\tau)\Big\{\lambda V(\alpha)\Big\}\psi_V(\alpha,\tau|x,s) \\
&\text{LI}\ \psi_V(y,t|x,s) = B(y,t|x,s) \\
&\quad - \int_s^t d\tau \int_{\mathbb{R}^d} d\alpha\ \psi_V(y,t|\alpha,\tau)\Big\{\lambda V(\alpha)\Big\}B(\alpha,\tau|x,s)
\end{aligned}\right.$$

Thus we see it is *tempting* to define an absorbing/reflecting potential as follows:

$$V(\alpha) := \mp\frac{\sigma^2}{2}\nabla_\alpha^2 \mathbb{1}_{\alpha\in D}$$

This potential manages both to replicate the desired physical situation (namely reflect or absorb), while also allowing for an easily interpretable and computable perturbation series. We can make sense of this seemingly ill-defined function either by 1) a limiting procedure, or by 2) using partial integrations (or Green's theorem) as if everything is well-behaved. The function V above has, to the author's best knowledge, never been defined before.

Finally, in Proposition 6, we saw that we can associate 2 integral equations with each Feynman-Kac functional, i.e.

$$\left.\begin{array}{c}\psi_V(y,t|x,s) = \\ \mathbb{E}_{(x,s)}\Big(\delta(B_t-y)\mathrm{Exp}\Big[-\lambda\int_s^t V(B_\tau)d\tau\Big]\Big)\end{array}\right\} = \left\{\begin{array}{ll}\text{FI} & \psi_V(y,t|x,s) = B(y,t|x,s) \\ & \quad -\int_s^t d\tau\int_{\mathbb{R}^d} d\alpha\; B(y,t|\alpha,\tau)\Big\{\lambda\, V(\alpha)\Big\}\psi_V(\alpha,\tau|x,s) \\ \text{LI} & \psi_V(y,t|x,s) = B(y,t|x,s) \\ & \quad -\int_s^t d\tau\int_{\mathbb{R}^d} d\alpha\; \psi_V(y,t|\alpha,\tau)\Big\{\lambda\, V(\alpha)\Big\}B(\alpha,\tau|x,s)\end{array}\right.$$

Combining all the above, we now come to the final theorem of this paper:

Theorem 3. ABM and RBM as Feynman-Kac exponentials. *For all $x, y \in D$, for all regular boundary coordinates β, and for all domains allowing Green's theorem (1.3.1), the following problem formulations are equivalent:*

$$\left.\begin{array}{r}\Big(\partial_t - \frac{\sigma^2}{2}\nabla_y^2\Big)A(y,t|x,s) = 0 \\ \Big(\partial_s + \frac{\sigma^2}{2}\nabla_x^2\Big)A(y,t|x,s) = 0 \\ A(\beta,t|x,s) = 0 \\ A(y,t|\beta,s) = 0 \\ \lim_{s\nearrow t} A(y,t|x,s) = \delta(|y-x|) \\ \lim_{t\searrow s} A(y,t|x,s) = \delta(|y-x|)\end{array}\right\} = \left\{\begin{array}{r}\Big(\partial_t - \frac{\sigma^2}{2}\nabla_y^2 + V(y)\Big)\psi_V(y,t|x,s) = 0 \\ \Big(\partial_s + \frac{\sigma^2}{2}\nabla_x^2 - V(x)\Big)\psi_V(y,t|x,s) = 0 \\ -\frac{\sigma^2}{2}\nabla_\alpha^2\mathbb{1}_{\alpha\in D} = V(\alpha) \\ \\ \lim_{s\nearrow t}\psi_V(y,t|x,s) = \delta(|y-x|) \\ \lim_{t\searrow s}\psi_V(y,t|x,s) = \delta(|y-x|)\end{array}\right. \tag{5.5.1}$$

where the left-hand problem is defined for x and y in the interior of D and all regular boundary points β. The right-hand problem is defined for all x and y in $\mathbb{R}^d$. But the claim is that

$$A(y,t|x,s) = \psi_V(y,t|x,s)\;\forall x,y\in D \tag{5.5.2}$$

where

$$A(y,t|x,s) = \mathbb{E}_x\left(\delta(B_t-y)\; e^{\frac{\sigma^2}{2}\int_s^t \nabla_u^2\mathbb{1}_{u\in D}(B_\tau)d\tau}\right). \tag{5.5.3}$$

Similarly, we conclude for the reflected transition density R that

$$\left.\begin{array}{r}\Big(\partial_t - \frac{\sigma^2}{2}\nabla_y^2\Big)R(y,t|x,s) = 0 \\ \Big(\partial_s + \frac{\sigma^2}{2}\nabla_x^2\Big)R(y,t|x,s) = 0 \\ n_\beta\cdot\overrightarrow{\nabla}_\beta R(\beta,t|x,s) = 0 \\ R(y,t|\beta,s)\overleftarrow{\nabla}_\beta\cdot n_\beta = 0 \\ \lim_{s\nearrow t} R(y,t|x,s) = \delta(|y-x|) \\ \lim_{t\searrow s} R(y,t|x,s) = \delta(|y-x|)\end{array}\right\} = \left\{\begin{array}{r}\Big(\partial_t - \frac{\sigma^2}{2}\nabla_y^2 + V(y)\Big)\psi_V(y,t|x,s) = 0 \\ \Big(\partial_s + \frac{\sigma^2}{2}\nabla_x^2 - V(x)\Big)\psi_V(y,t|x,s) = 0 \\ \frac{\sigma^2}{2}\nabla_\alpha^2\mathbb{1}_{\alpha\in D} = V(\alpha) \\ \\ \lim_{s\nearrow t}\psi_V(y,t|x,s) = \delta(|y-x|) \\ \lim_{t\searrow s}\psi_V(y,t|x,s) = \delta(|y-x|)\end{array}\right. \tag{5.5.4}$$

where again the problem on the left-hand side is defined for x and y in the interior of D and all regular boundary points β, while the right-hand side is defined for all x and y in $\mathbb{R}^d$. The claim is that

$$R(y,t|x,s) = \psi_V(y,t|x,s) \ \forall x,y \in D \tag{5.5.5}$$

where

$$R(y,t|x,s) = \mathbb{E}_x\left(\delta(B_t - y)\ e^{-\frac{\sigma^2}{2}\int_s^t \nabla_u^2 \mathbb{1}_{u\in D}(B_\tau)d\tau}\right). \tag{5.5.6}$$

In conclusion, the absorbing (reflecting) potential is identified as

$$V(u) := \mp\frac{\sigma^2}{2}\nabla_u^2 \mathbb{1}_{u\in D}.$$

The Taylor expansion of the Feynman-Kac exponential converges as follows:

convergence of expansion	absorbed BM	reflected BM
convex domain	*alternating*	*monotone*
concave domain	*monotone*	*alternating*

While we admit that the path integral cannot be calculated exactly, and that the potential $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$ looks ill defined, at least we can say we have obtained an elegant *short-hand* for the expansion of the Green function. In their book on random walks and path integrals, [48] write

> a clear indication of one of the advantages of the generating function, is that it represents a prescription for the construction of the special function that it generates. [...] In this sense, the generating function encapsulates all information with regard to the function that it generates. Furthermore, it contains this information in an extremely compact form.

We conclude that our expression does exactly the above — for the transition density of absorbed or reflected Brownian motion. In the words of theoretical physicist and chemist [49]:

> One of the principal objects of theoretical research in my department of knowledge is to find the point of view from which the subject appears in its greatest simplicity.

In this view, the main contribution of this paper is that it provides a solution — to the heat kernel with boundary conditions and by extension for the (modified) Dirichlet problem, as pioneered by [3] — that is 1) new and 2) very compact. If one were to communicate the solution in the least possible number of bits, then this would be a very good candidate. It is remarkable that the potential V above has — to the author's best knowledge — never been defined before.

6 Conclusion

This paper considered the modified Dirichlet and Neumann boundary value problems for the heat and Laplace equations, where the value and normal derivative are prescribed on the boundary. Our approach was probabilistic in nature, interpreting the heat kernel as the absorbed or reflected transition density of a Brownian motion.

We contrasted our approach with that of classical potential theory and its *ansatz* of single and double boundary layers. We find that 1) single and double boundary layers need not be based on an ansatz, but follow from the first- and last-interaction (or reflection) decompositions 2) either problem may be solved with either method and their distinction is thus arbitrary, and 3) they may be useful for irregular as well as regular domains, by virtue of Green's theorem. We also showed that all the information in the problem may be represented by two integral equations with the potential is taken to be $\mp\frac{\sigma^2}{2}\nabla_u^2\mathbb{1}_{u\in D}$ (Theorem 1). Furthermore we showed that the series converge in an alternating or monotone fashion depending on the geometry of the domain (convex or concave).

Second, we considered a new approach to the Feynman-Kac functional. We introduced the first- and last-interaction decompositions as analogous to the first- and last-passage decompositions, and we again derived two integral equations rather than one. The perturbation of this series was shown to be in agreement with the Taylor expansion of the Feynman-Kac exponential, as it expected. But we proposed a *new* set of Feynman rules (Theorem 2) to explain why the series converges in an alternating or monotone fashion, which differs from what can be found in for example [45]. The proposed Feynman rules involve the Pascal matrix, which to date has not been connected with the Feynman-Kac exponential, and are as follows:

$$\begin{pmatrix} \text{free term} \\ \lambda^1 \text{ term} \\ \lambda^2 \text{ term} \\ \lambda^3 \text{ term} \\ \lambda^4 \text{ term} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 2 & 3 & 4 & \cdots \\ 0 & 0 & 1 & 3 & 6 & \cdots \\ 0 & 0 & 0 & 1 & 4 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \cdot \begin{pmatrix} \text{paths with 0 interactions} \\ \text{paths with 1 interaction} \\ \text{paths with 2 interactions} \\ \text{paths with 3 interactions} \\ \text{paths with 4 interactions} \\ \vdots \end{pmatrix}$$

where each λ^i-term is positive. It implies that the propagation with any number i of interactions occurs with the following probability:

$$\mathbb{P}\Big(B_t = y;\ \text{exactly } i \text{ interactions}\Big|B_s = x\Big) = \mathbb{E}_x\left(\delta(B_t - y)\,\frac{1}{i!}\left[\int_s^t V(B_\tau)d\tau\right]^i e^{-\int_s^t V(B_\tau)d\tau}\right)$$

where $\lambda = 1$, and where substituting $i = 0$ immediately returns the Feynman-Kac formula with a δ-function as the initial condition. In the presence of an annihilating potential, only

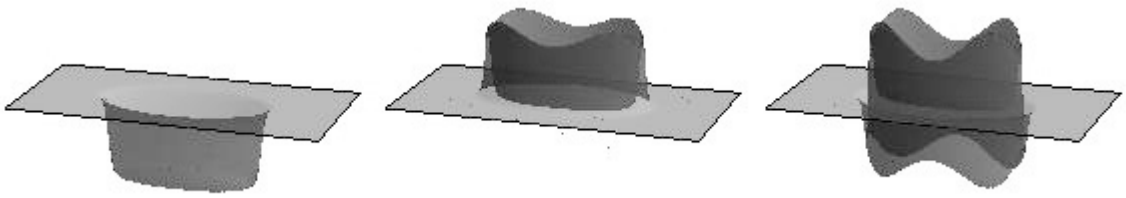


Figure 3. In this graph, the domain is taken to be an ellipse in $d = 2$. On the left we have a smoothed out (negative) indicator of this ellipse, where the smoothing depends on ϵ. For $\epsilon \searrow 0$ we get $-\mathbb{1}_{x\in D}$, where the domain D is the ellipse. For any $\epsilon > 0$, all three functions plotted above are continuously differentiable to all orders. But in the limit where $\epsilon \searrow 0$, then we get $-\mathbb{1}_{x\in D}$ (left-hand graph), $-n \cdot \nabla_x \mathbb{1}_{x\in D}$ (middle graph) and $-\nabla_x^2 \mathbb{1}_{x\in D}$ (right-hand graph). The singular quantity $-n \cdot \nabla_x \mathbb{1}_{x\in D}$ goes to positive ∞ on the boundary of the ellipse, while $-\nabla_x^2 \mathbb{1}_{x\in D}$ goes to $-\infty$ infinitesimally on the outside, and $+\infty$ infinitesimally on the inside. Both singular quantities are zero anywhere else, just like the Dirac δ- and δ'-functions.

paths with $i = 0$ interactions survive. This situation in probability theory corresponds to the Schrödinger equation in quantum mechanics (by a rotation to/from imaginary time).

Finally, in Theorem 3, we proposed the synthesis of classical potential theory and the theory of path integrals by postulating the following seemingly ill-defined potential:

$$V(x) := \mp \frac{\sigma^2}{2} \nabla_x^2 \mathbb{1}_{x\in D}$$

and by showing that the transition density of absorbed or reflected Brownian motion can be written as

$$\mathbb{E}_x \left(\delta(B_t - y)\ e^{\pm \frac{\sigma^2}{2} \int_s^t \nabla_u^2 \mathbb{1}_{u\in D}(B_\tau) d\tau} \right).$$

This connects, as a by-product, potential theory to the study of Brownian local time. The potential can be viewed as the 'acceleration' of the time spent in D by the Brownian particle when the boundary points of D move outwards in the normal direction. The function V above has — to the author's best knowledge — never been defined before. We can make sense of this seemingly ill-defined function either by 1) a limiting procedure, or by 2) using partial integrations (or Green's theorem) as if everything is well-behaved.

The potential also shows, for the first time, that the Dirichlet and Neumann problems are very closely related: the potential generating the absorbed/reflected density differs only by a *sign*. We have noted that positive potentials destroy paths while negative potentials create paths. Through the one dimensional analogy, we see that the Laplacian of the Heaviside step-function, $\pm\nabla_x^2 \mathbb{1}_{x\in D}$, is equally positive and negative, as in the right-most graph in Figure 3.

As a result the proposed potential conserves *particle number*. If a particle reaches the boundary of the domain, it is both copied (by the negative peak) and destroyed (by the positive peak). But these actions happen at slightly different places. If the copying happens just inside the domain, and the destroying just outside, then the boundary is reflecting from the inside: every time it hits the boundary it is destroyed just outside the domain and put back just on the inside. But if the destroying happens just inside the domain while the copying happens just outside, then the particle can get out but it can never get back in. Seen from the inside, therefore, the boundary acts as an absorbing barrier. This intuition explains why the potential for the Dirichlet and Neumann problems differ only by a sign: $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$ is reflecting from one side, and absorbing from the other. In one dimension this can easily be verified (see subsection 1.8).

This result is new, and we believe that this is the first time that a boundary value problem has been turned into a potential problem. While we admit that the path integral cannot be calculated exactly and that the potential $\mp\frac{\sigma^2}{2}\nabla_x^2\mathbb{1}_{x\in D}$ looks ill defined, we can at least say we have obtained a compact *short-hand* for the expansion of the Green function, where the convergence of its Taylor series is as follows:

mode of convergence	absorbed BM	reflected BM
convex domain	*alternating*	*monotone*
concave domain	*monotone*	*alternating*

and where the smoothness requirement on the domain is *only* that it allows Green's identity (1.3.1) — in contrast with all series solutions that are based on an *ansatz* — and thus all piecewise smooth domains in $d \geq 2$ are included for the first time.

In each case, two series are possible: one where all the differential operators work towards the right, and one where all the differential operators work towards the left. While those series are identical, term by term, in classical potential theory one of them would be classified as a double boundary layer and the other as a single boundary layer. Here we have shown that their equivalence follows from the equivalence of the first- and last-passage decompositions (if no passage is allowed, neither first nor last passages may happen). Therefore we propose that there is no fundamental difference between single and double boundary layers either, and neither is there a need for either to be presented as an *ansatz*.

Furthermore, this is the first time that Feynman-Kac path integrals have been used to study boundary value problems, whereas previously these were considered separate fields of study. The fundamental reason for the complexity of using path integrals for boundary value problems is that path integrals assume the possibility of movement throughout the whole of space. Boundary value problems, however, confine the particle to a particular region of space. It is tempting to postulate an infinite potential outside of the allowed

region, such that every path is annihilated there, but as a result all terms in the Taylor series become infinite.

Merging the subjects of path integrals and boundary value problems has thus been difficult. Either the potential does not correspond to the desired physical situation (as it does not contain the particle), or it does, but its perturbation expansion contains terms that are all infinite. It seems impossible to reconcile the two, and this is the first time that a potential is presented that corresponds to the physical situation *while also* allowing an easily interpretable and computable perturbation expansion.

The main contribution of this paper is that it provides a solution to the heat kernel with boundary conditions — and by extension for the (modified) Dirichlet problem, as pioneered by Gauss in 1840 [3] — that is 1) new and 2) very compact. If one aimed to communicate the solution in the least possible number of bits, then the solution provided in this paper would be a good candidate.

References

[1] O. Kellogg, *Foundations of potential theory.* Dover Pubns, 1929. 3, 4, 5, 7, 10, 56

[2] J. L. Doob, *Classical Potential Theory and Its Probabilistic Counterpart (Classics in Mathematics).* Springer, Mar., 2001. 3, 7, 10, 14

[3] C. Gauss, *Allgemeine Lehrsätze in Beziehung auf die im verkehrten Verhältnisse des Quadrates der Entfernung wirkenden Anziehungs-und Abstossungs-Kräfte.* K. Gesellsch. d. Wiss. sch., 1840. 5, 31, 117, 121

[4] S. Zaremba, *Sur le principe de Dirichlet*, *Acta Mathematica* **34** (1911), no. 1 293–316. 5, 13

[5] H. Lebesgue, *Sur des cas dimpossibilité du probleme de Dirichlet*, *Comptes Rendus de la Société Mathématique de France* **41** (1913) 17. 5, 13

[6] H. Poincaré, *Théorie du potentiel Newtonien.* G. Carré et C. Naud, 1899. 5

[7] N. Wiener, *The Dirichlet problem.* 1924. 5

[8] G. Green, *An essay on the application of mathematical analysis to the theories of electricity and magnetism.* Printed for the author by T. Wheelhouse, 1828. 5, 6, 7, 8, 16

[9] F. Knight, *Essentials of Brownian motion and diffusion, volume 18 of Mathematical Surveys.* 1981. 7, 10, 41, 50

[10] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus (Graduate Texts in Mathematics).* Springer, 2nd ed., Aug., 1991. 7, 10, 21, 39, 41, 50

[11] R. Bass, *Probabilistic techniques in analysis.* Springer, 1995. 7

[12] P. Mörters and Y. Peres, *Brownian Motion.* Cambridge University Press, 1 ed., Mar., 2010. 7, 11, 12, 13, 21, 23, 26, 40, 49, 103, 105

[13] N. Muskhelishvili, *Singular integral equations. Translated from the second Russian edition.* Groningen: Wolters-Noordhoff Publishing, 1967. 7, 17, 50, 62, 80

[14] S. C. Port and C. J. Stone, *Brownian Motion and Classical Potential Theory.* Academic Press, New York, 1978. 7, 9, 14, 40, 50, 97

[15] K. Itô and H. P. McKean, *Diffusion Processes and their Sample Paths.* Springer, Feb., 1996. 7, 10, 21, 41, 50

[16] W. Jurkat and D. Nonnenmacher, *The general form of Green's theorem*, in *Proc. Amer. Math. Soc*, vol. 109, pp. 1003–1009, 1990. 8

[17] R. Bartle, *A modern theory of integration.* American Mathematical Society, 2001. 8

[18] L. Bachelier, *Théorie de la spéculation*, in *Annales Scientifiques de lÉcole Normale Supérieure*, vol. 17, pp. 21–86, 1900. 8

[19] A. Einstein, *On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat*, *Annalen der Physik* **17** (1905) 549–560. 8

[20] N. Wiener, *Differential space*, *Journal of Mathematical Physics* **2** (1923) 131–174. 9

[21] K. Itô, *On a formula concerning stochastic differentials*, *Nagoya Mathematical Journal* **3** (1951) 55–65. 9

[22] M. Baxter and A. Rennie, *Financial calculus: an introduction to derivative pricing*. Cambridge University Press, 1996. 9

[23] A. Etheridge and M. Baxter, *A course in financial calculus*. Cambridge University Press, 2002. 9

[24] J. Norris, *Markov chains*. No. 2008. Cambridge University Press, 1998. 9

[25] J. D. Jackson, *Classical Electrodynamics Third Edition*. Wiley, Third ed., Aug., 1998. 10, 14, 16, 26, 48, 53

[26] K. Chung, *Green, Brown, and probability*. World Scientific Publishing Company, 1995. 11, 13, 14, 45, 48

[27] S. Kakutani, *Two-dimensional Brownian motion and harmonic functions*, *Proceedings of the Japan Academy, Series A, Mathematical Sciences* **20** (1944), no. 10 706–714. 12

[28] G. Brosamler, *A probabilistic solution of the Neumann problem*, *Math. Scand* **38** (1976), no. 1 137–147. 14

[29] P. Hsu, *Probabilistic approach to the neumann problem*, *Communications on Pure and Applied Mathematics* **38** (1985), no. 4 445–472. 14

[30] P. Hsu, *On Excursions of Reflecting Brownian Motion*, *Transactions of the American Mathematical Society* **296** (1986), no. 1 239–264. 14, 19, 39, 49

[31] R. Bass and P. Hsu, *Some potential theory for reflecting Brownian motion in Hölder and Lipschitz domains*, *The Annals of Probability* **19** (1991), no. 2 486–508. 14

[32] R. Balian and C. Bloch, *Distribution of eigenfrequencies for the wave equation in a finite domain I. Three-dimensional problem with smooth boundary surface*, *Annals of Physics* **60** (Oct., 1970) 401–447. 16, 18, 19, 62, 64

[33] W. Hackbusch, *Integral equations: theory and numerical treatment*. Birkhauser Verlag, Basel, Switzerland, 1995. 17, 56, 62, 64, 80

[34] A. Polyanin, A. Manzhirov, and A. Polianin, *Handbook of integral equations*. CRC press, 1998. 17, 50, 56, 60, 62, 68

[35] T. H. Hansson and R. L. Jaffe, *Cavity quantum chromodynamics*, *Phys. Rev. D* **28** (Aug, 1983) 882–907. 19, 62

[36] T. H. Hansson and R. L. Jaffe, *The multiple reflection expansion for confined scalar, Dirac, and gauge fields*, *Annals of Physics* **151** (1983), no. 1 204 – 226. 19, 62, 63

[37] M. Bordag and D. V. Vassilevic, *Heat kernel expansion for semitransparent boundaries*, *Journal of Physics A: Mathematical and General* **32** (1999), no. 47 8247. 19

[38] M. Bordag, D. Vassilevich, H. Falomir, and E. M. Santangelo, *Multiple reflection expansion and heat kernel coefficients*, *Phys. Rev. D* **64** (Jul, 2001) 045017. 19, 56, 62, 63, 74

[39] M. Kac, *Can One Hear the Shape of a Drum?*, *The American Mathematical Monthly* **73** (1966), no. 4 1–23. 20, 81

[40] K. Stewartson and R. Waechter, *On hearing the shape of a drum: further results*, in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 69, pp. 353–363, Cambridge University Press, 1971. 20, 81

[41] M. Protter, *Can One Hear the Shape of a Drum? Revisted*, *Siam Review* **29** (1987), no. 2 185–197. 20

[42] O. Giraud and K. Thas, *Hearing shapes of drums: Mathematical and physical aspects of isospectrality*, *Reviews of modern physics* **82** (2010), no. 3 2213–2255. 20

[43] R. Feynman, *Space-time approach to non-relativistic quantum mechanics*, *Reviews of Modern Physics* **20** (1948), no. 2 367. 21

[44] M. Kac, *On distributions of certain Wiener functionals*, *Trans. Amer. Math. Soc* **65** (1949), no. 1 1–13. 21, 102

[45] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*. McGraw-Hill Companies, June, 1965. 23, 24, 107, 118

[46] L. H. Ryder, *Quantum Field Theory*. Cambridge University Press, 2 ed., June, 1996. 23, 24, 103, 105, 107, 108

[47] W. Janke and H. Kleinert, *Summing Paths for a Particle in a Box*, *Lettere Al Nuovo Cimento (1971–1985)* **25** (1979), no. 10 297–300. 25

[48] J. Rudnick and G. Gaspari, *Elements of the random walk: an introduction for advanced students and researchers*. Cambridge University Press, 2004. 30, 117

[49] J. Gibbs, *The Collected Works of J. Willard Gibbs: Thermodynamics*, vol. 1. Longmans, Green and Co., 1928. 31, 117

[50] D. Geman and J. Horowitz, *Occupation densities*, *The Annals of Probability* **8** (1980), no. 1 1–67. 33

[51] P. Lévy, *Processus stochastiques et mouvement brownien. Suivi d'une note de M. Loève*, . 33

[52] J. Dodziuk, *Eigenvalues of the Laplacian and the Heat Equation*, *The American Mathematical Monthly* **88** (1981), no. 9 686–695. 39, 81

[53] S. Port and C. Stone, *Classical potential theory and Brownian motion*, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 3, pp. 143–176, 1972. 46

[54] M. Kac, *On some connections between probability theory and differential and integral equations*, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. July 31-August 12, 1950. Statistical Laboratory of the University of California, Berkeley. Berkeley, Calif.: University of California Press, 1951. 666 pp. Editor: Jerzy Neyman*, vol. 1, pp. 189–215, 1951. 48

[55] K. Chung, *Probabilistic approach in potential theory to the equilibrium problem*, *Ann. Inst. Fourier* **23** (1973), no. 3 313–322. 48

[56] D. Porter and D. Stirling, *Integral equations: a practical treatment, from spectral theory to applications*. Cambridge University Press, 1990. 60, 68, 80, 81

[57] G. Peskir, *On integral equations arising in the first-passage problem for Brownian motion*, *J. Integral Equations Appl* **14** (2002) 397–423. 68

[58] I. Wolfram Research, "Mathematica Edition: Version 8.0." Wolfram Research, Inc., Champaign, Illinois, 2010. 87

[59] K. Yajima, *Existence of solutions for Schrödinger evolution equations*, *Communications in Mathematical Physics* **110** (1987), no. 3 415–426. 95

[60] R. Shankar, *Principles of quantum mechanics*. Springer, 1994. 97

[This page was intentionally left blank]

TO BE SUBMITTED

Part II — The problem of alternatives

Rutger-Jan Lange[1]

University of Cambridge,
792 King's College, Cambridge, CB2 1ST, United Kingdom

E-mail: rjl63@cam.ac.uk

ABSTRACT: We consider parallel investment in several alternative technologies or drugs that are developed over time, where only the *one* project with the best final performance goes to market. While simultaneous investment in parallel lines of research increases the probability of success of at least one project, it is also costly. As the 'performance' of each project changes stochastically over time, we are presented with a d-dimensional performance space. We assume that the performance of each project is given by some function of one or more Brownian motions, and of time. At each point in time it must be determined which candidates show sufficient performance and/or potential to justify further investment. We propose a new integral equation that all boundary points of the continuation domain must simultaneously satisfy. This integral equation can be solved numerically in some cases — but not in all cases, yet. We provide a numerical example, where only one of two options may be exercised.

KEYWORDS: problem of alternatives, multidimensional optimal stopping, MOS, optimal stopping, sequential investment, free-boundary problem, boundary value problem, Dirichlet problem, Neumann problem, integral equation, Brownian motion

[1]Research support from the Electricity Policy Research Group in Cambridge is gratefully acknowledged (http://www.eprg.group.cam.ac.uk).

Contents

1 Introduction

This paper considers parallel investment in several alternative technologies or drugs that are developed over time, and where there can only be *one* winner. At each point in time it must be determined which candidates show sufficient performance and/or promise to justify further investment. We call this the *problem of alternatives*.

The problem of alternatives can be viewed in the wider context of *multidimensional optimal stopping* (MOS) *problems*. Although the theory for 1-dimensional optimal stopping problems is well developed — see for example the extensive literature on the American option — far fewer results can be found regarding MOS.

The explanation of the problem of alternatives lends itself well to an informal discussion with examples. But to treat the problem of alternatives in an MOS setting, we need to extend the theory of 1-dimensional optimal stopping to higher dimensions, which is a relatively technical endeavour. The introduction of this paper will therefore review the *whole* of the paper in an expository manner: including the results, intuition and outline of the proof, but excluding the proof itself. In particular, this introduction will

1. explain the problem of alternatives, and why it is relevant (subsection 1.1),
2. explain how it can be formulated as an MOS problem (1.2 and 1.3),
3. present its relationship with Brownian motion and free-boundary problems (1.4 and 1.5),
4. present the main result and its intuition (1.6),
5. discuss the relationship with the literature (1.7).

Section 2 discusses d-dimensional Brownian motion in domain $D(\cdot)$ with boundary conditions, and provides the necessary mathematical tools for the proof in Section 3. Section 4 provides a numerical example, and Section 5 concludes.

1.1 The problem of alternatives

It is generally true that uncertainty plays a large role in investment decisions, especially in complex and changing environments, and when consequences play out over long timeframes. Within any firm multiple projects can be developed, each with its own risk and reward profile, and often these projects compete for resources. Furthermore, their success may be mutually exclusive. There may initially be many candidates for any one goal, such as in new drug development. Many compounds are initially considered, but only one is ultimately used in a new medicine. It may therefore be sensible for a pharmaceutical company to invest in a *portfolio of alternatives*.

When betting on an externally organised race that can only have one winner, such as a horse race, the optimal (risk-neutral) strategy is straightforward: bet everything (i.e. go all in) on what presently looks like the best candidate, i.e. the one that has the highest expected value. But in-house investment in a portfolio of alternatives is different in at least two respects:

1. First, unlike in a horse race, the number of projects that are running is a *decision variable* on the part of the investor. The investor could run fewer candidates at a smaller cost, but also with a smaller probability that one of them will be a blockbuster. Or the investor could develop more projects, where by definition more effort will turn

out to have been in vain because only one project can win, but with a larger likelihood that the winner performs well.

2. And second, real projects, unlike financial bets, often have a *natural size* and cannot easily be scaled up or down: these are what the literature calls *real options*. They can either be part of the race, or not at all — and thus the decision to include them is a digital one.

We thus conclude that a firm could face any given number of optional projects — where each can be part of the race or not, with its own intrinsic size and cost — and where the project with the best final performance wins. Investing in a large portfolio of alternatives is costly but also increases the expected performance of the winner, and thus even a risk-neutral strategy could do well to diversify.

