Stability and Dimensionality Reduction in Nonlinear Filtering

Eliana Fausti

The Queen's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

To Andreas,
for all the wine and the Pink Rabbits,
and making me love my life in the rain.

Acknowledgements

They say that what matters is the journey, not the destination. Now that I have finally reached the end of this four-and-a-bit-years long tumble down the rabbit hole of mathematics, I cannot help but think that this saying has never rung truer. The past four years have been such a roller-coaster of emotions and experiences that the ending feels quite anticlimatic. Looking back, it is of course the people who come to mind, and give it all meaning in the end. This is to thank all those who shared bits of this journey with me, making the tough life of a poor aspiring mathematician so much better for it.

First and foremost, this journey would never even have started without my supervisor Sam Cohen, steering me down the path of filtering and geometry and a bunch of other things that sounded way cooler when I did not know anything about them. I am extremely thankful for his generosity in sharing his brilliant mathematical intuitions (half of which I still do not understand—through no fault of his), and his patience in sorting me out or following along whenever, out of sheer stubborness, I tended to wander off track. The good in this thesis is thanks to his guidance; the bad is (mostly) in spite of it.

Sam has been to the beginning of my PhD what my examiners Ben Hambly and John Armstrong have been to its end. I am very grateful for their helpful comments and input on my work, both to improve the presentation of this thesis and in view of future publications.

The other fundamental character in this journey, and in fact in any journey I have undertaken so far, is my family. Firstly, my parents Vittoria and Pier Luigi, who are not always good at hiding that they would rather have me closer to home, but always advise and support me in all my choices nonetheless, even those that lead me further away. Then, my brothers Luigi and Giacomo, who are a constant source of strength, and are possibly the two people in the world whose opinions I care the most

about. My cousin Vittoria, who is the warmest amd most fun friend one could hope to have. And all the other cousins and aunts and uncles who know that asking about my work at family gatherings is bound to lead to awkward and painful conversation, but ask anyway. Thanks for caring. A special thanks should go to my grandma Mariella, for bringing liveliness and a bit of crazyness anywhere she goes. And for letting me hide in her house in Bormio during the second Covid lockdown, which saved me from what would have been a very depressing winter otherwise.

And now, on to the people of my everyday life who made my time in Oxford memorable (and bearable). In the Maths Institute, and in no particular order, thanks to my brothers-in-arms from the 5th cohort of the PDE CDT, to Tommaso for all the coffees, to my officemates Catalina and Vicki for all the chats; thanks to the enlarged OxPDE group, to André for the swimming, Alexei for the salsa, to Jeremy, Antonio, Rafa, Duncan and Alejandro for making me believe that academia can be a friendly place. Thanks to the Stochastic Analysis and Maths Finance group, to Cris for the parties, to James and Patrick for the chess, to Ian for Catalina, to Filippo for giving me a place to sleep when I was homeless, to my maths-siblings Lingyi and Saad for sharing Sam with me.

Thanks to Angela, Alberto and Giulia for building a home-away-from-home at 35 Howard Street. Thanks to all my friends, in England, around the world and back at home in Bergamo for always being up for having fun. To my Italian friends in particular, thanks for always welcoming me back with open arms, despite my disappearing for months at a time. And thanks especially to those who came to visit: to Isotta, Vittoria and Federica for risking their lives cycling around London with me, to Marco for forcing me to wake up for May Morning, to Pietro and Stefano for the beers and the laughs and the brilliant idiocy.

And lastly, if I had not met Andreas right about at the start of my PhD studies, undoubtedly I would had far fewer distractions from my research. But I would have missed out on the best journey of my life so far. So, to the one journey-mate who was there for almost all the ups and downs of this bumpy road: thanks for that beer, four years ago. It turned out pretty well.

Abstract

The focus of this thesis is the analysis of the stability and robustness of continuous-time, finite state-space nonlinear filters, in order to provide new and practically relevant quantitative error bounds for a general class of approximate filters. This analysis is carried out through the use of the Hilbert projective metric.

We begin by providing a self-contained introduction to the Hilbert metric and its fundamental properties, with a particular focus on the space of probability measures. We then derive and study various dual formulations, and exploit these to obtain a contraction result for linear operators on convex cones with respect to a new distance, the hyperbolic tangent of the Hilbert metric. This general observation directs us naturally towards a range of new results on stability and robustness in nonlinear filtering.

Specifically, we turn to the problem of estimating the state of a continuous-time Markov chain from noisy observations. As regards stability, our key contribution is a proof that the corresponding optimal filter, called the Wonham filter, is contracting pathwise in the aforementioned distance given by the hyperbolic tangent of the Hilbert metric. Moreover, we give explicit deterministic and pathwise rates of convergence. By utilising these results, we are able to take an alternative approach to the study of the robustness of the Wonham filter, thereby improving on known error estimates and deriving rigorous, computable error bounds of theoretical and practical relevance as concerns the analysis and implementation of approximate filters.

Finally, we consider the problem of reducing the dimensionality of the Wonham filter via geometric projections, with a view towards defining an optimal projection filter. Building on the intuition provided by our error bounds, we find a natural submanifold for the Wonham filter such that the error of the projection filter is minimized.

Contents

List of Figures

Chapter 1

Introduction

1.1 Nonlinear stochastic filtering

Estimating a random hidden process from incomplete, noisy observations is a common problem arising in engineering, signal processing, finance, and many other disciplines. The general setting consists of a *signal (or state) process* X evolving in time (typically taken to be Markovian), which cannot be measured directly, but rather needs to be estimated using the information given by an *observation process* Y, whose dynamics depend on the signal X. Computing and analyzing the optimal solution to this problem is the main objective of the theory of stochastic filtering.

Stochastic filtering is a classical topic in stochastic analysis, and optimal filters have been derived in various contexts, e.g. in continuous or discrete time, with finite or infinite state-space, and under different structural assumptions. The setting of linear underlying dynamics, giving rise to the famous Kalman–Bucy filter [53, 54], was the first to be considered in continuous time, and is by now well understood. Nonlinear filtering, on the other hand, still presents challenges, from both the theoretical and practical perspective. We refer to Bain and Crisan [10] or Liptser and Shiryaev [66] for a comprehensive exposition of the classsical theory of nonlinear stochastic filtering.

The optimal nonlinear filter is the solution to a *nonlinear stochastic* (depending on the context, *partial*) *differential equation* called the *Kushner–Stratonovich* equation, attributed to both Kushner [60] and Stratonovich [80] following a mix-up due partly to the use of Itô's calculus in the first case, while the corresponding notion of Stratonovich integration was used in the second case (see Crisan [32] for a historical account of stochastic filtering). In the majority of practical applications, however, the nonlinear filter cannot be computed directly: for example, the model for X and Y, which the filtering equations explicitly depend on, might have misspecified param-

eters, or be completely unknown. Moreover, even when the true model is available, solving the filtering equations numerically can be intractable due to the high (in many cases, infinite) dimensional and non-local nature of the problem. Therefore, more often than not, approximate filters, rather than the optimal filter, are employed. This begs the questions of whether or not these approximations are reliable, and how we can quantify their error with respect to the optimal filter. These questions are central in the study of the *robustness* of the nonlinear filter, which is one of the key topics to be explored in this thesis.

More specifically, this thesis was born out of the desire to obtain a better understanding of the error arising from a specific class of approximate filters called *projection filters*, as introduced in Brigo, Hanzon and Le Gland [19]. The projection filter is one among several approximations of the Kushner–Stratonovich equation that attempts to provide a low-dimensional solution to the nonlinear filtering problem. In the next section, we shall briefly discuss approximate filters in general and the projection filter in particular.

1.2 Approximate filters

The filter is a measure-valued stochastic process which provides, at each time t, the conditional law of the signal X given the information accumulated from observing Y up to time t. Depending on the dimension of the state-space of X, the filter might be very high dimensional. If, for example, X is valued in $\mathbb{R}^d$, for $d \geq 1$, which is the case for many standard applications, then the nonlinear filter is a probability measure on $\mathbb{R}^d$, which is generally an infinite-dimensional object. In this sense, the Kushner–Stratonovich equation suffers from the so-called "curse of dimensionality".

To avoid infinite-dimensionality, engineers (and mathematicians) have come up with approximate filters, which are both finite-dimensional and easy to compute. The most well-known of these approximations is the *Extended Kalman Filter (EKF)*, based on linearization around the current state estimate, and it is still the most widely used in practice. Famously, the early development of the Extended Kalman Filter is due to research carried out at the NASA Ames Research Center for applications to aerospace engineering in the 1960's (see McGee, Schmidt and Smith [68]). Variation of the EKF are given for example by the augmented (or extended) EKF, which retains higher order terms in the Taylor expansion around the current estimate, the *Unscented Kalman Filter (UKF)* (see Julier and Uhlmann [51]), the *Assumed Density Filter (ADF)* (see Maybeck [67, Ch. 12]) and the *Ensemble Kalman Filter EnKF*

(introduced by Evensen [41]), particularly useful for very high-dimensional systems (such as in geophysical models and weather prediction).

The EnKF is related to another class of filtering approximations called *particle filters*. Particle filters approximate the filter by evolving randomly in time a large number of particles, whose empirical measure follows closely the distribution of the filter. They are also known as sequential Monte Carlo methods, since they employ sequential sampling and resampling of the particles to adaptively concentrate them in regions of high posterior probability. Particle filter methods are in general very flexible, and can be easily used to approximate any filtering density—however, to be effective in high dimensions with a reasonable rate of convergence of the error, they need to be finely tuned to suit each specific problem, and this can sometimes be a difficult task. For an introduction to particle filters, we refer to Doucet, de Freita and Gordon [36].

1.2.1 Dimensionality reduction via geometric projections

As already mentioned, in this thesis we are especially interested in the so-called *projection filter*, which is another example of a finite-dimensional approximate filter. Applying methods from differential geometry, this was first introduced by Hanzon [43] in 1987, and a comprehensive treatment was then given by Brigo, Hanzon and Le Gland [19] in the 90s. Very recently, the analysis of the projection filter has been taken further by Brigo and collaborators, see in particular Brigo and Armstrong [4] and Brigo, Armstrong and Rossi Ferrucci [6]. We will not give too many details here, since we review this filter extensively in Chapter 5.

Conceptually, the projection filter can be seen as closely related to the EKF, the UKF and the ADF briefly mentioned above. All of these approaches proceed by approximating the true filtering dynamics, to give an approximate filter living in a finite-dimensional subset of the space of probability measures. The differences between the approaches principally relate to how general this finite-dimensional space is allowed to be, and how the approximation is chosen. In projection filtering, a particularly elegant formulation based on the geometry of the space of probability measures is utilized.

The main idea is to view the Kushner–Stratonovich SPDE as describing a vector field on the space of probability densities. If the optimal filtering distribution is 'close' (in some suitable sense) to a statistical family M parametrized by a finite (and possibly low) dimensional parameter, then it should be possible to compute an approximate filter 'close' to the optimal filter by solving a suitable SDE for the

parameter determining M. To obtain the dynamics of this approximate filter which belongs to M, we can project the dynamics of the Kushner–Stratonovich SPDE onto the tangent space of M, and obtain a new differential equation, which now describes a vector field on M: it is the solution of this equation that is called *the projection filter*. Since M is chosen to be finite dimensional, the projection filter is also finite dimensional. Numerical experiments presented in Brigo, Hanzon and Le Gland [18] show remarkably good results when an appropriate statistical family is selected, even when dealing with filtering distributions that are known to not have a finite dimensional representation (such as in the case of the cubic sensor problem, see Hazewinkel, Marcus and Sussmann [44]).

Nevertheless, just as in the case of the EKF, the UKF and the ADF, and in fact in the case of most other approximate filters, no general convergence result for the projection filter is known, nor do we have precise estimates of the error between the projection filter and the solution to the Kushner–Stratonovich SPDE. (We exclude particle filters from this statement, as convergence results for particle filters do exist, and the challenges in implementing them consist mostly in finding ways around the very high number of particles that are needed—in theory—for a precise approximation of the filtering process; see for example [10, Section 8.6]). It is the lack of such results that motivates our analysis in this thesis.

1.2.2 A heuristic discussion of approximate filtering

To be concrete, let us consider the question of how one can go about computing error estimates for approximate filters in general, and the projection filter in particular. Focusing on the projection filter, we provide a sketch of this in Figure 1.1 below. This sketch and the appertaining discussion will serve well to illustrate the intuitive reasoning behind our approach to stability and robustness, both in general and for the projection filter in particular. Moreover, it also serves to motivate the specific development of the different chapters in this thesis.

In Figure 1.1 below, we have sketched some realized paths, discretized in time, for the optimal filter, denoted by π_t, and a projection filter, denoted by $\tilde{\pi}_t$, when the state-space of the signal X is taken to consist of exactly 3 states. Consequently, π_t (and therefore also its approximation $\tilde{\pi}_t$) are 3-dimensional probability vectors. The space of 3-dimensional probability vectors is the 2-dimensional probability simplex $\mathcal{S}^2 \subset \mathbb{R}^3$. In Figure 1.1 we have drawn $\mathcal{S}^2$ as a flat triangle for illustration purposes.

Denote by $\varphi : \mathcal{S}^n \mapsto \mathcal{S}^n$ the (discretized and idealized) flow of the Kushner–Stratonovich equation on $\mathcal{S}^n$ over 1 time step. In other words, we have $\pi_{t_{n+1}} = \varphi(\pi_{t_n})$,

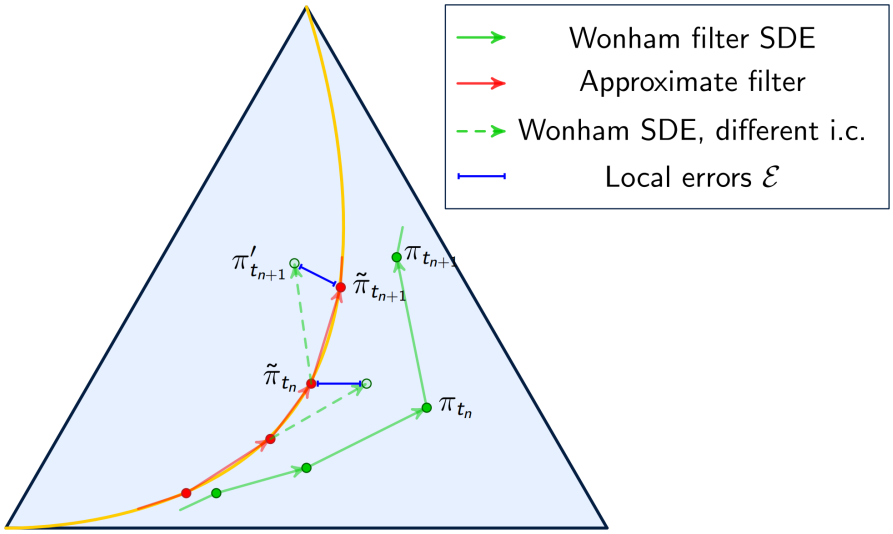

Figure 1.1: A sketch of the discretized dynamics of the optimal filter $\pi_t \in \mathcal{S}^2$ and the projection filter $\tilde{\pi}_t \in \gamma$

and the 1-step flow $\varphi(\pi_{t_n}) - \pi_{t_n}$ is represented by the green vectors in Figure 1.1. Similarly, denote by $\tilde{\varphi}$ the flow of the projection filter $\tilde{\pi}_t$ over 1 time step. The 1-step flow $\tilde{\varphi}(\tilde{\pi}_{t_n}) - \tilde{\pi}_{t_n}$ is represented by the red vectors. The projection filter $\tilde{\pi}$ over time stays on a 1-dimensional subspace of $\mathcal{S}^2$, which we call γ and which is represented by the yellow line in Figure 1.1.

Now, let D be a distance function on $\mathcal{S}^2$, and consider the error $D(\pi_{t_{n+1}}, \tilde{\pi}_{t_{n+1}})$ between the optimal filter π_t and the projection filter $\tilde{\pi}_t$ at the given time step. The triangle inequality yields

$$D(\pi_{t_{n+1}}, \tilde{\pi}_{t_{n+1}}) = D(\varphi(\pi_{t_n}), \tilde{\varphi}(\tilde{\pi}_{t_n})) \leq D(\varphi(\pi_{t_n}), \varphi(\tilde{\pi}_{t_n})) + D(\varphi(\tilde{\pi}_{t_n}), \tilde{\varphi}(\tilde{\pi}_{t_n})). \quad (1.1)$$

Denote by $\mathcal{E}_{t_n}$ the second error term $\mathcal{E}_{t_n} = D(\varphi(\tilde{\pi}_{t_n}), \tilde{\varphi}(\tilde{\pi}_{t_n}))$ on the right-hand side of (1.1). These are the local errors of the approximation, given in this case by the 'projection errors' of the flow φ onto the line γ. In Figure 1.1 we illustrate these projection errors in blue. At this point, if we have chosen γ and our projection operator wisely, then we have some hope that the projection errors $\mathcal{E}_{t_n}$ are small, so that the second term on the right of (1.1) is controlled (in other words, we have chosen an approximation with small local error).

The question that remains is the following: what can we say about the first term on the right-hand side of (1.1). Given the existing literature, the answer to this question is, sadly, 'not much'. The error $D(\varphi(\pi_{t_n}), \varphi(\tilde{\pi}_{t_n}))$ can only be quantified if we

turn to the analysis of the *stability* of the filtering dynamics. The study of the stability of the nonlinear filter is concerned with understanding the behaviour in time of the error of the optimal filter π_t when the initial conditions of the Kushner–Stratonovich SPDE are misspecified. Turning back to our sketched argument above, $\varphi(\pi_{t_n})$ and $\varphi(\tilde{\pi}_{t_n})$ represent respectively the 1-step dynamics of the Kushner–Stratonovich equation started at π_{t_n} and at $\tilde{\pi}_{t_n}$. Thus, understanding the stability of the nonlinear filter will tell us if $D(\varphi(\pi_{t_n}), \varphi(\tilde{\pi}_{t_n}))$ expands, contracts, or stays stationary: that is, whether $D(\varphi(\pi_{t_n}), \varphi(\tilde{\pi}_{t_n}))$ is greater than, smaller than, or equal to $D(\pi_{t_n}, \tilde{\pi}_{t_n})$. From there, a simple recursive application of the triangle inequality shows that unless the stability error contracts, we have no hope of controlling the error between π_t and $\tilde{\pi}_t$ in the medium to long time horizon.

These simple arguments motivate why, in our pursuit to understand the error of the projection filter, we must first turn our attention to the problem of understanding the stability of the nonlinear filter.

1.3 Stability and robustness of the Wonham filter

In this thesis, we focus on *finite state-space nonlinear filtering* in *continuous time*. The optimal filter in this case is the solution to an SDE, first derived by Wonham [87], and it is often referred to as the *Wonham filter*. We describe this filtering setting in detail in Chapter 3. The reason why we focus on the finite-dimensional setting is mostly practical: as we will see, analyzing the stability of the nonlinear filtering equations is already very challenging in this simpler setting—without adding to it the full generality of the infinite dimensional Kushner-Stratonovich SPDE.

Stability and robustness of the Wonham filter are ultimately two sides of the same coin. Stability, as we have already mentioned, is concerned with the error due to misspecification of the initial conditions of the filtering equations. Robustness, on the other hand, is the analysis of the error when the misspecification extends to the model parameters. Our goal of quantifying the errors of approximate filters goes one step further: we want to allow for misspecification of the model itself.

As we have tried to convey in the heuristics in the previous section, everything relies on obtaining a 'good stability estimate'. If the error between the true filter and the 'wrongly initialized' filter decays to zero as time passes, then the filter 'forgets' the initial error and is called *asymptotically stable*. If the nonlinear filter is stable, using an approximate filter is essentially the same as using the optimal filter, but introducing an approximation error at each time step. If all the approximation errors

are bounded, stability ensures that they are also 'forgotten' as time goes on, thus ensuring that the total error stays bounded, and we then recover robustness-type estimates.

In discrete-time, finite state-space nonlinear filtering, an argument of this kind is indeed how robustness estimates have been derived (see Budhiraja and Kushner [21], Le Gland and Mevel [62] and Le Gland and Oudjane [63]). The main difference with the continuous-time setting is that the stability estimates available in the literature for the Wonham filter are not strong enough to directly apply this kind of methodology. Indeed, to pull off this argument in continuous time, one would need exponential (or similar) contraction of the stability error. The first goal of this thesis will be to establish such a contraction result for the Wonham filter. The second objective is to use our stability estimates to provide meaningful, computable error bounds for general approximate filters.

We shall review extensively the literature for the stability of the Wonham filter in Chapter 3. For now, we limit ourselves to describing briefly two of the main contributions in this field, due to Atar and Zeitouni [8] and more recently Chigansky and Van Handel [26]. The works by Chigansky and Van Handel [26] and Van Handel [83] made significant progress on the robustness of the Wonham filter, as measured by the L^1-error between the true Wonham filter and the Wonham filter with misspecified model parameters. Following an approach that relies on computing bounds for suitable derivatives of the filter, they prove that the error stays finite over an infinite time horizon, and vanishes as the misspecified parameters are sent to the true ones. The method in [26] could potentially be used to compute error bounds for more general approximate filters, and not only those given by misspecification in the underlying model. However, the estimates in [26] are not tight enough to provide useful quantitative bounds (see [26, Remark 2.8] and [83, Remark 3.3.8]) so these results remain primarily of qualitative interest.

The earlier results of Atar and Zeitouni [8], also for the Wonham filter, concern asymptotic rates of decay for the stability error of the filter. These asymptotic results are derived using suitable time discretizations and properties of the Hilbert projective metric together with Birkhoff's ergodic theorem in order to pass to the limit. Arguably, the fundamental contribution of Atar and Zeitouni was the idea of exploiting the Hilbert projective metric as a tractable notion of distance on the space of probability measures. This idea is also at the heart of our approach in this thesis. However, the methodological similarities end there. In particular, as we just discussed, Atar and

Zeitouni go through time discretizations, while we derive our results in a differential way based on a helpful coordinate transformation linked to information geometry.

Before finishing this introduction with an outline of the thesis and its main contributions, we first take a little moment to discuss the history of the Hilbert projective metric and some prominent examples of its use.

1.4 The Hilbert projective metric

In [14], Garrett Birkhoff[1] proved that positive linear operators on a positive convex cone contract in the Hilbert projective (pseudo-)metric (introduced by Hilbert in [45]). Starting from this result, he easily derives a proof of the famous Perron theorem[2], by essentially reducing it to a special case of the Banach fixed point theorem. Similarly, he also immediately proves several of its generalizations: first, the extension of the Perron theorem to non-negative matrices due to Frobenius, then the extensions to infinite dimensional function spaces and positive integral and compact operators, originally due to Jentzsch [50] and Krein and Rutman [58].