In the simplest example, we may suppose that there are only two projects. Suppose that these projects have random revenues given by the normally distributed random variables $N_i(\mu_i, \sigma_i^2)$ with means μ_i and variance σ_i^2, and known development costs, given by c_i. In this table we indicate the difference between a financial portfolio, a real portfolio and a portfolio of alternatives:

Invest in...	Financial portfolio	Real portfolio	Portfolio of alternatives
Project 1 only	$N_1 - c_1$	$N_1 - c_1$	$N_1 - c_1$
Project 2 only	$N_2 - c_2$	$N_2 - c_2$	$N_2 - c_2$
Both projects	$a(N_1 - c_1) + b(N_2 - c_2)$ $a + b = 1$	$(N_1 - c_1) + (N_2 - c_2)$	$\max(N_1, N_2) - c_1 - c_2$

Every time we write the phrase 'portfolio of alternatives', the mathematical formulation involves a max-function, as in the table above. The expectation of the max-function is, in this case, driven by four parameters: the means and variances of both N_1 and N_2. With a little work we can derive that

$$\mathbb{E}\big[\max(N_1, N_2)\big] = \frac{{\sigma_1}^2 + {\sigma_2}^2}{\sqrt{2\pi}\sqrt{{\sigma_1}^2 + {\sigma_2}^2}} e^{\frac{-(\mu_1-\mu_2)^2}{2({\sigma_1}^2+{\sigma_2}^2)}} + \frac{1}{2}\mu_1 \mathrm{Erfc}\left[\frac{\mu_2 - \mu_1}{\sqrt{2}\sqrt{{\sigma_1}^2 + {\sigma_2}^2}}\right] + \frac{1}{2}\mu_2 \mathrm{Erfc}\left[\frac{\mu_1 - \mu_2}{\sqrt{2}\sqrt{{\sigma_1}^2 + {\sigma_2}^2}}\right]$$

where 'Erfc' is the complementary error function Erfc = 1− Erf.[1] It can be shown that the expectation of $\max(N_1, N_2)$ is increasing in all four parameters μ_1, μ_2, σ_1, σ_2. This contrasts starkly with the paradigm of financial portfolio theory, in which the maximisation of risk-neutral expectation amounts to putting all eggs in the basket with the highest expected return — and only risk aversion provides an incentive for diversification. For a portfolio of alternatives, the effects may be diagonally opposite: maximising expected value may lead to investing in several projects such that the expected performance of the winner

[1] $\mathrm{Erf}(x) = 2/\sqrt{\pi} \int_0^x e^{-t^2} dt$

is higher, whereas risk aversion may lead to investing in one project only, so as to avoid a scenario in which multiple projects fail to produce an acceptable result. Whereas the values of financial and real portfolios are *only* affected by the averages of the distributions involved — e.g. μ_1 and μ_2 in this case — the value of a portfolio of alternatives is affected by all the (entire) distributions, and in particular by their right-end tail behaviour — e.g. as determined in this case by σ_1 and σ_2. In other words, the expectation of a max-function 'feels' the variance of its constituents. The expectation of a sum is *not* sensitive to the variance of its components, and that is why a portfolio of alternatives behaves differently from a 'normal' portfolio.

As a result, it is possible (but not necessary) that, amongst the three portfolios of alternatives above, the portfolio 'invest in both' has the highest expectation: for finite c_1 and c_2 and high enough σ_1 and σ_2, the portfolio containing both alternatives will dominate the value of either alone — and therefore a risk-neutral investor, who maximises expectation, would invest in both projects. We see that the required actions for a portfolio of alternatives may be diagonally opposite from the actions that would be taken for a financial portfolio:

	Financial portfolio	Real portfolio	Portfolio of alternatives
Increase expectation	narrow down	possibly diversify	possibly diversify
Decrease risk	diversify	possibly narrow down	possibly narrow down

Real portfolios and portfolios of alternatives are similar in the sense that risk increases with the number of projects that are executed, i.e. with the amount of money on the table. But the expectation of the portfolio of alternatives is driven by distributions rather than expected values. In this paper, we deal only with the risk-neutral case.

Apart from the non-additivity of revenues of alternative projects, a further difference with standard (financial or real) portfolio theory is that the problem of alternatives is dynamic. Investment in any one project will change the performance of that project stochastically — while altering, at the same time, the potential for *other* projects to win or lose, which is a departure from a dynamic (additive!) portfolio theory. Furthermore, projects can be discontinued at any time. When simultaneously developing several competing projects, therefore, the question arises whether the option value of having multiple projects at any point in time still outweighs the cost. In the early stages of development the option value is expected to be dominant, but it is equally clear that not all projects should be pursued to the end.

Both a non-additive revenue structure (e.g. when projects are alternatives) and an explicit time element are necessary ingredients for the problem of alternatives. But there is a third ingredient: when several projects compete for the best result, the decision maker may choose to invest either sequentially or in parallel. In a sequential strategy, and when

i projects have already been performed, the decision to be made is whether or not to accept profit $= \max(N_1, \ldots, N_i) - \sum_{j=1}^{i} c_j$ or to add project $i+1$ to the list. Adding another project raises the total cost by c_{i+1} and may raise the maximum, but *only if* the $(i+1)$-th project outperforms all the previous ones. The advantage of a sequential strategy is that the decision on whether or not to invest in the next project is taken once all the previous results are already known, which allows for a better informed decision. The disadvantage, however, is that a sequential search strategy pushes possible revenues further into the discounted future. There are several versions of this problem, such as the 'Secretary problem' (or 'Marriage problem'); see e.g. [1]. In a similar style, Weitzman [2] discusses 'Pandora's problem': in what order to execute different alternative projects, where only the winner determines the revenues, and when to accept the current winner and stop searching? Weitzman's sequential strategy is optimal when discounting is insignificant, and when there is an infinite time-horizon. But when the horizon is finite, or when discounting of future revenues is significant, then parallel investment can outperform sequential investment.

Therefore it is *conceivable* that it is optimal to develop several alternative technologies or drugs in parallel, even when there can only be one winner. Early on, the option value is expected to be dominant, but it is clear that not all projects should be pursued to the end. Finding the right balance, however, is surprisingly tricky — especially analytically. The resulting problem can be classified as a *multidimensional optimal stopping* (MOS) *problem*.

One might be tempted to ask for conditions under which parallel development is more profitable than sequential investment, before we try to solve the problem of alternatives. But we need to solve the problem of alternatives first, before we can write down the conditions under which it is profitable. However, it is clear that *some* set of conditions exists under which parallel development is optimal — and we know that it is driven by increased discounting.

Concluding, we need the following four ingredients for an interesting MOS problem:

1. The different projects need to be *alternatives*, as exemplified by the use of 'max' in the mathematical formulation. Or more generally: there needs to be some structure ensuring that the total revenue is a *non-linear* function of the project revenues. When profits are additive, such as for a standard portfolio, then one never needs to make 'multidimensional decisions', as each project can be optimised in isolation.

2. There needs to be a *time* element, and it needs to be possible to *stop* investing. In the 1-dimensional case this field is known as *optimal stopping* and hence our generalisation to *multidimensional optimal stopping* (MOS).

3. There needs to be a *finite time horizon* or *sufficiently high discounting* such that sequential investment is not optimal.

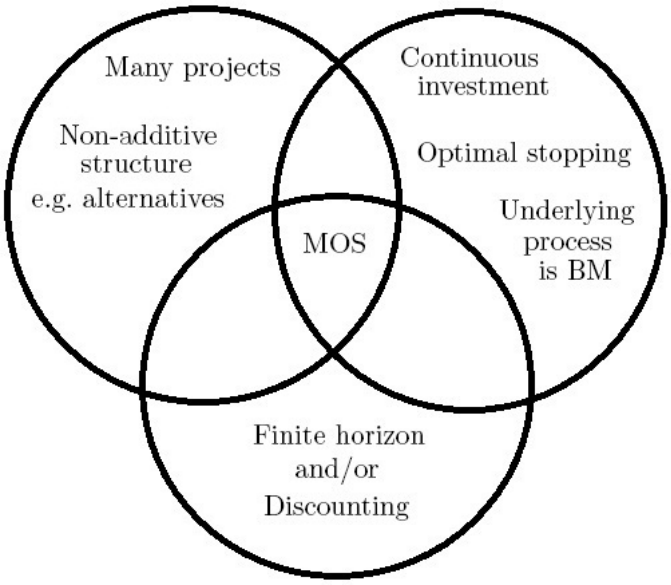


Figure 1. A non-additive revenue structure, an explicit time element and a finite horizon (or significant discounting) are all necessary conditions for a non-trivial multidimensional optimal stopping (MOS) problem. A fourth assumption that we will make is that all projects start at the same time, and cannot be restarted once abandoned.

4. Lastly, we assume that all projects start at the same time and any project can be abandoned at any time. Once abandoned, however, projects cannot be restarted.[2]

If the revenues of different projects are additive, each can be considered individually and there is no need for a multidimensional analysis. If there is no time element or possible stopping, it becomes a one-shot problem. If the problem poses neither discounting nor maturity, a sequential strategy will outperform a parallel strategy. These key ingredients are indicated in Figure 1.

As an illustration of the problem of alternatives, consider the rescue of the 33 Chilean miners who were trapped 700 metres underground in a copper-gold mine for 69 days from 5 August 2010. The three tunnels being dug by the rescuers were alternatives (i.e. only one was needed), it was possible to review the progress of the three tunnels over time (tunnels A and C were abandoned once B was almost complete), and it was highly desirable that at least one tunnel should be finished quickly (significant discounting). While in this case the decision for a parallel strategy was hardly the result of a cost-benefit analysis, the example does illustrate clearly that there are indeed situations when a parallel strategy is optimal. The progress of the three different tunnels after 64 (out of the 69) days is indicated in Figure 2, when plan B had only 90 more metres to go, and which appeared on the BBC

[2]Projects are not required to have the same end-date.

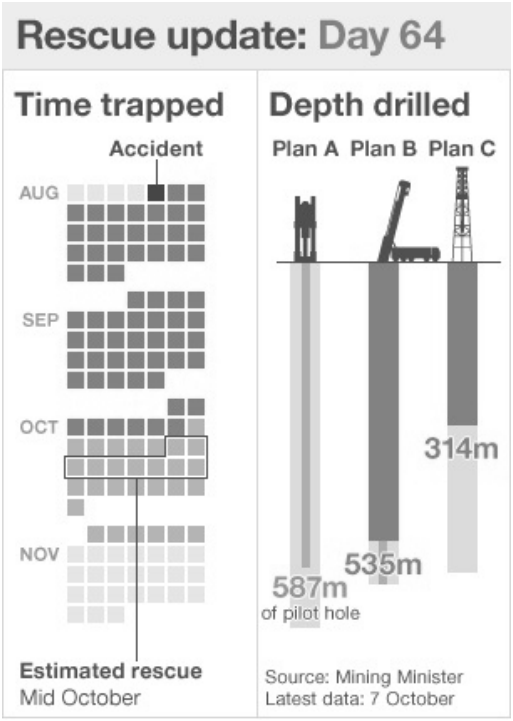


Figure 2. The progress of the three different tunnels after 64 (out of the 69) days, when tunnel B had only 90 more metres to go. From the BBC website of 8 October 2010.

website on 8 October 2010.[3]

For the purposes of this paper, the discounting of (or deadlines for) the success of different projects can be either 1) caused by competitive pressure, or 2) self-imposed. In the case of drug development, for example, it may be desirable to find a successful candidate quickly, e.g. to combat a new and contagious disease, even if that means spending money on many alternatives in the beginning. In the case of public spending on several alternative green technologies, the government may wish to impose a deadline for success, so that failing technologies are not supported indefinitely. For instance, the German government has announced that the solar feed-in tariffs will decrease by 9% per annum. This implies, by extrapolation, that solar energy must become economical around 2020, when the feed-in tariff matches predicted industrial market prices. Implicitly, solar energy competes not only against the incumbent technology that sets the 2020 target, but also against other low-carbon technologies such as wind — where the extent to which different low-carbon technologies are alternatives is debatable.

[3] http://www.bbc.co.uk/news/world-latin-america-11497394

1.2 A portfolio of two alternatives as an MOS problem

Consider once more the fictional projects 1 and 2 introduced above. The final performance, i.e. at maturity $T = 1$, is still given by the normally distributed random variables N_1 and N_2 — except now the performance develops *gradually* and stochastically. In particular, the performance at time $t < T$ could be determined by

$$P_1(t) = \mu_1\, t + \sigma_1\, B_{1,t}$$
$$P_2(t) = \mu_2\, t + \sigma_2\, B_{2,t}$$

where $B_{1,t}$ and $B_{2,t}$ are independent Brownian motions, satisfying $\mathbb{E}\, B_{i,t} = 0$ and $\mathbb{E}\, B_{i,t}^2 = t$. As a result, the final performances $P_1(T)$ and $P_2(T)$ are distributed as before:

$$P_1(T) \sim N_1(\mu_1 T, \sigma_1^2 T)$$
$$P_2(T) \sim N_2(\mu_2 T, \sigma_2^2 T)$$

where the $\sim$ sign means 'is distributed like' and where the end-date of each project is taken to be $T = 1$. We have shown that it may be profitable to develop both projects simultaneously, even if only one of them can win. The value of *unconditionally* completing both projects (i.e. without any intermediate revision) equals:

$$e^{-r\,T}\ \mathbb{E}\left[\max(P_1(T), P_2(T))\right] - \int_0^T d\tau\, e^{-r\,\tau}(c_1 + c_2)$$

where c_1 and c_2 are now interpreted as the expenditure on projects 1 and 2 per infinitesimal unit of time, i.e. (infinitesimal) continuation cost. We have shown that for high enough σ_1 and σ_2, this value *could* be higher than 1) the value of either project alone, and 2) the value of a sequential strategy. But it is obvious that *unconditionally* completing both projects is suboptimal. It may become clear, for example, at some time *earlier* than T that project 1 is much more likely to win than project 2. In this case, it may be optimal to abandon project 2 and continue only with project 1. From then onwards, the remaining project 1 will be developed optimally, and in isolation. We assume that the abandoned project 2 cannot be restarted. Therefore, the value of optimal parallel investment in projects 1 and 2 is given by $V_{1,2}$ as follows:

$$V_{1,2}(\{x,y\}, s) := \max_{s \le \tau \le T} \mathbb{E}_{(x,y)} \left[e^{-r\,(\tau - s)} \max\Big\{ V_1(B_{1,\tau}, \tau), V_2(B_{2,\tau}, \tau) \Big\} - \int_s^\tau d\theta\, (c_1 + c_2)\, e^{-r(\theta - s)} \right]$$

where V_1 and V_2 are the optimal values of projects 1 and 2, if they were continued optimally and in isolation, where the maximisation is over stopping time τ, and where the conditioning in the subscript of $\mathbb{E}$ is on the values of $B_{1,s} = x$ and $B_{2,s} = y$. The above says that the optimal value of having two alternative projects at time s equals 1) the expectation of the (optimal) value of the single project that is chosen at time τ, 2) minus the cost

to get there. In this particular example, the 'continuation gain' is negative and equal to $-(c_1 + c_2)$ per unit of time that both projects are continued, and the 'stopping gain' (i.e. when the choice is made) equals $\max\{V_1(B_{1,\tau}, \tau), V_2(B_{2,\tau}, \tau)\}$. The optimal value defined as such is an 'optimal value to-go', i.e. all costs already paid (and sunk) are not included. $V_{1,2}$ thus gives the optimal value *from now on*. The optimal values of projects 1 and 2 are given by:

$$V_1(x,s) := \max_{s \le \tau \le T} \mathbb{E}_x \left[e^{-r\,(T-s)}\,(\mu_1\,T + \sigma_1\,B_{1,T})\,\mathbb{1}_{\tau=T} - \int_s^\tau d\theta\,c_1\,e^{-r\,(\theta-s)} \right],$$

$$V_2(x,s) := \max_{s \le \tau \le T} \mathbb{E}_x \left[e^{-r\,(T-s)}\,(\mu_2\,T + \sigma_2\,B_{2,T})\,\mathbb{1}_{\tau=T} - \int_s^\tau d\theta\,c_2\,e^{-r\,(\theta-s)} \right].$$

Here $\mathbb{1}$ is the indicator function, which equals 1 if the condition in its subscript is satisfied and 0 otherwise, and the maximisation is over all stopping times τ. The optimal value of either project in isolation equals an expectation of the performance at maturity, if and only if the project is not abandoned before that time, minus an expectation of the continuation cost c_i which is to be paid at each unit of time when the project is not stopped. Again, V_i is an optimal value 'to-go', i.e. it only takes into account *future* costs and revenues, because everything received and paid so far is already sunk.

1.3 A portfolio of multiple alternatives as an MOS problem

We will now show how to extend the two-project portfolio of alternatives to the situation with many projects. In a general MOS problem, we may face three possible 'gains', where a gain can be positive, if it is received, or negative, if it is to be paid:

- A continuation gain G_C, to be received/paid at every infinitesimal unit of time dt when the process is not stopped.
- A stopping gain G_S, to be received/paid when the process is stopped before maturity.
- A maturity gain G_M, to be received/paid when the process reaches maturity (i.e. without being stopped).

Occasionally, we will write 'continuation cost' instead of 'continuation gain' when G_C is negative. Each of the three possible gains may depend both on state and time, i.e. $G_S = G_S(y,t)$, $G_C = G_C(y,t)$ and $G_M = G_M(y,T)$. Not every combination of the three gains adds up to a non-trivial optimal stopping problem. If only the maturity gain is non-zero, for example, and positive, then one would always wait until time T and collect G_M. Such cases can be solved by inspection, and thus we assume that the problem has a non-trivial solution and proceed from there.

For the parallel investment in projects 1 and 2, we see that the continuation gain equals $-(c_1 + c_2)$, which is to be paid at every infinitesimal instant of time when both projects are continued. The stopping gain equals the gain from choosing one project over the other and continuing that project optimally, i.e. $\max\{V_1(B_{1,\tau}, \tau), V_2(B_{2,\tau}, \tau)\}$. Once the decision is taken to continue only one project, the situation changes: the continuation gain now becomes the cost to keep that project alive for a small unit of time (i.e. c_i), the stopping gain goes to zero (as the project can be abandoned without cost), and the maturity gain goes to $P_i(T)$. This is summarised in the following table:

	$V_{1,2}$	V_1	V_2
Continuation gain	$-c_1 - c_2$	$-c_1$	$-c_2$
Stopping gain	$\max\left\{V_1(B_{1,\tau}, \tau), V_2(B_{2,\tau}, \tau)\right\}$	0	0
Maturity gain		$P_1(T)$	$P_2(T)$

We expect that the continuation region of $V_{1,2}$ — corresponding to continued investment in both projects — shrinks to zero as time goes to T, because an optimal policy would never allow both projects to be completed. It would be better, for example, to abandon the project that is almost certainly going to lose a short time before completion.

If we assume that we can solve 1-dimensional optimal stopping problems, then V_1 and V_2 are known functions. We may therefore focus on exclusively on the $V_{1,2}$ column in the above table, and try to determine $V_{1,2}$ given the continuation, stopping and maturity gains in that column. Finding $V_{1,2}$ amounts to solving an MOS problem.

Supposing that we can solve the 2-dimensional optimal stopping problem, we may proceed and add a *third* project to the mix. The optimal value of developing three projects in parallel is as follows:

$$V_{1,2,3}(\{x, y, z\}, s) := \max_{s \le \tau \le T} \mathbb{E}_{(x,y,z)} \left[e^{-r\,(\tau - s)} \max\left\{V_{1,2}, V_{1,3}, V_{2,3}\right\} - \int_s^\tau d\theta\, (c_1 + c_2 + c_3)\, e^{-r(\theta - s)} \right]$$

where the 'continuation gain' equals $-(c_1 + c_2 + c_3)$ for every unit of time that all three projects are pursued in parallel, and the 'stopping gain' is equal to the value of continuing the chosen pair optimally, and where the maximisation is over all stopping times τ. If we can solve the 2-dimensional optimal stopping problem above, then the $V_{i,j}$ are all known functions. Therefore, we obtain a 3-dimensional optimal stopping problem with known continuation and stopping gains. We can extend this to the situation of four alternative projects, and so on.

We conclude, therefore, that if we can solve a d-dimensional optimal stopping problem — with $d \ge 1$ and any *given* set of G_C, G_S and G_M — we can iteratively solve *all* these problems. First we would solve the optimal development of all i projects in isolation: we would need to solve i 1-dimensional optimal stopping problems. (If all projects were

identical, then only *one* 1-dimensional optimal policy would apply to all projects.) Second, we would solve the 2-dimensional problem for all possible pairs chosen from i projects, where the stopping gain is equal to continuing the single chosen project optimally. Third, we would solve the 3-dimensional problem, where the stopping gain is equal to continuing the one chosen pair optimally, and so on. Therefore, if we can solve a general optimal stopping problem in d dimensions — with given G_C, G_S and G_M — then we can iteratively build the solution to the entire problem. If all projects are identical, we need to solve only d problems: the 1-d problem once, the 2-d problem once, etc.

1.4 Brownian motion

In this paper we will show how to solve multidimensional optimal stopping (MOS) problems, where $d \geq 1$. We will take the underlying stochastic space to be a d-dimensional Brownian motion of unit variance in each spatial direction. We will call this a standard (d-dimensional) Brownian motion, or simply Brownian motion. Using a Brownian motion as the underlying process may be slightly unusual, because many authors take a geometric Brownian motion as the underlying stochastic process, for example. There is some arbitrariness in choosing the underlying process, but we will find it *useful* to choose a standard Brownian motion, and we lose no generality by doing so. But the fact that we choose a standard Brownian motion as the underlying stochastic process does imply something about the pay-off. To model an American option, for example, one usually takes a geometric Brownian motion GBM_t as the stochastic process and $\max\{K - GBM_t, 0\}$ as the pay-off. Instead we would take as the stochastic process the standard Brownian motion B_t and as pay-off $\max\{K - GBM_0\, e^{\mu\, t + \sigma\, B_t}, 0\}$. It is clear that both formulations are equivalent. In cases where only one of two American options may be exercised, as in [3], the stopping gain equals

$$G_S(GBM_{1,t}, GBM_{2,t}) = \max\left[\max\left\{GBM_{1,t}, GBM_{2,t}\right\} - K, 0\right]$$

where the GBMs are correlated geometric Brownian motions. The stopping gain G_S in our framework would be

$$G_S(B_{1,t}, B_{2,t}) = \max\left[\max\left\{GBM_{1,0}\, e^{\mu_1\, t + \sigma_1\, B_{1,t}}, GBM_{2,0}\, e^{\mu_2\, t + \sigma_2\, (\rho\, B_{1,t} + (1-\rho^2)^{1/2}\, B_{2,t})}\right\} - K, 0\right]$$

because the processes $B_{1,t}$ and $(\rho\, B_{1,t} + (1-\rho^2)^{1/2}\, B_{2,t})$ each have unit variance and are correlated with correlation ρ, where $B_{1,t}$ and $B_{2,t}$ are truly independent Brownian motions; see e.g. [4], p. 171.

Thus we allow all problems to be solved that are, ultimately, based on a standard Brownian motion. If the underlying process is truly different, for example when it allows jumps, as is the case for Lévy processes, then it cannot be modelled with the methods presented in this paper. But any process that is a function of Brownian motion can indeed be dealt with using these methods.

1.5 Free-boundary problems

Optimal stopping problems are closely related to boundary value problems. Boundary value problems take the boundary as *given*, and prescribe one boundary condition: the value at the boundary can be prescribed (Dirichlet problem), the normal derivative can be prescribed (Neumann problem), or a linear combination of the value and derivative can be prescribed (third boundary value problem).

Free-boundary problems in physics and optimal stopping problems in finance originate from different disciplines and have different objects of study (e.g. Stefan's ice-melting problem vs American options), but mathematically they are equivalent. For these free-boundary problems, as the name suggests, the boundary of the domain is not given, but instead two boundary conditions are specified: both the value and normal derivative are prescribed. The task, then, is to find the unique domain that allows both boundary conditions to be satisfied. In any dimension, the domain to be found corresponds to the region of continued investment in all alternatives. As soon as the stochastic process reaches the boundary, one project is terminated, the continuation space reduces to $d-1$ dimensions, and $d-1$ projects are continued optimally.

For optimal stopping problems with a finite horizon (i.e. maturity), the continuation domain and its boundary are in general time-dependent. Intuitively we would expect the optimal continuation domain to *shrink* over time, forcing a decision before time T. We may summarise as follows:

	domain $D(\cdot)$	value at $\partial D(\cdot)$	derivative at $\partial D(\cdot)$
Dirichlet problem	given	prescribed	to be found
Neumann problem	given	to be found	prescribed
MOS problem	to be found	prescribed	prescribed

Here and elsewhere, the dynamic domain and its boundary are indicated by $D(\cdot)$ and $\partial D(\cdot)$, and where these at a *specific* time t are indicated by $D(t)$ and $\partial D(t)$.

In classical potential theory, the Dirichlet problem was posed for the Laplace operator and for a static domain D. Parabolic potential theory poses the same problem for the heat operator. But in both cases the *value* is prescribed at the boundary of the static domain. We allow for a dynamic domain, but the fact that the value at the boundary is prescribed still justifies that we view it as a Dirichlet problem. This holds, too, for the classical Neumann problem and the parabolic Neumann problem, both of which prescribe the normal derivative at the boundary of a static domain. We allow for a dynamic domain and in that sense the terms 'Dirichlet problem' and 'Neumann problem' are used loosely: they refer only to what kind of boundary condition is satisfied.

The optimal value V in any dimension d should satisfy four conditions. First there is a partial differential equation to be satisfied in the interior of $D(\cdot)$. Then there are two

boundary conditions: both the value and normal derivative at the dynamic boundary are prescribed. Lastly there is a 'boundary condition' at maturity T, which prescribes the value of V at maturity. The task, then, is to find 1) the optimal value V and 2) the optimal dynamic domain $D(\cdot)$, where these must be determined simultaneously. The equations satisfied by an optimal V are as follows:

$$\begin{aligned}
&\text{The value is unbiased} && \left(\frac{1}{2}\nabla_x^2 + \frac{\partial}{\partial s} - r\right) V(x,s) = -G_C(x,s) && x \in D(s), \\
&\text{Value-matching condition} && V(\beta,s) = G_S(\beta,s) && \beta \in \partial D(s), \\
&\text{Smooth-pasting condition} && \partial_\beta V(\beta,s) = \partial_\beta G_S(\beta,s) && \beta \in \partial D(s), \\
&\text{Value at maturity} && V(x,T) = G_M(x,T) && x \in D(T).
\end{aligned} \tag{1.1}$$

The dynamic domain and its boundary at any particular time s are indicated by $D(s)$ and $\partial D(s)$, and the Laplacian in d dimensions is defined by

$$\nabla_x^2 := \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}.$$

Also, β is a regular boundary coordinate, and ∂_β is the inward normal derivative at the regular boundary coordinate β. Regular boundary points are defined as those allowing a tangent plane, at each point in time, such that the normal direction unambiguously exists. The optimal value at space-time coordinate (x, s) is indicated by $V(x, s)$, where x is a vector of positions, i.e. the space-time coordinate (x, s) equals $(\{B_{1,s}, \cdots, B_{d,s}\}, s)$. The constant discount rate is indicated by r. These four conditions are indicated for a 1-dimensional and a 2-dimensional token problem in Figure 3. The continuation regions that are drawn are only indicative. The differential equation is written as $\Delta V = -G_C$ where

$$\Delta := \frac{1}{2}\nabla_x^2 + \frac{\partial}{\partial s} - r.$$

It should also be noted that in the problem formulation (1.1), the three possible gains $G_\cdot$ all appear, and in different places. Apart from their intuitive appeal, therefore, their existence is also suggested by the mathematical formulation.

We have established that we can iteratively solve the problem of alternatives that we set out to solve, if we can solve (1.1) for any dimension d. We will now discuss each of these four conditions in (1.1).

1. The first condition says that the value is unbiased. When x is inside $D(s)$, a passage from the continuation region into the stopping region cannot happen immediately. Thus, as x progresses to $x+dB$ and s to $s+ds$, where both x and dB are d-dimensional vectors, the continuation gain $G_C(x, s)\, ds$ is received, and the value function goes to $V(x + dB, s + ds)$. The new value is obtained after time ds and must be discounted,

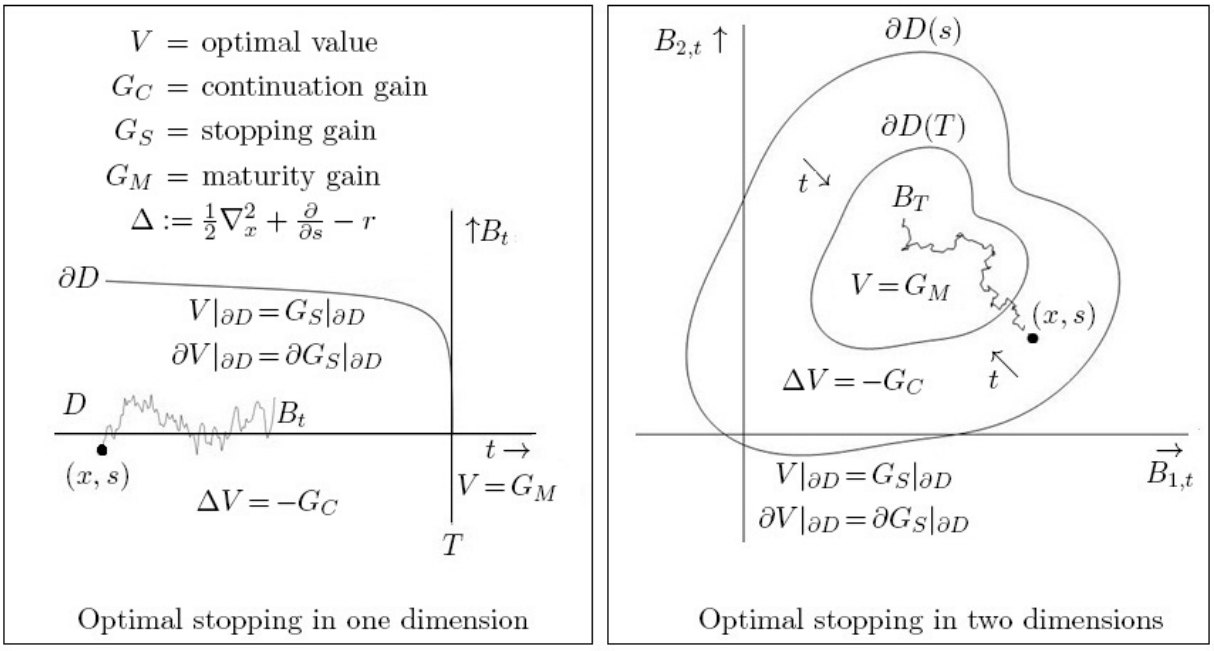


Figure 3. Optimal stopping problems in one and two spatial dimensions, with four conditions on the optimal value V, in each case. Together they determine the optimal value V and the domain $D(\cdot)$ uniquely. The situation in dimensions $d \geq 3$ is analogous — and in fact the four conditions appear unchanged. This is because the definition for the Laplacian ∇^2 and the inward normal derivative ∂ hold for any dimension d. In $d = 1$ we get simply that the Laplacian equals $\partial^2/\partial x^2$ and the inward normal derivative $\pm\partial/\partial x$, depending on the inward direction.

i.e. we get $e^{-r\,ds}V(x + dB, s + ds) + G_C(x, s)\,ds$. The new value will almost surely be different from the old value $V(x, s)$. But *on expectation*, it must be the same. We can thus write a Taylor expansion to first order in ds as follows:

$$\begin{aligned} V(x,s) &= \mathbb{E}\Big[V(x+dB, s+ds)\,e^{-r\,ds}\Big] + G_C(x,s)\,ds \\ &= \mathbb{E}\Big[\Big(V(x,s) + \nabla_x V(x,s)\cdot dB + \tfrac{1}{2}\nabla_x^2 V(x,s) dB\cdot dB + \tfrac{\partial}{\partial s}V(x,s)\,ds\Big)\Big(1 - r\,ds\Big)\Big] + G_C(x,s)\,ds \\ &= V(x,s) + \Big[\tfrac{1}{2}\nabla_x^2 + \tfrac{\partial}{\partial s} - r\Big]V(x,s)\,ds + G_C(x,s)\,ds \\ &= V(x,s) + \Big[\Delta V(x,s) + G_C(x,s)\Big]\,ds \end{aligned}$$

where we have used Itô's lemma. It should not be surprising that V is unbiased. Ultimately V is an expectation over all possible continuation gains, stopping gains and maturity gains, and thus — upon progressing a short time ds — the expectation should not be expected to change! In fact, this property also holds for *non-optimal* boundaries. As long as the value is defined as an expectation over all future gains until

some stopping time determined by the first exit out of the continuation region, then it must be unbiased. For continuous time optimal stopping problems, therefore, the equation that is often referred to as Bellman's dynamic programming equation has nothing to do with optimality, but only with the fact that expectations are unbiased.