In the same year as Birkhoff, Samelson [77] also published a proof of the Perron–Frobenius theorem using projective geometry, and similarly a few years later Hopf [46] presented a contraction result for positive linear integral operators and an alternative proof of Jentzsch's theorem (apparently without being aware of Birkhoff's previous results). The combination of the Hilbert projective metric with the contraction mapping theorem has inspired a vast amount of further work on the Perron–Frobenius theorem and its various extensions, see Lemmens and Nussbaum [65] and references therein for a detailed overview of the topic. We highlight the early works of Thompson [82], Bushell [22], and Kohlberg and Pratt [57], as well as the recent works by Rugh [76] and Dubois [37] on operators on complex cones which have strongly influenced our presentation in Chapter 2.

While the impact on Perron–Frobenius theory is certainly the most significant consequence of Birkhoff's work, we became interested in the Hilbert metric and Birkhoff's contraction result due to a different (although related) application: the study of the ergodicity of non-negative linear operators. Work in this direction was presented by Birkhoff himself in [15, Chapter XVI, Sec.7-8], Seneta and Sheridan [78, 79] and Le Gland and Mevel [61]. Evidently, the ergodic theory of Markov processes with linear

[1] The son of George D. Birkhoff, who is better known in probabilistic circles for his proof of the ergodic theorem.

[2] For a statement of the theorem, see [88, Theorem 5.25]; for the original work by Perron, see [72].

transition kernels can be reduced to a particular case of this broader topic. Nevertheless, the Hilbert metric is not strictly necessary for developing the standard ergodic theory of Markov processes. For example, in [78, Chapter 3] Seneta uses the Hilbert metric to study the ergodicity of inhomogenous products of non-negative matrices; on the other hand, in the following chapter [78, Chapter 4], which treats specifically the ergodicity of discrete Markov chains, the Hilbert metric is not used, since the analysis simplifies by virtue of the generators being stochastic matrices.

Perhaps as a consequence of this, one of the main avenues of application for the Hilbert projective metric in probability theory appears to be relatively unknown. This might in part explain why, despite being rather powerful tools, neither the Hilbert metric nor Birkhoff's contraction results seem to have found widespread use in the probability community[3]. There are of course a few notable exceptions: Birkhoff's contraction theorem was employed, for example, in an elegant proof of the linear convergence of the Sinkhorn algorithm by Franklin and Lorenz [42]. Moreover, of particular relevance to the present thesis, it has proved fundamental in the qualitative and quantitative analysis of the stability of hidden Markov processes, where the use of the Hilbert metric was introduced by Atar and Zeitouni [8, 9] as discussed in the previous section. Following their work, it has become a rather well-established approach to the problem, see e.g. [12, 62, 63]. More recently, the Hilbert metric has found computational applications in entropic interpolation [23] and nonlinear embeddings [70].

1.5 Outline and main contributions

The contents of this thesis consists of three main components, which form the bulk of my work during my DPhil studies in Oxford. Chapter 2 closely follows the treatment of the Hilbert projective metric from the preprint [30]; Chapters 3 and 4, taken together, present the results of the preprint [29] on the stabililty and robustness of the Wonham filter; Chapter 5 contains my research on the projection filter in finite state-space which has not yet been uploaded to the arXiv. All of this has been done with collaboration from my DPhil supervisor, Sam Cohen.

- In Chapter 2, we begin our study with a comprehensive treatment of the Hilbert projective (pseudo-)metric and its contractivity properties. The overarching goal is to provide a clear, self-contained guide to this powerful tool and its merits

[3]To illustrate this point, no mention is made of this approach in any of the books by Ethier and Kurtz [40], Meyn and Tweedie [69], Brémaud [17], Kallenberg [52], or Jacod and Shiryaev [49].

in probability theory. Whereas most available treatments are chiefly analytic in nature, here we instead have in mind applied probabilists and statisticians who find themselves curious, as we were, about the merits and limitations of working with the Hilbert projective metric in a probabilistic context. In particular, our main focus is on studying the space of probability measures when equipped with the Hilbert projective metric. This may be viewed as a subspace of the positive cone in the space of signed measures, which of course comes with a natural Banach space structure, as would be the point of view in analytic treatments. In order to illustrate the general principles at work, we therefore place our results within the context of locally convex topological vector spaces.

While we believe the results of Chapter 2 could be of independent interest, their role in the present thesis is first and foremost to set the stage for how we will approach the analysis of stability and robustness for the Wonham filter in the chapters that follow. The central result is the fact that linear operators on convex cones contract in a new distance given by the hyperbolic tangent of the Hilbert metric, which in particular implies Birkhoff's classical contraction result. This will relate closely to our stability analysis for the Wonham filter in the subsequent Chapter 3. Moreover, in the space of probability measures, we analyse dual formulations of the metric and, in the special case of discrete probability measures, we explore the resulting geometry of the probability simplex. The latter relates to our treatment of the projection filter in Chapter 5. Finally, we address comparisons of the Hilbert metric with the total variation norm, p-Wasserstein distance, as well as any f-divergence, and we derive a novel sharp bound for the total variation between two probability measures in terms of their Hilbert distance.

- In Chapter 3, we introduce the filtering setting that has been at the center of our work, namely the continuous-time nonlinear filter on finite state-space. In particular, we show that the corresponding optimal filter, called the *Wonham filter*, is strictly contracting pathwise with respect to the distance given by hyperbolic tangent of the Hilbert projective metric as discussed above. Moreover, we give explicit deterministic and pathwise rates of convergence. These results will allows us to give alternative proofs of the robustness of the Wonham filter in the subsequent Chapter 4, improving on known error estimates and yielding rigorous, computable error bounds for approximate filters.

The main contribution of Chapter 3 is an exponential contraction estimate for the stability error of the Wonham filter, in Hilbert distance (see Theorem 3.2.1). In fact, our statement is stronger, as we can prove contractivity of the hyperbolic tangent of the Hilbert distance, which directly implies the former. Both of these are, to the best of our knowledge, new results, which improve significantly on the quantitative estimates for the error available in the literature. We also present an alternative way to study the stability error of the continuous-time filter in Hilbert distance, which does not rely on Atar and Zeitouni's arguments. Instead, inspired by Amari [3], we will introduce a change of coordinates from the probability simplex to $\mathbb{R}^n$, and study the evolution of the Wonham filter in the new coordinate system. As we will see, our arguments will present some similarities with the proof of [12, Theorem 4.3], despite a different approach.

- In Chapter 4 we shall exploit the contraction results established in Chapter 3 in order to investigate the error when approximate filters, rather than the optimal filter, are employed. The first main contribution of Chapter 4 is a robustness-type estimate for the Wonham filter (see Theorem 4.1.1). Compared to [26], we state our error bounds for a general approximate filter, and in terms of the Hilbert distance. Since the Hilbert distance is stronger than the L_1-norm, which is used in [26], the error bounds we provide are tighter (although still not optimal, as we will discuss in Section 3.3.4 and Section 4.4). We also believe our proof methodology to be interesting in its own right, being far simpler than the arguments in [26]: it relies only on standard stochastic analysis tools, while in [26] the authors need Malliavin calculus to deal with anticipative stochastic integrals.

  Finally, both our work in Chapter 2 and our findings in Theorem 3.2.1 suggest that the hyperbolic tangent of the Hilbert distance (instead of simply the Hilbert distance) might have advantages as a metric for studying the error of approximate filters. In the particular case when the approximate filter is chosen so that the stochastic term of the Wonham SDE is matched exactly, this yields tighter, pathwise bounds for the error, which we prove in Theorem 4.1.2.

- In Chapter 5 we focus on the projection filter. We adapt the three types of SDE projections defined in [6] to our setting, and compare the errors of the resulting projection filters in the expected Hilbert metric, applying Theorem 4.1.1. We then take a leaf out of our analysis in Chapters 3 and 4, and move our analysis once more from the probability simplex to $\mathbb{R}^n$, employing the same coordinate

transformation. Considering projection filters in the transformed space, we find a primary submanifold for the projection filter such that the stochastic error terms of Theorem 4.1.1 vanish, and so tighter error bounds along the lines of Theorem 4.1.2 apply.

Chapter 2

The Hilbert projective metric on probability measures

The main goal of this chapter is to provide a clear, self-contained guide to the Hilbert projective (pseudo-)metric and its merits and limitations in a probabilistic context. Consequently, our main focus is the study of the space of probability measures when equipped with the Hilbert projective metric. Nevertheless, it is natural to start from general convex cones in locally convex topological vector spaces. Of course, the space of probability measures is itself a subspace of the positive cone in the vector space of signed measures.

2.1 Outline

We shall start by defining the Hilbert projective (pseudo-)metric on a proper convex cone in a locally convex topological vector space. From there, our main contribution in Section 2.2 is the introduction of a new (pseudo-)metric, the hyperbolic tangent of the Hilbert metric (which we call the $\mathcal{T}$-distance), under which we prove linear operators also contract (see Definition 2.2.5 and Theorem 2.2.6). The advantage of using the $\mathcal{T}$-distance compared to the Hilbert projective metric is that $\mathcal{T}$ stays bounded, while the Hilbert metric (easily) diverges to infinity. As far as we are aware, the formulation of this contraction result in Theorem 2.2.6 has not been given before (but we note that our proof is inspired by Dubois' proof of [37, Thm. 2.3], so we do not wish to claim full credit for the result).

When introducing the Hilbert projective metric, we especially insist on its definition through duality, which we first came across in [76]. This 'dual' definition, in particular through a predual space, is natural in the context of probability measures,

where distances are often defined by considering measures as integrators dual to particular classes of functions. In Section 2.3 we provide a careful analysis of the Hilbert metric and the topology it induces on the space of probability measures using duality, and we show that convergence of measures in the Hilbert metric (or in the $\mathcal{T}$-distance) is stronger than convergence in total variation (as was already shown in [8, Lemma 1]), convergence in p-Wasserstein distance and convergence in any f-divergence.

In Section 2.4 we study the geometry of the probability simplex under the Hilbert metric. Using our dual approach, we give an easy derivation of the explicit formula for the contraction rate of a linear operator (Proposition 2.4.1), which can also be extended to infinite settings (Proposition 2.4.2). As the probability simplex is finite dimensional, it has a natural manifold structure, however the Hilbert metric does not induce a hyperbolic (in the sense of Gromov) geometry on the simplex (since its boundary is not differentiable, see [13]), nor a Riemannian structure. The Hilbert geometry is far more curious: we find an explicit characterizations of Hilbert balls as 'hexagonal' convex polytopes, extending to the n-dimensional case work by Phadke [73] and de la Harpe [34]. Finally, in Section 2.5 we use these geometric observations to prove a sharp bound for the total variation distance between two probability measures in terms of their $\mathcal{T}$-distance (see Theorem 2.5.1 and Cor. 2.5.2.1).

2.2 The Hilbert projective metric: definitions and contractivity

Let us start with the definition of the Hilbert projective distance, in the sense of Birkhoff [14, 15], on a cone in a (real) locally convex topological vector space (LCS). Note that Birkhoff works more specifically with cones in real Banach spaces (lattices). However, as we will see shortly, the definition of the metric does not require the space to be Banach, so for the sake of generality we work with an LCS.

Let X be an LCS. Let $C \subset X$ be a proper closed convex cone, meaning that C is closed and satisfies

$$C + C \subseteq C, \qquad \mathbb{R}_+ C = C, \qquad C \cap -C = \{0\}.$$

Following [76, Sec. 4] (or extrapolating directly from [14] or [15, Chapter XVI]), we give Birkhoff's definition of the Hilbert projective distance.