2. The second condition in (1.1) says that stopping becomes imminent as x in $V(x, s)$ approaches a regular boundary coordinate β on $\partial D(s)$. We can allow a finite number of singular boundary points, because there is zero probability that a Brownian path will hit any of them. At regular boundary points β it is obvious that the Brownian motion immediately exits the continuation domain and thus the only value that is obtained, from then onwards, equals the immediate stopping value $G_S(\beta, s)$. Again it should be noted that this condition has nothing to do with optimality: if the value is defined as an expectation over all future gains and possible first exits out of a non-optimal domain, then the 'immediate stopping' condition will hold also at the boundary of the non-optimal domain.

3. We skip the third condition in (1.1) for the moment, and jump to the fourth. The fourth condition says that as s approaches the maturity date T, the only remaining contribution is the maturity gain G_M. Again this condition holds automatically by defining V as an expectation over all future gains, even when the domain is non-optimal: if the Brownian path never leaves the continuation domain, then it survives until time T and picks up G_M.

4. Given the above, we must conclude that the condition that actually defines optimality is the smooth-pasting condition in (1.1). The smooth-pasting condition (also known as 'smooth-fit' and 'high-order contact') may appear mysterious, but is very widely used and quoted; see for example [5] and [6]. The intuition is as follows: if smooth-pasting holds, then the value of V can be approximated, on both sides of the boundary, by one and the same hyperplane that is unambiguously defined. If instead there is a 'kink' in V at the boundary, then it can be approximated by two hyperplanes and thus locally V is either convex or concave — depending on how the two hyperplanes meet. At the boundary of the continuation domain the decision maker is supposed to be indifferent between continuing for an infinitesimally small time and stopping. Therefore, the value at the boundary (i.e. stopping immediately) must be equal to the average of continuing for a short time. If this average is taken over a strictly convex or concave future value V, then the continuation value will either be strictly higher or lower than the immediate stopping value — contradicting the supposed indifference. Kinks are thus not allowed and smooth-pasting must hold; for

a 1-dimensional explanation see e.g. [5] (p. 130), or for the multi-dimensional case see e.g. [3] (p. 249).

1.6 Assumptions and main result

Multidimensional optimal stopping is a relatively new field, in which little is known. To proceed, we will need to make the following (heroic!) assumptions:

1. The solution to the MOS problem (1.1), given by the pair $D(\cdot)$ and $V(x, s)$, exists and is unique.

2. The solution $D(\cdot)$ that solves the MOS problem (1.1) allows Green's theorem at each point in time: at each time it only has a finite number of edges, corners and cusps.

3. The solution $D(\cdot)$ that solves the MOS problem (1.1) allows Reynold's theorem: it moves with an integrable velocity, for all regular boundary points and at all times.

The question of whether or not a solution exists depends on the smoothness of the three gains. If G_S is smooth everywhere, then smooth-pasting should hold everywhere, and therefore the dynamic boundary *can* be smooth everywhere. If G_S is smooth everywhere and if the boundary itself is continuous, then smooth-pasting will hold everywhere except where the boundary has cusps or corners.

If G_S is non-smooth on some subset of $\mathbb{R}^d$, for example on the diagonal, as is the case for $\max\{x, y\}$, then smooth-pasting may not hold when the boundary crosses the diagonal. Often, in such cases, the optimal boundary in fact *never* crosses the diagonal and therefore smooth-pasting *still* holds for all boundary points. The 1-dimensional American option, for example, has a stopping gain G_S that is not smooth (it involves a max-function). But the boundary never crosses the level where there is a kink in G_S, and therefore smooth-pasting still holds everywhere. Even when the dynamic boundary *does* in fact cross the non-smooth subset a finite number of times, then we could hypothesise that the dynamic boundary would still be piecewise smooth — which is allowed by Green's theorem. Assuming the validity of Green's theorem, therefore, does not seem to drastically limit the set of problems we can solve.

As far as Reynold's theorem is concerned, we allow that the underlying stopping gain G_S depends on time. If it has a finite time-derivative at each spatial location, then there is no reason to expect any boundary element to have an infinite speed, except possibly at maturity. To see why it could have an infinite speed at maturity, consider again the well-known example of the 1-dimensional American option. The exercise boundary has an infinite slope *at* the horizon, but the 'speed' of the boundary at $t \to T$ is integrable because the distance travelled by the boundary is finite. In the multidimensional case we allow for

the same situation: where the location of the boundary is assumed to be integrable. Again, we do not find this assumption too stringent.

Our approach will be to assume, simply, that the abovementioned assumptions are satisfied. While our assumptions do not seem *overly* restrictive, it is not clear that they should hold in all cases. Furthermore, it is possible that our assumptions are not independent: one would be hard pressed to come up with a problem, for example, with a unique solution but including an *infinite* number of singular boundary points. Existence and uniqueness assumptions may therefore be equivalent to certain smoothness assumptions, but here we are guessing. Thus we allow our intuition (and Green's and Reynold's theorems) to inspire the assumptions we need, and we will see where this leads us. If the assumptions above hold, we obtain the following theorem:

Theorem 1. *If the solution to the MOS problem* (1.1) *exists and is unique, and if the optimal domain $D(\cdot)$ allows both Green's theorem and Reynold's theorem at all times, then the optimal value V is given by:*

$$\begin{aligned} V(x,s) = {} & G_S(x,s) \\ & + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} B(\alpha,\tau|x,s) \\ & \quad + \int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} B(\alpha,T|x,s) \end{aligned} \tag{1.2}$$

where $D(\cdot)$ is the optimal domain. The optimal domain can be found by applying either *value-matching* or *smooth-pasting to the optimal value, at all boundary coordinates $\beta \in \partial D(s)$, $\forall s \leq T$. If G_S is only piecewise smooth, then ΔG_S should be interpreted as involving distributional derivatives.*

Here $B(y,t|x,s)$ indicates the free Brownian density as defined in (2.1). This theorem is new, to the author's best knowledge — and it seems to be one of the first more general results in the field of MOS. The book *Optimal stopping and free boundary problems* by [6], for example, only tangentially touches upon the multidimensional case. [7] confine themselves to the 2-dimensional American option: $d = 2$ and only G_S is non-zero and required to be convex. [8] consider a portfolio of savings and stocks where rearranging occurs a transaction cost, and solve a 2-dimensional free-boundary problem with an infinite horizon. In [9] time is discretised. But there appears to be a lack of more general results regarding MOS.

The proof of Theorem 1 consists of two parts, i.e. we need to show that:

1. *Given* that $D(\cdot)$ is optimal, *then* V can be written as proposed.

2. *Given* this V, $D(\cdot)$ can be determined by demanding *either* value-matching *or* smooth-pasting for all boundary locations and times $\beta \in \partial D(\cdot)$.

We also provide some intuition for Theorem 1. The optimal value V is separated in an 'immediate-stopping' value G_S (the first term) and an 'option value' (the second and third term). The option value consists of a weighted integration over the continuation region $D(\cdot)$ at all times τ for $s \leq \tau \leq T$ (second term), as well as an integration over the continuation region at maturity (third term). We may define

- effective continuation gain $:= G_C(\alpha, \tau) + \Delta G_S(\alpha, \tau)$,
- effective maturity gain $:= G_M(\alpha, T) - G_S(\alpha, T)$.

At a location and time where the effective continuation gain is positive, one would always continue a short time dt. With these definitions, we see that the 'option value' consists of the expected sum of 'effective continuation gains' and 'effective maturity gains', as collected by a *free* Brownian motion, in all of the *future* continuation region $D(\cdot)$. As a result, the 'effective continuation gains' and 'effective maturity gain' are weighted by the free propagator $B(\alpha, \tau | x, s)$ and $B(\alpha, T | x, s)$. For $x \to \partial D(s)$ value-matching must hold. The value V must equal the stopping gain G_S, and therefore the 'option value' at the optimal boundary must be zero. Thus we have:

Corollary 1. *For x on the optimal boundary, the* expected *value of all effective continuation and effective maturity gains, as collected by the* free *Brownian path during its time in the optimal continuation domain $D(\cdot)$, equals zero. This holds true for* all *boundary locations $x \in \partial D(\cdot)$, i.e.*

$$0 = \mathbb{E}_x \left[\int_s^T d\tau \left(G_C(B_\tau, \tau) + \Delta G_S(B_\tau, \tau) \right) \mathbb{1}_{B_\tau \in D(\tau)} \, e^{-r\,(\tau - s)} + \left(G_M(B_T, T) - G_S(B_T, T) \right) \mathbb{1}_{B_T \in D(T)} \, e^{-r\,(T-s)} \right] \quad \forall x \in \partial D(s), \, \forall s \leq T. \tag{1.3}$$

This interpretation holds for all $d \geq 1$ and, to the author's best knowledge, is new. It follows that the optimality of any *single* boundary location is dependent on *all* future boundary locations, and therefore the *entirety* of all dynamic boundary locations must be determined at once. In general, we cannot determine any boundary location without knowing all others.

Although we can write the optimality equation for a general situation as in Corollary 1 — as an integration over $D(\cdot)$ — in any specific situation we will need to parametrise the boundary, in one way or another. The parametrisation of a volume, such as the domain $D(\cdot)$, allows one to identify the set of coordinates that lie within the volume.

Parametrisations are not generally unique: one can often parametrise a volume in either Cartesian or spherical coordinates, for example. A cube is more easily parametrised in Cartesian coordinates, and a sphere is more easily parametrised in spherical coordinates. Of course, the intrinsic properties of a geometric object (such as length, volume or surface area) do *not* depend on its parametrisation.

For the integral equation in question, it is a priori unclear what the solution $D(\cdot)$ will look like, and therefore it is a priori unclear what parametrisation would be convenient. If the parametrisation chosen is sufficiently general then it can never be wrong. But if some properties of the optimal solution are known (or expected) beforehand, then it might be possible to choose a more specific parametrisation that allows for easier calculation.

In a relatively general case, a closed curve could be parametrised by a certain parameter ϕ, i.e.

$$x = x(\phi),$$
$$y = y(\phi).$$

If, moreover, the shape of the curve is time-dependent, then both Cartesian coordinates may also depend on time, i.e.

$$x = x(\phi, t),$$
$$y = y(\phi, t).$$

In general, therefore, to specify a curve we would need to specify two functions. The same logic holds in the other direction: to find the 2-dimensional domain $D(\cdot)$ we must generally find two functions.

If some properties of the domain to be found are known beforehand, then it may be possible to choose a more convenient parametrisation. If it is a priori known, for example, that the domain to be found is convex at all times, then it should be possible to express the radius ρ of the domain as a function of the polar angle ϕ, and of time — in which case we call the domain *radial*. It is implied that $0 \leq \phi \leq 2\pi$ and $t \leq T$. The advantage would be that we would only need to find *one* function, rather than two.

Focusing on the 2-dimensional case specifically, and on the case when the domain is known to be radial, we investigate the optimality condition of Corollary 1 in detail. We will see that, even if only *one* function $\rho(\phi, t)$ is to be found, the task is still formidable. The optimality equation that must be satisfied by all boundary locations can be classified as a *multidimensional non-linear homogeneous Volterra integral equation of the first kind*, with the following distinguishing features:

1. The unknown function $\rho(\phi, t)$ determines the domain of the integration over $D(\cdot)$; therefore it is a Volterra-type equation. Although Volterra equations normally have the *variable* in the limit of the integration, rather than the unknown *function* as in

this case, we argue that Volterra equations are still more applicable than Fredholm equations, which have a fixed and known domain of integration.

2. Apart from its appearance in the domain $D(\cdot)$, the unknown function $\rho(\phi, t)$ also appears *under* the integral sign, because the location of the boundary coordinate x depends on $\rho(\phi, t)$. Because $\rho(\phi, t)$ appears under the integral sign but not outside the integral, it is an equation of the first kind.

3. The unknown function $\rho(\phi, t)$ appears under the integral sign as a function of the free Brownian density B; therefore it is non-linear.

4. The expected value of all effective gains equals zero; therefore it is a homogeneous equation.

For 1-dimensional Volterra equations, where the integration extends over a variable linear interval, many known methods exist; see e.g. [10], [11], [12] or [13]. Unfortunately this is not the case for multidimensional Volterra integral equations. For the existence and uniqueness of solutions to Volterra equations, see e.g. [12] (p. 25).

In the single asset case, [7] have shown that the numerical procedure based on the integral method can compete with the standard binomial procedure. In the multidimensional case, unfortunately, no analogous result exists. Therefore, we will provide our own (possibly very inefficient) numerical procedure in section 4 for the case when one of two options may be exercised. For the parallel investment in two alternative projects, we are as yet unable to provide a numerical example, but we are able to provide some intuition in Corollary 2:

Corollary 2. *For investment in two alternative projects with continuation costs c_1 and c_2, the expected time spent by a free Brownian motion, from each optimal boundary location β, in the continuation region where project 1 is in the lead — as weighted by c_2 — plus the expected time spent in the continuation region where project 2 is in the lead — as weighted by c_1 — must equal the expectation of time spent on the curve $V_1 = V_2$, where both projects are equally valuable — as weighted by $\frac{1}{2}\left(\frac{\partial V_1}{\partial x}\right)^2 + \frac{1}{2}\left(\frac{\partial V_2}{\partial y}\right)^2$.*

Intuitively, this means that the total expected 'loss' — defined as the total amount of money spent on projects when they are not in the lead — is allowed to be greater if many switches are expected in which project is leading.

So far we have said nothing about the proof of Theorem 1. It turns out that the proof can follow one of three possible routes:

	starting point	based on	satisfies	to be imposed
proof A	Dirichlet value	absorbed BM	value-matching	smooth-pasting
proof B	Neumann value	reflected BM	smooth-pasting	value-matching
proof C	a smart guess	free BM	neither	both

where 'BM' stands for Brownian motion. Here we encounter a classic catch-22: it was unclear, to the author of this paper, how to distribute effort between the three alternatives proofs, such that at least one proof is completed before the PhD is handed in, and the answer to the problem of alternatives is provided. Thus the *result* of this research would have been useful to formulate a strategy to complete it.[4] It is obvious for any theorem that only *one* proof is sufficient (i.e. different proofs are alternatives), the progress and potential of each route can be reviewed over time, and there is a clear deadline before which at least one proof must be completed (i.e. the end of the PhD time window). Parallel investment allows for more exploration, but is also costly, resulting in the classic trade-off between exploration on the one hand, and exploitation of the current best candidate on the other.

One of the main contributions of this paper, however, is *not* the solution to the problem of alternatives (although that problem originally inspired this work), but the general *method* for solving MOS problems. MOS problems include all problems based on Brownian motion and with a non-additive revenue structure, of which the problem of alternatives is an example. To show the validity and versatility of the main idea — i.e. the interplay between boundary value problems and free-boundary value problems — we will present *all three* proofs A, B and C; these are different but analogous, as the reader will quickly discover. There is a certain symmetry connecting these different approaches, which would be lost if they were presented in isolation. A second advantage of our approach is its mere reliance only on:

1. Green's theorem (see e.g. [14]), allowing a finite number of edges, corners and cusps — as explained in many classic reference works, such as [15] (p. 118-119).

2. Reynold's transport theorem, allowing domains that are piecewise smooth at each point in time, and where all regular boundary elements have integrable speeds at all times — as in e.g. [16].

Section 2 relies heavily on both theorems. It also builds on part I of the thesis, as will become clear from the formulae.

1.7 Comparison with the literature

Finally we will discuss how the current paper deviates from the literature. Option theory has largely focused on options within projects, rather than on options where either project may succeed. Portfolio theory considers multiple assets and allows for correlation, but usually assumes that profits are additive and there is (at least classically) no time element or adjustment (e.g. [17] or [18]). Practitioners do of course rebalance portfolios, and modern literature such as [19] allows for uncertainty in the parameters, but portfolio theory

[4] Although, technically, 'effort' is divisible, whereas the problem of alternatives relates to real projects.

assumes an additive revenue structure which makes it not completely suited to the problem of alternatives.

Search theory considers the trade-off between exploiting the current best candidate and sampling further, and it also allows for learning; see e.g. [2], [20], [21], [22] or [23]. But sampling is usually discrete and sequential, making search theory not wholly suitable to the problem of alternatives.

The theory on multi-armed bandits involves multiple projects (slot-machines) and is very well developed; see e.g. [24] and [25]. The theory is not fully applicable to the problem of alternatives, however, because play is usually discrete and sequential, revenues of different machines are additive and the time-horizon is normally infinite. [26] generalise this situation and allow for any subset of machines to be activated at any decision epoch and for any distribution of effort (i.e. the resource is divisible). Still, the bandits evolve independently and revenues of different bandits are additive. In all likelihood it is possible to tackle the problem of alternatives using the theory of multi-armed bandits, as the number of decision epochs goes to infinity, when only the bandit with the best state generates revenues at some predetermined finite time T, and when that bandit has been selected for play at all preceding decision epochs (i.e. it has not been abandoned). In the bandit literature, the paper *Stoppable families of alternative bandit processes* [27] is probably closest related to the issue discussed here. The original aim of multi-armed bandit theory, however, quoting Gittins, is to 'decide which arm to pull next at each stage so as to maximise the total expected reward from an infinite sequence of pulls' (see [28]). This shows that multi-armed bandit theory may not be the most natural starting point for the problem of alternatives.

The theory of optimal stopping includes the time element very explicitly, but by and large it concerns single projects. In the finance literature it is mainly the American option that has attracted much interest; see e.g. [29], [30] and [31]. While our problem of alternatives contains real projects rather than financial projects, it is equally possible to have financial options on more underlying assets. An example is [7], where only one of two American options may be exercised — and the problem thus requires a 2-dimensional continuation region. The approach in our paper can indeed be used to solve the American-max option as described by [7]. The theory of optimal stopping is in many ways a natural starting point for the problem of alternatives, except that it needs to be formulated 1) not just for financial options, and 2) for $d \geq 2$. We do exactly this, and we note that the extension from one to more dimensions suggests itself: the four conditions of (1.1) are unchanged for any d.

We conclude that the question of parallel and continuous investment in alternatives has not been considered fully and systematically, at least not analytically. We propose that

this is due not to a perceived lack of relevance, but to the lack of an adequate MOS theory.

This paper is organised as follows. Section 2 discusses d-dimensional Brownian motion in domain $D(\cdot)$ with boundary conditions, and provides the necessary mathematical tools for the proofs in Section 3. Section 4 provides a numerical example, and Section 5 concludes.

2 Mathematical prerequisites

2.1 Brownian motion

In d dimensions, the transition density of a standard Brownian motion is as follows:

$$B(y,t|x,s) = \frac{1}{[2\pi(t-s)]^{d/2}}\ e^{-\frac{|y-x|^2}{2(t-s)}} \tag{2.1}$$

where $B(y,t|x,s)$ is equal to the (marginal) probability that a Brownian particle moves to space-time coordinate (y,t) given that it started at (x,s). Formally, a Brownian motion is defined as the continuous process, with independent increments, and such that the increment during dt is normally distributed with mean zero and variance dt. The explicit representation (2.1) shows that the density B safisfies

$$\begin{aligned}
\text{forward PDE} \quad & \left(\frac{\partial}{\partial t} - \frac{1}{2}\nabla_y^2\right) B(y,t|x,s) \;=\; 0, \\
\text{backward PDE} \quad & \left(\frac{\partial}{\partial s} + \frac{1}{2}\nabla_x^2\right) B(y,t|x,s) \;=\; 0, \\
\text{forward STC} \quad & \lim_{s\nearrow t} B(y,t|x,s) \;=\; \delta(|y-x|), \\
\text{backward STC} \quad & \lim_{t\searrow s} B(y,t|x,s) \;=\; \delta(|y-x|).
\end{aligned} \tag{2.2}$$

Here and elsewhere, PDE stands for 'partial differential equation', STC stands for 'short-time condition' and where δ is the Dirac δ-function.

2.2 Absorbed Brownian motion

The domain and its boundary at a specific time t are indicated by $D(t)$ and $\partial D(t)$. The transition density of absorbed Brownian motion (ABM) in the dynamic domain $D(\cdot)$ is indicated by $A(y,t|x,s)$, with forward and backward space-time coordinates (y,t) and (x,s). The absorbed transition density $A(y,t|x,s)$ satisfies the following set of equations:

$$\begin{aligned}
\text{forward PDE} \quad & \left(\frac{\partial}{\partial t} - \frac{1}{2}\nabla_y^2\right) A(y,t|x,s) \;=\; 0 && x\in D(s) \quad y\in D(t), \\
\text{backward PDE} \quad & \left(\frac{\partial}{\partial s} + \frac{1}{2}\nabla_x^2\right) A(y,t|x,s) \;=\; 0 && x\in D(s) \quad y\in D(t), \\
\text{forward BC} \quad & A(\beta,t|x,s) \;=\; 0 && x\in D(s) \quad \beta\in \partial D(t), \\
\text{backward BC} \quad & A(y,t|\beta,s) \;=\; 0 && \beta\in \partial D(s) \quad y\in D(t), \\
\text{forward STC} \quad & \lim_{s\nearrow t} A(y,t|x,s) \;=\; \delta(|y-x|) && x\in D(t) \quad y\in D(t), \\
\text{backward STC} \quad & \lim_{t\searrow s} A(y,t|x,s) \;=\; \delta(|y-x|) && x\in D(s) \quad y\in D(s).
\end{aligned} \tag{2.3}$$

The boundary conditions hold only for all *regular* (i.e. non-singular) boundary points β if the boundary is only piecewise smooth. BC stands for 'boundary condition'. It can be

proved that the absorbed transition density 1) exists, 2) is unique and 3) is determined by the above conditions; see for example [32], [33] or [34]. The definition of a 'regular' boundary point is one allowing a tangent plane. The PDEs are satisfied because the transition density is unbiased. The BCs are satisfied because no Brownian particle can move to or from a regular boundary point without being absorbed, and the STCs are satisfied for x and y in the interior because in the short-time limit the absorbed transition density must behave like the free transition density.

Because paths are absorbed at the boundary $\partial D(\cdot)$, the density of all paths that are 'alive' is decreasing. The probability that the first passage occurs at time τ is equal to the 'proportion' of paths that disappear at time τ. Therefore

$$\begin{aligned}
\mathbb{P}\left(\tau^{\text{FP}} \in d\tau | B_s = x\right) &= -\frac{\partial}{\partial \tau} \int_{D(\tau)} d\alpha\, A(\alpha, \tau | x, s) \\
&= -\int_{D(\tau)} d\alpha\, \frac{1}{2} \nabla_\alpha^2 A(\alpha, t | x, s) \\
&= -\oint_{\partial D(\tau)} d\beta\, \frac{1}{2} n_\beta \cdot \nabla_\beta A(\beta, t | x, s) \\
&= \frac{1}{2} \oint_{\partial D(\tau)} d\beta\, \overrightarrow{\partial}_\beta A(\beta, t | x, s)
\end{aligned}$$

where the third line follows by the divergence theorem, where n_β is the outward normal at β, and where $\overrightarrow{\partial}_\beta$ is the inward normal derivative, differentiating towards its right. It is a positive operator when working on the absorbed density A, because A is zero on the boundary but positive in the interior. Because probability can only disappear at the boundary, the joint probability for the first-passage time and first-passage location is

$$\mathbb{P}\left(\tau^{\text{FP}} \in d\tau;\, B_{\tau^{\text{FP}}} \in d\beta | B_s = x\right) = \frac{1}{2} \overrightarrow{\partial}_\beta A(\beta, \tau | x, s) \quad \forall \beta \in \partial D(\tau) \tag{2.4}$$

and this holds for all regular boundary coordinates β if the boundary is only piecewise smooth. The original research on this topic starts here. We start by writing down the following identities:

$$\begin{aligned}
\text{FP}\quad A(y, t | x, s) &= B(y, t | x, s) - \int_s^t d\tau \left(-\frac{\partial}{\partial \tau}\right) \int_{D(\tau)} d\alpha\, B(y, t | \alpha, \tau) A(\alpha, \tau | x, s), \\
\text{LP}\quad A(y, t | x, s) &= B(y, t | x, s) - \int_s^t d\tau \left(\frac{\partial}{\partial \tau}\right) \int_{D(\tau)} d\alpha\, A(y, t | \alpha, \tau) B(\alpha, \tau | x, s).
\end{aligned} \tag{2.5}$$

These indenties hold by virtue of the fundamental theorem of calculus and the STCs satisfied by the free density B and the absorbed density A. The nomenclature of first-passage (FP) and last-passage (FP) decomposition is discussed extensively in Part I of this

thesis. Next we will use the PDEs that are satisfied by A. Because A disappears on the boundary, differentiation under the integral sign is allowed and we can use the PDEs of (2.3) and (2.2), to obtain

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \int_{D(\tau)} d\alpha\; B(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} A(\alpha,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \int_{D(\tau)} d\alpha\; A(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2\right\} B(\alpha,\tau|x,s).
\end{aligned} \tag{2.6}$$

The arrows indicate the *direction* of the differentiation, and we feel this notation makes expressions more readable. Then we use Green's second identity — which is valid for domains with a finite number of edges, corners and cusps — to obtain

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) + \frac{1}{2}\int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; B(y,t|\beta,\tau)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \frac{1}{2}\int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; A(y,t|\beta,\tau)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta\right\} B(\beta,\tau|x,s).
\end{aligned} \tag{2.7}$$

The operator ∂_β is again the inward normal derivative, operating in the direction of the arrow. The BCs of (2.3) require that A is zero on the boundary, and thus we must have that ∂_β points towards A, so we obtain:

Proposition 1. FP & LP decomposition for ABM in $D(\cdot)$. *The absorbed Brownian density $A(y,t|x,s)$ in the time-dependent domain $D(\cdot)$, which allows both Green's theorem and Reynold's transport theorem at each point in time, is determined by (2.3), or, equivalently, by the following pair of integral equations:*

$$\begin{aligned}
\text{FP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; B(y,t|\beta,\tau)\left\{\frac{1}{2}\overrightarrow{\partial}_\beta\right\} A(\beta,\tau|x,s),\\
\text{LP}\quad A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; A(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial}_\beta\right\} B(\beta,\tau|x,s).
\end{aligned} \tag{2.8}$$

This proposition is believed to be new. In fact, its derivation mirrors exactly the derivation of Proposition 1 in Part I of this thesis, except that the domain $D(\cdot)$ here is time-dependent, and therefore this result is more general. We see that a *positive* term is subtracted from the free density to obtain the absorbed density.

Also, we recognise that we have now used all 6 PDEs, STCs and BCs of (2.3) in the derivation of these two integral equations, i.e. all the conditions that are supposed to specify A uniquely have now been used — along with Green's second identity on the domain. The first- and last-passage decompositions relate the value of A to its boundary behaviour and can be used to obtain a series solution for the absorbed density, as in Part I of this thesis, but we will not pursue this here.

2.3 Reflected Brownian motion

The reflected transition density is indicated by $R(y,t|x,s)$, with forward and backward space-time coordinates (y,t) and (x,s). The reflected transition density $R(y,t|x,s)$ satisfies the following set of equations:

$$\begin{aligned}
&\text{forward PDE} && \left(\frac{\partial}{\partial t} - \frac{1}{2}\nabla_y^2\right) R(y,t|x,s) &&= 0 && x \in D(s) \quad y \in D(t),\\
&\text{backward PDE} && \left(\frac{\partial}{\partial s} + \frac{1}{2}\nabla_x^2\right) R(y,t|x,s) &&= 0 && x \in D(s) \quad y \in D(t),\\
&\text{forward BC} && \left(\overrightarrow{\partial_\beta} - 2\dot{\beta}(t)\cdot n_\beta(t)\right) R(\beta,t|x,s) &&= 0 && x \in D(s) \quad \beta \in \partial D(t),\\
&\text{backward BC} && R(y,t|\beta,s)\overleftarrow{\partial_\beta} &&= 0 && \beta \in \partial D(s) \quad y \in D(t),\\
&\text{forward STC} && \lim_{s\nearrow t} R(y,t|x,s) &&= \delta(|y-x|) && x \in D(t) \quad y \in D(t),\\
&\text{backward STC} && \lim_{t\searrow s} R(y,t|x,s) &&= \delta(|y-x|) && x \in D(s) \quad y \in D(s).
\end{aligned} \tag{2.9}$$

Here and elsewhere, PDE stands for 'partial differential equation', BC stands for 'boundary condition' and STC stands for 'short-time condition'. The outward normal is indicated by n_β, $\dot{\beta}(t)$ indicates the *velocity* (i.e. a vector) of the boundary element $\beta(t)$, and ∂_β indicates the inward normal derivative at β. The boundary conditions hold only for all *regular* boundary points β if the boundary is only piecewise smooth. The STCs are satisfied for x and y in the interior because in the short-time limit the reflected transition density must behave like the free transition density. The reflected transition density 1) exists, 2) is unique and 3) is determined by the above conditions; see for example [33] and [34]. The PDEs are satisfied because the reflected transition density is unbiased. The BCs are satisfied because a Brownian particle is reflected in the normal direction, at any regular boundary point β, and because the moving and reflecting boundary also 'drags' along some density. The backward BC here is equal to the backward BC for a static domain D, as in Part I of this thesis. The forward BC, however, is different, but can be derived as follows. By way of the Chapman-Kolmogorov equation, we have

$$R(y,t|x,s) = \int\limits_{D(\tau)} d\alpha\, R(y,t|\alpha,\tau)R(\alpha,\tau|x,s). \tag{2.10}$$

The left-hand side does not depend on τ. Differentiating with respect to τ gives:

$$0 = \frac{\partial}{\partial \tau} \int\limits_{D(\tau)} d\alpha\, R(y,t|\alpha,\tau)R(\alpha,\tau|x,s). \tag{2.11}$$

We use Reynold's transport theorem, as in [16], to differentiate the *limit* of integration (i.e. the changing domain), and we also integrate under the integral sign and use the PDEs to

obtain the following:

$$\begin{aligned} 0 = & \oint\limits_{\partial D(\tau)} d\beta\, R(y,t|\beta,\tau)\left\{n_\beta\cdot\dot\beta(\tau)\right\}R(\beta,\tau|x,s) \\ & -\frac{1}{2}\int\limits_{D(\tau)} d\alpha\; R(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}^2_\alpha-\overrightarrow{\nabla}^2_\alpha\right\}R(\alpha,\tau|x,s), \\ = & \oint\limits_{\partial D(\tau)} d\beta\, R(y,t|\beta,\tau)\left\{n_\beta\cdot\dot\beta(\tau)\right\}R(\beta,\tau|x,s) \\ & +\frac{1}{2}\oint\limits_{\partial D(\tau)} d\beta\; R(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta}-\overrightarrow{\partial_\beta}\right\}R(\beta,\tau|x,s). \end{aligned} \tag{2.12}$$

The last equality follows from Green's theorem, which holds as long as $D(\tau)$ is piecewise smooth. As usual, the arrows on the differential operators indicate the *direction* of their applicability. Using the backward BC, and given that this should hold for *each* domain D, it follows that the forward BC must hold at each boundary location β.

We will now proceed as we did for A. By virtue of the fundamental theorem of calculus, and by the STCs satisfied by both B and R, we can write down two identities:

$$\begin{aligned} \text{FR}\quad R(y,t|x,s) &= B(y,t|x,s)+\int_s^t d\tau\left(-\frac{\partial}{\partial\tau}\right)\int\limits_{D(\tau)} d\alpha\, R(y,t|\alpha,\tau)B(\alpha,\tau|x,s), \\ \text{LR}\quad R(y,t|x,s) &= B(y,t|x,s)+\int_s^t d\tau\left(\frac{\partial}{\partial\tau}\right)\int\limits_{D(\tau)} d\alpha\, B(y,t|\alpha,\tau)R(\alpha,\tau|x,s). \end{aligned} \tag{2.13}$$

The abbreviations FR and LR indicate the first- and last-reflection decompositions, and the nomenclature is discussed in Part I of this thesis. It is clear in either case that both decompositions hold as identities, following directly from the fundamental theorem of calculus and the STCs. The reflected density does not disappear at the boundary and therefore we must differentiate the limits of the spatial integration over $D(\tau)$, as well as under the integral sign. By using Reynold's transport theorem (see e.g. [16]), applicable to piecewise smooth domains which deform at finite (or integrable) speeds, and by using the PDEs of (2.9) under the integral sign, we get

$$\begin{aligned} \text{FR}\quad R(y,t|x,s) = B(y,t|x,s) & -\int_s^t d\tau\oint\limits_{\partial D(\tau)} d\beta\, R(y,t|\beta,\tau)\left\{n_\beta\cdot\dot\beta(\tau)\right\}B(\beta,\tau|x,s) \\ & +\frac{1}{2}\int_s^t d\tau\int\limits_{D(\tau)} d\alpha\; R(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}^2_\alpha-\overrightarrow{\nabla}^2_\alpha\right\}B(\alpha,\tau|x,s), \\ \text{LR}\quad R(y,t|x,s) = B(y,t|x,s) & +\int_s^t d\tau\oint\limits_{\partial D(\tau)} d\beta\, B(y,t|\beta,\tau)\left\{n_\beta\cdot\dot\beta(\tau)\right\}R(\beta,\tau|x,s) \\ & -\frac{1}{2}\int_s^t d\tau\int\limits_{D(\tau)} d\alpha\; B(y,t|\alpha,\tau)\left\{\overleftarrow{\nabla}^2_\alpha-\overrightarrow{\nabla}^2_\alpha\right\}R(\alpha,\tau|x,s). \end{aligned} \tag{2.14}$$

As before, $\dot{\beta}(\tau)$ indicates the velocity (i.e. a vector) of the boundary coordinate β, n_β denotes the outward normal, and ∂_β denotes the inward normal derivative at boundary coordinate β. Using Green's second identity — which is valid for domains with a finite number of edges, corners and cusps — we obtain

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\, R(y,t|\beta,\tau)\left\{n_\beta\cdot\dot{\beta}(\tau)\right\} B(\beta,\tau|x,s) \\
&\quad -\frac{1}{2}\int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; R(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta}-\overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s), \\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\, B(y,t|\beta,\tau)\left\{n_\beta\cdot\dot{\beta}(\tau)\right\} R(\beta,\tau|x,s) \\
&\quad +\frac{1}{2}\int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; B(y,t|\beta,\tau)\left\{\overleftarrow{\partial_\beta}-\overrightarrow{\partial_\beta}\right\} R(\beta,\tau|x,s).
\end{aligned} \tag{2.15}$$

The operator ∂_β is again the inward normal derivative. Using the BCs of (2.9) we obtain:

Proposition 2. FR & LR decomposition for RBM in $D(\cdot)$. *The reflected Brownian density $R(y,t|x,s)$ in the time-dependent domain $D(\cdot)$, which allows both Green's theorem and Reynold's transport theorem at each point in time, is determined by (2.9), or, equivalently, by the following pair of integral equations:*

$$\begin{aligned}
\text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; R(y,t|\beta,\tau)\left\{-n_\beta\cdot\dot{\beta}(\tau)+\frac{1}{2}\overrightarrow{\partial_\beta}\right\} B(\beta,\tau|x,s), \\
\text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; B(y,t|\beta,\tau)\left\{\frac{1}{2}\overleftarrow{\partial_\beta}\right\} R(\beta,\tau|x,s).
\end{aligned} \tag{2.16}$$

This proposition is believed to be new. As in the previous subsection, we note that the derivation relies on Part I of this thesis. In this case, the derivation mirrors Proposition 2 in Part I of this thesis, except that the domain $D(\cdot)$ here is time-dependent. This result is thus more general. Finally, we note, again, that we could use this proposition to construct a series solution for R, as in Part I of this thesis, but we shall not pursue that line of enquiry here.

3 Multidimensional optimal stopping (MOS)

In this section we consider the following *multidimensional optimal stopping* (MOS) *problem*:

$$\begin{aligned}
&\text{The value is unbiased} && \left(\frac{1}{2}\nabla_x^2 + \frac{\partial}{\partial s} - r\right) V(x,s) = -G_C(x,s) && x \in D(s), \\
&\text{Value-matching condition} && V(\beta,s) = G_S(\beta,s) && \beta \in \partial D(s), \\
&\text{Smooth-pasting condition} && \partial_\beta V(\beta,s) = \partial_\beta G_S(\beta,s) && \beta \in \partial D(s), \\
&\text{Value at maturity} && V(x,T) = G_M(x,T) && x \in D(T).
\end{aligned} \tag{3.1}$$

The dynamic domain and its boundary at any particular time s are indicated by $D(s)$ and $\partial D(s)$, and the Laplacian in d dimensions is defined by

$$\nabla_x^2 := \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}.$$

Also, β is a regular boundary coordinate, ∂_β is the inward normal derivative at the regular boundary coordinate β, and regular boundary coordinates are defined as those allowing a tangent plane, such that the normal direction unambiguously exists. The optimal value at space-time coordinate (x,s) is indicated by $V(x,s)$, where x is a vector. The constant discount rate is indicated by r. The intuition for each of the four equations was discussed in the introduction on page 14. We make the following assumptions:

1. The solution to the MOS problem (3.1), given by the pair $D(\cdot)$ and $V(x,s)$, exists and is unique.

2. The solution $D(\cdot)$ that solves the MOS problem (3.1) allows Green's theorem at each point in time: at each time it only has a finite number of edges, corners and cusps.

3. The solution $D(\cdot)$ that solves the MOS problem (3.1) allows Reynold's theorem: it moves with an integrable velocity, for all regular boundary points and at all times.

And we will prove Theorem 1:

Theorem 1. Optimal solution to MOS problem. *If the solution to the MOS problem (3.1) exists and is unique, and if the optimal domain $D(\cdot)$ allows both Green's theorem and Reynold's theorem at all times, then the optimal value V is given by:*

$$\begin{aligned}
V(x,s) = {} & G_S(x,s) \\
& + \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau)\right) e^{-r(\tau-s)}\, B(\alpha,\tau|x,s) \\
& \quad + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T)\right) e^{-r(T-s)}\, B(\alpha,T|x,s)
\end{aligned} \tag{3.2}$$

The optimal domain can be found by applying either *value-matching* or *smooth-pasting to the optimal value, at all boundary coordinates* $\beta \in \partial D(s)$, $\forall s \leq T$. *If* G_S *is only piecewise smooth, then* ΔG_S *should be interpreted as involving distributional derivatives.*

This theorem is new, to the author's best knowledge — and it seems to be one of the first more general results in the field of MOS, as discussed in the introduction (page 18). Its proof consists of two parts — we need to show that:

1. *Given* that $D(\cdot)$ is optimal, *then* V can be written as proposed.
2. *Given* this V, $D(\cdot)$ can be determined by demanding *either* value-matching *or* smooth-pasting for all boundary locations and times $\beta \in \partial D(\cdot)$.

The theorem also invites a new interpretation, namely:

Corollary 1. *For* x *on the optimal optimal boundary, the* expected *value of all effective continuation and effective maturity gains, as collected by the* free *Brownian path during its time in the optimal continuation domain* $D(\cdot)$*, equals zero. And this holds true for* all *boundary locations* $x \in \partial D(\cdot)$*, i.e.*

$$0 = \mathbb{E}_x \left[\int_s^T d\tau \, \Big(G_C(B_\tau, \tau) + \Delta G_S(B_\tau, \tau) \Big) \, \mathbb{1}_{B_\tau \in D(\tau)} \, e^{-r\,(\tau - s)} \right.$$
$$\left. + \Big(G_M(B_T, T) - G_S(B_T, T) \Big) \, \mathbb{1}_{B_T \in D(T)} \, e^{-r\,(T-s)} \right] \quad \forall x \in D(s), \forall s \leq T \tag{3.3}$$

Here, the 'effective' continuation and maturity gains are defined as follows:

- effective continuation gain $:= G_C(\alpha, \tau) + \Delta G_S(\alpha, \tau)$,
- effective maturity gain $:= G_M(\alpha, T) - G_S(\alpha, T)$.

At a location and time where the effective continuation gain is positive, one would always continue a short time dt. We discussed in the introduction that there are three possible routes to the proof of this theorem, namely:

	starting point	based on	satisfies	to be imposed
proof A	Dirichlet value	absorbed BM	value-matching	smooth-pasting
proof B	Neumann value	reflected BM	smooth-pasting	value-matching
proof C	a smart guess	free BM	neither	both

The next three subsections will each present a proof. For readers uninterested in the relationship between the Dirichlet and Neumann problems on the one hand, and free-boundary problems on the other, we suggest that they jump straight to proof C. Although proof C may appear rather ad-hoc (it starts off with a somewhat arbitrary identity), and although it does not provide the insight that the first two proofs provide, it requires no knowledge of either absorbed or reflected Brownian motion.

3.1 Proof A: The Dirichlet route

For some arbitrary continuation domain $D(\cdot)$, the *Dirichlet value* is defined by

$$\begin{aligned} V^D(x,s) := \quad & \mathbb{E}_x\left[\mathbb{1}_{\tau_{\text{FP}}<T}\, G_S(B_{\tau_{\text{FP}}},\tau_{\text{FP}})\, e^{-r(\tau_{\text{FP}}-s)}\right] \\ & + \mathbb{E}_x\left[\int_s^T d\tau\ \mathbb{1}_{\tau<\tau_{\text{FP}}}\, G_C(B_\tau,\tau)\, e^{-r(\tau-s)}\right] \\ & + \mathbb{E}_x\left[\mathbb{1}_{\tau_{\text{FP}}=T}\, G_M(B_T,T)\, e^{-r(T-s)}\right]. \end{aligned} \tag{3.4}$$

The first-passage time τ_{FP} is defined as the first passage over the (not necessarily optimal) domain $D(\cdot)$, or T, whichever occurs first, i.e.

$$\tau_{\text{FP}} := \min\left\{\text{first-passage time over } \partial D(\cdot), T\right\}. \tag{3.5}$$

The expectation in the Dirichlet value is conditional on the starting coordinate x, indicated by the subscript. The superscript of V^D refers to the Dirichlet value — not to be confused with the domain $D(\cdot)$.

The intuition for the Dirichlet value is as follows: the stopping gain G_S is collected *at* the first-passage time τ_{FP} if and only if $\tau_{\text{FP}} < T$. The continuation gain G_C is collected *until* the first-passage time τ_{FP}. Lastly, the maturity gain G_M is collected *at* the first-passage time τ_{FP} if and only if $\tau_{\text{FP}} = T$.

This is called the Dirichlet value because it satisfies the value-matching condition, as will be shown in Proposition 3 below. Its usefulness derives from the fact that we can define the Dirichlet value for *any* boundary, and not just the optimal one — as long as we suppose that the boundary has *some* regularity: singular boundary points are allowed as long as there are a finite number of them, and infinite boundary speeds are allowed as long as they are integrable.

In Section 2 we introduced the absorbed transition density $A(y,t|x,s)$ of a Brownian motion. In terms of the absorbed transition density, we can write the *boundary representation* of the Dirichlet value as follows:

$$\begin{aligned} V^D(x,s) = \quad & \int_s^T d\tau \int\limits_{\partial D(\tau)} d\beta\ G_S(\beta,\tau)\, \tfrac{1}{2}\overrightarrow{\partial_\beta} A(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ & + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha\ G_C(\alpha,\tau)\ A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ & + \int\limits_{D(T)} d\alpha\ G_M(\alpha,T)\ A(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.6}$$

The name 'boundary representation' follows from the fact that the stopping gain G_S is integrated over the boundary $\partial D(\cdot)$. In the expression above, the stopping gain G_S is

multiplied with the probability of a first passage at that location, i.e. ∂A, then discounted, and then integrated over all future boundary locations β and times τ.

Instead of this boundary representation of the Dirichlet value, we can also provide the following *interior representation*:

$$\begin{aligned} V^D(x,s) = G_S(x,s) \\ + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ + \int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) A(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.7}$$

The nomenclature stems from the fact that all integrations are now over the interior of $D(\cdot)$. This interior representation of the Dirichlet value has one technical and one intuitive advantage:

1. Technical advantage: the limit $x \to \partial D(s)$ commutes with the volume integrals of the interior representation, but not with the boundary integral of the boundary representation.

2. Intuitive advantage: the interior representation decomposes the Dirichlet value into an 'immediate-stopping gain' G_S (the first term) and an 'option value' (second and third term), that invites us to define the 'effective continuation gain' and 'effective maturity gain' — as mentioned before.

To show that the interior representation follows from the boundary representation, we rewrite the boundary term in (3.7) as follows:

$$\begin{aligned} & \int_s^T d\tau \int\limits_{\partial D(\tau)} d\beta\, G_S(\beta,\tau) \frac{1}{2} \overrightarrow{\partial}_\beta A(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &= -\frac{1}{2} \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta\, G_S(\beta,\tau) \left\{ \overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta \right\} A(\beta,\tau|x,s)\, e^{-r(\tau-s)}. \end{aligned}$$

This follows from the fact that the absorbed propagator A disappears on the boundary. Then, using Green's identity, which is valid for piecewise smooth domains, and by the PDE satisfied by A, we get

$$\begin{aligned} &= \frac{1}{2} \int_s^T d\tau \int\limits_{D(\tau)} d\alpha\, G_S(\alpha,\tau) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} A(\alpha,\tau|x,s)\, e^{-r(\tau-s)}, \\ &= \int_s^T d\tau \int\limits_{D(\tau)} d\alpha\, G_S(\alpha,\tau) \left\{ \frac{1}{2} \overleftarrow{\nabla}_\alpha^2 - r - \frac{\partial}{\partial \tau} \right\} \left(A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \right). \end{aligned}$$

With a partial integration (in time) and using the definition for the differential operator $\Delta := \frac{1}{2}\nabla_\alpha^2 + \frac{\partial}{\partial\tau} - r$, we get

$$= \int_s^T d\tau \int_{D(\tau)} d\alpha\ \Delta G_S(\alpha,\tau)\, A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} - \int_s^T d\tau \frac{\partial}{\partial\tau} \int_{D(\tau)} d\alpha\ G_S(\alpha,\tau) A(\alpha,\tau|x,s)\, e^{-r(\tau-s)}.$$

Using the fundamental theorem of calculus, we get

$$= \int_s^T d\tau \int_{D(\tau)} d\alpha\ \Delta G_S(\alpha,\tau)\, A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} - \left(\lim_{\tau\nearrow T} - \lim_{\tau\searrow s}\right) \int_{D(\tau)} d\alpha\ G_S(\alpha,\tau) A(\alpha,\tau|x,s)\, e^{-r(\tau-s)}.$$

By the STC satisfied by A, we get

$$= \int_s^T d\tau \int_{D} d\alpha\ \Delta G_S(\alpha,\tau)\, A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} + G_S(x,s) - \int_{D(T)} d\alpha\ G_S(\alpha,T) A(\alpha,T|x,s)\, e^{-r(T-s)}.$$

Plugging this expression back into the boundary representation (3.7) gives the interior representation for the Dirichlet value. Using the interior representation of the Dirichlet value, we derive the following proposition:

Proposition 3. The Dirichlet value. *The Dirichlet value for the arbitrary domain $D(\cdot)$ is defined by:*

$$\begin{aligned} V^D(x,s) = G_S(x,s) &\\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau)\right) A(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T)\right) A(\alpha,T|x,s)\, e^{-r(T-s)} \end{aligned}$$

and satisfies three of the four conditions of optimality in (3.1), namely:

$$\begin{aligned} \text{The value is unbiased} \quad \left(\frac{\partial}{\partial s} + \frac{1}{2}\nabla_x^2 - r\right) V^D(x,s) &= -G_C(x,s) \quad x \in D(s), \\ \text{Value-matching} \quad V^D(\beta,s) &= G_S(\beta,s) \quad \beta \in \partial D(s), \\ \text{Value at maturity} \quad V^D(x,T) &= G_M(x,T) \quad x \in D(T). \end{aligned}$$

Proof. The proof of the partial differential equation follows by using the PDE in (2.3) and by differentiation the integration limit s, of the second term, and taking into account the STC of (2.3). The boundary condition follows immediately from the fact that the limit $x \to D(s)$ commutes with the integration and that A is zero for x on the boundary. The maturity condition follows from the fact that the second term disappears in the limit $s \to T$, and A in the third term involves a Dirac δ-function by the STC in (2.3). □

What Proposition 3 says, in other words, is that the value V^D that is defined as a sum of all the expected gains until some stopping time over an *arbitrary* domain D satisfies three out of the four conditions for optimality of V. This should not be overly surprising: of course an expectation should be unbiased, of course it should only pick up G_S as the starting point x moves to the boundary, and of course it should only pick up G_M as the time goes to the maturity T. The only condition that is not satisfied automatically is the 'smooth-pasting' condition. If we impose this last condition on the as yet arbitrary domain $D(\cdot)$, we obtain a condition on the optimal domain $D(\cdot)$ that should specify it uniquely. Smooth-pasting requires that

$$\text{Smooth-pasting } \partial_\beta V(\beta, s) = \partial_\beta G_S(\beta, s) \;\; \beta \in D(s), \forall s < T.$$

From the interior representation, we see that we must have

$$\begin{aligned} 0 = \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) A(\alpha,\tau|\beta,s) \overleftarrow{\partial_\beta}\, e^{-r(\tau-s)} \\ + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) A(\alpha,T|\beta,s) \overleftarrow{\partial_\beta}\, e^{-r(T-s)} \end{aligned} \qquad \forall \beta \in D(s),\, \forall s < T. \tag{3.8}$$

The optimality conditions thus demands that the normal derivative of the 'option value', as one approaches the stopping boundary, goes to zero. As a result, the derivative of the total value (i.e. immediate stopping gain plus 'option value') equals the derivative of the immediate stopping gain, as requested. We will now use this optimality condition in the interior representation of the Dirichlet value. First, recall that we have found in Proposition 1 that

$$\begin{aligned} \text{FP } A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; B(y,t|\beta,\tau) \left\{ \frac{1}{2} \overrightarrow{\partial_\beta} \right\} A(\beta,\tau|x,s), \\ \text{LP } A(y,t|x,s) &= B(y,t|x,s) - \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\; A(y,t|\beta,\tau) \left\{ \frac{1}{2} \overleftarrow{\partial_\beta} \right\} B(\beta,\tau|x,s). \end{aligned}$$

Next, substitute the last-passage (LP) decomposition of A into the Dirichlet value V^D to

obtain:

$$
\begin{aligned}
V^D(x,s) = \; & G_S(x,s) \\
& + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} \, B(\alpha,\tau|x,s) \\
& - \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} \int_s^\tau d\theta \oint\limits_{\partial D(\theta)} d\gamma \; A(\alpha,\tau|\gamma,\theta) \left\{ \frac{1}{2} \overleftarrow{\partial_\gamma} \right\} B(\gamma,\theta|x,s) \\
& \quad + \int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} \, B(\alpha,T|x,s) \\
& \quad - \int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} \int_s^T d\theta \oint\limits_{\partial D(\theta)} d\gamma \; A(\alpha,T|\gamma,\theta) \left\{ \frac{1}{2} \overleftarrow{\partial_\gamma} \right\} B(\gamma,\theta|x,s)
\end{aligned}
$$

This looks unwieldy, but, in fact, it will allow a great simplification. First we note that any time-ordered integration can be written in one of two ways, i.e.

$$
\int\limits_{s \le \theta \le \tau \le T} \!\!\! d\tau \, d\theta = \int_s^T d\tau \int_s^\tau d\theta = \int_s^T d\theta \int_\tau^T d\tau.
$$

We use this to rewrite the third term

$$
- \int_s^T d\theta \oint\limits_{\partial D(\theta)} d\gamma \left[\int_\theta^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-\theta)} \, A(\alpha,\tau|\gamma,\theta) \left\{ \frac{1}{2} \overleftarrow{\partial_\gamma} \right\} \right] e^{-r(\theta-s)} B(\gamma,\theta|x,s),
$$

as well as the fifth term

$$
- \int_s^T d\theta \oint\limits_{\partial D(\theta)} d\gamma \left[\int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-\theta)} \, A(\alpha,T|\gamma,\theta) \left\{ \frac{1}{2} \overleftarrow{\partial_\gamma} \right\} \right] e^{-r(\theta-s)} B(\gamma,\theta|x,s).
$$

Now we notice that the optimality condition (3.8) demands that the sum of these two terms equals zero! (Pay attention to the terms in square brackets.) Therefore we must have that the *optimal* value V equals the Dirichlet value with the absorbed density A replaced by the free density B:

$$
\begin{aligned}
V(x,s) = \; & G_S(x,s) \\
& + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} \, B(\alpha,\tau|x,s) \\
& \quad + \int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} \, B(\alpha,T|x,s).
\end{aligned}
\tag{3.9}
$$

We see that for the optimal value V, the absorbed density A (which is a complicated quantity in itself) is miraculously replaced by the free density B (which we know). It is thus easier to calculate the value corresponding to the optimal domain, when the optimal

domain is given, than it is to calculate the Dirichlet value corresponding to any other domain. It may seem remarkable that for the optimal domain $D(\cdot)$, we have

$$V(x,s) = V^D(x,s) \quad \text{only when } D(\cdot) \text{ is optimal}$$

even though absorbed propagator A is *everywhere* smaller than the free propagator B. It would appear that the expectation over all effective gains by the *free* and *absorbed* Brownian motions must be different — because the free and absorbed Brownian motions only coincide before the first passage, and not thereafter. But this is where we forget that from *any* optimal boundary location, the expectation over all future effective gains equals zero. Therefore, as far as the expectation is concerned, *it makes no difference* whether the Brownian motion is stopped upon the first-passage or left to proceed as a free Brownian motion. The only way, therefore, that the absorbed and free expectation over the future domain can be equal, is when the domain is optimal — such that from *every* boundary location, the expectation over all future effective gains equals zero.

We have now shown that we can write the optimal value as a function of the *free* propagator B, if the optimal domain $D(\cdot)$ is known. But of course $D(\cdot)$ is not known yet; we will discuss how to find it in subsection 3.4.

3.2 Proof B: The Neumann route

For some arbitrary continuation domain $D(\cdot)$, allowing Green's theorem and Reynold's theorem, the *Neumann value* is defined by

$$\begin{aligned} V^N(x,s) := \quad & -\mathbb{E}_x\left[\int_s^T d\tau\, G_S(\beta,\tau)\frac{1}{2}\overleftarrow{\partial}_\beta\, \mathbb{1}_{R_\tau\in\beta}\, e^{-r(\tau-s)}\right] \\ & + \mathbb{E}_x\left[\int_s^T d\tau\;\, G_C(R_\tau,\tau)\, e^{-r(\tau-s)}\right] \\ & + \mathbb{E}_x\Big[G_M(R_T,T)\, e^{-r(T-s)}\Big]. \end{aligned} \tag{3.10}$$

Here R_τ represents a reflected Brownian motion, and β is a boundary element, i.e. $\beta \in \partial D(\cdot)$. The expectation in the Neumann value is conditional on the starting coordinates x, indicated by the subscript. The superscript of V^N refers to the Neumann value.

The intuition for the Neumann value is as follows: the value $-1/2\partial G_S$ is collected during the time that the reflected Brownian motion spends at the boundary of the domain. (The minus sign, the factor of $1/2$ and the ∂ operator in front of G_S are because of symmetry reasons which will become clear.) The continuation gain G_C is collected by the Brownian motion during its entire time in the interior of $D(\cdot)$. Lastly, the maturity gain G_M is collected at the maturity time T. Notice that the maturity gain will *always* be obtained by a reflected Brownian motion, since a reflected Brownian motion always reaches maturity (it is not absorbed, for example).

The reason for the name 'Neumann value' is that it satisfies the smooth-pasting condition, as we will shown in Proposition 4. Its usefulness derives from the fact that we can define the Neumann value for *any* boundary, and not just for the optimal one — as long as we suppose that the boundary has *some* regularity: singular boundary points are allowed as long as there are a finite number of them, and infinite boundary speeds are allowed as long as they are integrable.

In Section 2 we introduced the reflected transition density $R(y,t|x,s)$ of a Brownian motion. In terms of the reflected transition density, we find the following *boundary representation* of the Neumann value:

$$\begin{aligned} V^N(x,s) = &- \int_s^T d\tau \int\limits_{\partial D(\tau)} d\beta\; G_S(\beta,\tau)\, \tfrac{1}{2}\overleftarrow{\partial}_\beta\, R(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha\; G_C(\alpha,\tau)\; R(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha\; G_M(\alpha,T)\; R(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.11}$$

The name 'boundary representation' follows from the fact that the stopping gain is integrated over the time spent at the boundary $\partial D(\cdot)$: i.e. the stopping gain is multiplied with the probability R of being at the boundary, discounted, and then integrated over all future boundary locations β at times τ. Under our usual assumptions, we can also provide the following *interior representation* of the Neumann value:

$$\begin{aligned} V^N(x,s) = \;& G_S(x,s) \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \Big(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \Big)\, R(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \Big(G_M(\alpha,T) - G_S(\alpha,T) \Big)\, R(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.12}$$

The nomenclature stems from the fact that all integrations are now over the interior of $D(\cdot)$. This interior representation of the Neumann value has one technical and one intuitive advantage:

1. Technical advantage: the limit $\lim\limits_{x\to\beta} n_\beta \cdot \nabla_x$, where $\beta \in \partial D(\cdot)$, commutes with the volume integrals of the interior representation, but not with the boundary integral of the boundary representation.

2. Intuitive advantage: the interior representation decomposes the Neumann value into an 'immediate-stopping gain' G_S (the first term) and an 'option value' (second and

third term), that invite us to define the 'effective continuation gain' and 'effective maturity gain' — as mentioned before.