Definition 2.2.1 (Hilbert projective pseudo-metric)**.** For $x, y \in C \setminus \{0\}$, where C is a proper closed convex cone in a real LCS, let $\beta(x, y) \in (0, \infty]$ be given by

$$\beta(x, y) = \inf\{r > 0 \,:\, rx - y \in C\} = \sup\{r > 0 \,:\, rx - y \notin C\}.$$

Then the Hilbert projective distance is defined by

$$\mathcal{H}(x,y) = \log\big(\beta(x,y)\beta(y,x)\big) \in [0,\infty], \qquad \forall x,y \in C \setminus \{0\}. \tag{2.1}$$

It is worth spending a few moments to properly understand these definitions. Since C is closed, and $-y \notin C$ for all $y \in C \setminus \{0\}$, we have $\beta(x,y) > 0$. It is then straightforward to see that $\beta(x,y)\beta(y,x) \geq 1$, by noting that

$$\begin{aligned}\frac{1}{\beta(x,y)} &= \inf\{1/r > 0 \,:\, rx - y \notin C\} = \inf\{r > 0 \,:\, x - ry \notin C\} \\ &= \inf\{r > 0 \,:\, ry - x \notin -C\} \leq \inf\{r > 0 \,:\, ry - x \in C\} = \beta(y,x), \end{aligned} \tag{2.2}$$

and $\beta(x,y) = \frac{1}{\beta(y,x)}$ if and only if $y = cx$ for some $c \in \mathbb{R}_+$ (in this case, x and y are said to be *collinear*). Hence (2.1) is well-defined with $\mathcal{H}(x,y) = 0$ if and only if x and y are collinear. Symmetry of $\mathcal{H}$ is clear from the definition, and one can verify that the triangle inequality is also satisfied (see Remark 2.2): then, $\mathcal{H}$ is a pseudo-metric for C (see also [15, Chapter XVI]).

Example 2.2.2. The finite non-negative measures on $[0,1]$ form a proper convex (non-negative) cone in the space of the signed measures on $[0,1]$. Since the signed measures on $[0,1]$ equipped with the total variation distance form a Banach space (and therefore an LCS), the definition above applies. The Hilbert pseudo-metric can then be restricted to the probability measures on $[0,1]$, to give a true metric. We will explore in detail the Hilbert distance on the space of measures in Section 2.3.

Remark 2.1. The only role of the specific choice of LCS topology on X, in defining the $\mathcal{H}$-metric, is to have the notion of C being closed in X. Two different topologies on X for which C is closed in X will both give rise to the same Hilbert projective metric. In this sense, the Hilbert metric on C is independent of the topology on the ambient space X.

By closure of C in X, if $\beta(x,y) < \infty$, we must have that $\beta(x,y)x - y \in \partial C$, where ∂C is the boundary of C. If $x \in \partial C$ and $y \in \mathring{C}$, where $\mathring{C}$ denotes the interior of C, then $\beta(x,y) = \infty$. However, if both $x, y \in \partial C$, then $\beta(x,y)$ might be finite. Using Birkhoff's choice of terminology [15, Chapter XVI], $x, y \in C$ are *comparable* if $\beta(x,y)$ and $\beta(y,x)$ are both finite. We observe that two boundary elements $x, y \in \partial C$ might still be comparable.

Especially when X is infinite dimensional, it can be useful to understand the Hilbert projective distance through duality (see [37, 76], where this is exploited in the

analysis of complex cones). Let X^* be the (topological) dual of X, and let $\langle \cdot, \cdot \rangle$ be the natural bilinear form $X^* \times X \to \mathbb{R}$. One can define the dual cone C^* as

$$C^* = \big\{ f \in X^* \,:\, f|_C \geq 0 \big\}. \tag{2.3}$$

Proposition 2.2.3. *An equivalent definition of the Hilbert pseudo-metric is given by*

$$\mathcal{H}(x,y) = \sup_{\substack{f,g \in C^* \\ \langle f,x\rangle, \langle g,y\rangle \neq 0}} \left\{ \log \frac{\langle f,y\rangle \langle g,x\rangle}{\langle f,x\rangle \langle g,y\rangle} \right\}. \tag{2.4}$$

Proof. Let $\tilde{C} := \big\{ x \in X \,:\, \langle f,x\rangle \geq 0,\ \forall f \in C^* \big\}$. One can confirm that $\tilde{C}$ is a proper closed convex cone. We clearly have that $C \subseteq \tilde{C}$. Now consider $x \notin C$. Since C is convex and closed, by the Geometric Hahn–Banach theorem (see e.g. [31, Thm. IV.3.9 & Cor. IV.3.10]) there exists a continuous linear functional $g \in X^*$ and an $\alpha \in \mathbb{R}$ such that $\langle g,x\rangle < \alpha$ and $\langle g,y\rangle \geq \alpha$ for all $y \in C$. Since the image of the cone C under the functional g can only be one of $\mathbb{R}_+$, $\mathbb{R}_-$, $\mathbb{R}$ or $\{0\}$, we must have $\alpha = 0$. Hence $g \in C^*$ but $\langle g,x\rangle < 0$, which implies that $x \notin \tilde{C}$. Therefore $\tilde{C} \subseteq C$ also, and so $C = \tilde{C}$.

Take $x, y \in C$ and let $r \in \mathbb{R}_+$ with $rx - y \notin C$. Since $C = \tilde{C}$, there exists some $f \in C^*$ such that $\langle f, rx - y\rangle < 0$. Consequently,

$$\begin{aligned}
\beta(x,y) &= \sup\{r > 0 \,:\, rx - y \notin C\} = \sup\{r > 0 \,:\, \langle f, rx - y\rangle < 0 \text{ for some } f \in C^*\} \\
&= \sup\{r > 0 \,:\, r\langle f,x\rangle < \langle f,y\rangle \text{ for some } f \in C^*\} \\
&= \sup\left\{ \frac{\langle f,y\rangle}{\langle f,x\rangle} \,:\, f \in C^*,\ \langle f,x\rangle \neq 0 \right\},
\end{aligned} \tag{2.5}$$

and similarly for $\beta(y,x)$. Then (2.4) is indeed equivalent to (2.1). □

Remark 2.2. Given $\log \frac{\langle f,y\rangle\langle g,x\rangle}{\langle f,x\rangle\langle g,y\rangle} = \log \frac{\langle f,z\rangle\langle h,x\rangle}{\langle f,x\rangle\langle h,z\rangle} + \log \frac{\langle h,z\rangle\langle g,x\rangle}{\langle h,x\rangle\langle g,z\rangle}$, for any $f, g, h \in C^*$, it is easy to verify the triangle inequality for $\mathcal{H}$ using the representation (2.4).

Remark 2.3. Once more, since the topology on X does not change the Hilbert metric on C, when working with (2.4) one can choose the topology, and therefore the dual space, cleverly: a coarser topology, with a smaller corresponding dual space, will almost always be preferable.

2.2.1 Contraction properties

Positive linear operators on a positive closed convex cone contract in the Hilbert projective distance: this result is again due to Birkhoff [14]. More generally, for proper closed convex cones $C \subset X$, Birkhoff's contraction theorem can be stated as follows:

Theorem 2.2.4 (Birkhoff)**.** *Let X be a LCS, take $L : X \to X$ to be a linear transformation, and suppose that $L(C \setminus \{0\}) \subseteq C \setminus \{0\}$. If the diameter $\Delta(L) = \sup_{x,y \in C\setminus\{0\}} \mathcal{H}(Lx, Ly)$ is finite, then we have*

$$\mathcal{H}(Lx, Ly) \leq \tau(L)\mathcal{H}(x, y), \qquad \forall x, y \in C \setminus \{0\}, \tag{2.6}$$

and $\tau(L) = \tanh\left(\frac{\Delta(L)}{4}\right)$ is called the Birkhoff contraction coefficient.

The theorem holds equivalently if one drops finiteness of $\Delta(L)$ as a condition and extends the definition of the contraction coefficient to $\tau(L) = 1$ when $\Delta(L) = \infty$. In other words, any bounded linear operator L is non-expansive in C under the Hilbert distance, but it is *strictly contracting* if and only if the diameter $\Delta(L)$ of C under L in the Hilbert metric is finite, i.e. $\tau(L) < 1$.

We will now show that a stronger result than Theorem 2.2.4 is possible.

Definition 2.2.5 ($\mathcal{T}$-distance)**.** For $x, y \in C \setminus \{0\}$, where C is a proper closed convex cone in a real LCS, we define the *hyperbolic tangent of the Hilbert pseudo-metric* as

$$\mathcal{T}(x, y) := \tanh\Big(\frac{\mathcal{H}(x, y)}{4}\Big), \tag{2.7}$$

where $\tanh(\infty) := 1$. For simplicity we refer to (2.7) as the *$\mathcal{T}$-distance*.

Note that the $\mathcal{T}$-distance is a pseudo-metric for C: one can easily check that the triangle inequality and symmetry properties are inherited from the Hilbert distance. However, the metric $\mathcal{T}$ makes the cone C into a *bounded* space, while $\mathcal{H}$ gives an infinite distance between any points in $\mathring{C}$ and ∂C. Borrowing ideas from the proof of [37, Thm. 2.3], we obtain the following theorem.

Theorem 2.2.6. *Let X be a LCS, and $L : X \to X$ a linear transformation. Suppose that $L(C \setminus \{0\}) \subseteq C \setminus \{0\}$. Then we have*

$$\mathcal{T}(Lx, Ly) \leq \tau(L)\mathcal{T}(x, y), \qquad \forall x, y \in C \setminus \{0\}, \tag{2.8}$$

where $\tau(L) = \sup_{x,y \in C\setminus\{0\}} \mathcal{T}(Lx, Ly)$ is the diameter of C under L in the $\mathcal{T}$-distance, and is equal to the Birkhoff contraction coefficient.

Proof. For all $x, y \in C \setminus \{0\}$, define the set

$$E_C(x, y) := \{r > 0 \,:\, rx - y \notin C\}.$$

Using the same notation as in (2.1), we have $\mathcal{H}(x, y) = \log(\beta(x, y)\beta(y, x))$, where

$$\beta(x, y) := \sup E_C(x, y) \in (0, \infty], \qquad \beta(y, x) := \sup E_C(y, x) \in (0, \infty].$$

Fix $x, y \in C \setminus \{0\}$. If Lx and Ly are collinear, then $\mathcal{T}(Lx, Ly) = 0$ and the claim holds trivially. Similarly, if $\mathcal{H}(x, y) = \infty$, then $\mathcal{T}(x, y) = 1$; as $\mathcal{T}(Lx, Ly)$ is certainly less than its supremum over x and y, the claim holds. It remains to consider the case $\mathcal{H}(x, y) < \infty$ and Lx and Ly not collinear (so $\mathcal{H}(Lx, Ly) \neq 0$).