To show that the interior representation follows from the boundary representation, we consider the term in (3.12) with the boundary integral. We realise that we may add two terms that disappear by the BC on R, i.e.

$$
\begin{aligned}
&-\int_s^T d\tau \int\limits_{\partial D(\tau)} d\beta\; G_S(\beta,\tau)\frac{1}{2}\overleftarrow{\partial}_\beta R(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\
&= -\frac{1}{2}\int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta\; G_S(\beta,\tau)\left\{\overleftarrow{\partial}_\beta - \overrightarrow{\partial}_\beta + 2n_\beta\cdot\dot{\beta}(\tau)\right\} R(\beta,\tau|x,s)\, e^{-r(\tau-s)}.
\end{aligned}
$$

We go through the same steps that we used to write the interior respresentation of the Dirichlet value, as in subsection 3.1: i.e. use Green's theorem, use the PDEs to obtain a differentiation with respect to τ on $R\,e^{-r(\tau-s)}$, perform a partial integration in τ to obtain Δ working on G_S, use Reynold's theorem to place $\partial/\partial\tau$ outside the integration over $D(\tau)$, notice that this disposes of the term with $\dot{\beta}$, and, finally, use the STCs. Going through these steps carefully, the result appears as follows:

$$
= G_S(x,s) + \int_s^T d\tau \int_D d\alpha\; \Delta G_S(\alpha,\tau)\, R(\alpha,\tau|x,s)\, e^{-r(\tau-s)} - \int\limits_{D(T)} d\alpha\; G_S(\alpha,T) R(\alpha,T|x,s)\, e^{-r(T-s)}.
$$

By plugging this expression back into the boundary representation (3.12), we obtain the promised interior representation. Using this interior representation, we can easily show:

Proposition 4. The Neumann value. *The Neumann value for an arbitrary domain $D(\cdot)$ is defined by:*

$$
\begin{aligned}
V^N(x,s) = G_S(x,s)&\\
&+\int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau)+\Delta G_S(\alpha,\tau)\right) R(\alpha,\tau|x,s)\, e^{-r(\tau-s)}\\
&\quad+\int\limits_{D(T)} d\alpha \left(G_M(\alpha,T)-G_S(\alpha,T)\right) R(\alpha,T|x,s)\, e^{-r(T-s)}
\end{aligned}
$$

and satisfies three out of the four conditions of optimality in (3.1), namely:

$$
\begin{aligned}
\text{The value is unbiased} \quad \left(\frac{\partial}{\partial s}+\frac{1}{2}\nabla_y^2 - r\right) V^N(x,s) &= -G_C(x,s) \quad x\in D(s),\\
\text{Smooth-pasting} \quad \partial_\beta V^N(\beta,s) &= \partial_\beta G_S(\beta,s) \quad \beta\in\partial D(s),\\
\text{Value at maturity} \quad V^N(x,T) &= G_M(x,T) \quad x\in D(T).
\end{aligned}
$$

Proof. The proof follows in the same way as the proof of Proposition 3, and by noting that the operator $\lim_{x\to\beta} n_\beta \cdot \nabla_x$ commutes with the integration over the interior such that smooth-pasting immediately follows. □

Value-matching, however, is not satisfied. Note that value-matching is satisfied if we have, in addition, that

$$\begin{aligned} 0 = \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} R(\alpha,\tau|\beta,s) \\ + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} R(\alpha,T|\beta,s) \end{aligned} \qquad \forall \beta \in D(s),\ \forall s < T. \tag{3.13}$$

We will use this optimality condition in the Neumann value shortly. But first we recall that we found in Proposition 2 that

$$\begin{aligned} \text{FR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\, R(y,t|\beta,\tau) \left\{ -n_\beta \cdot \dot\beta(\tau) + \frac{1}{2}\overrightarrow{\partial}_\beta \right\} B(\beta,\tau|x,s) \\ \text{LR}\quad R(y,t|x,s) &= B(y,t|x,s) + \int_s^t d\tau \oint_{\partial D(\tau)} d\beta\, B(y,t|\beta,\tau) \left\{ \frac{1}{2}\overleftarrow{\partial}_\beta \right\} R(\beta,\tau|x,s) \end{aligned}$$

Now substitute the first-reflection decomposition (i.e. FR) into the Neumann value V^N to obtain

$$\begin{aligned} V^N(x,s) &= G_S(x,s) \\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} B(\alpha,\tau|x,s) \\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} \int_s^\tau d\theta \oint_{\partial D(\theta)} d\gamma\, R(\alpha,\tau|\gamma,\theta) \left\{ -n_\gamma \cdot \dot\gamma(\theta) + \frac{1}{2}\overrightarrow{\partial}_\gamma \right\} B(\gamma,\theta|x,s) \\ &+ \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} B(\alpha,T|x,s) \\ &+ \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} \int_s^\tau d\theta \oint_{\partial D(\theta)} d\gamma\, R(\alpha,T|\gamma,\theta) \left\{ -n_\gamma \cdot \dot\gamma(\theta) + \frac{1}{2}\overrightarrow{\partial}_\gamma \right\} B(\gamma,\theta|x,s). \end{aligned} \tag{3.14}$$

Using the optimality condition (3.13) we can show that the third and fifth term add to zero if the optimality condition is satisfied. Therefore, we find once more that the optimal value must satisfy that:

$$\begin{aligned} V(x,s) = G_S(x,s) \\ + \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} B(\alpha,\tau|x,s) \\ + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} B(\alpha,T|x,s). \end{aligned} \tag{3.15}$$

For the optimal value V, the reflected density R (which is a complicated quantity in itself) in the Neumann value is replaced by the free density B (which we know). It is thus easier to calculate the value corresponding to the optimal domain, when the optimal domain is given, than it is to calculate the Neumann value corresponding to any other domain. Again it may appear as quite remarkable that for the optimal domain $D(\cdot)$, we have

$$V(x,s) = V^N(x,s) \quad \text{only when } D(\cdot) \text{ is optimal}$$

because, when the domain is convex and shrinking, for example, the absorbed propagator R is *everywhere* larger than the free propagator B. Thus it would appear that the expectation over all effective gains as collected by the *free* and *reflected* Brownian motions must be different — because the free and reflected Brownian motions only coincide before the first reflection, and not thereafter. But, on expectation, the contribution after the first reflection equals zero! The only way, therefore, that the reflected and free expectation over the future domain can be equal is when the domain is optimal.

We have found the optimal value in two ways: either by imposing smooth-pasting on the Dirichlet value or by imposing value-matching on the Neumann value. We have seen that the Dirichlet value automatically satisfies three out of the four conditions of optimality, including value-matching. The Neumann value also automatically satisfies three out of the four conditions of optimality, including smooth-pasting. The optimal value must satisfy all four conditions, and therefore there must be exactly *one* domain for which the Dirichlet and Neumann values coincide, and thus allowing all four conditions to be satisfied. The optimality condition can thus be reinterpreted as follows:

$$V(x,s) = V^D(x,s) = V^N(x,s) \quad \text{if and only if } D(\cdot) \text{ is optimal.} \tag{3.16}$$

We have now shown that we can write the optimal value as a function of the *free* propagator B, if the optimal domain $D(\cdot)$ is known. But of course $D(\cdot)$ is not known yet; we will discuss how to find it in subsection 3.4.

3.3 Proof C: A smart guess

Compared to the above, proof C may appear rather ad-hoc. It will start off with a somewhat arbitrary identity, and will then use all the conditions that are supposed to specify the optimal value V, and — seemingly out of the blue — arrives at the correct result. Although it does not provide the insight that the first two proofs provide, it requires no knowledge of either absorbed or reflected Brownian motion. We start the above-mentioned ad-hoc

identity, i.e. consider that the following holds by definition:

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \left(\frac{\partial}{\partial \tau}\right) \int\limits_{D(\tau)} d\alpha \, V(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \, G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.17}$$

The reason that this holds by definition is that we get, by the fundamental theorem of calculus, that

$$\begin{aligned} V(x,s) = &\left(\lim_{\tau \searrow s} - \lim_{\tau \nearrow T}\right) \int\limits_{D(\tau)} d\alpha \, V(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \, G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)} \end{aligned} \tag{3.18}$$

and using the STC satisfied by B and $V(\alpha,T) = G_M(\alpha,T)$ the above is an identity. Returing to our identity, we perform the differentiation and by virtue of Reynold's transport theorem and the PDEs satisfied by V and B, we obtain

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta \, V(\beta,\tau) \left\{ \dot{\beta}(\tau) \cdot n_\beta \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \frac{1}{2} \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \, V(\alpha,\tau) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \, G_C(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \, G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.19}$$

Using Green's identity for the second term, we get that

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \oint\limits_{D(\tau)} d\beta \, V(\beta,\tau) \left\{ \dot{\beta}(\tau) \cdot n_\beta \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &- \frac{1}{2} \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta \, V(\beta,\tau) \left\{ \overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta} \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \, G_C(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \, G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.20}$$

Now using *both* value-matching and smooth-pasting in the first and second terms, we get

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta \; G_S(\beta,\tau) \left\{ \dot{\beta}(\tau) \cdot n_\beta \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &- \frac{1}{2} \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta \; G_S(\beta,\tau) \left\{ \overleftarrow{\partial_\beta} - \overrightarrow{\partial_\beta} \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \; G_C(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \; G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.21}$$

We use Green's identity *again*, in the second term, to re-obtain an integration over the interior:

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta \; G_S(\beta,\tau) \left\{ \dot{\beta}(\tau) \cdot n_\beta \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \frac{1}{2} \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \; G_S(\alpha,\tau) \left\{ \overleftarrow{\nabla}_\alpha^2 - \overrightarrow{\nabla}_\alpha^2 \right\} B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \; G_C(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \; G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.22}$$

With the PDE satisfied by B, we get that

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \oint\limits_{\partial D(\tau)} d\beta \; G_S(\beta,\tau) \left\{ \dot{\beta}(\tau) \cdot n_\beta \right\} B(\beta,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \; G_S(\alpha,\tau) \left\{ \frac{1}{2} \overleftarrow{\nabla}_\alpha^2 - r - \frac{\partial}{\partial \tau} \right\} \left(B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \right) \\ &+ \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \; G_C(\alpha,\tau)\, B(\alpha,\tau|x,s)\, e^{-r(\tau-s)} \\ &+ \int\limits_{D(T)} d\alpha \; G_M(\alpha,T) B(\alpha,T|x,s)\, e^{-r(T-s)}. \end{aligned} \tag{3.23}$$

With a partial integration in the second term, we get

$$\begin{aligned} V(x,s) = &- \int_s^T d\tau \oint_{\partial D(\tau)} d\beta \, G_S(\beta,\tau) \left\{ \dot{\beta}(\tau) \cdot n_\beta \right\} B(\beta,\tau|x,s) \, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \, \Delta G_S(\alpha,\tau) \, B(\alpha,\tau|x,s) \, e^{-r(\tau-s)} \\ &- \int_s^T d\tau \int_{D(\tau)} d\alpha \, \frac{\partial}{\partial \tau} \left(G_S(\alpha,\tau) \, B(\alpha,\tau|x,s) \, e^{-r(\tau-s)} \right) \\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \, G_C(\alpha,\tau) \, B(\alpha,\tau|x,s) \, e^{-r(\tau-s)} \\ &+ \int_{D(T)} d\alpha \, G_M(\alpha,T) B(\alpha,T|x,s) \, e^{-r(T-s)}. \end{aligned} \tag{3.24}$$

Using Reynold's theorem, we get

$$\begin{aligned} V(x,s) = &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \, \Delta G_S(\alpha,\tau) \, B(\alpha,\tau|x,s) \, e^{-r(\tau-s)} \\ &- \int_s^T d\tau \left(\frac{\partial}{\partial \tau} \right) \int_{D(\tau)} d\alpha \, G_S(\alpha,\tau) \, B(\alpha,\tau|x,s) \, e^{-r(\tau-s)} \\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \, G_C(\alpha,\tau) \, B(\alpha,\tau|x,s) \, e^{-r(\tau-s)} \\ &+ \int_{D(T)} d\alpha \, G_M(\alpha,T) B(\alpha,T|x,s) \, e^{-r(T-s)}, \end{aligned} \tag{3.25}$$

where the first term has conveniently cancelled. With the STC satisfied by B, we obtain — once more — that the optimal value must satisfy that:

$$\begin{aligned} V(x,s) = \; & G_S(x,s) \\ &+ \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} \, B(\alpha,\tau|x,s) \\ &\quad + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} \, B(\alpha,T|x,s). \end{aligned} \tag{3.26}$$

In this derivation we used 1) the PDE on B, 2) the STC on B, 3) the PDE satisfied by the optimal value V, 4) the maturity condition on V, 5) both boundary conditions on the optimal value V, and 6) Green's second identity on the continuation domain at each time, requiring a piecewise smooth boundary, and 7) Reynold's transport theorem on the domain, requiring a finite velocity (or integrability) of all regular boundary points. We may therefore expect that the last expression should determine the optimal continuation

domain uniquely: all conditions that are supposed to specify the optimal value have been used, and only those relatively mild assumptions on the smoothness of the domain have been used that are specified by Green's identity and Reynold's theorem.

3.4 The integral equation for the boundary

We have shown that the optimal value can be written as a function of the free density B — if and only if the continuation domain is optimal. In other words, the fact that the optimal value can be written as a function of B is a necessary condition, or can be seen as *resulting* from optimality, but it is not a sufficient condition. This can easily be shown: for any given domain, one *could* calculate the supposedly 'optimal value' using B, as if the domain was optimal when in fact it is not. For any such domain, the PDE and condition at maturity of (3.1) would hold, but neither value-matching nor smooth-pasting would be satisfied. If we want to find the optimal domain, therefore, we must impose value-matching and smooth-pasting on the expression for the optimal value. But which one should we impose first? It turns out that this is irrelevant; we only need to impose one of them, and either will do. To show why this is the case, let us extend the range of x from $D(s)$ to all of $\mathbb{R}^d$, i.e. we define

$$\begin{aligned} V(x,s) := \; & G_S(x,s) \\ & + \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} \, B(\alpha,\tau|x,s) \\ & \quad + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} \, B(\alpha,T|x,s) \end{aligned} \qquad \forall x \in \mathbb{R}^d, \tag{3.27}$$

where x is now allowed in *all* of $\mathbb{R}^d$, i.e. also outside the continuation domain $D(\cdot)$. As can easily be seen, we *automatically* have that:

- The value of V, defined as above, is everywhere continuous for x in $\mathbb{R}^d$ — and in particular the value of V is continuous across $\partial D(s)$ for all $s < T$, for *any* domain.
- The gradient of V, defined as above, is everywhere continuous for x in $\mathbb{R}^d$ — and in particular the gradient (and thus also the normal derivative!) is continuous across $\partial D(s)$ for all $s < T$, for *any* domain.
- Far away from the continuation region $D(\cdot)$, we get that B decays exponentially, and thus we *automatically* get $V = G_S$ for x far into the stopping region.

Furthermore,

- The value of V defined as above *automatically* satisfies $\Delta V = -G_C$ inside of $D(s)$, for *any* domain.

- The value of V defined as above *automatically* satisfies $\Delta V = \Delta G_S$ outside of $D(s)$, for *any* domain.

- The second derivative with respect to x is *not* continuous across the boundary $\partial D(s)$, but the second derivative is not required to be continuous, and so this shall not bother us.

Because the value far into the stopping region is fixed, and because a given second order differential equation is satisfied outside of $D(\cdot)$ as well as inside of $D(\cdot)$, it suffices to impose *either* the value at $\partial D(\cdot)$ *or* the normal derivative at $\partial D(\cdot)$. We choose to impose that the value option value is zero for all space-time coordinates (x, s) on ∂D. As a result, the equation that should define the optimal continuation domain uniquely is as follows:

$$0 = \int_s^T d\tau \int_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau-s)} B(\alpha,\tau|\beta,s) + \int_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)} B(\alpha,T|\beta,s). \qquad \forall \beta \in \partial D(s),\ \forall s \leq T. \tag{3.28}$$

In other words: the expected value of the sum of all effective gains as collected in the continuation region by a *free* Brownian motion must be zero, from each and every boundary location and time.

The authors of [35] and [6] have obtained this equation in $d = 1$, with $G_C = 0$ and $G_M = G_S$. For higher dimensional problems this result is new, to the author's best knowledge. It follows that the optimality of any *single* boundary location is dependent on *all* future boundary locations, and therefore the *entirety* of all dynamic boundary locations must be determined at once. In general, we cannot determine any boundary location without knowing all others.

Focusing on the 2-dimensional case specifically, and on the case when the domain is known to be radial, we investigate the optimality condition in detail. We will see that, even if only *one* function $\rho(\phi, t)$ is to be found, the task is still formidable. The optimality equation that must be satisfied by all boundary locations can be classified as a *multidimensional non-linear homogeneous Volterra integral equation of the first kind*, with the distinguishing features that:

1. The unknown function $\rho(\phi, t)$ determines the domain of the integration over $D(\cdot)$; therefore it is a Volterra-type equation. Although Volterra equations normally have the *variable* in the limit of the integration, rather than the unknown *function* as in this case, we argue that Volterra equations are still more applicable than Fredholm equations, which have a fixed and known domain of integration.

2. Apart from its appearance in the domain $D(\cdot)$, the unknown function $\rho(\phi, t)$ also appears *under* the integral sign, because the location of the boundary coordinate x depends on $\rho(\phi, t)$. Because $\rho(\phi, t)$ appears under the integral sign but not outside the integral, it is an equation of the first kind.

3. The unknown function $\rho(\phi, t)$ appears under the integral sign as a function of the free Brownian density B; therefore, it is non-linear.

4. The expected value of all effective gains equals zero; therefore, it is a homogeneous equation.

For 1-dimensional Volterra equations, where the integration extends over a variable linear interval, many known methods exist; see e.g. [10], [11], [12] or [13]. Unfortunately, this is not the case for multidimensional Volterra integral-equations. For existence and uniqueness of solutions of Volterra equations, see e.g. [12] p. 25.

In the single asset case, [7] have shown that the numerical procedure based on the integral method is competitive with the standard binomial procedure. In the multidimensional case, unfortunately, no analogous result exists. Therefore, we will provide our own (possibly very inefficient) numerical procedure in Section 4, to illustrate an example of a max-option.

4 Examples

4.1 The max-option

Consider possibly the simplest non-trivial MOS problem:

$$V_{1,2}(\{x,y\},s) = \max_{s\leq\tau\leq T} \mathbb{E}_{\{x,y\}}\left[e^{-r\,(\tau-s)}\,\max\left\{B_{1,\tau},B_{2,\tau},0\right\}\right].$$

The conditioning in the subscript is on $B_{1,s} = x$ and $B_{2,s} = y$ and we set $r = 1$. The stopping gain G_S and the maturity gain G_M are equal, since it does not matter if the process is stopped before or at maturity. We do expect, however, that the continuation region shrinks over time, such that no gain will ever be collected at maturity. We write G_S as follows:

$$G_S(\{y,x\}) = \max\left\{y,x,0\right\} = \mathbb{1}_{y>x}\,\mathbb{1}_{y>0}\,y + \mathbb{1}_{x>y}\,\mathbb{1}_{x>0}\,x$$

where $\mathbb{1}$ is the indicator function, which equals 1 if the condition in its subscript is satisfied and zero otherwise. A visualisation of the stopping gain is shown in Figure 4. The indicator function $\mathbb{1}$ is not differentiable, but by a limiting procedure we may find that (see e.g. [36] p. 26, or [37] p. 54):

$$\frac{\partial}{\partial x}\mathbb{1}_{x>a} = \delta(x-a),$$
$$\frac{\partial^2}{\partial x^2}\mathbb{1}_{x>a} = \delta'(x-a),$$

where δ is the Dirac delta-function. This is what is meant by a 'distributional derivative' in Theorem 1. Furthermore, the theory of generalised functions suggests (again as in [37] p. 54) that

$$(x-a)\,\delta(x-a) = 0,$$
$$(x-a)\,\delta'(x-a) = -\delta(x-a).$$

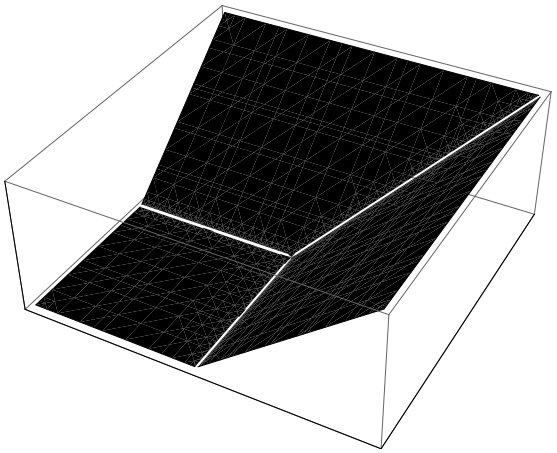


Figure 4. Visualisation of stopping gain $G_S(\{x,y\}) = \max\left\{y,x,0\right\}$

Recall that $\Delta = \frac{1}{2}\nabla^2 + \partial_\tau - r$, but where the derivative with respect to time is redundant in this case, as G_S is independent of time. Also recall that $G_C = 0$. We obtain

$$\begin{aligned}\Delta G_S(\{y,x\}) &= \left(\frac{1}{2}\frac{\partial^2}{\partial x^2} + \frac{1}{2}\frac{\partial^2}{\partial y^2} - r\right) G_S(\{y,x\}) \\ &= \delta(y-x)\mathbb{1}_{x>0} + \frac{1}{2}\delta(x)\mathbb{1}_{x>y} + \frac{1}{2}\delta(y)\mathbb{1}_{y>x} - r\,\max\{x,y,0\}\end{aligned}$$

where δ functions show up wherever G_S has a kink. As a result, the optimality equation (3.28) can be written as:

$$\begin{aligned}0 = \int_s^T d\tau \int\!\!\int_{D(\tau)} dx\,dy \left(\delta(y-x)\mathbb{1}_{x>0} + \frac{1}{2}\delta(x)\mathbb{1}_{x>y} + \frac{1}{2}\delta(y)\mathbb{1}_{y>x} - r\,\max\{x,y,0\}\right) \\ \times e^{-r(\tau-s)}\,B(\{x,y\},\tau|\beta,s).\end{aligned} \tag{4.1}$$

The coordinate β is a 2-dimensional boundary coordinate, i.e. $\beta = \{\beta_1, \beta_2\}$. Intuitively, we see that for an optimal boundary coordinate β, the weighted expected time spent on those lines where G_S has a kink (i.e. the negative x axis, the negative y axis, and the positive diagonal $x = y$) must equal the appropriately weighted time spent in the entire continuation region.

Far away from the diagonal, i.e. when one Brownian motion is much more likely to win than the other, we expect that the optimal policy should only depend on the level of the *leading* one. In those regions, therefore, the problem is a 1-dimensional one; i.e. when to exercise the Brownian motion that is leading. Suppose that $B_{2,t}$ is very negative, such that $B_{1,t}$ is leading. Only the level of $B_{1,t}$ is relevant for the exercise policy, and thus the boundary of the continuation domain should appear as a vertical line in the x, y-plane. In this case, the relevant problem is

$$V_1(x,s) = \max_{s\le\tau\le T} \mathbb{E}_x\left[e^{-r\,(\tau-s)}\,\max\left\{B_{1,\tau},0\right\}\right]$$

and this problem is discussed in for example [35], p. 13. The speed with which the 2-dimensional boundary moves, far away from the diagonal, is determined by the 1-dimensional problem. To solve the 1-dimensional problem, we may use the *same* machinery that we developed for MOS, as our approach is valid for $d \ge 1$. The 1-dimensional problem has

$$G_S(x) = \max(x,0) = \mathbb{1}_{x>0}\,x.$$

And thus

$$\Delta G_S(x) = \frac{1}{2}\delta(x) - r\,\mathbb{1}_{x>0}\,x.$$

Therefore the optimal value, as given by Theorem 1, reads

$$\begin{aligned}V(x,s) &= G_S(x,s) \\ &\quad + \int_s^T d\tau \int_{D(\tau)} d\alpha \quad \left(\tfrac{1}{2}\delta(x) - r\,\mathbb{1}_{x>0}\,x\right) e^{-r(\tau-s)}\,B(\alpha,\tau|x,s).\end{aligned}$$

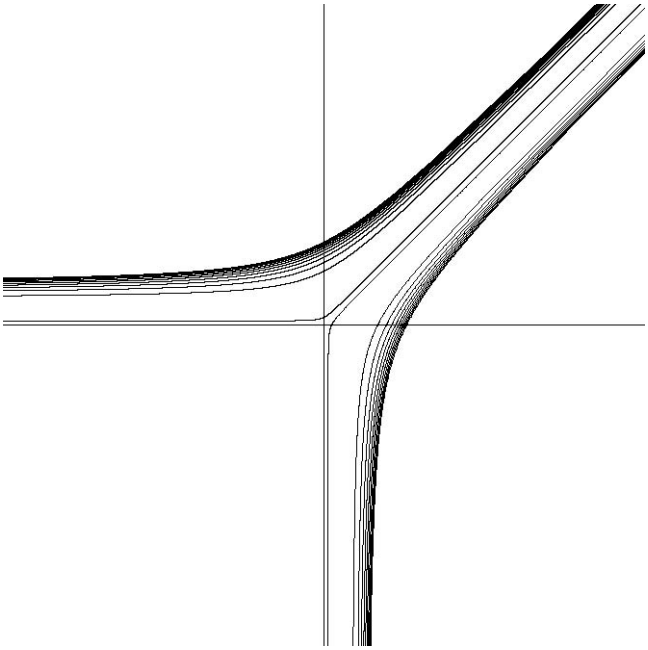


Figure 5. Expected behaviour of the boundary, where 1) the boundary is straight when one Brownian motion is leading the other by a long way, 2) stopping never occurs when the process is on the diagonal, and 3) the continuation region shrinks over time. The domain is radial, where the radius ρ is a function of the polar angle ϕ and time t, where $\rho(\phi, t)$ is given in (4.3), and where ρ goes to ∞ for $\phi \to \pi/4$, $\phi \to -\pi/2$ and $\phi \to \pi$. Whenever the lines in the plot are dense, the speed of the boundary is low. The speed of the boundary increases as time approaches maturity and the boundary at time $t = 0.99$ is seen to approach the axes and the diagonal.

To determine optimality we may apply value-matching. Value-matching implies, intuitively, that, from an optimal boundary location, half the expected time spent where G_S has a kink is equal to the total expected time spent in the continuation region as weighted by $r\,G_S$. This must hold for every optimal boundary location. In [35] the solution is estimated using a discrete approximation of the integral equation (i.e. a sum). We solve the integral equation similarly, but to limit the number of data points that our numerical procedure must remember, we estimate the boundary in the following form:

$$g(t) = \alpha\,(1-t)^{1/6} + \beta\,(1-t)^{1/5} + \gamma\,(1-t)^{1/4} + \delta\,(1-t)^{1/3} + \epsilon\,(1-t)^{1/2} + \zeta\,(1-t). \quad (4.2)$$

It is known from e.g. [35] that the boundary has infinite slope at $t = 1$ and the parametrisation above is chosen to capture this. The parameters are estimated to be $\{\alpha, \beta, \gamma, \delta, \epsilon, \zeta\} = \{0.20, 0.19, 0.17, 0.14, 0.08, 0.12\}$, where we have only indicated the first two decimal places of actual estimates, which involve 6 decimal places. Based on this estimate of the 1-dimensional optimal stopping curve g, we can calculate the violation of value-matching (which is extremely small) to convince ourselves that this is indeed an accurate solution.

Given that the solution to the 1-dimensional problem is given by $g(t)$, we use the program *Mathematica* [38] to *guess* the 2-dimensional boundary as follows:

```
ρ[φ_, t_] := g[t] ((1/Sin[φ] + 1/(Sin[φ] - Cos[φ])) If[Pi / 4 ≤ φ ≤ Pi, 1, 0] +
    (1/Cos[φ] - 1/(Sin[φ] - Cos[φ])) If[-Pi / 2 ≤ φ ≤ Pi / 4, 1, 0])
```
(4.3)

The radius ρ of the 2-dimensional domain is given as a function of the angle ϕ and time t. This particular guess supposes that the radius ρ is separable in time t and angle ϕ: it is given by the product of a function that depends only on t, and a function that depends only on ϕ. The polar angle ϕ runs from $-\pi/2$ (south) counter-clockwise to π (west). The diagonal is at $\pi/4$ (north-east). The dependence on t is inspired by the 1-dimensional problem, because we suspect that the 2-dimensional boundary is straight for ϕ in the directions south or west. In those areas, the boundary should move with a speed that is dictated by the 1-dimensional problem. The proposed dependence on the polar angle is such that the boundary becomes a straight for either $\phi \to -\pi/2$ and $\phi \to \pi$, which is suspected by intuition. The dependence on ϕ is different on either side of the diagonal (where $\phi = \pi/4$): to the right of the diagonal we guess that it looks like $1/\cos(\phi) + 1/(\sin(\phi) + \cos(\phi))$, such that becomes infinite when ϕ points south or north-east. And similarly for ϕ in the range from $\pi/4$ to π. The resulting boundary is plotted in Figure 5. Whenever the lines are dense, the speed of the proposed boundary is relatively low. Towards maturity, the speed of the proposed boundary increases and at maturity the speed is infinite. It is still integrable, however, as can be seen from the definition of g in (4.2).

This initial guess is is not necessarily optimal, even though we do expect that it is optimal for $\phi \to -\pi/2$ and $\phi \to \pi$. We will try and improve on our initial guess by the following procedure in *Mathematica*:

1. We make a table of our initial guess $\rho(\phi, t)$, where ϕ runs from $-\pi/2 + \epsilon$ to $\pi - \epsilon$ (where ϵ is small, to avoid having to deal with infinite ρ), and where t runs from 0 to $T = 1$. Each point in the table represents a boundary location, where the radius is given as a function of the angle ϕ and time t. We prescribe that $\rho(\pi/4, t) = 20\, g(t)$, because we know that ρ should be infinite on the diagonal:

```
s = 0; T = 1; AnglePoints = 101; TimePoints = 5; ∈ = 1 / 100;
InterpolationTable = Table[If[φ ≠ Pi / 4, Round[ρ[φ, t], 0.001], 20 g[t]],
  {φ, -Pi / 2 + ∈, Pi - ∈, (3 / 2 Pi - 2 ∈)/(AnglePoints - 1)}, {t, s, T, (T - s)/(TimePoints - 1)}]
```

2. Second, we do an interpolation of order 1 in both angle and time to obtain, once more, a function ρ that is defined for all ϕ and t, i.e.

```
InterpolOrder = 1;
Interpolatedρ[table_] := ListInterpolation[table,
  {{-Pi / 2 + ∈, Pi - ∈}, {0, T}}, InterpolationOrder → {InterpolOrder, InterpolOrder}]
```

3. Based on this 'interpolated ρ', we can calculate the (supposedly optimal) value V. First we define

```
r = 1;
WorkingPrec = 4;
B[y_, t_, x_, s_] := 1 / (√(2 π (t - s))) e^(-((y-x).(y-x)) / (2 (t-s)))
Gs[ρ_, φ_] := Max[ρ Sin[φ], ρ Cos[φ], 0]
InteriorCoord[ρ_, φ_] := {ρ Cos[φ], ρ Sin[φ]}
SurfaceCoord[table_, φ_, t_] :=
 {Interpolatedρ[table][φ, t] Cos[φ], Interpolatedρ[table][φ, t] Sin[φ]}
```

And then the value V is given by

```
V[table_, phi_, s_] :=
 Gs[Interpolatedρ[table][phi, s], phi]
+NIntegrate[1/2 e^(-r (τ-s)) B[{x, 0}, τ, SurfaceCoord[table, phi, s], s],
  {τ, s, T}, {x, -∞, 0}, WorkingPrecision → WorkingPrec]
+NIntegrate[1/2 e^(-r (τ-s)) B[{0, y}, τ, SurfaceCoord[table, phi, s], s],
  {τ, s, T}, {y, -∞, 0}, WorkingPrecision → WorkingPrec]
+NIntegrate[e^(-r (τ-s)) B[{z, z}, τ, SurfaceCoord[table, phi, s], s],
  {τ, s, T}, {z, 0, ∞}, WorkingPrecision → WorkingPrec]
+With[{R = Interpolatedρ[table]}, NIntegrate[
   - r Gs[ρ, φ] ρ e^(-r (τ-s)) B[InteriorCoord[ρ, φ], τ, SurfaceCoord[table, phi, s], s],
   {τ, s, T}, {φ, 0, 2 Pi}, {ρ, 0, R[φ, τ]}, WorkingPrecision → WorkingPrec]]
```

where the value V at boundary coordinate (ϕ, t) is equal to a sum of the immediate stopping value (first term), an integration is over the negative x-axis (second term), over the negative y-axis (third term), over the positive diagonal $x = y$ (fourth term), and over the continuation region (fifth term). The radius is involved in the space-time boundary coordinate

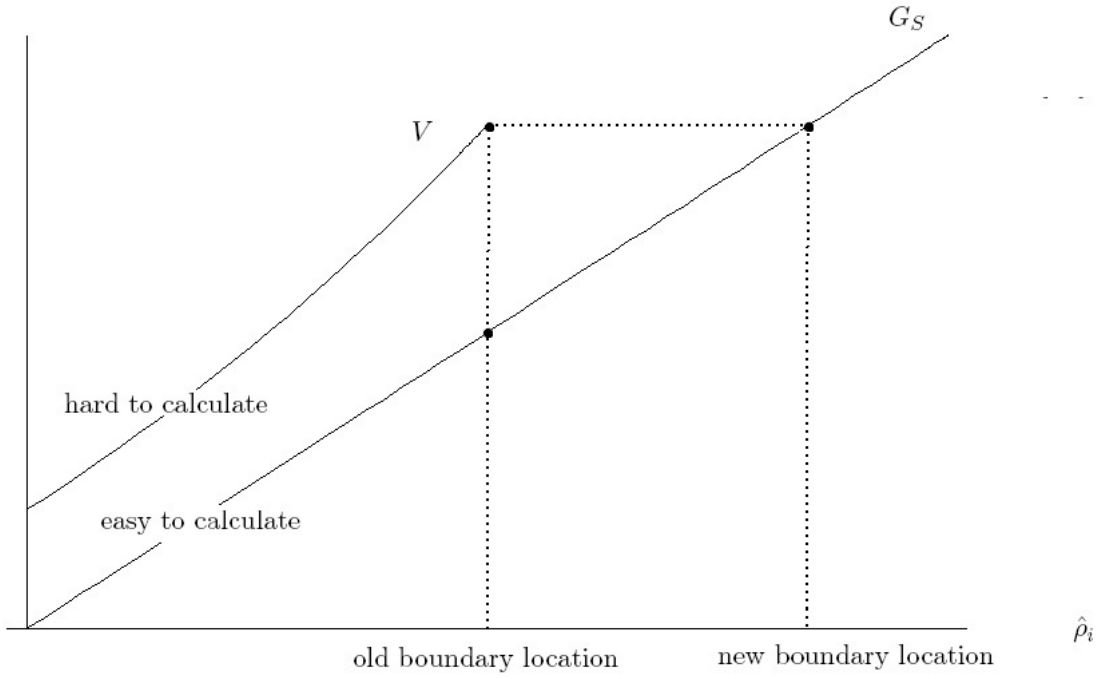
$$(\{x, y\}, s) = (\{\rho(\phi, s)\ \cos(\phi), \rho(\phi, s)\ \sin(\phi)\}, s)$$

as well the integration over the continuation region as

$$\int_0^T d\tau \int_{-\pi/2}^{\pi} d\phi \int_0^{\rho(\phi,\tau)} d\rho\, \rho$$

where $\rho(\phi, \tau)$ is based on an interpolation of a table, wherever it appears.

4. The value at the boundary must satisfy value-matching, i.e. the second to fifth terms in V are supposed to add to zero. If the value V at some grid-point is *higher* than the immediate stopping gain G_S, then the domain needs to grow bigger. If the value V at some grid-point is *smaller* than the immediate stopping gain G_S, then the domain needs to grow smaller. This observation suggests the following procedure to update the boundary, when value-matching is not satisfied:



Where we rely on the fact that the slope of G_S is often greater than the slope of V, so that the new boundary location is 'close' to the old one. When ρ is used as the parameter on the horizontal axis, as we do, then the slope of G_S can be very small when ϕ is close to $-\pi/2$ or π. To correct for this, we may add a linear curve to both curves drawn above — and as a result the correction becomes smaller. For some ϕ and t, the new $\rho(\phi, t)$ can be found as follows:

```
FR[φ_, s_] := Module[{ρ}, Re[
   ρ //. FindRoot[V[InterpolationTable2, φ, s] + Interpolatedρ[InterpolationTable2][φ, s] ==
       Gs[ρ, φ] + ρ, {ρ, 1}, MaxIterations → 5, WorkingPrecision → 3]]]
```

where we have added ρ to both sides of the equation to make sure the slope of the right-hand-side is large enough, and where we allow a maximum of 5 find-root iterations. The new radius, for which value-matching would be satisfied if V was

calculated based on the old one, is used as the new grid-point. We estimate a new value of $\rho(\phi, \tau)$ for all points in the table that defines $\rho(\phi, \tau)$, i.e.:

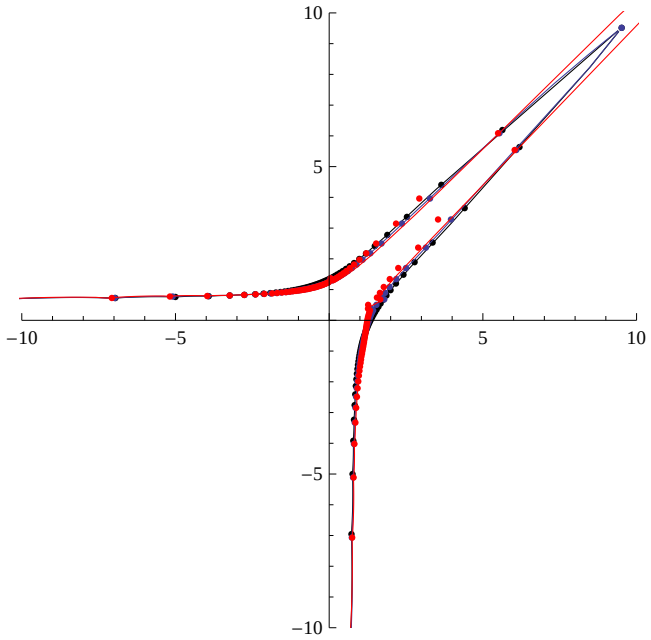
```
InterpolationTable2 = Table[If[φ ≠ Pi / 4, FR[φ, t], 20 g[t]],
  {φ, -Pi / 2 + ϵ, Pi - ϵ, (3 / 2 Pi - 2 ϵ) / (AnglePoints - 1)}, {t, s, T, (T - s) / (TimePoints - 1)}]
```

5. Based on this new table, we do a new interpolation:

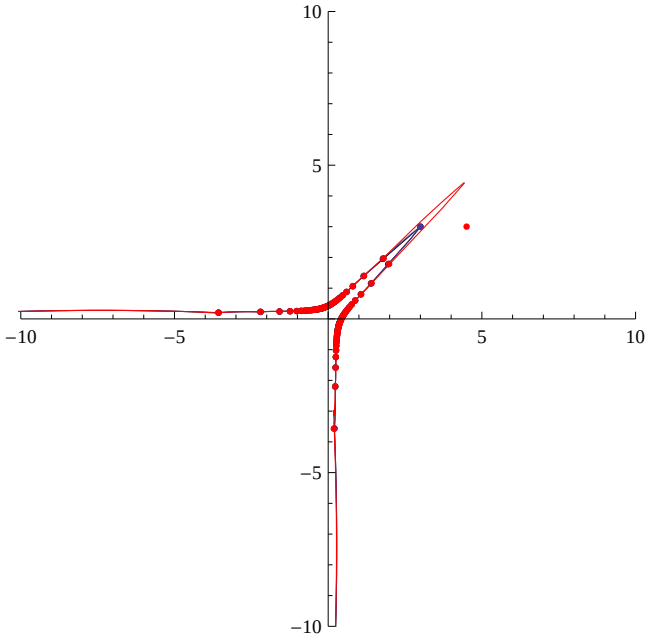
```
InterpolationTable2 = Table[If[φ ≠ Pi / 4, FR[φ, t], 20 g[t]],
  {φ, -Pi / 2 + ϵ, Pi - ϵ, (3 / 2 Pi - 2 ϵ) / (AnglePoints - 1)}, {t, s, T, (T - s) / (TimePoints - 1)}]
```

and continue with this procedure until the table defining $\rho(\phi, \tau)$ no longer changes up to some small positive tolerance.

This procedure is just heuristic — there is no proof that it will or should converge, and we try it in *Mathematica* [38]. It might not be very accurate, but we simply show it here as a proof of principle. After two iterations, we get



where the black equals the interpolation of our initial guess, the blue dots indicate the first iteration and the red dots the second one. For $t = .9$, i.e. very near maturity, we get with the same colour-coding:



We see that the continuation domain shrinks over time, and in both cases we see that subsequent iterations do not change the boundary points much, and therefore we conclude that the boundary must be near-optimal — at least with the current working precisions as specified. The reliance on the initial guess, however, is substantial, especially as far as the *movement* of the boundary is concerned. Although for each point in time we specified 100 points on the boundary, we only specified 5 points in time. Therefore the speed of the boundary in this solution might not be very accurate, but it does give us some idea of its shape.

4.2 Parallel investment in two alternatives

Here we consider the problem of parallel investment in two alternatives, where the winner produces revenues at maturity, which are linear in its performance at maturity. The performance develops stochastically as follows

$$P_1(t) = \mu_1\, t + \sigma_1\, B_{1,t}$$
$$P_2(t) = \mu_2\, t + \sigma_2\, B_{2,t}$$

where $B_{1,t}$ and $B_{2,t}$ are independent Brownian motions, satisfying $\mathbb{E}\, B_{i,t} = 0$ and $\mathbb{E}\, B_{i,t}^2 = t$, and where the end date of each project is taken to be $T = 1$. The value of optimal parallel investment in projects 1 and 2 is given by $V_{1,2}$:

$$V_{1,2}(\{x,y\},s) := \max_{s\le\tau\le T} \mathbb{E}_{(x,y)} \left[e^{-r\,(\tau-s)} \max\Big\{V_1(B_{1,\tau},\tau), V_2(B_{2,\tau},\tau)\Big\} - \int_s^\tau d\theta\,(c_1+c_2)\, e^{-r(\theta-s)} \right]$$

where V_1 and V_2 are the optimal values of projects 1 and 2, if they were continued optimally and in isolation, where the maximisation is over stopping time τ, and where the

conditioning in the subscript of $\mathbb{E}$ is on the values of $B_{1,s} = x$ and $B_{2,s} = y$. The optimal values of projects 1 and 2 are given by:

$$V_1(x,s) := \max_{s \le \tau \le T} \mathbb{E}_x \left[e^{-r\,(T-s)} \left(\mu_1\, T + \sigma_1\, B_{1,T}\right) \mathbb{1}_{\tau = T} - \int_s^\tau d\theta\, c_1\, e^{-r\,(\theta - s)} \right]$$

$$V_2(x,s) := \max_{s \le \tau \le T} \mathbb{E}_x \left[e^{-r\,(T-s)} \left(\mu_2\, T + \sigma_2\, B_{2,T}\right) \mathbb{1}_{\tau = T} - \int_s^\tau d\theta\, c_2\, e^{-r\,(\theta - s)} \right]$$

where the maximisation is over all stopping times τ and where the optimal value of either project in isolation equals an expectation of the performance at maturity, if and only if the project is not abandoned before that time, minus an expectation of the continuation cost c_i which is to be paid at each unit of time when the project is not stopped. To summarise the 2-dimensional problem, we have

Continuation gain	$-c_1 - c_2$
Stopping gain	$\max\left\{V_1(B_{1,\tau},\tau), V_2(B_{2,\tau},\tau)\right\}$
Maturity gain	

We expect that the continuation region of $V_{1,2}$ — corresponding to continued investment in both projects — shrinks to zero as time goes to T, because an optimal policy would never allow *both* projects to be completed. It would be better, for example, to abandon the project that is almost certainly going to lose a small time before completion, and therefore the maturity gain will never be obtained. In general, the optimal value reads as follows:

$$\begin{aligned} V(x,s) = {} & G_S(x,s) \\ & + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau - s)}\, B(\alpha,\tau | x,s) \\ & \quad + \int\limits_{D(T)} d\alpha \left(G_M(\alpha,T) - G_S(\alpha,T) \right) e^{-r(T-s)}\, B(\alpha,T|x,s). \end{aligned}$$

Because G_M and G_S at $t = T$ are equal, we get

$$\begin{aligned} V(x,s) = {} & G_S(x,s) \\ & + \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left(G_C(\alpha,\tau) + \Delta G_S(\alpha,\tau) \right) e^{-r(\tau - s)}\, B(\alpha,\tau | x,s). \end{aligned}$$

The continuation gain equals

$$G_C(\{x,y\},s) = -c_1 - c_2$$

and for the stopping gain we have that

$$G_S(\{x,y\},s) = \max\left\{V_1(x,s), V_2(y,s)\right\} = \mathbb{1}_{V_1(x,s) > V_2(y,s)}\, V_1(x,s) + \mathbb{1}_{V_2(x,s) > V_1(y,s)}\, V_2(y,s)$$

and we need to calculate $\Delta G_S(\{x,y\},s)$, which is not very hard but a little tedious. After a lot of bookkeeping, and using the same δ-function identities as in the previous subsection, we get that

$$G_C(\{x,y\},s)+\Delta G_S(\{x,y\},s) = -c_1\,\mathbb{1}_{V_2(x,s)>V_1(y,s)} - c_2\,\mathbb{1}_{V_1(x,s)>V_2(y,s)} \\ +\delta\big(V_1(x,s)-V_2(y,s)\big)\left(\frac{1}{2}\left(\frac{\partial V_1}{\partial x}\right)^2+\frac{1}{2}\left(\frac{\partial V_2}{\partial y}\right)^2\right).$$

Here we have also used that the 1-dimensional values V satisfy

$$\Delta V_i(x,s) = c_i.$$

The optimal value reads

$$V(x,s) = G_S(x,s) \\ +\int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left[-c_1\,\mathbb{1}_{V_2(x,s)>V_1(y,s)} - c_2\,\mathbb{1}_{V_1(x,s)>V_2(y,s)} \right. \\ \left. +\,\delta\big(V_1(x,s)-V_2(y,s)\big)\left(\frac{1}{2}\left(\frac{\partial V_1}{\partial x}\right)^2+\frac{1}{2}\left(\frac{\partial V_2}{\partial y}\right)^2\right)\right] e^{-r(\tau-s)}\,B(\alpha,\tau|x,s)$$

and value-matching requires

$$0 = \int_s^T d\tau \int\limits_{D(\tau)} d\alpha \left[-c_1\,\mathbb{1}_{V_2(x,s)>V_1(y,s)} - c_2\,\mathbb{1}_{V_1(x,s)>V_2(y,s)} \right. \\ \left. +\,\delta\big(V_1(x,s)-V_2(y,s)\big)\left(\frac{1}{2}\left(\frac{\partial V_1}{\partial x}\right)^2+\frac{1}{2}\left(\frac{\partial V_2}{\partial y}\right)^2\right)\right] e^{-r(\tau-s)}\,B(\alpha,\tau|\beta,s)$$

for all boundary coordinates β. Interpreting this optimality equation, we find that

Corollary 2. *For investment in two alternative projects with continuation costs c_1 and c_2, from each optimal boundary location β, the expected time spent, by a free Brownian motion, in the continuation region where project 1 is in the lead — as weighted by c_2 — plus the expected time spent in the continuation region where project 2 is in the lead — as weighted by c_1 — must equal the expectation of time spent on the curve $V_1 = V_2$, where both projects are equally valuable — as weighted by $\frac{1}{2}\left(\frac{\partial V_1}{\partial x}\right)^2+\frac{1}{2}\left(\frac{\partial V_2}{\partial y}\right)^2$.*

The total 'loss' is given by the expected total value of all money spent on either project while the other project is in the lead. This is something you would want to minimise in general. The total 'gain' is equal to a weighted expectation of the time spent on the curve $V_1 = V_2$, where both projects are equally valuable. If the expected time on the curve given by $V_1 = V_2$ is high, then many switches are expected in which project is leading, and thus the option value to wait is valuable. Theorem 1 applied to this situation shows that from each boundary point, the expected total loss and expected total gain must equal.

If we want to allow the situation where the performances of both projects are correlated, then we need to solve

$$V_{1,2}(\{x,y\},s) := \max_{s\le\tau\le T} \mathbb{E}_{(x,y)} \left[e^{-r\,(\tau-s)} \max\Big\{V_1(W_{1,\tau},\tau), V_2(W_{2,\tau},\tau)\Big\} - \int_s^\tau d\theta\,(c_1+c_2)\,e^{-r(\theta-s)} \right]$$

where

$$\begin{aligned} W_{1,t} &= B_{1,t} \\ W_{2,t} &= (\rho\,B_{1,t} + (1-\rho^2)^{1/2}\,B_{2,t}) \end{aligned}$$

where $B_{1,t}$ and $B_{2,t}$ are truly independent Brownian motions, and where $W_{1,t}$ and $W_{1,t}$ are processes of unit variance and correlation ρ, see e.g. [4], p. 171. This shows that the treatment where the different stochastic processes are correlated is not fundamentally different from the case where they are independent.

However, it is not clear that the optimal domain should be radial. When V_1 and V_2 represent identical projects, the domain is symmetric in the diagonal, i.e. $x = y$, and it may well be radial. But when the two alternative projects have different variances σ_i^2, for example, and if their correlation is furthermore non-zero, then it is not clear a priori if the domain to be found will be radial. The reduction from two to one unknown functions is therefore not guaranteed to be valid, and therefore we cannot solve this problem numerically yet.

5 Conclusion

In this paper we have considered the *problem of alternatives*, where several competing technologies or drugs are developed over time, and where there can be only *one* winner.

We have viewed this problem in the wider context of *multidimensional optimal stopping* (MOS) *problems*. We have noted that far fewer results are known than in the case of 1-dimensional optimal stopping problems, where results are abundant — see for example the extensive literature on the American option.

We have established the relationship between boundary value problems on the one hand, in particular the Dirichlet and Neumann problems, and free-boundary problems on the other. Both boundary value problems can be solved for any domain $D(\cdot)$, as long as it has a finite number of singular boundary points and the boundary has an integrable speed. In each case the solution automatically satisfies 3 out of the 4 conditions that would be imposed on an optimal value. The Dirichlet value satisfies value-matching but not smooth-pasting, and the Neumann value satisfies smooth-pasting but not value-matching.

The optimal stopping problem, or free-boundary problem, prescribes both the value and derivative, and the *one* domain is optimal for which the Dirichlet and Neumann values are equal. Optimality can be obtained either by imposing smooth-pasting on the Dirichlet value, or by imposing value-matching on the Neumann value. A third proof was added, which arrives at the correct result without referring to either absorbed or reflected Brownian motion, with the disadvantage that it may appear rather ad-hoc.

Using the integral formulae derived in Section 2, it was shown in Section 3 that by imposing optimality in either of the described ways, the absorbed/reflected density in the Dirichlet/Neumann value is replaced by the free Brownian density — which is a known quantity. The derivation assumes that the boundary has a finite number of singular points at each time, such that Green's theorem is allowed, as well as that boundary points have integrable speeds, such that Reynold's theorem is allowed. While it is not obvious, a priori, that these conditions should be satisfied, they do not seem overly restrictive.

While the optimal value can be written as a function of the free Brownian density if the optimal domain is known, the optimal domain is a priori unknown. We show that the optimal domain can be found by imposing *either* value-matching *or* smooth-pasting on the derived optimal value. The result is an integral equation, where the optimality of any *single* boundary location depends on the *entire* future continuation domain $D(\cdot)$ and thus no boundary location is independent of others. Instead, the entire boundary must be found at once.

Specifically, we found in Corollary 1 that for x on the optimal boundary, the expected value of all effective continuation and effective maturity gains, as collected by the free Brownian path during its time in the optimal continuation domain $D(\cdot)$, equals zero. This

holds true for all boundary locations $x \in \partial D(\cdot)$.

Theorem 1 and Corollary 1 are new, to the author's best knowledge — and this paper seems to provide one of the first more general results in the field of MOS. While some specific multidimensional problems have been discussed in the literature (as mentioned on p. 18), there appears to be a lack of more general results.

To solve the integral equation for any specific problem, the boundary needs to be parametrised in one way or another. For a general 2-dimensional domain, a parametric representation requires both Cartesian coordinates x and y to be specified as functions of a certain parameter, and of time. Instead we examine a problem for which we know a priori that the domain can be specified by providing the *radius* of the boundary as a function of the polar angle ϕ and of time t (and we call this a *radial domain*). This reduces the number of unknown functions from two to one, but finding it is still a formidable task. The integral equation is classified, and Section 4 proposes a (possibly inefficient) method of improving upon some initial estimate $\rho(\phi, t)$. The approach is heuristic and depends on a good first guess. It is more a proof of principle rather than a rigorous algorithm for solving multidimensional integral equations, and does not deal with matters of convergence and accuracy.

Even if we cannot yet solve the problem of alternatives numerically with great accuracy, the application of Theorem 1 does provide some new intuition. Corollary 2, for example, states that for investment in two alternative projects with continuation costs c_1 and c_2, optimality demands that from each optimal boundary point β, the expected time spent by a free Brownian motion in the continuation region where project 1 is in the lead — as weighted by c_2 — plus the expected time spent in the continuation region where project 2 is in the lead — as weighted by c_1 — is equal to the expectation of the (appropriately weighted) time spent on the curve $V_1 = V_2$, where both projects are equally valuable. Intuitively, this means that the total expected 'loss' — defined as the total amount of money spent on projects when they are not in the lead — is allowed to be greater if many switches are expected in which project is leading.

We conclude that the problem of alternatives — or more generally problems with 1) a non-additive revenue structure regarding different project, 2) stochastic development of the 'performance' of each project and 3) significant discounting and/or finite maturity — can be viewed in the wider context of MOS problems. The MOS literature is limited and this paper takes a first step in developing the theory. Much remains to be explored, however, and in particular we need a method for solving multidimensional integral equations efficiently and accurately, which would allow our optimality equation to be used by practitioners for real-world problems.

References

[1] M. DeGroot, *Optimal statistical decisions.* McGraw-Hill Series in Probability and Statistics, 1970. 6

[2] M. Weitzman, *Optimal search for the best alternative*, *Econometrica: Journal of the Econometric Society* **47** (1979), no. 3 641–654. 6, 23

[3] M. Broadie and J. Detemple, *The valuation of American options on multiple assets*, *Mathematical Finance* **7** (1997), no. 3 241–286. 12, 17

[4] S. Shreve, *Stochastic Calculus Models for Finance II: Continuous Time Models.* Springer Finance, 2005. 12, 59

[5] A. Dixit, R. Pindyck, and G. Davis, *Investment under uncertainty*, vol. 15. Princeton University Press Princeton, NJ, 1994. 16, 17

[6] G. Peskir and A. Shiryaev, *Optimal stopping and free-boundary problems*, vol. 10. Birkhauser, 2006. 16, 18, 47

[7] M. Broadie and J. Detemple, *American option valuation: new bounds, approximations, and a comparison of existing methods*, *Review of Financial Studies* **9** (1996), no. 4 1211–1250. 18, 21, 23, 48

[8] M. Davis and A. Norman, *Portfolio selection with transaction costs*, *Mathematics of Operations Research* **15** (1990), no. 4 676–713. 18

[9] V. Bally and G. Pags, *A Quantization Algorithm for Solving Multi-Dimensional Discrete-Time Optimal Stopping Problems*, *Bernoulli* **9** (2003), no. 6 1003–1049. 18

[10] N. Muskhelishvili, *Singular integral equations. Translated from the second Russian edition.* Groningen: Wolters-Noordhoff Publishing, 1967. 21, 48

[11] D. Porter and D. Stirling, *Integral equations: a practical treatment, from spectral theory to applications.* Cambridge University Press, 1990. 21, 48

[12] W. Hackbusch, *Integral equations: theory and numerical treatment.* Birkhauser Verlag, Basel, Switzerland, 1995. 21, 48

[13] A. Polyanin, A. Manzhirov, and A. Polianin, *Handbook of integral equations.* CRC press, 1998. 21, 48

[14] G. Green, *An essay on the application of mathematical analysis to the theories of electricity and magnetism.* Printed for the author by T. Wheelhouse, 1828. 22

[15] O. Kellogg, *Foundations of potential theory.* Dover Pubns, 1929. 22

[16] T. Lorenz, *Reynold's transport theorem for differential inclusions*, *Set-Valued Analysis* **14** (2006), no. 3 209–247. 22, 28, 29

[17] H. Markowitz, *Portfolio selection*, *Journal of Finance* **7** (1952), no. 1 77–91. 22

[18] M. Rubinstein, *Markowitz's portfolio selection: A fifty-year retrospective*, *The Journal of Finance* **57** (2002), no. 3 1041–1045. 22

[19] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, *Mathematics of Operations Research* **28** (2003), no. 1 1–38. 22

[20] F. Kelly, *A remark on search and sequencing problems*, *Mathematics of Operations Research* **7** (1982), no. 1 154–157. 23

[21] U. Lössner and I. Wegener, *Discrete sequential search with positive switch cost*, *Mathematics of Operations Research* **7** (1982), no. 3 426–440. 23

[22] E. Denardo, U. Rothblum, and L. Van der Heyden, *Index policies for stochastic search in a forest with an application to R&D project management*, *Mathematics of Operations Research* **29** (2004), no. 1 162–181. 23

[23] J. Sethuraman and J. Tsitsiklis, *Stochastic search in a forest revisited*, *Mathematics of Operations Research* **32** (2007), no. 3 589–593. 23

[24] R. Weber, *On the Gittins index for multiarmed bandits*, *The Annals of Applied Probability* **2** (1992), no. 4 1024–1033. 23

[25] K. Glazebrook and D. Wilkinson, *Index-based policies for discounted multi-armed bandits on parallel machines*, *Annals of Applied Probability* **10** (2000), no. 3 877–896. 23

[26] K. Glazebrook and R. Minty, *A Generalized Gittins Index for a Class of Multiarmed Bandits with General Resource Requirements*, *Mathematics of Operations Research* **34** (2009), no. 1 26–44. 23

[27] K. Glazebrook, *Stoppable families of alternative bandit processes*, *Journal of Applied Probability* (1979) 843–854. 23

[28] J. Gittins, *Bandit processes and dynamic allocation indices*, *Journal of the Royal Statistical Society. Series B (Methodological)* **41** (1979), no. 2 148–177. 23

[29] I. Karatzas, *On the pricing of American options*, *Applied mathematics and optimization* **17** (1988), no. 1 37–60. 23

[30] S. Jacka, *Optimal stopping and the American put*, *Mathematical Finance* **1** (1991), no. 2 1–14. 23

[31] P. Carr, R. Jarrow, and R. Myneni, *Alternative characterizations of American put options*, *Mathematical Finance* **2** (1992), no. 2 87–106. 23

[32] J. Dodziuk, *Eigenvalues of the Laplacian and the Heat Equation*, *The American Mathematical Monthly* **88** (1981), no. 9 686–695. 26

[33] P. Hsu, *On Excursions of Reflecting Brownian Motion*, *Transactions of the American Mathematical Society* **296** (1986), no. 1 239–264. 26, 28

[34] K. Burdzy, Z. Chen, and J. Sylvester, *The heat equation and reflected Brownian motion in time-dependent domains*, *Annals of probability* **32** (2004), no. 1B 775–804. 26, 28

[35] J. Pedersen and G. Peskir, *On nonlinear integral equations arising in problems of optimal stopping*, *Proc. Funct. Anal. VII (Dubrovnik 2001)* **46** (2002) 159–175. 47, 50, 51

[36] I. Gel'fand and G. Shilov, *Generalized Functions: Properties and operations, by I. Gel'fand and G. Shilov, translated by E. Saletan*, vol. 1. Academic Press, 1964. 49

[37] W. Steeb, *Hilbert spaces, wavelets, generalised functions, and modern quantum mechanics*, vol. 451. Kluwer Academic Pub, 1998. 49

[38] I. Wolfram Research, "Mathematica Edition: Version 8.0." Wolfram Research, Inc., Champaign, Illinois, 2010. 52, 55

[This page was intentionally left blank]

[This page was intentionally left blank]

TO BE SUBMITTED

Part III — Optimal support for renewable deployment

Rutger-Jan Lange[1]

University of Cambridge,
792 King's College, Cambridge, CB2 1ST, United Kingdom

E-mail: rjl63@cam.ac.uk

Abstract: This paper considers the role of government subsidies for renewable technologies, such as feed-in tariffs, which have become a topic of controversy. Rapidly diminishing levels of subsidy are cheaper to the taxpayer, but could equally kill an otherwise successful technology prematurely. Sustaining high levels of subsidy, on the other hand, is not only expensive but also keeps alive the worst-case scenario of a technology being supported indefinitely without ever becoming economical. Our first contribution is that we find an optimal feed-in tariff for German photo-voltaics, by formulating the trade-off as an optimal stopping problem. We use the method of (multidimensional) optimal stopping as developed in Part II of the thesis, and apply it to the 1-dimensional case of Part III. A second, and independent, contribution of this paper is that it provides a new model for technology learning with endogenous market growth. Other analyses of learning curves have taken capacity expansion to be the exogenous variable. We recognise, similarly, that investment drives cost reductions, through the learning curve, but our model recognises *also* that cost reductions, in their turn, drive profitability and (further) investment. This feedback loop is new and adds to the literature on mean-reverting and runaway processes.

Keywords: renewable energy, optimal stopping, feed-in tariff

[1]Research support from the Electricity Policy Research Group in Cambridge is gratefully acknowledged (http://www.eprg.group.cam.ac.uk).

Contents

1 Introduction

This paper considers the role of government subsidies for renewable technologies, such as feed-in tariffs, which have become a topic of controversy. Rapidly diminishing levels of subsidy are cheaper to the taxpayer, but could equally kill an otherwise successful technology prematurely. Sustaining high levels of subsidy, on the other hand, is not only expensive but also keeps alive the worst-case scenario of a technology being supported indefinitely without ever becoming economical.

Within this introduction, subsection 1.1 introduces the questions that arise from large-scale deployment of renewable energy, in the context of both technology policy and climate change policy. Subsection 1.2 then discusses how technologies learn from increased deployment by progress on their 'learning curve'. Subsection 1.3 challenges the traditional view that capacity expansion drives cost reduction, and proposes that the reverse is also true: cost reductions also drive capacity expansion. Finally, subsection 1.4 discusses how to formulate the problem of feed-in tariffs as an optimal stopping problem, and suggests that we use the machinery of multidimensional optimal stopping problems, as developed in Part II of this thesis, to tackle the 1-dimensional problem. This paper makes two contributions:

1. It formulates an optimal policy for the feed-in tariff for German photo-voltaic energy.
2. And, independently, it proposes a model for technology learning that makes market growth endogenous.