For $r > 0$, consider $rx - y \in C$. By linearity of L, we also have $rLx - Ly \in C$, and in particular $E_C(Lx, Ly) \subset E_C(x, y)$. This implies that

$$\beta(x, y) \geq \beta(Lx, Ly) > \frac{1}{\beta(Ly, Lx)} \geq \frac{1}{\beta(y, x)},$$

where the strict inequality in the middle is due to (2.2) and the assumption that $\mathcal{H}(Lx, Ly) \neq 0$. We now approximate $\beta(x, y)$ and $\beta(y, x)$ from above (since $\mathcal{H}(x, y) < \infty$, also $\beta(x, y), \beta(y, x) < \infty$), and $\beta(Lx, Ly)$ and $\beta(Ly, Lx)$ from below, i.e. take $M, m > 0$ and $M', m' > 0$ such that

$$M > \beta(x, y), \quad m > \beta(y, x), \quad \frac{1}{\beta(Ly, Lx)} < \left\{ \frac{1}{m'}, M' \right\} < \beta(Lx, Ly).$$

By definition of $E_C(Lx, Ly)$ and $E_C(Ly, Lx)$, we have $M'Lx - Ly \notin C$, and $m'Ly - Lx \notin C$. Similarly, $Mx - y \in C$ and $my - x \in C$. For $r > 0$, note that

$$rL(Mx - y) - L(my - x) = (rM + 1)Lx - (r + m)Ly \in C \iff \frac{rM + 1}{r + m} Lx - Ly \in C.$$

Letting $h_1(r) = (rM + 1)/(r + m)$, this implies in particular that

$$\begin{aligned} E_C(L(Mx - y), L(my - x)) &= \left\{ r > 0 \,:\, h_1(r)Lx - Ly \notin C \right\} \\ &= \left\{ h_1^{-1}(w) > 0 \,:\, wLx - Ly \notin C \right\} \\ &= h_1^{-1}\left(\left\{ \frac{1}{m} < w < M \,:\, wLx - Ly \notin C \right\} \right) \\ &\subset h_1^{-1}\Big(E_C(Lx, Ly) \Big), \end{aligned}$$

where $h_1^{-1}(r) = (rm - 1)/(M - r)$. Since $M' \in \left(\frac{1}{\beta(Ly, Lx)}, \beta(Lx, Ly) \right) \subset \left(\frac{1}{m}, M \right)$ by assumption, and $M'Lx - Ly \notin C$, we have that $h_1^{-1}(M') \in E_C(L(Mx - y), L(my - x))$.

Analogously, for $r > 0$,

$$rL(my - x) - L(Mx - y) \in C \iff \frac{rm + 1}{r + M} Ly - Lx \in C.$$

Letting $h_2(r) = (rm + 1)/(r + M)$ and $h_2^{-1}(r) = (rM - 1)/(m - r)$, this implies that

$$\begin{aligned} E_C(L(my - x), L(Mx - y)) &= h_2^{-1}\left(\left\{ \frac{1}{M} < w < m \,:\, wLy - Lx \notin C \right\} \right) \\ &\subset h_2^{-1}\Big(E_C(Ly, Lx) \Big). \end{aligned}$$

Since $\frac{1}{m'} \in (\frac{1}{m}, M)$ by assumption, which implies that $m' \in (\frac{1}{M}, m)$, and $m'Ly - Lx \notin C$, we have that $h_2^{-1}(m') \in E_C(L(my-x), L(Mx-y))$.

We know $Mx - y, my - x \in C$ by definition of $\beta(x,y)$ and $\beta(y,x)$, and $\Delta(L) = \sup_{x,y \in C\setminus\{0\}} \mathcal{H}(Lx, Ly)$. Therefore,

$$\begin{aligned} h_1^{-1}(M')h_2^{-1}(m') &\leq \beta\big(L(Mx-y), L(my-x)\big)\beta\big(L(my-x), L(Mx-y)\big) \\ &\leq \sup_{\tilde{x},\tilde{y} \in C\setminus\{0\}} \beta(L\tilde{x}, L\tilde{y})\beta(L\tilde{y}, L\tilde{x}) = e^{\Delta(L)}, \end{aligned}$$

which yields the inequality

$$\frac{(M'm-1)}{(M-M')}\frac{(m'M-1)}{(m-m')} \leq e^{\Delta(L)}.$$

Now let $D = \log(Mm)$ and $d = \log(M'm')$. Note that since $M' \in (\frac{1}{m}, M)$ and $m' \in (\frac{1}{M}, m)$, $d \leq D$. Substituting $m' = \frac{e^d}{M'}$ in the above yields

$$f(M') := \frac{(M'm-1)}{(M-M')}\frac{(e^dM - M')}{(mM' - e^d)} \leq e^{\Delta(L)}. \tag{2.9}$$

Noting that $M' = \frac{e^d}{m'} \in (\frac{e^d}{m}, e^dM)$, intersecting this set with $(\frac{1}{m}, M)$ yields $M' \in (\frac{e^d}{m}, M)$. Differentiating the left-hand side of (2.9), we find that the minimum of $f(M')$ within these constraints for M' is attained at $M'^* = e^{d/2}\sqrt{\frac{M}{m}}$. Substituting into the expression above, we get

$$f(M'^*) = \frac{\left(e^{d/2}\sqrt{Mm} - 1\right)}{\left(1 - e^{d/2}\frac{1}{\sqrt{Mm}}\right)}\frac{\left(e^d - e^{d/2}\frac{1}{\sqrt{Mm}}\right)}{\left(e^{d/2}\sqrt{Mm} - e^d\right)} = \frac{\sinh^2\left(\frac{D+d}{4}\right)}{\sinh^2\left(\frac{D-d}{4}\right)}.$$

Taking square-roots yields

$$\frac{\sinh\left(\frac{D+d}{4}\right)}{\sinh\left(\frac{D-d}{4}\right)} \leq \sqrt{f(M')} \leq e^{\Delta(L)/2}.$$

Using the identity $\sinh(a \pm b) = \sinh(a)\cosh(b) \pm \sinh(b)\cosh(a)$ and the fact that $\frac{x-1}{x+1}$ is increasing for $x > 0$, we obtain the final expression

$$\tanh\left(\frac{d}{4}\right) \leq \tanh\left(\frac{\Delta(L)}{4}\right)\tanh\left(\frac{D}{4}\right).$$

Taking limits as $M, m \to \beta(x,y), \beta(y,x)$ and $M', m' \to \beta(Lx, Ly), \beta(Ly, Lx)$, we are done. □

Remark 2.4. Birkhoff's Theorem 2.2.4 is immediate from concavity and monotonicity of $\tanh(x)$ for $x \geq 0$. The advantage of using $\mathcal{T}$ instead of $\mathcal{H}$ is negligible when the distances are small, since $\mathcal{T}(x,y)$ is equivalent to $\mathcal{H}(x,y)$ asymptotically as $\mathcal{H}(x,y)$ approaches 0. However, we can find points $x, y \in C$ such that $\mathcal{H}(x,y) = \infty$, such as when comparing an element $x \in \mathring{C}$ with an element $y \in \partial C$. In these cases, the $\mathcal{T}$-distance is preferable, since $\mathcal{T}(x,y)$ stays finite and the inequality (2.8) remains meaningful.

Example 2.2.7. As we mentioned in the introduction, an immediate application of Birkhoff's theorem is in the ergodic theory of Markov processes, since transition operators are positive linear operators that map probability distributions to probability distributions, and so the assumptions of Theorem 2.2.4 are satisfied. We discuss explicit forms of Birkhoff's contraction coefficient for stochastic matrices and a class of transition kernels in Section 2.4.

2.3 $\mathcal{H}$-metric on the space of probability measures

From general LCS we now move to the space of probability measures, and consider the Hilbert projective distance in this context specifically. In [15, Chapter XVI] Birkhoff works with positive cones in a Banach lattice. Since the probability measures are a subset of the positive measures, which form the positive cone in the space of signed measures, the definition of the Hilbert distance on probability measures can be easily deduced from Birkhoff's work (see e.g. [8, Eq. 9] and [63, Def. 3.3]). In this section we choose to derive the Hilbert distance in the framework of duality instead, drawing a parallel with the works on convex cones [37,76]. The purpose of this exercise is to gain an understanding of the Hilbert metric in terms of functions acting on probability measures, and to investigate how a change in the test-functions affects the distance itself.

Notation. For any σ-algebra $\mathcal{F}$ and space F, let $L^0(\mathcal{F}, F)$ denote the space of $\mathcal{F}$-measurable functions, valued in F, and let $B(\mathcal{F}, F)$ be the subspace of bounded $\mathcal{F}$-measurable functions. For any two spaces E, F, let $C_b(E,F)$ denote the space of bounded continuous functions $E \to F$. If E and F are metric spaces, let $C_{bL}(E,F)$ be the space of bounded F-valued Lipschitz functions. By $\|f\|_\infty$ we denote the L^∞-norm of f and by $\|f\|_{Lip}$ its Lipschitz coefficient.

Let $(E, \mathcal{F})$ be a measurable space, and consider the space $\mathcal{M}(E)$ of finite signed measures on $(E, \mathcal{F})$. A natural approach is to make $\mathcal{M}(E)$ into a Banach space by equipping it with the total variation norm $\|\cdot\|_{\mathsf{TV}}$. The total variation norm is defined, as usual, by

$$\|\mu\|_{\mathsf{TV}} := |\mu|(E) = \mu^+(E) + \mu^-(E), \quad \text{for } \mu \in \mathcal{M}(E), \tag{2.10}$$

where $\mu = \mu^+ - \mu^-$ is the Hahn–Jordan decomposition of μ. It can be expressed equivalently in terms of μ acting on elements of $B(\mathcal{F}, \mathbb{R})$ as

$$\|\mu\|_{\mathsf{TV}} := \sup\Big\{ \int_E f \,\mathrm{d}\mu \,:\, f \in B(\mathcal{F}, \mathbb{R}),\, \|f\|_\infty \le 1 \Big\}. \tag{2.11}$$

Now, the subset of positive measures $\mathcal{M}_+(E)$ is a proper closed convex cone in $\mathcal{M}(E)$. The probability measures on $(E, \mathcal{F})$, denoted by $\mathcal{P}(E)$, are a subset of $\mathcal{M}_+(E)$. In (2.4), following ideas from [37,76], we provided an equivalent definition of the Hilbert metric using duality. This is not a convenient approach when viewing $\mathcal{M}(E)$ as the Banach space $\big(\mathcal{M}(E), \|\cdot\|_{\mathsf{TV}}\big)$: for one thing, when dealing with signed measures, one usually prefers to work with a *predual* space instead of the dual.

Taking the predual point of view, we could consider $\mathcal{M}(E)$ as a subset of $X = C_b(E, \mathbb{R})^*$, which is a real Banach space under the operator norm. When E is a Polish space with the Borel σ-algebra $\mathcal{B}(E)$, this amounts to a linear isometric embedding that is weak*-dense. Equipping X with the weak*-topology, rather than the operator norm, we get that X is a LCS and $X^* = C_b(E, \mathbb{R})$, so one expects a predual characterisation of the cone of positive measures in terms of $C_b(E, \mathbb{R})$. This, however, does not immediately follow from the procedure that led to (2.4). Instead, we give here a direct argument for the desired characterization (2.12), where one can think of $C_b(E, \mathbb{R}_+)$ as the 'predual cone', in analogy with (2.3). In fact, we can further restrict the space to bounded Lipschitz functions.

Proposition 2.3.1. *Let E be a Polish space with Borel σ-algebra $\mathcal{B}(E)$ and let $C_{bL}(E, \mathbb{R}_+)$ be the space of bounded Lipschitz functions $E \to \mathbb{R}_+$. Then we can characterize the space of positive measures $\mathcal{M}_+(E)$ as*

$$\mathcal{M}_+(E) = \big\{\mu \in \mathcal{M}(E) : \langle \mu, f\rangle \ge 0,\ \ \forall f \in C_{bL}(E, \mathbb{R}_+)\big\}. \tag{2.12}$$

If E is only metrizable, then replace $C_{bL}(E, \mathbb{R}_+)$ with $C_b(E, \mathbb{R}_+)$.