1.1 The problem

This paper considers government subsidies for the deployment of renewables, such as feed-in tariffs. Some governments oblige electricity utilities to buy all electricity generated by renewable sources at a high, predetermined price — the feed-in tariff — which may be up to 6 times the market rate, and which decreases annually such that energy generated by newer capacity gets paid at a lower rate, even though the tariff stays fixed for 20 years for all electricity generated by a given unit of capacity.

In general, four arguments — of variable validity — can be made in support of low-carbon technologies (relying heavily on [1]):

1. The EU ETS does not generate a carbon price that is high enough to stimulate renewable investment. Further, its timeframe is too short to secure sustained private commitment. Feed-in tariffs can indeed be credible in the long term, but effectively they a) lower the carbon price and b) do not reduce total carbon emissions.

2. A flourishing renewable industry is supposed to increase the security of supply, but renewable energy appears to displace secure LNG imports in particular, rather than imports from unstable regions such as Russia and Algeria.

3. Subsidies in renewable deployment help technologies move down the 'experience curve' (or 'learning curve'), such that current subsidies are in effect an investment in reduced future cost. The validity of this argument depends on a) the potential of future cost reductions, b) whether large-scale deployment is a better method of achieving this than e.g. R&D, and c) whether or not the technology can be deployed globally. The answer to these questions is technology specific.

4. Renewables may seem to be the only politically and publicly acceptable road towards a low-carbon future — but attitudes may change.

The 'experience curve' argument requires some extra conditions (is there enough potential, why deployment rather than R&D, and is it scalable?) to justify spending large sums of public money. But in fact it may well be the *only* argument, out of the four presented, that withstands proper scrutiny. Solar energy, for example, is currently still expensive; but it is operationally clean, abundant, and has the potential for large cost reductions as well as large-scale deployment. In Germany, solar power already produces up to 10% of the energy on a sunny day. Solar energy is in principle so attractive that it justifies taking some risks. Furthermore, announcing a receding feed-in tariff does not commit society to investing in solar energy for the next 20 years, as the solar industry will grow *if and only if* it can continue be profitable under the (receding) feed-in tariff.

Feed-in tariffs are technology specific (i.e. different for wind and solar), can be up to 6 times the market rate (up to 0.29€/kWh for solar roof panels), depend on the year of installation (but are then fixed for 20 years) and are lower for more recently installed panels (9% decrease p.a. for solar in Germany). The main driver of feed-in tariffs has been the expectation that as the (total historical) cumulative installed capacity increases, the cost per unit will decline.

The author of this paper is of the opinion that a sensible policy should, at least, address these four questions:

1. When should government subsidies be discontinued? It is generally agreed that this should happen when an energy source has captured a sizeable portion of the market. 'Sizeable' is considered to be 5% by some, but that number is debatable. Wind energy has already reached this level — but it is unclear whether the wind sector would stagnate without further subsidies. The photovoltaic industry is still in its development phase; its penetration in the German market, for example, is around 2%. Currently German EEG law stipulates that the decrease in the tariff for solar energy 'shall be 9% from the year 2010 onwards', but no end date is specified. This implies, by extrapolation, that solar energy must become economical around 2020, when the feed-in tariff is expected to match industrial market prices. But if current German growth rates (around 40%) are anything to go by, the market size of solar energy in 2020 will far exceed 5%.

2. How quickly should feed-in tariffs recede? Rapidly diminishing levels of subsidy are cheaper to the taxpayer and presumably incentivise innovation, but could equally kill an otherwise successful technology prematurely. Sustaining high levels of subsidy, on the other hand, is not only expensive but also keeps alive the worst-case scenario of a technology being supported indefinitely without ever becoming economical.

3. Should feed-in tariffs be predetermined or performance dependent? Predetermined (but decreasing) feed-in tariffs may encourage higher private commitment, but they tie governments' hands when costs come down quicker than expected. The wrong choice can be expensive. Spain's renewable energy sector is suing its government over promises that prompted heavy investment but that were subsequently broken. Likewise, the unexpected reduction of solar subsidies in France is said to have damaged investor confidence. On 30 June 2011, the Deutsche Bundestag passed changes to the feed-in tariff, making them dependent on the *amount* of aggregate newly installed capacity of the previous year, thereby making the feed-in tariff a performance-dependent variable. This allows for flexibility on the part of the government, but provides pri-

vate investors with less certainty about future returns and may thus disincentivise investment.

4. Implicitly, solar energy competes not only against the incumbent technology that sets the 2020 target, but also against other low-carbon technologies such as wind — where the extent to which different low-carbon technologies are truly alternatives is debatable. Feed-in tariffs, it can be argued, are ultimately not technology policies but climate policies. It is highly desirable to have a) competition between technologies, and b) parallel paths towards a sustainable mix of energy sources, *even if* wind energy and solar energy are considered as alternatives. But it is unclear if the feed-in tariffs for wind and solar energy should be considered in isolation, or if there should be some interdependency.

As far as the last point is concerned, it is clear that the world should not release all the carbon currently stored in its fossil fuels, in particular another 500 giga-tonnes of carbon — on top of the 500 giga-tonnes that have already been emitted — at a 50% chance that the rise in global temperatures will exceed 2 degrees Celsius; see e.g. [2].

Regarding the second question, it is clear that the speed of the receding tariff should depend on how quickly the technology is expected to improve. The idea that costs come down as total installed capacity increases is captured by the 'experience curve'. Indeed, the main driver of feed-in tariffs in Europe has been the expectation that as volumes increase, marginal costs will decline, such that the tariff is an investment in future reduced cost.

1.2 Learning curves

As Moselle [1] has argued, the experience curve is the only viable argument for the extraordinary support of renewables through the feed-in tariff. We argued on page 4 that the speed with which the subsidies should be reduced depends on the expected speed of learning, which is technology specific. The experience curve is therefore crucial, and we examine the concept in some more detail here. Experience curves made their first appearance in the forecasting of aircraft production rates in WWII; see e.g. [3]. In 1972, the Boston Consulting Group applied it to entire industries to analyse competitive positions; see e.g. [4]. In the operations research literature, experience curves have been discussed by [5–10] and many others. Key elements in these analyses are:

1. an expected decline of production costs with cumulative (i.e. total historical) output,
2. an element of *uncertainty* regarding future margins due to stochastically decreasing costs and fixed output prices, or vice versa,
3. the recognition that investment in capacity expansion, driving further cost reduction, is an *option* and needs to be valued as such.

In [8] there are stochastically varying output prices and a deterministic learning curve. This results in an optimal 'bang-bang' solution: at any given time, it is optimal either to produce nothing at all, or to produce at the maximum rate. [6] allow a competitive setting where learning is a strategic choice. [10] explicitly allow for a stochastic learning curve. The optionality is recognised by all. Majd and Pindyck [8], for example, write (p. 332):

> A firm facing a stochastic output price can be thought of as having a set of call options on future production at every instant of time. Each call option has an exercise price equal to the production cost, which in our case decreases with cumulative output. [...] When a firm faces a learning curve, part of its production is an fact an investment expenditure: the firm is investing in reduced future costs. This is an irreversible investment, i.e., the expenditure is sunk.

It is this optionality that makes R&D programmes hard to value. In the literature quoted above, the decision variable is normally the amount (and timing) of capacity growth. This results, for example, in bang-bang solutions. Through the stochastic learning curve, we obtain new (and hopefully lower) marginal costs. In the case of a single decision maker, such as one company or factory, this is perfectly satisfactory. But when the 'decision maker' is given by the community of potential German buyers of solar panels, however, then it is not clear how this 'decision maker' should come to his/her decision regarding capacity expansion. We expect that, at least to some extent, the decision by the German consumer is driven by return on investment, and this leads us to the feedback loop considered in the next subsection.

1.3 Endogenous market growth

The literature on experience curves, almost without exception, considers marginal cost as endogenously determined by the exogenously determined total capacity. But what determines the decision on capacity expansion? *The Economist*, for example, wrote as recently as 26 July 2011[1] that

> ... German home-owners, propelled by generous government subsidies, installed more solar panels on their rooftops in 2010 than the entire planet had managed in the year before. [...] But curiously, this building frenzy coincides with a wave of cuts to solar subsidy schemes across Europe. [...] In fact, the positive and negative trends share a cause — the steep drop in solar photo-voltaic (PV) panel prices. That has left many fixed-price incentive schemes looking absurdly generous, prompting enormous spikes in investment.

[1] http://www.economist.com/node/21524449

In this paper we use the word 'predetermined' rather than 'fixed' tariffs, because the announced German tariffs are, in fact, decreasing. We realise, from the comments above, that capacity expansion not only drives lower cost, but lower cost also drives capacity expansion. A model that takes this endogeneity into account is thus needed, and we present a model for solar learning with endogenous capacity growth in Section 2.

1.4 Feed-in tariff as an optimal stopping problem

Regarding the announcement of feed-in tariffs, governments may seem to be caught in a classic catch-22. If the government announces a relatively high tariff, then it stimulates more investment, causing the cost of solar energy to fall more quickly than expected, and making the tariff seem extraordinarily generous and ill-informed, in hindsight. If, on the other hand, the government announces a tariff that solar energy finds hard to beat, causing both investment and learning to stifle, then this will, indeed, be self-fulfilling. Therefore it may seem that there will inevitably be regret.

To this 1-dimensional problem we apply the machinery of (multidimensional) optimal stopping as developed in Part II of this thesis; methodologically, there is no new content here. While there are many existing ways to solve 1-dimensional optimal stopping problems (see e.g. [11–13]), we feel that this application is new and addresses an important problem.

This paper is organised as follows. Section 2 introduces a stylised model that makes capacity growth endogenous. Section 3 discusses data for the German photo-voltaic industry and estimates how capacity expansion drives learning, and how learning drives capacity expansion. Section 4 formulates an optimal stopping problem and derives a policy under which there is no regret. Section 5 concludes.

2 Technology learning with endogenous capacity growth

In this section we introduce a model for technology learning with endogenous capacity growth; capacity expansion drives cost reductions, and cost reductions drive capacity expansion. Our notation will be as follows:

$f(t) :=$ feed-in tariff, in cents per kWh, at time t. For each consumer this is fixed for 20 years, once a solar panel is bought
$C(t) :=$ cost (to the consumer) of solar energy, in cents per kWh, at time t. For each consumer this is fixed for 20 years, once a solar panel is bought
$Q(t) :=$ total size of the market, in MWp, at time t
$\hat{g}_0, \hat{g}_1 :=$ parameters that determine the speed of *growth* of Q (see below)
$\hat{lr} :=$ parameter that determines the speed of *learning* in C (see below)
$\delta :=$ end-user's discount rate

where f, C and Q depend on the continuous time parameter t, and where stochastic quantities have capital letters. The feed-in tariff f has a lower case because it is announced by the government for the years to come, and thus deterministic. All 'hatted' parameters (growth parameters $\hat{g}_0$, $\hat{g}_1$ and learning rate $\hat{lr}$) are estimated from that data. 'Unhatted' parameters (like the consumer discount rate δ) are assumed to be known. Our simple model thus consists of 1 known parameter, 3 estimated parameters and the dynamics (to be postulated below) between the 3 quantities which depend on time: the deterministic feed-in tariff f (the decision variable), and the stochastically developing $C(t)$ and $Q(t)$.

To explain what we mean by $Q(t)$, we should mention that the capacity of a solar panel is measured in kilo Watt peak (kWp). A solar panel of 1 kWp produces 1 kWh in a sunny hour. When operated in Germany it produces only 8% of that on average, i.e. $365 * 24 * 0.08 \approx 700$kWh per year or $14,000$kWh over 20 years. Future energy production should be discounted by δ. Using that $\sum_{i=1}^{n} \delta^i \approx (\delta^n - 1)/(\log[\delta])$ if $d < 1$, the cost $C(t)$ of solar energy per kWh can be defined as:

$$\begin{aligned} C(t) &:= \frac{\text{price of a solar panel of 1 kWp, at time } t}{\text{sum of discounted future production in kWh}} \\ &= \frac{\text{price of a solar panel of 1 kWp, at time } t}{365 * 24 * 0.08 * (\delta^{20} - 1)/(\log[\delta])} \end{aligned} \tag{2.1}$$

where we will take the consumer discount rate to be $\delta = 0.97$. As for the dynamics between $f(t)$, $C(t)$ and $Q(t)$, we propose the following model:

1. The relative cost reduction over time, as measured by $\log\big[C(t)/C(0)\big]$, is proportional to the relative market growth over that same period, as measured by $\log\big[Q(t)/Q(0)\big]$,

plus noise. In particular we propose

$$\log\left[\frac{C(t)}{C(0)}\right] = -\hat{lr}\,\log\left[\frac{Q(t)}{Q(0)}\right] + \sigma\,B(t),$$

where $B(t)$ is a standard Brownian motion and σ its volatility. In this model, the volatility σ is just a parameter, but later it will be estimated, and thus hatted. It can be re-written to read:

$$\frac{C(t)}{C(0)} = \left(\frac{Q(t)}{Q(0)}\right)^{-\hat{lr}} \exp\left[\sigma\,B(t)\right]. \tag{2.2}$$

This is the classic 'learning curve' or 'experience curve' that is very common in the operations research literature; see e.g. [5–10] and others. In the empirical literature, there is both plenty of evidence and criticism regarding the learning curve, see e.g. [14–16]. We adopt the learning curve as given, because it is analytically practical and because it is known, generally, to explain cost data quite accurately.

2. We assume that the market grows with 'base' growth rate $\hat{g}_0$, but that this rate increases if $f(t)/C(t)$ is larger than 1. Specifically, we propose:

$$\frac{dQ(t)}{Q(t)} = \hat{g}_0\,dt + \log\left[\frac{f(t)}{C(t)}\right]\hat{g}_1\,dt \approx \hat{g}_0\,dt + \left(\frac{f(t)}{C(t)} - 1\right)\hat{g}_1\,dt \tag{2.3}$$

where the approximation holds when $f(t)/C(t)$ is close to 1, and where $\hat{g}_1$ measures the strength of the feedback effect. Learning accelerates if $\hat{g}_1$ is larger than zero and if $f(t)$ exceeds $C(t)$[2]. In the next section, the parameters $\hat{g}_0$ and $\hat{g}_1$ are estimated to be 0.44 and 1.13. If $f(t)/C(t)$ is much larger than 1, the market still responds to increases in $f(t)/C(t)$, but slower than linearly. This could be explained by assuming that extremely high growth rates are not physically possible.

While assumption 1, the experience curve, is widely debated, documented and used, the second assumption is new and crucial for the feedback loop that we intend. Let us investigate assumption 2 in more detail:

- Suppose that the feed-in tariff $f(t)$ exceeds the cost to the consumer $C(t)$, such that capacity expansion is profitable. People invest and the installed capacity $Q(t)$ goes up. Through the learning curve, $C(t)$ is expected to fall in the next period. If the decrease in $C(t)$ is larger than the decrease in $f(t)$ — in percentage terms — then the ratio $f(t)/C(t)$ increases. Profitability for consumers goes up as the 'rate of return' $\log\left[\frac{f(t)}{C(t)}\right]$ increases. Thus, by assumption 2, the speed of market growth goes up, i.e. $Q(t)$ accelerates, driving further (and faster) cost reductions.

[2] If $\hat{g}_1$ were negative it would drive reversion towards the moving curve $f(t)$.

- If capacity expansion reduces the cost $C(t)$, but if $f(t)$ falls *faster* than $C(t)$ — in percentage terms — then the ratio $f(t)/C(t)$ decreases. Profitability for the consumer decreases and capacity expansion slows down. Capacity $Q(t)$ decelerates, driving further (but slower) cost reductions. If we think the base growth rate $\hat{g}_0$ is zero, then capacity growth is *only* caused by profitability; by $f(t) > C(t)$. In that case, when $C(t)$ can no longer beat the feed-in tariff, the capacity $Q(t)$ decelerates further such that $Q(t)$ comes to a halt exactly when $f(t) = C(t)$. (If we take the base growth rate $\hat{g}_0 \geq 0$, then the market would keep growing regardless, and thus the model cannot be trusted for $C(t)$ exceeding $f(t)$.)

In other words, $\log\left[\frac{f(t)}{C(t)}\right]$ may be viewed as the return on investment (RoI) to the consumer of solar energy. The rate of investment goes up (down) as the RoI goes up (down). When the feed-in tariff $f(t)$ is only a little bit larger than the cost to the consumer $C(t)$, then $\log\left[\frac{f(t)}{C(t)}\right]\hat{g}_1\,dt \approx \left(\frac{f(t)}{C(t)} - 1\right)\hat{g}_1\,dt$. Thus for small RoI, the speed of capacity expansion (as measured by $dQ(t)/Q(t)$) increases linearly in the RoI. For larger RoI, the speed of capacity expansion increases slower than linearly in the RoI. This could be explained by assuming that extraordinarily large growth rates are not physically possible, even if they are economically desirable. We choose this model because it seems that it is

- not unreasonable — i.e. one would expect that investment in capacity expansion is at least partly driven by the return on investment
- broadly consistent with the data on e.g. German photo-voltaic market growth, as will be shown in the next section
- analytically tractable — and this is the first model with the desired properties that has this property, as discussed below
- consistent with a known vanilla model in the limit where $\hat{g}_1 \to 0$; such that $Q(t)$ grows exponentially at the base rate $\hat{g}_0$ and $C(t)$ follows a geometric Brownian motion with drift $-\hat{lr}\,\hat{g}_0$.

Models with self-reinforcing effects are related to models with mean-reversion. While mean-reversion drives the process back to the equilibrium, a run-away effect could be obtained by driving the process away from what used to be the equilibrium position. Therefore, a mean-reverting model with a *negative* parameter gives rise to situation with a repulsive force away from (what used to be) the equilibrium position.

Mean-reverting models have been discussed extensively in the financial literature, mainly to model interest rates. But many mean reverting models allow negative interest rates, while others are not analytically solvable. The Vasicek interest rate model can be expressed in closed form, but it mean-reverts to a constant level and may take negative

values. The Cox-Ingersoll-Ross interest rate model, on the other hand, has the attractive feature that its interest rate is always positive, but it cannot be solved analytically. We refer to the overview paper by [17] for an overview of mean-reverting processes in finance.

We are unaware of an(other) analytical model for a process that 1) allows us to specify an attractive force to (or repulsive force away from) a given time-dependent curve, while 2) enforcing that the process stays positive, and 3) is analytically solvable. Therefore, we propose such a model here:

Theorem 1. Technology learning with endogenous capacity growth. *We assume that technology learning and market growth are specified by*

$$\begin{aligned} C(t) &= C(0) \left(\frac{Q(t)}{Q(0)} \right)^{-\hat{lr}} \exp\left[\sigma\, B(t) \right], \\ \frac{dQ(t)}{Q(t)} &= \hat{g}_0\, dt + \log\left[\frac{f(t)}{C(t)} \right] \hat{g}_1\, dt, \end{aligned} \tag{2.4}$$

where $B(t)$ is a standard Brownian motion of unit variance. Then the pair $C(t)$ and $Q(t)$ can be provided in closed form as follows:

$$\begin{aligned} C(t) &= \exp\left[\log\left[C(0)\right] e^{\gamma_1 t} - e^{\gamma_1 t} \int_0^t e^{-\gamma_1 \tau} \left(\gamma_1 \log\left[f(\tau) \right] + \gamma_0 \right) d\tau + \sigma\, e^{\gamma_1 t} \int_0^t e^{-\gamma_1 \tau} dB(\tau) \right], \\ Q(t) &= Q(0) \exp\left[\hat{g}_0\, t + \hat{g}_1 \int_0^t \log\left[\frac{f(\tau)}{C(\tau)} \right] d\tau \right]. \end{aligned} \tag{2.5}$$

The parameters γ_0 and γ_1 are defined as $\gamma_0 := \hat{lr}\, \hat{g}_0$ and $\gamma_1 := \hat{lr}\, \hat{g}_1$. Further, $f(t)$ is a given function of time; the feed-in tariff in this case.

When we let $\hat{g}_1 \searrow 0$, the feedback loop is turned off and we return to a vanilla model regarding market growth. In that case, we obtain

$$\begin{aligned} C(t) &= C(0) \exp\left[-\gamma_0\, t + \sigma\, B_t \right], \\ Q(t) &= Q(0) \exp\left[\hat{g}_0\, t \right]. \end{aligned} \tag{2.6}$$

We see that in this limit, the capacity $Q(t)$ grows exponentially and the cost $C(t)$ follows a downward sloping geometric Brownian motion. Finding the optimal tariff even for this most vanilla market growth model, however, is a non-trivial exercise. We will discuss this problem in Section 4. In the next section, we will apply this model to the German photovoltaic market and estimate the required parameters. The proof of Theorem 1 follows here:

Proof. Consider the following sequence of equalities:

$$\begin{aligned}
\frac{dQ(t)}{Q(t)} &= \hat{g}_0\, dt + \hat{g}_1\, \log\left[\frac{f(t)}{C(t)}\right]\, dt, \\
d\log\left[Q(t)\right] &= \hat{g}_0\, dt + \hat{g}_1\, \log\left[\frac{f(t)}{C(t)}\right]\, dt, \\
\log\left[\frac{Q(t)}{Q(0)}\right] &= \hat{g}_0\, t + \hat{g}_1 \int_0^t \log\left[\frac{f(\tau)}{C(\tau)}\right]\, d\tau, \\
\frac{Q(t)}{Q(0)} &= \exp\left[\hat{g}_0\, t + \hat{g}_1 \int_0^t \log\left[\frac{f(\tau)}{C(\tau)}\right]\, d\tau\right].
\end{aligned}$$

Substitute this into the expression

$$C(t) = C(0)\left(\frac{Q(t)}{Q(0)}\right)^{-\hat{lr}} \exp\left[\sigma\, B(t)\right]$$

to obtain

$$\begin{aligned}
C(t) &= C(0)\, \exp\left[- \hat{lr}\hat{g}_0\, t - \hat{lr}\hat{g}_1 \int_0^t \log\left[\frac{f(\tau)}{C(\tau)}\right]\, d\tau\right] \exp\left[\sigma\, B(t)\right] \\
C(t) &= C(0)\, \exp\left[- \gamma_0\, t - \gamma_1 \int_0^t \log\left[\frac{f(\tau)}{C(\tau)}\right]\, d\tau + \sigma\, B(t)\right],
\end{aligned}$$

where $\gamma_1 := \hat{lr}\, \hat{g}_1$ and $\gamma_0 := \hat{lr}\, \hat{g}_0$. Rewriting, we get

$$\log\left[\frac{C(t)}{C(0)}\right] = -\gamma_0\, t - \gamma_1 \int_0^t \log\left[\frac{f(\tau)}{C(\tau)}\right]\, d\tau + \sigma\, B(t).$$

Differentiate this expression to obtain

$$\frac{dC(t)}{C(t)} = -\gamma_0\, dt - \gamma_1\, \log\left[\frac{f(t)}{C(t)}\right]\, dt + \sigma\, dB(t).$$

We propose that this recursive equation is satisfied by

$$C(t) = \exp\left[\log\left[C(0)\right] e^{\gamma_1\, t} - e^{\gamma_1\, t} \int_0^t e^{-\gamma_1\, \tau}\left(\gamma_1\, \log\left[f(\tau)\right] + \gamma_0\right) d\tau + \sigma\, e^{\gamma_1\, t} \int_0^t e^{-\gamma_1\, \tau} dB(\tau)\right],$$

which can be verified by calculating $dC(t)$:

$$dC(t) = C(0)\left[\gamma_1 \log\left[C(t)\right]\, dt - \left(\gamma_1 \log\left[f(t)\right] + \gamma_0\right)\, dt + \sigma dB(t)\right].$$

The result follows. □

3 The German market for photo-voltaics

Although it allows for endogenous capacity growth, the proposed model in the previous section is stylised. For an application to the German photo-voltaic market, it requires us to estimate only 4 parameters: the learning rate $\hat{lr}$, the growth parameters $\hat{g}_0$, $\hat{g}_1$ and the volatility $\hat{\sigma}^2$. Germany has operated a large-scale feed-in policy since 2004, and our data set is consequently limited:

Year	FiT (c/kWh)
2004	57.40
2005	54.53
2006	51.80
2007	49.21
2008	46.75
2009	43.01
2010	35.40
2011	28.74

Year	Cost (c/kWh)
2004	50.49
2005	53.35
2006	47.64
2007	43.15
2008	40.93
2009	31.01
2010	26.2
2011	23.82

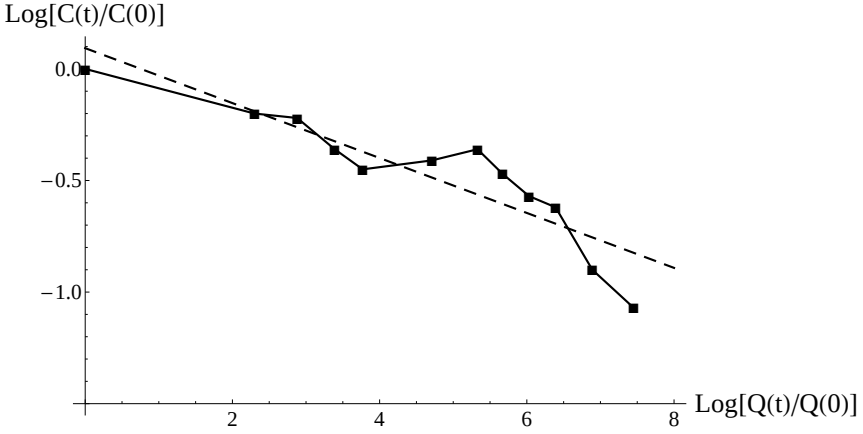
Year	Capacity (MWp)
2004	1105
2005	2056
2006	2899
2007	4170
2008	5979
2009	9785
2010	17 193

where c/kWh stands for € cents/kWh. These data are publicly available.[3] The cost of solar energy per kWh is calculated as in (2.1) and with $\delta = 0.97$. On the basis of this (limited) set of data, we estimate the parameters in the following model:

$$C(t) = C(0)\left(\frac{Q(t)}{Q(0)}\right)^{-\hat{lr}} \exp\left[\hat{\sigma}\, B(t)\right],$$
$$\frac{dQ(t)}{Q(t)} = \hat{g}_0\, dt + \log\left[\frac{f(t)}{C(t)}\right]\, \hat{g}_1\, dt.$$

First, for the learning rate we find

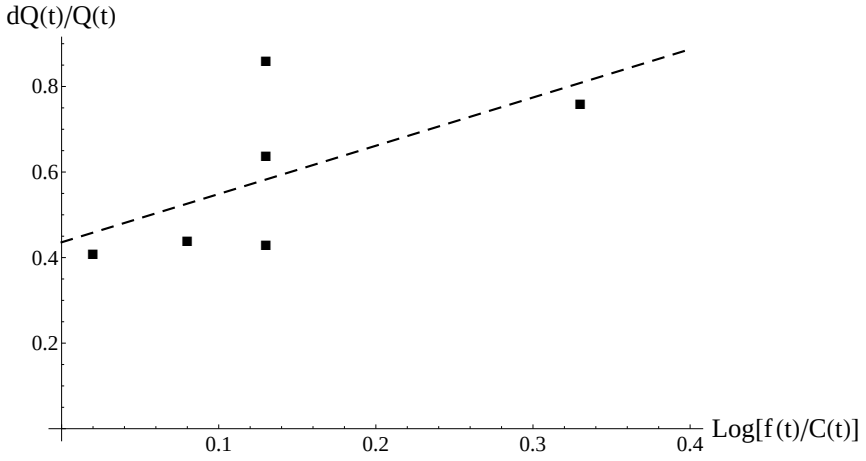
	Estimate	Standard Error	t Statistic	P-Value
1	0.0929534	0.0925845	1.00398	0.339061
lr	−0.12307	0.0184494	−6.67065	0.0000556417



[3] The data are taken from http://www.solarwirtschaft.de/.

and the slope $\hat{lr} = -0.12$ is highly significant (standard error = 0.02). The variance is estimated as 0.0176, such that we have $\hat{\sigma} = 0.133$. As far as the market growth is concerned, the feed-in tariff $f(t)$ has only been effective since 2004. We wish to explain the market growth based on the ratio $f(t)/C(t)$, and therefore our estimates $\hat{g}_0$, $\hat{g}_1$ will be based on only six data points, as follows:

	Estimate	Standard Error	t Statistic	P-Value
1	0.43581	0.121476	3.58761	0.0230109
g	1.12822	0.729442	1.54669	0.196842



Of course, no significance can be expected. There may seem to be a lot of uncertainty regarding the value $\hat{g}_0$ and $\hat{g}_1$, but simply declaring $\hat{g}_0 = 0.44$, $\hat{g}_1 = 1.13$ seems to result in good estimates, as we will discuss below. Concluding, we have that:

parameters	learning rate $\hat{lr}$	growth parameter $\hat{g}_0$	growth parameter $\hat{g}_1$	std. dev. $\hat{\sigma}$
estimated value	0.12	0.44	1.13	0.0177
standard error	0.02	0.12	0.73	

where the variance is estimated by the mean square error, and where we have added historic data going back to 1995 to estimate the learning rate $\hat{lr}$. This results in a highly significant estimate. The estimates of $\hat{g}_0$ and $\hat{g}_1$ are based on six data points, because the feed-in tariff has only been effective since 2004, and no significance can be expected. Therefore, we proceed with care but we note that:

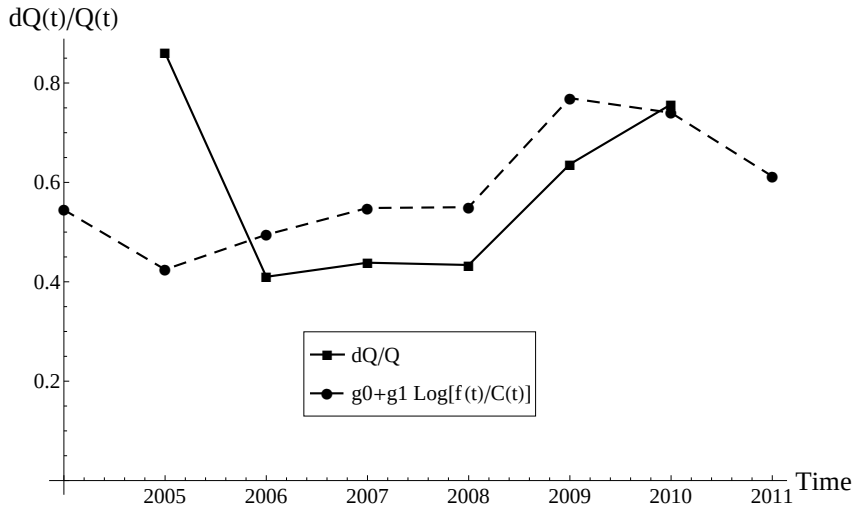
1. even if the estimates of $\hat{g}_i$ are uncertain, the model still generates qualitative insight.