Proof. Let M denote the right-hand side of (2.12). We want to show that $\mathcal{M}_+(E) = M$. By non-negativity of the elements of $\mathcal{M}_+(E)$ and $C_{bL}(E, \mathbb{R}_+)$ (resp. $C_b(E, \mathbb{R}_+)$),

it is obvious that $\mathcal{M}_+(E) \subseteq M$. For the opposite inclusion, suppose $\mu \notin \mathcal{M}_+(E)$. By the Hahn–Jordan decomposition, there exist disjoint sets $P, N \subset E$ such that $P \cup N = E$, and (Borel) measures μ^+ and μ^- such that μ^+ is supported on P and μ^- is supported on N. Since $\mu \notin \mathcal{M}_+$, we have $\mu^-(N) > 0$. Recall that μ, μ^+ and μ^- are regular, as they are Borel measures on a metric space E (see e.g [16, Thm. 7.1.7]). Take $0 < \varepsilon < \mu^-(N)/4$. By regularity of μ^-, we can find a closed set $A_\varepsilon^- \subset N$ such that $\mu^-(N \setminus A_\varepsilon^-) < \varepsilon$. Likewise, there exists a closed set $A_\varepsilon^+ \subset P$ such that $\mu^+(P \setminus A_\varepsilon^+) < \varepsilon$. Note that $A_\varepsilon^+ \cap A_\varepsilon^- = \emptyset$, since they are respectively the subsets of disjoint sets P and N. For E Polish, we can take the sets $A_\varepsilon^-, A_\varepsilon^+$ to be compact (again, [16, Thm. 7.1.7]), so the Lipschitz version of Urysohn's Lemma [27, Prop. 2.1.1] (resp. Urysohn's Lemma [27, Thm. 1.2.10]) guarantees that there exists $f \in C_{bL}(E, \mathbb{R}_+)$ (resp. $C_b(E, \mathbb{R}_+)$) taking values in $[0, 1]$ with

$$f(x) = \begin{cases} 0 & \text{for } x \in A_\varepsilon^+, \\ 1 & \text{for } x \in A_\varepsilon^-. \end{cases}$$

Integrating f against μ we have

$$\begin{aligned} \int_E f \, \mathrm{d}\mu = \int_E f \, \mathrm{d}\mu^+ - \int_E f \, \mathrm{d}\mu^- &\leq \mu^+(P \setminus A_\varepsilon^+) - \mu^-(A_\varepsilon^-) \\ &\leq 2\varepsilon - \mu^-(N) \leq -\frac{\mu^-(N)}{2} < 0. \end{aligned}$$

Consequently, we have $f \in C_{bL}(E, \mathbb{R}_+)$ (resp. $C_b(E, \mathbb{R}_+)$), but $\langle \mu, f \rangle < 0$, so $\mu \notin M$. Therefore $M \subseteq \mathcal{M}_+(E)$, and the two sets are equal. □

When E is a Polish space, the above proposition is all we need to define the Hilbert projective (pseudo-)metric on $\mathcal{M}_+(E)$ in terms of bounded positive Lipschitz functions in $C_{bL}(E, \mathbb{R}_+)$, analogously to (2.4). On the other hand, if we do not want to assume E to be Polish, or even metric, we need to enlarge the set of test-functions for the construction of the Hilbert metric to still make sense. Similar to Proposition 2.3.1, we find the following (trivial) characterization of $\mathcal{M}_+(E)$ in terms of bounded positive measurable functions.

Proposition 2.3.2. *Let $(E, \mathcal{F})$ be a measurable space and let $B(\mathcal{F}, \mathbb{R}_+)$ be the space of $\mathcal{F}$-measurable bounded functions taking values in $\mathbb{R}_+$. Then we have*

$$\mathcal{M}_+(E) = \big\{\mu \in \mathcal{M}(E) : \langle f, \mu \rangle \geq 0, \ \ \forall f \in B(\mathcal{F}, \mathbb{R}_+)\big\}. \tag{2.13}$$

Proof. Let M' be the right-hand side of (2.13). By non-negativity of the functions in $B(\mathcal{F}, \mathbb{R}_+)$, $\mu \in \mathcal{M}_+(E)$ implies $\mu \in M'$. Conversely, assume $\mu \notin \mathcal{M}_+(E)$. Using

the Hahn–Jordan decomposition, take $N \in \mathcal{F}$ such that $\mu(N) = -\mu^-(N) < 0$. Let $f := \mathbb{1}_N \in B(\mathcal{F}, \mathbb{R}_+)$. Then $\langle f, \mu\rangle < 0$, but $\langle f, \nu\rangle \geq 0$ for all $\nu \in \mathcal{M}_+(E)$. Hence $\mu \notin M'$, and we are done. □

Proposition 2.3.3. *Let $(E, \mathcal{F})$ be a measurable space. Write $\mathcal{S} = C_{bL}(E, \mathbb{R}_+)$ if E is Polish (with $\mathcal{F}$ the corresponding Borel σ-algebra), $\mathcal{S} = C_b(E, \mathbb{R}_+)$ if E is metrizable, or $\mathcal{S} = B(\mathcal{F}, \mathbb{R}_+)$ otherwise. Then the Hilbert projective pseudo-metric can be written as follows: for $\mu, \nu \in \mathcal{M}_+(E) \setminus \{0\}$,*

$$\mathcal{H}(\mu, \nu) = \sup_{\substack{f,g \in \mathcal{S} \\ \langle f,\mu\rangle, \langle g,\nu\rangle \neq 0}} \left\{ \log \frac{\langle f, \nu\rangle \langle g, \mu\rangle}{\langle f, \mu\rangle \langle g, \nu\rangle} \right\}. \tag{2.14}$$

Proof. Consider any $\mu, \nu \in \mathcal{M}_+(E)$, and take $r \in \mathbb{R}_+$ such that $r\mu - \nu \notin \mathcal{M}_+(E)$. Using Proposition 2.3.1 and Proposition 2.3.2, there is an $f \in \mathcal{S}$ such that $\langle f, r\mu - \nu\rangle < 0$. Then a calculation analogous to (2.5) gives equivalence between (2.14) and the original definition of the Hilbert metric (2.1). □

Now, a natural question to ask is under which conditions $\mathcal{H}(\mu, \nu)$ is finite. For $\mu, \nu \in \mathcal{M}_+(E)$, let

$$\beta(\mu, \nu) = \sup_{\substack{f \in B(\mathcal{F}, \mathbb{R}_+) \\ \langle f,\mu\rangle \neq 0}} \left\{ \frac{\langle f, \nu\rangle}{\langle f, \mu\rangle} \right\}. \tag{2.15}$$

Clearly, $H(\mu, \nu) < \infty$ if and only if $\beta(\mu, \nu), \beta(\nu, \mu) < \infty$. We see immediately that if there exists an unbounded measurable function $h \in L^0(\mathcal{F}, \mathbb{R}_+)$ such that $\langle h, \mu\rangle < \infty$ but $\langle h, \nu\rangle = \infty$, then we can take a sequence of bounded functions $h_n \in B(\mathcal{F}, \mathbb{R}_+)$ such that $h_n \to h$, and the right-hand side of (2.15) is infinite. Consequently, if, for example, ν has a strictly smaller number of finite moments than μ, then $\beta(\mu, \nu) = \infty$, and conversely if ν has (strictly) more finite moments, then $\beta(\nu, \mu) = \infty$. Thus, to have $\mathcal{H}(\mu, \nu) < \infty$, we need a condition on μ, ν that is quite a lot stronger than simple equivalence of measures (which we denote as usual by $\sim$). This condition, which we call *comparability* again, in accordance with Birkhoff, and denote by $\overset{\text{cmp}}{\sim}$, has already been stated in [63, Def. 3.1] and [8]. Here we derive it directly from the 'predual' formulation (2.14).

Let $\mu \sim \nu$, with Radon-Nikodym derivatives $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. For all $\varphi \in B(\mathcal{F}, \mathbb{R}_+)$ we have $\langle \varphi, \nu\rangle = \langle \varphi \frac{\mathrm{d}\nu}{\mathrm{d}\mu}, \mu\rangle \leq \| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \|_{L^\infty(\mu)} \langle \varphi, \mu\rangle$, so

$$M_\varphi := \sup_{\substack{\varphi \in B(\mathcal{F}, \mathbb{R}_+) \\ \langle \varphi,\mu\rangle \neq 0}} \left\{ \frac{\langle \varphi, \nu\rangle}{\langle \varphi, \mu\rangle} \right\} \leq \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_{L^\infty(\mu)}.$$

On the other hand,

$$\left\langle \varphi \frac{\mathrm{d}\nu}{\mathrm{d}\mu}, \mu \right\rangle = \langle \varphi, \nu \rangle \leq \langle \varphi, \mu \rangle M_\varphi,$$

so $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \leq M_\varphi$ μ-a.e., and in particular $\left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_{L^\infty(\mu)} \leq M_\varphi$. Therefore, $\beta(\mu, \nu) = \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_{L^\infty(\mu)}$, and hence we can state our comparability condition as follows:

Definition 2.3.4. Let $(E, \mathcal{F})$ be a measurable space. Two positive measures $\mu, \nu \in \mathcal{M}_+(E)$ are *comparable* if $\mu \sim \nu$ and their Radon-Nikodym derivatives $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ are essentially bounded, i.e. $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in L^\infty(\nu)$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty(\mu)$. Equivalently, μ and ν are comparable if there exists scalars $q, r > 0$ such that

$$q\mu(A) \leq \nu(A) \leq r\mu(A), \quad \forall A \in \mathcal{F}. \tag{2.16}$$

Then the Hilbert projective pseudo-metric for $\mu, \nu \in \mathcal{M}_+(E)$ can be defined as

$$\mathcal{H}(\mu, \nu) = \log \left(\left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_\infty \right) = \sup_{\substack{A,B \in \mathcal{F} \\ \nu(A)>0,\, \mu(B)>0}} \left\{ \log \frac{\nu(B)\mu(A)}{\mu(B)\nu(A)} \right\}, \quad \text{if } \mu \overset{\mathsf{cmp}}{\sim} \nu, \tag{2.17}$$

and $\mathcal{H}(\mu, \nu) = \infty$ otherwise, and these definitions are equivalent to (2.14). The right-most formulation of (2.17) is the definition chosen, for example, by Le Gland and Oudjane in [63, Def. 3.3] and Atar and Zeitouni in [8]. Note that if $\mu, \nu \in \mathcal{P}(E)$, then $\left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty, \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_\infty \geq 1$ since μ and ν must integrate to 1, and hence also $\left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty, \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_\infty \leq e^{\mathcal{H}(\mu,\nu)}$ by (2.17). Then, for $\mu, \nu \in \mathcal{P}(E)$, $\mu \sim \nu$, and an arbitrary $f \in B(\mathcal{F}, \mathbb{R})$ such that $\|f\|_\infty \leq 1$, we have

$$\begin{aligned}
|\langle f, \mu - \nu \rangle| &\leq \int_{\left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \geq 1 \right\}} |f| \Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1 \Big) \, \mathrm{d}\nu + \int_{\left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\nu} < 1 \right\}} |f| \Big(1 - \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \Big) \, \mathrm{d}\nu \\
&\leq \Big(\left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty - 1 \Big) \nu\big(\big\{ \tfrac{\mathrm{d}\mu}{\mathrm{d}\nu} \geq 1 \big\} \big) + \Big(1 - \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_\infty^{-1} \Big) \nu\big(\big\{ \tfrac{\mathrm{d}\mu}{\mathrm{d}\nu} < 1 \big\} \big) \\
&\leq e^{\mathcal{H}(\mu,\nu)} - 1.
\end{aligned}$$

Together with the fact that the total variation distance between two probability measures is at most 2, this yields the following bound, first shown by Atar and Zeitouni [8, Lemma 1]:

$$\|\mu - \nu\|_{\mathsf{TV}} \leq \frac{2}{\log 3} \mathcal{H}(\mu, \nu). \tag{2.18}$$

This bound is clearly not sharp, since the right-hand side can easily be much larger than 2. We will improve it in Corollary 2.5.2.1. For now, we use Atar and Zeitouni's result to prove the following Lemma.