2. simply declaring $\hat{g}_0 = 0.44$, $\hat{g}_1 = 1.13$ results in relatively good estimates, at least for historic market growth, as we will show below.

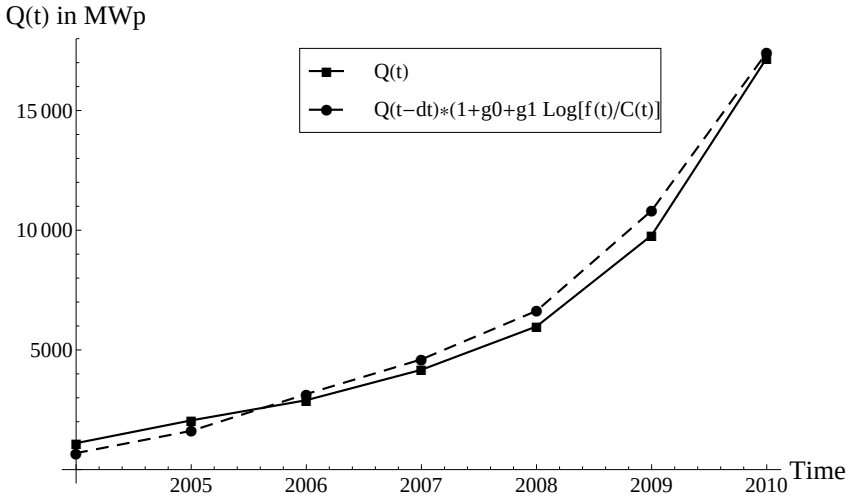
For example, we may predict dQ/Q on the basis of the tariff $f(t)$ and cost $C(t)$ by the following equation

$$\frac{dQ(t)}{Q(t)} = \hat{g}_0\, dt + \log\left[\frac{f(t)}{C(t)}\right]\, \hat{g}_1\, dt.$$

Actual growth rates dQ/Q have ranged wildly; from 41% to 86% between 2005 and 2010. We compare the predicted and actual growth rates in the following graph:



Apart from 2005, this is not so bad. It seems that market growth is indeed affected by return on investment for the consumer. Thus next year's capacity can be predicted with this year's return on investment as follows:

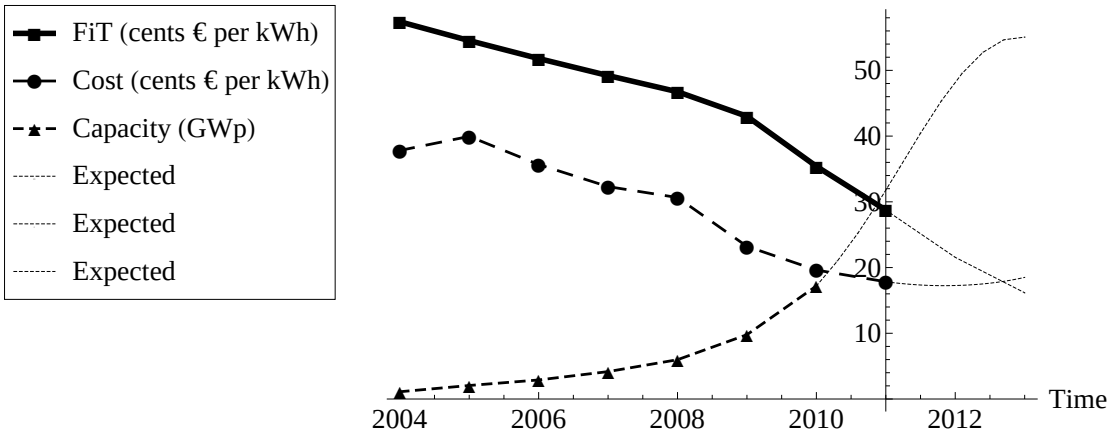


We see that even though there is uncertainty around the values of $\hat{g}_i$, taking $\hat{g_0} = 0.44$, $\hat{g_1} = 1.13$ results in relatively good predictions, at least of *past* capacity expansions.

We are mainly interested in developing a new *model* for capacity expansion $dQ(t)/Q(t)$, and a new *methodology* for deciding on the feed-in tariff $f(t)$. For convenience, therefore, all estimated parameters will be assumed to be fixed and known, from this point onwards. But it should be noted that the estimates may be wrong, and therefore the policy conclusions are

not necessarily robust. More work will have to be done elsewhere, either on 1) estimating the parameters better, or on 2) including Bayesian learning in the model, regarding the estimated parameters. Since the exact numerical value of the parameters is not relevant for the methodology proposed here, we will take the parameters as known and fixed at their current estimates, and see where the model leads us. The inevitable consequence is that policy implications will have to be taken with a grain of salt, even if one does believe the dynamics proposed by the model.

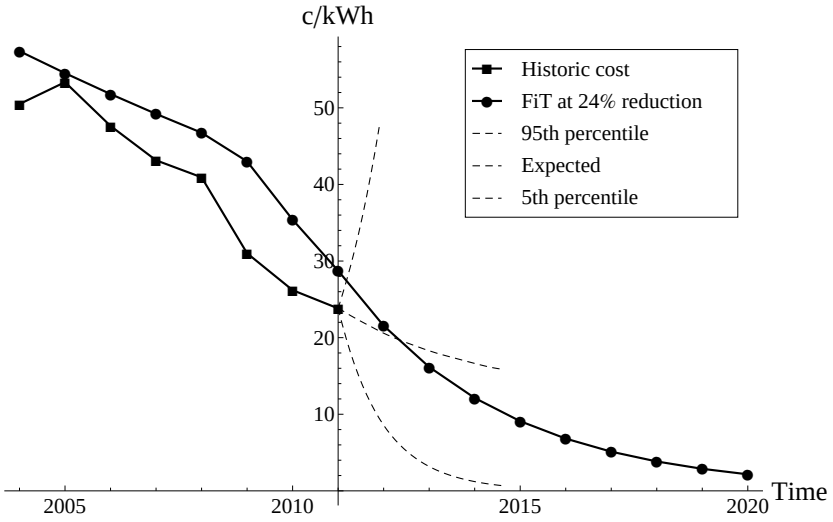
The German government has announced that if the market keeps growing at its current rate (i.e. adding over 7,500 MWp p.a.), then the feed-in tariff will be reduced by 24% p.a. instead of the previously announced 9%. When the feed-in tariff $f(t)$ is given, then we may use the set of equations in Theorem 1 to predict the path of future cost reductions and capacity expansions. With the parameters as estimated, and with a 25% reduction p.a. in the feed-in tariff, the capacity $Q(t)$ and cost $C(t)$ are predicted to develop as follows:



where the three dotted lines indicate the expected development of $C(t)$, $Q(t)$ and $f(t)$. We note that the capacity growth is expected to halt around 2013. This S-shaped curve is one of the main attractive features of the model in Theorem 1. If, for example, one were to use a constant growth rate, then it would be unclear what growth rate to choose, since the growth rate has varied between 41% and 86% in the last 6 years alone. And even if one were able to decide on a growth rate, then it would be clear that the suggested growth cannot be sustained indefinitely — due to both physical and economic constraints. Our model sidesteps this issue by internalising the growth rate, and we take the economic incentives (but not the physical constraints) into account. However, it is expected that in the short to medium term, the economic factors will be of greater significance as German feed-in tariffs may drop by as much as 24%. The market will still grow, but — in percentage terms — not as quickly as the tariff will fall. As a result, both investment and learning will be stifled. When the cost can no longer beat the tariff, $\log\left[f(t)/C(t)\right]$ is negative and, therefore, $\hat{g}_1$

will drive a mean reverting force rather than a run-away force. Rather than away from the tariff $f(t)$, costs will be driven towards it, but it is clear that the model cannot be trusted for $f(t) < C(t)$.

Although the cost is expected to meet the tariff in 2012, given the expected 24% reductions, it is not certain that this will, in fact, happen. The model of Theorem 1 incorporates uncertainty, and we may plot the expected progress of $C(t)$, along with its 5^{th} and 95^{th} percentiles, as follows:



where $f(t)$ is taken as the input, and where $C(t)$ is an output of the model. From the current year (2011), three dotted lines are drawn: the middle one is the expected path, and the outer ones are the paths in the best and worst 5% of cases (i.e. the 5^{th} and 95^{th} percentiles). It is clear that the *variance* in possible outcomes is huge. The feedback allows the technology to obtain critical mass and beat the tariff by a long way (the best 5% of cases), or crash into the tariff relatively soon if it is unable to gain and sustain an early lead (worst 5% of cases). In terms of the sample paths generated by the model, we note that

- paths either gain an early cost advantage, and sustain this lead to move further away from the tariff (such as in the best 5% of cases), or
- paths obtain an early setback and do not develop enough critical mass, and meet the tariff quickly (such as in the worst 5% of cases). After the paths of $C(t)$ meet the tariff $f(t)$, the model can no longer be trusted.

Focusing on the middle dotted line, we see that, on expectation, the tariff falls quicker than the cost, and investment slows down and comes to a halt as cost and tariff meet in 2012. Although this is what happens on expectation, the feedback look in the process

makes it hard to predict what will happen. Because of the runaway effect that may occur in either direction, extremes are more likely than one would initially think. This feature of the model may be worth noting in itself. Also, this bandwidth of outcomes may be more realistic than that given by e.g. geometric Brownian motion (i.e. $\hat{g}_1 \searrow 0$), which is relatively narrow. One disadvantage of the experience curve, as it is generally formulated, is that it does not formulate a minimum cost — and, in the model, the paths in the best 5% of cases exploit this caveat to the full extent, reducing the cost of photovoltaics to near-zero levels (as can be seen in the graph above). But, one could argue, this is still better than e.g. the Ornstein-Uhlenbeck model, which would allow the run-away process to take negative values. We could adopt a framework with a minimum cost without major alterations, if its estimate was informed by physical considerations, independent of observed cost data.

We conclude that the model of technology learning with endogenous capacity growth, as developed in the previous section, is to some extent applicable to the German photovoltaic market, although 1) we cannot determine the growth parameters $\hat{g}_i$ with any confidence, and 2) making predictions is extremely hard, given the intrinsic structure of the proposed stochastic process.

4 Technology learning as an OS problem

In this section we will again consider the model of technology learning, and taking a government's point of view. The government announces the feed-in tariff for the years to come, and both $C(t)$ and $Q(t)$ will develop as by the endogenous growth model. If the cost $C(t)$ manages to beat the tariff $f(t)$ for all time, then growth continues. If, however, the tariff $f(t)$ recedes faster than the capacity grows, in percentage terms, then growth slows down and may come to a halt. In this section we will present our own (admittedly stylised) model of solar learning. Our model is is indicated in Figure 1:

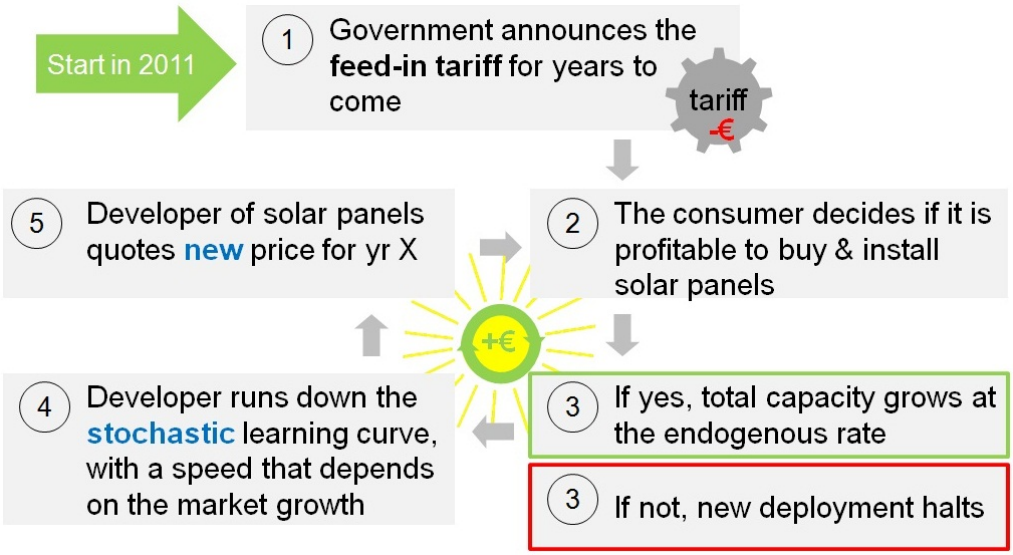


Figure 1. A simple model for solar learning.

Regarding technology learning with endogenous growth, governments are caught in a classic catch-22. If the government announces a relatively high tariff, then it stimulates more investment, causing the cost of solar energy to fall more quickly than expected, and making the tariff seem extraordinarily generous and ill-informed, in hindsight. There is some evidence that this happened in Europe, as discussed in subsection 1.3. If, on the other hand, the government announces a tariff that solar energy finds hard to beat, stifling both investment and learning, then this will, indeed, be self-fulfilling. There are some indications that this backlash may occur in the future, judging from our predictions in Section 3. Therefore, it may seem that there will inevitably be regret.

Here we propose that the problem may be formulated as an *optimal stopping* (OS) problem or *optimal control* (OC) problem — depending on one's view of the world:

- If we believe that the market for solar panels will grow at the rate $\hat{g}_0$, irrespective of economic incentives, then we believe that $\hat{g}_1$ is really zero, i.e. no feedback loop.

In this world, the market grows at the exogenously determined rate $\hat{g}_0$, which could stem e.g. from concerns about climate change. Consumers will invest at the rate $\hat{g}_0$, as long as it is profitable to do so. But they will not invest more (less) if it is more (less) profitable to do so. In this world, the government does not influence the market growth $Q(t)$ by announcing $f(t)$ directly: it grows at the rate $\hat{g}_0$ until $C(t)$ can no longer beat $f(t)$. Thus, although the government does not influence the process $Q(t)$ directly, it does influence when it *stops*: namely when it is no longer profitable for the market to invest. In this case, deciding on the feed-in tariff $f(t)$ is a problem of *optimal stopping*. Regarding question 3 as posed on page 4, the feed-in tariff $f(t)$ in this world should be predetermined (and not performance dependent).

- If we believe that the market growth depends *only* on return on investment, then we should set $\hat{g}_0 = 0$ and let $\hat{g}_1 > 0$. In this world, the announced feed-in tariff $f(t)$ influences the growth of $Q(t)$ directly: higher return on investment leads to higher growth. No return on investment leads to zero growth. Of course tariff $f(t)$ also determines the moment when the market stops investing. Because the feed-in tariff $f(t)$ now influences the stochastic process directly, as well as when the market stops investing, deciding on $f(t)$ has become a problem of *optimal control*. Regarding question 3 as posed on page 4, the feed-in tariff $f(t)$ in this world should be performance-dependent (and not predetermined).

Common to both sets of problems is that the government pays the feed-in tariff with the idea that it invests in possible future savings. While those savings may or may not be realised, by choosing an appropriate receding tariff, society should at least explore the *option* of those savings — and preferably for several technologies in parallel. We have established that the announced feed-in tariff influences the cost to society in two ways:

1. First, it influences the amount of investment and learning (but only if $\hat{g}_1 > 0$).

2. Second, it determines when the market stops investing in the option of solar energy.

New about this paper is that it has made capacity expansion endogenous. With endogenous capacity growth (i.e. $\hat{g}_1 > 0$), however, the resulting problem regarding the determination of $f(t)$ is one of optimal control, and not of optimal stopping. Unfortunately, the methods developed in Part II of this thesis relate only to optimal stopping. Therefore, we will henceforth set $\hat{g}_1 = 0$ and solve the resulting optimal stopping problem. We hope to solve the resulting optimal control problem in a separate paper.

Regarding the optimal stopping problem, it is clear that savings are realised when the new technology beats the market price that is set by the incumbent technology. What about the cost? The bill of the feed-in programme in Germany is paid by the consumers of

electricity: a small amount is added to each electricity bill. Not everyone has a solar panel on their roof, but, in sum, the cost of the programme is spread over the same population that receives its handouts. There is a redistribution of wealth from those who have not invested in solar panels to those who have, but the net total is zero as far as the whole society is concerned. There have been private investments in solar panels by consumers, but we do not take these into account. Generally these are profitable investments, given the the redistribution of wealth that follows. The redistribution itself, however, is a zero-sum game.

The *real* cost to society, therefore, derives from the fact that electricity is produced in an expensive way. This is a learning investment, and it is sunk once it has been made. Presumably the tariff $f(t)$ exceeds the cost of solar energy $C(t)$, which exceeds the cost of the incumbent technology that sets the market price $p(t)$; i.e. $f(t) \geq C(t) \geq p(t)$. The price of the mature incumbent technology is assumed known, and hence the small letter p. It may increase due to foreseeable taxation, for example, and hence the time-dependence. The difference $f(t) - C(t)$ presents no real cost to society: this difference is paid to those with solar panels, by all electricity users. This redistribution rewards those who have made private investments in solar panels and does not represent a real cost to society. But the difference $C(t) - p(t)$ does present a real cost: this difference is a learning investment which, once made, is immediately sunk. We can consider this expenditure as an investment, because it may lead to reduced future costs.

Supposing that $dQ(t)$ panels are installed in year t, then for 20 years the feed-in tariff $f(t)$ will be paid for electricity generated by those units. Therefore, the total cost (over 20 years) associated with the capacity expansion $dQ(t)$ is as follows:

$$\text{total amount paid by government per } dt = \sum_{i=1}^{20} \delta^i \,(365 * 24 * 0.08)\big(f(t) - p(t)\big)\frac{dQ(t)}{dt}$$

$$\text{profit made by consumers per } dt = \sum_{i=1}^{20} \delta^i \,(365 * 24 * 0.08)\big(f(t) - C(t)\big)\frac{dQ(t)}{dt}$$

$$\text{true cost to society per } dt = \sum_{i=1}^{20} \delta^i \,(365 * 24 * 0.08)\big(C(t) - p(t)\big)\frac{dQ(t)}{dt}$$

where δ is the discount-rate. For brevity, we can rewrite this as

$$\text{true cost to society per } dt = \alpha\big(C(t) - p(t)\big)\frac{dQ(t)}{dt}$$

where $\alpha := (\delta^n - 1)/(\log[\delta])(365 * 24 * 0.08)$ and $n = 20$. In Part II of this thesis we introduced the 'continuation gain' G_C for optimal stopping problems as follows: it was the amount to be paid/received during time dt when the process was not stopped. For this particular problem, we obtain

$$\text{continuation gain} = G_C = \alpha\big(p(t) - C(t)\big)\frac{dQ(t)}{dt}.$$

If $C(t)$ is higher than $p(t)$, then it is costly to run the programme and the continuation gain G_C will be negative. But its running does induce future cost reductions. Given the current tariff, one may ask whether it is in the interest of society that the exploration of the solar option continues. If the answer is 'yes' and $C(t) < f(t)$, then there is no problem. If the answer is 'no' and $C(t) > f(t)$, then this, too, is unproblematic. But if it is socially desirable that the exploration of the option should continue, while $C(t) > f(t)$, then a problem arises. The market only grows when $C(t) < f(t)$, and therefore we should choose a policy that is consistent with that desire. The optimal stopping problem can be formulated as follows:

$$V = \max_{s\le\tau\le T} \mathbb{E} \int_s^\tau \alpha\Big(p(t) - C(t)\Big)\, e^{-r(t-s)}\, \frac{dQ(t)}{dt}\, dt \tag{4.1}$$

where the maximisation is over all stopping times τ and where the stochastic process $C(t)$ is given by Theorem 1 with $\gamma_1 = 0$. We can indicate the dependence on the initial state $B(s) = x$ and time s to write the problem as

$$\begin{aligned}
V(x,s) &= \max_{s\le\tau\le T} \mathbb{E}_{(x,s)} \int_s^\tau \alpha\bigg(p(t) - C(s)\exp\bigg[-\int_s^t \gamma_0\, d\tau + \sigma\int_s^t dB(\tau)\bigg]\bigg)\, e^{-r(t-s)}\, \frac{dQ(t)}{dt}\, dt \\
&= \max_{s\le\tau\le T} \mathbb{E}_{(x,s)} \int_s^\tau \alpha\bigg(p(t) - C(s)\exp\Big[-\gamma_0\,(t-s) + \sigma(B_t - B_s)\Big]\bigg)\, e^{-r(t-s)}\, \frac{dQ(t)}{dt}\, dt \\
&= \max_{s\le\tau\le T} \mathbb{E}_{(x,s)} \int_s^\tau \alpha\bigg(p(t) - C(s)\exp\Big[-\gamma_0\,(t-s) + \sigma(B_t - B_s)\Big]\bigg)\, e^{-r(t-s)}\, \hat{g}_0 Q(t)\, dt \\
&= \max_{s\le\tau\le T} \mathbb{E}_{(x,s)} \int_s^\tau \alpha\bigg(p(t) - C(s)\exp\Big[-\gamma_0\,(t-s) + \sigma(B_t - x)\Big]\bigg)\, e^{-r(t-s)}\, \hat{g}_0 Q(s) e^{\hat{g}_0\,(t-s)}\, dt \\
&= \hat{g}_0\, Q(s) \max_{s\le\tau\le T} \mathbb{E}_{(x,s)} \int_s^\tau \alpha\Big(p(t)e^{\hat{g}_0(t-s)} - C(s)e^{-(\gamma_0-\hat{g}_0)(t-s)+\sigma(B_t-x)}\Big)\, e^{-r(t-s)}\, dt.
\end{aligned} \tag{4.2}$$

This problem cannot necessarily be solved for an infinite time-horizon. If the growth-rate $\hat{g}_0$ exceeds the interest rate r, then infinite gains can be obtained. But for finite maturity the problem is always solvable. From Part II of this thesis, we know that the value V of an optimal stopping problem, where only the 'continuation gain' G_C is non-zero, must satisfy the following set of conditions:

$$\begin{aligned}
&\text{The value is unbiased} && \big(\tfrac{1}{2}\nabla_x^2 + \frac{\partial}{\partial s} - r\big)V(x,s) = -G_C(x,s) && x \in D(s), \\
&\text{Value-matching condition} && V(\beta,s) = 0 && \beta \in \partial D(s), \\
&\text{Smooth-pasting condition} && \partial_\beta V(\beta,s) = 0 && \beta \in \partial D(s), \\
&\text{Value at maturity} && V(x,T) = 0 && x \in D(T).
\end{aligned} \tag{4.3}$$

Here, the dynamic domain and its boundary at any particular time s are indicated by $D(s)$ and $\partial D(s)$, and we have shown that that the optimal value V is given by:

$$V(x,s) = \int_s^T d\tau \int_{D(\tau)} d\alpha\; G_C(\alpha,\tau) e^{-r(\tau-s)}\, B(\alpha,\tau|x,s), \tag{4.4}$$

where $D(\cdot)$ is the optimal domain, and where $G_C(\alpha, \tau)$ is given by

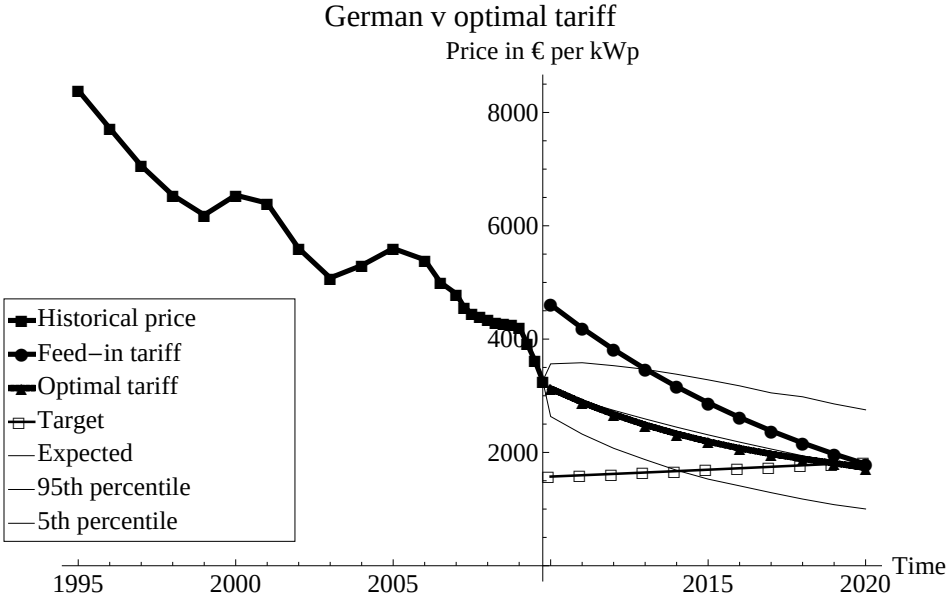
$$G_C(x,t) = \hat{g}_0\, Q(s)\, \alpha \left(p(t) e^{\hat{g}_0 (t-s)} - C(s) e^{-(\gamma_0 - \hat{g}_0)(t-s) + \sigma(B_t - x)} \right). \tag{4.5}$$

At time equal to s, the gain of running the programme is

$$G_C(x,0) = \hat{g}_0\, Q(s)\, \alpha \left(p(s) - C(s) \right). \tag{4.6}$$

If the cost of solar $C(s)$ exceeds the incumbent price $p(s)$, then it is costly to run the programme. Thus G_C is negative. The cost, per unit of time, equals the newly installed quantity, which is $\hat{g}_0\, Q(s)$, times the cost to pay for that capacity for 20 years, which is $\alpha(C(s) - p(s))$.

The optimal domain can now be found by applying either value-matching or smooth-pasting to the optimal value, at all boundary coordinates $\beta \in \partial D(s), \forall s \leq T$. For x on the optimal boundary, the expected value of all continuation gains, as collected by the free Brownian path during its time in the optimal continuation domain $D(\cdot)$, equals zero. And this holds true for all boundary locations $x \in \partial D(\cdot)$. When we apply this theory to the problem in this section, the optimal tariff is as follows:



where it has been assumed that $T = 2020$, $\hat{g}_1 = 0$, i.e. no endogeneity, and $\delta = 0.97$. With the feed-in tariff as given, there is no regret: it is socially optimal that the market should invest in solar energy if it can beat the tariff, and it is socially optimal that it should stop once it can no longer beat the tariff. The 'target' $p(t)$ which we are trying to beat is set by industrial electricity prices, with a yearly increase of 2%.

From the graph we see that the optimal feed-in tariff, as measured in the total amount of subsidy on a solar panel of 1kWp, discounted to net present value, is only slightly higher

than the current price of a solar panel, making it still a profitable investment (but only just). If solar energy can no longer beat the tariff, then the tariff should not be reconsidered; the tariff was chosen such that the market will automatically make the decision that is socially optimal.

5 Conclusion

The first contribution of this paper is the finding of an optimal feed-in tariff for German photo-voltaics. The trade-off concerning feed-in tariffs is that rapidly diminishing levels of subsidy are cheaper to the taxpayer, but could equally kill an otherwise successful technology prematurely. Sustaining high levels of subsidy, on the other hand, is not only expensive but also keeps alive the worst-case scenario of a technology being supported indefinitely without ever becoming economical. By formulating the trade-off as an optimal stopping problem, we are able to find a policy under which the market automatically makes the socially optimal decisions. We have used the methods of (multidimensional) optimal stopping, as developed in Part II of this thesis, to solve a 1-dimensional problem. Methodologically, nothing new was done.

A second contribution of this paper is that it provides a new model for technology learning with endogenous market growth. While other analyses of learning curves have taken capacity expansion to be an exogenous variable, we recognise that new investment drives cost reductions, while cost reductions, in their turn, drive new investment. This adds to the literature on mean-reverting models such as the Vasicek interest rate model, as an example of an Ornstein-Uhlenbeck process, or the Cox-Ingersoll-Ross interest rate model. We thus fill a gap in the literature by providing the closed-form solution of a stochastic process with feedback that stays positive and can be driven either towards or away from a given time-dependent curve.

References

[1] B. Moselle, J. Padilla, and R. Schmalensee, *Harnessing Renewable Energy in Electric Power Systems: Theory, Practice, Policy*. RFF Press, 2010. 3, 5

[2] M. Allen, D. Frame, C. Huntingford, C. Jones, J. Lowe, M. Meinshausen, and N. Meinshausen, *Warming caused by cumulative carbon emissions towards the trillionth tonne*, *Nature* **458** (2009), no. 7242 1163–1166. 5

[3] L. Yelle, *The learning curve: Historical review and comprehensive survey*, *Decision Sciences* **10** (1979), no. 2 302–328. 5

[4] C. Stern and M. Deimler, *The Boston Consulting Group on strategy*. Wiley, 2006. 5

[5] G. Harpaz, W. Lee, and R. Winkler, *Learning, Experimentation, and the Optimal Output Decisions of a Competitive Firm*, *Management Science* **28** (1982), no. 6 589–603. 5, 9

[6] D. Fudenberg and J. Tirole, *Learning-by-doing and market performance*, *The Bell Journal of Economics* **14** (1983), no. 2 522–530. 6

[7] R. Hiller and J. Shapiro, *Optimal capacity expansion planning when there are learning effects*, *Management Science* **32** (1986), no. 9 1153–1163.

[8] S. Majd and R. Pindyck, *The Learning Curve and Optimal Production under Uncertainty*, *The Rand Journal of Economics* **20** (1989), no. 3 331–343. 6

[9] J. Mazzola and K. McCardle, *A Bayesian approach to managing learning-curve uncertainty*, *Management science* **42** (1996), no. 5 680–692.

[10] J. Mazzola and K. McCardle, *The stochastic learning curve: optimal production in the presence of learning-curve uncertainty*, *Operations research* **45** (1997), no. 3 440–450. 5, 6, 9

[11] I. Karatzas, *On the pricing of American options*, *Applied mathematics and optimization* **17** (1988), no. 1 37–60. 7

[12] S. Jacka, *Optimal stopping and the American put*, *Mathematical Finance* **1** (1991), no. 2 1–14.

[13] P. Carr, R. Jarrow, and R. Myneni, *Alternative characterizations of American put options*, *Mathematical Finance* **2** (1992), no. 2 87–106. 7

[14] K. Neuhoff, *Large-scale deployment of renewables for electricity generation*, *Oxford Review of Economic Policy* **21** (2005), no. 1 88. 9

[15] L. Coulomb and K. Neuhoff, *Learning curves and changing product attributes: the case of wind turbines*, *Cambridge Working Papers in Economics* (2006).

[16] S. Alberth, *Forecasting technology costs via the Learning Curve–Myth or Magic?*, *Cambridge Working Papers in Economics* (2007). 9

[17] K. Chan, G. Karolyi, F. Longstaff, and A. Sanders, *An empirical comparison of alternative models of the short-term interest rate*, *Journal of Finance* **47** (1992), no. 3 1209–1227. 11

[This page was intentionally left blank]

[This page was intentionally left blank]