Lemma 2.3.5. *Let $(E, \mathcal{F})$ be a measurable space. Then $(\mathcal{P}(E), \mathcal{H})$ is a complete metric space.*

Proof. Note that two probability measures $\mu, \nu \in \mathcal{P}(E)$ which are collinear must be necessarily equal, so $\mathcal{H}$ is a metric on $\mathcal{P}(E)$. Let $(\mu_n) \in \mathcal{P}(E)$ be a Cauchy sequence for the Hilbert metric $\mathcal{H}$. Then (μ_n) is also Cauchy in total variation norm by (2.18), so $\mu_n \to \mu \in \mathcal{P}(E)$ in $\|\cdot\|_{\mathsf{TV}}$ since $\mathcal{P}(E)$ is complete as it is a closed subset of $\mathcal{M}(E)$. Since $\langle f, \mu_n \rangle \to \langle f, \mu \rangle$ for $f \in B(E, \mathbb{R})$ if $\mu_n \to \mu$ in total variation norm, (2.14) gives that $\mathcal{H}$ is lower semi-continuous with respect to $\|\cdot\|_{\mathsf{TV}}$, as a supremum over continuous functions. Hence $\mathcal{H}(\mu_n, \mu) \leq \liminf_{k\to\infty} \mathcal{H}(\mu_n, \mu_k)$, where the right-hand side goes to 0 as $n \to \infty$ by the Cauchy assumption. □

Corollary 2.3.5.1. *Let $(E, \mathcal{F})$ be a measurable space. Then $(\mathcal{P}(E), \mathcal{T})$ is a complete metric space.*

We have seen in Section 2.2 that many of the interesting properties of the Hilbert metric also hold for its transformation $\mathcal{T}$. The following gives a key reason why the classic Hilbert pseudo-metric is also of interest: $\mathcal{H}$ turns the space of probability measures comparable to a reference measure ρ into a normed vector space (with a modified algebra).

Proposition 2.3.6. *Let $(E, \mathcal{F})$ be a measurable space and fix a reference measure $\rho \in \mathcal{M}_+(E)$. We let essential infima and suprema be defined with respect to the nullsets of ρ, and consider*

(i) the space of measures comparable to ρ, namely $\mathcal{M}_\rho := \{\mu \in \mathcal{M}_+(E) : \frac{\mathrm{d}\mu}{\mathrm{d}\rho}, \frac{\mathrm{d}\rho}{\mathrm{d}\mu} \in L^\infty(\rho)\}$, and $\mathcal{P}_\rho := \mathcal{M}_\rho \cap \mathcal{P}(E)$;

(ii) the equivalence relation $\sim_{\mathrm{coll}}$ on $\mathcal{M}_\rho$ given by collinearity, that is $\mu \sim_{\mathrm{coll}} \nu \Leftrightarrow \mu = c\nu$ for some $c > 0$; note that $\mathcal{P}_\rho$ is isomorphic to $\mathcal{M}_\rho / \sim_{\mathrm{coll}}$ (as it is a selection of a unique element from each equivalence class);

(iii) the equivalence relation $\sim_{\mathrm{const}}$ on $L^\infty(\rho)$ given by $f \sim_{\mathrm{const}} g \Leftrightarrow f = g + c$ ρ-a.e. for some $c \in \mathbb{R}$; and the associated quotient space $\Theta_\rho := L^\infty(\rho) / \sim_{\mathrm{const}}$.

Then the map

$$\|\cdot\|_\Theta : L^\infty(\rho) \to \mathbb{R}; \quad f \mapsto \operatorname*{ess\,sup}_{x\in E} f(x) - \operatorname*{ess\,inf}_{x\in E} f(x),$$

defines a seminorm on $L^\infty(\rho)$, and a norm on Θ_ρ. Moreover, the map

$$\theta : \mathcal{M}_\rho \to L^\infty(\rho); \quad \mu \mapsto \log(\,\mathrm{d}\mu / \,\mathrm{d}\rho)$$

is an isomorphism of the pseudo-metric spaces $(\mathcal{M}_\rho, \mathcal{H})$ *and* $(L^\infty(\rho), \|\cdot\|_\Theta)$*, satisfying*

$$\mathcal{H}(\mu,\nu) = \|\theta(\mu) - \theta(\nu)\|_\Theta, \qquad \text{for all } \mu,\nu \in \mathcal{M}_\rho,$$

and it is an isomorphism of the metric spaces $(\mathcal{P}_\rho, \mathcal{H})$ *and* $(\Theta_\rho, \|\cdot\|_\Theta)$*. In particular,* $(\mathcal{P}_\rho, \mathcal{H})$ *is a normed vector space, when endowed with the algebra of (renormalized) addition and scalar multiplication of log-densities.*

Proof. It is easy to see that $\|\cdot\|_\Theta$ is absolutely homogeneous. From (2.17), we know that

$$\begin{aligned}
\mathcal{H}(\mu,\nu) &= \log\left(\left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_\infty\right) + \log\left(\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right\|_\infty\right) \\
&= \operatorname{ess\,sup}\left\{\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right\} - \operatorname{ess\,inf}\left\{\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\right\} \\
&= \operatorname{ess\,sup}\Big\{\theta(\mu) - \theta(\nu)\Big\} - \operatorname{ess\,inf}\Big\{\theta(\mu) - \theta(\nu)\Big)\Big\} \\
&= \|\theta(\mu) - \theta(\nu)\|_\Theta.
\end{aligned}$$

From this it follows that $\|\cdot\|_\Theta$ is sublinear (as $\mathcal{H}$ satisfies the triangle inequality), and is therefore a seminorm. It is easy to check that $\|\theta\|_\Theta = 0$ iff $\theta \sim_{\text{const}} 0$, so $\|\cdot\|_\Theta$ is a norm on the vector space $\Theta_\rho = L^\infty(\rho)/\sim_{\text{const}}$.

For $f \in \Theta_\rho$, the inverse of $\theta : \mathcal{M}_\rho \to L^\infty(\rho)$ is given by

$$\theta^{-1}(f)(A) = \int_A \exp(f(x))\,\mathrm{d}\rho, \qquad \forall A \in \mathcal{F},$$

so θ is clearly a bijection, and hence an isomorphism of $(\mathcal{M}_\rho, \mathcal{H})$ and $(L^\infty(\rho), \|\cdot\|_\Theta)$. Similarly, taking account of the equivalence relation, the inverse of $\theta : \mathcal{P}_\rho \to \Theta_\rho$ is given, for $f \in \Theta_\rho$, by

$$\theta^{-1}(f + c)(A) = \frac{\int_A \exp(f(x))\,\mathrm{d}\rho}{\int_E \exp(f(x))\,\mathrm{d}\rho}, \qquad \forall A \in \mathcal{F},$$

which clearly does not depend on the choice of $c \in \mathbb{R}$ (and hence is well defined on $\Theta_\rho = L^\infty(\rho)/\sim_{\text{const}}$). It follows that θ is an isomorphism of $(\mathcal{P}_\rho, \mathcal{H})$ and $(\Theta_\rho, \|\cdot\|_\Theta)$.

Finally, as $(\Theta_\rho, \|\cdot\|_\Theta)$ is a normed vector space, we simply observe that addition and scalar multiplication in Θ_ρ correspond to (renormalized) addition and scalar multiplication of log-densities. □

Remark 2.5. We will see in Section 2.4 that, when E is finite, we can avoid the equivalence relation above by selecting the unique representative $\theta_0(\mu)$ which satisfies $\theta_0(\mu)(x_0) = 0$ for a fixed $x_0 \in E$ (see Remark 2.10). This does not work as cleanly in infinite state spaces, as the value at a single point is typically not well defined when functions are only specified ρ-a.e.

Remark 2.6. Proposition 2.3.6 also helps us to understand the topology of $(\mathcal{P}(E), \mathcal{H})$. For every $\rho \in \mathcal{P}(E)$, we have the corresponding vector space $\mathcal{P}_\rho$ (and any $\rho' \in \mathcal{P}_\rho$ will yield $\mathcal{P}_{\rho'} = \mathcal{P}_\rho$). As they are normed vector spaces (with an appropriate algebra), these sets are both closed and open, and give a disconnected partition of $\mathcal{P}(E)$. In other words, $(\mathcal{P}(E), \mathcal{H})$ has the topology of a disjoint union of normed vector spaces (which may have different dimensions).

We conclude this (rather lengthy) section about the Hilbert metric on probability measures with a few important observations, which motivate why we started looking carefully at the 'predual' formulation of the Hilbert metric in the first place.

Remark 2.7. Take $\mu, \nu \in \mathcal{M}_+(E)$, with E Polish. Consider distances of the form

$$D(\mu, \nu) = \sup\Big\{ \Big| \int_E f \,\mathrm{d}(\mu - \nu) \Big| \,:\, f \in X \Big\}, \quad X \subseteq C_{bL}(E, \mathbb{R}_+).$$

Different conditions on $\|f\|_\infty$ and $\|f\|_{Lip}$ yield different metrics: the total variation norm (2.11) if one imposes $\|f\|_\infty \leq 1$, the bounded-Lipschitz distance by taking $\|f\|_\infty + \|f\|_{Lip} \leq 1$, or the 1-Wasserstein distance $\mathcal{W}_1$ (when μ and ν are additionally taken to have finite first moment) by imposing $\|f\|_{Lip} \leq 1$. This differentiation based on the choice of test-functions is *completely lost* when we work with the Hilbert metric. If we restricted our space to $\mathcal{P}_1(E)$ (the probability measures on E with finite first moment), for example, we could of course characterize our 'predual' cone of test-functions using (unbounded) positive Lipschitz functions $Lip(E, \mathbb{R}_+)$, analogously to the Kantorovich–Rubinstein dual formulation of $\mathcal{W}_1$. However, this would not yield a different metric from (2.14). Since any Lipschitz function can be approximated from below by bounded Lipschitz functions, if $\mu \overset{\text{cmp}}{\sim} \nu$ and $\mu, \nu \in \mathcal{P}_1(E)$, taking the supremum over $Lip(E, \mathbb{R}_+)$ or $C_{bL}(E, \mathbb{R}_+)$ does not change the Hilbert distance.

In the wake of the above remark, we deduce that the Hilbert metric is stronger than the p-Wasserstein distance $\mathcal{W}_p$. Let (E, d) be a metric space and $\{\mu_n\} \in \mathcal{P}_p(E)$ a sequence of probability measures with finite p^{th}-moment such that $\mu_n \to \mu$ in Hilbert metric. By (2.18), convergence in Hilbert metric implies convergence in total variation norm, which in turn implies $\mu_n \to \mu$ weakly. Moreover, $\mu \in \mathcal{P}_p(E)$, by definition of the Hilbert metric and comparability of measures (2.16). Fix an arbitrary $x_0 \in E$. Then an argument similar to the one that lead to (2.18) yields that, for all $q \leq p$,

$$\Big| \int_E d(x_0, x)^q \,\mathrm{d}(\mu_n - \mu) \Big| \leq K_q \big(e^{\mathcal{H}(\mu_n, \mu)} - 1 \big), \tag{2.19}$$

where $K_q < \infty$ is the q^{th}-moment of $\mu \in \mathcal{P}_p(E)$. So convergence of moments is preserved under convergence in the Hilbert metric, and thus convergence in the Hilbert metric implies convergence in $\mathcal{W}_p$.

The Kantorovich–Rubinstein dual formulation in particular yields the following bound for the $\mathcal{W}_1$ distance with respect to $\mathcal{H}$. For $\mu, \nu \in \mathcal{P}_1(E)$, and an arbitrary $x_0 \in E$, we have

$$\mathcal{W}_1(\mu, \nu) \leq \left(e^{\mathcal{H}(\mu,\nu)} - 1\right) \int_E d(x_0, x) \, \mathrm{d}\mu, \tag{2.20}$$

or analogously in terms of the first moment of ν.

Finally, our work so far, the definition of the Hilbert metric and of comparability of measures (2.16), all clearly emphasise that convergence in the Hilbert metric is a very strong form of convergence. The Hilbert projective metric not only dominates TV and $\mathcal{W}_p$, but also the Kullback–Leibler divergence (or relative entropy):

$$D_{KL}(\mu \| \nu) := \int_E \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \, \mathrm{d}\mu \leq \log \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|_\infty \leq \mathcal{H}(\mu, \nu). \tag{2.21}$$

In fact, one can show that the Hilbert metric dominates all f-divergences.

Definition 2.3.7 (f-divergence)**.** Let $f : \mathbb{R}_+ \to (-\infty, \infty]$ be a convex function with $f(1) = 0$, and $f(x) < \infty$ for all $x > 0$. Let $\mu, \nu \in \mathcal{P}(E)$, with $\mu \ll \nu$. Then the f-divergence of μ from ν, denoted by $D_f(\mu \| \nu)$, is given by

$$D_f(\mu \| \nu) = \int_E f\Big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\Big) \, \mathrm{d}\nu. \tag{2.22}$$

Remark 2.8. Total variation distance, Kullback–Leibler divergence, Jensen–Shannon divergence, squared Hellinger distance, α-divergence and χ^2-divergence are all examples of f-divergences.

Proposition 2.3.8. *Let $(E, \mathcal{F})$ be a measurable space, and let $\{\mu_n\} \in \mathcal{P}(E)$ be a sequence of probability measures converging to $\mu \in \mathcal{P}(E)$ in $\mathcal{H}$. Then, for any f-divergence D_f, $D_f(\mu_n \| \mu) \to 0$ and $D_f(\mu \| \mu_n) \to 0$ as $n \to \infty$ also.*

Proof. Let f be a convex function of the form specified in Definition 2.3.7, and let $D_f(\mu \| \nu)$ be the associated f-divergence of μ from ν, where $\mu, \nu \in \mathcal{P}(E)$ and $\mu \ll \nu$. Note that D_f is unchanged if we add a linear term to f, i.e. let $\bar{f}(u) = f(u) + c(u-1)$, then $D_{\bar{f}}(\mu \| \nu) = D_f(\mu \| \nu)$. Moreover, by taking $c \in -\partial f(1)$ (where by ∂ we denote the subgradient of f), we have $0 \in \partial \bar{f}(1)$, so without loss of generality we can restrict our attention to convex functions f such that $f(1) = 0 \in \partial f(1)$.

Consider a sequence $\{\mu_n\} \in \mathcal{P}(E)$ such that $\lim_{n\to\infty} \mathcal{H}(\mu_n, \mu) = 0$. Then there exists $N > 0$ such that for all $n \geq N$, $\mu_n \sim \nu$. Since f must be decreasing for $x < 1$ and increasing for $x > 1$ by virtue of being convex, we have, for all $n \geq N$,

$$\begin{aligned} D_f(\mu_n \| \mu) &= \int_{\left\{\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu} \leq 1\right\}} f\Big(\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\Big)\,\mathrm{d}\mu + \int_{\left\{\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu} > 1\right\}} f\Big(\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\Big)\,\mathrm{d}\mu \\ &\leq f\Big(\operatorname*{ess\,inf}_{x\in E} \frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\Big)\mu\big(\{\tfrac{\mathrm{d}\mu_n}{\mathrm{d}\mu} \leq 1\}\big) + f\Big(\operatorname*{ess\,sup}_{x\in E} \frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\Big)\mu\big(\{\tfrac{\mathrm{d}\mu_n}{\mathrm{d}\mu} > 1\}\big) \\ &\leq \max\left\{ f\Big(\Big\|\frac{\mathrm{d}\mu}{\mathrm{d}\mu_n}\Big\|_\infty^{-1}\Big),\, f\Big(\Big\|\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\Big\|_\infty\Big) \right\} \\ &\leq \max\left\{ f\big(e^{-\mathcal{H}(\mu_n,\mu)}\big),\, f\big(e^{\mathcal{H}(\mu_n,\mu)}\big) \right\}, \end{aligned}$$

where we have used that $\big\|\frac{\mathrm{d}\mu}{\mathrm{d}\mu_n}\big\|_\infty, \big\|\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\big\|_\infty \geq 1$, and $\big\|\frac{\mathrm{d}\mu}{\mathrm{d}\mu_n}\big\|_\infty^{-1} \geq e^{-\mathcal{H}(\mu_n,\mu)}$ and $\big\|\frac{\mathrm{d}\mu_n}{\mathrm{d}\mu}\big\|_\infty \leq e^{\mathcal{H}(\mu_n,\mu)}$. Since $f(1) = 0$ by assumption, the right-hand side of the above goes to 0 as $\mathcal{H}(\mu_n, \mu) \to 0$, so $D_f(\mu_n\|\mu)$ converges. The argument for $D_f(\mu\|\mu_n)$ is analogous by symmetry, and we are done. □

2.4 Hilbert projective geometry on the probability simplex

In this section we consider the Hilbert metric on the probability measures with finite state-space $E \cong \{0, \cdots, n\}$, which form the *probability simplex*. In this case, the form of the Hilbert metric simplifies, and there exists an explicit expression for Birkhoff's contraction coefficient (see Section 4 of [78, Chapter 3]). We briefly remark on this, and present a short derivation of Birkhoff's coefficient using duality. Then we move on to studying the geometry of the probability simplex under the Hilbert projective metric: using a coordinate transformation inspired by information geometry [2, 3], we describe the Hilbert balls as convex polytopes in the probability simplex, in an extension of the work in [73] to the n-dimensional case.

Let $E \cong \{0, \cdots, n\}$. The probability measures $\mathcal{P}(E)$ are represented by the set

$$\mathcal{P}(E) \cong \mathcal{S}^n = \left\{ x \in \mathbb{R}^{n+1} \,:\, \sum_{i=0}^{n} x_i = 1 \right\} \subset \mathbb{R}^{n+1},$$

and $\mathcal{S}^n$ is called the *n-dimensional probability simplex*. It is given by the intersection of the convex cone of non-negative vectors $\mathbb{R}^{n+1}_+$ with the plane $\sum_i x_i = 1$.

Consider the Hilbert distance on the non-negative orthant $C = \mathbb{R}^n_+$. Recall the duality expression for $\mathcal{H}$ given in (2.4) and the equality (2.5). The dual cone C^* is

again $\mathbb{R}^n_+$. Take $x, y \in \mathbb{R}^n_+ \setminus \{0\}$ such that $\beta(x,y) < \infty$. Note that $\beta(x,y) < \infty$ if and only if there exists a scalar $b > 0$ such that $y^i \leq bx^i$ for all $i = 1, \ldots, n$, which in particular implies that $x^i > 0$ whenever $y^i > 0$. Hence we have

$$\beta(x,y) = \sup_{\substack{w \in \mathbb{R}^n_+ \setminus \{0\} \\ w^\intercal x > 0}} \frac{w^\intercal y}{w^\intercal x} = \sup_{r \in [0,1]^{n+1} \setminus \{0\}} \sum_{j \colon x^j > 0} r^j \frac{y^j}{x^j} = \max_{j \colon x^j > 0} \frac{y^j}{x^j} = \max_{j \colon e_j^\intercal x > 0} \frac{e_j^\intercal y}{e_j^\intercal x}, \tag{2.23}$$

where the second equality follows by setting $r^j = w^j x^j / w^\intercal x$, $0 \leq r^j \leq 1$ for all $j = 1, \ldots, n$ and at least one $r^j > 0$, and $\{e_j\}_{j=1}^n$ are the basis vectors of $\mathbb{R}^n_+$. So $\sup_w w^\intercal y / w^\intercal x$ is attained when w is a basis vector. By symmetry, we have that $\beta(y,x) < \infty$ if there exists $b' > 0$ such that $x^i \leq b' y^i$ for all $i = 1, \ldots n$. Then two elements $x, y \in \mathbb{R}^n_+$ are *comparable* (denoted again by $\overset{\text{cmp}}{\sim}$) if there exist constants $a, b > 0$ such that $ax \leq y \leq bx$, where the inequalities hold component-wise (this is the definition of comparability given in [15, Chapter XVI]). Then the definition (2.1) of the Hilbert projective distance for $x, y \in \mathbb{R}^n_+$ simplifies to

$$\mathcal{H}(x,y) = \begin{cases} \log \left(\dfrac{\max_{i : y^i > 0} \frac{x^i}{y^i}}{\min_{j : y^j > 0} \frac{x^j}{y^j}} \right), & x \overset{\text{cmp}}{\sim} y, \\ \infty, & x \overset{\text{cmp}}{\not\sim} y. \end{cases} \tag{2.24}$$

Remark 2.9. In this finite-state context, the comparability condition $\overset{\text{cmp}}{\sim}$ reduces to *equivalence of measures* $\sim$ on $\mathcal{S}^n$. Recall that $(\mathcal{S}^n, \mathcal{H})$ is a complete metric space by Lemma 2.3.5.

Using (2.23) we can now easily derive an explicit expression for Birkhoff's contraction coefficient of a linear operator $\mathbb{R}^n_+ \setminus \{0\} \to \mathbb{R}^n_+ \setminus \{0\}$. Define a matrix $A = (A_{ij})$ to be *allowable* if A is non-negative (i.e. $A_{ij} \geq 0$ for all i, j) and if every row and every column of A has at least one strictly positive element (this definition is given by Seneta in [78, Def. 3.1]). Clearly any linear operator $\mathbb{R}^n_+ \setminus \{0\} \to \mathbb{R}^n_+ \setminus \{0\}$ can be represented as an allowable $n \times n$ matrix. We prove the following result, which was already stated by Birkhoff without proof in Corollary 2 of [15, Chapter XVI, Section 3] and obtained by Seneta in Section 4 of [78, Chapter 3], although the derivation there is significantly more involved.

Proposition 2.4.1. *Let $A = (A_{ij})$ be an allowable $n \times n$ matrix. Birkhoff's contraction coefficient $\tau(A)$ can be written as*

$$\tau(A) = \frac{1 - \sqrt{\phi(A)}}{1 + \sqrt{\phi(A)}}, \quad \textit{with } \phi(A) = \min_{i,j,k,l} \frac{A_{ik} A_{jl}}{A_{jk} A_{il}}, \tag{2.25}$$

(with the convention that $0/0 = 1$).

Proof. Consider the diameter of $\mathbb{R}^n_+$ under the matrix A, i.e.

$$\Delta(A) = \sup_{x,y\in\mathbb{R}^n_+\setminus\{0\}} \mathcal{H}(Ax, Ay).$$

Assume that $\Delta(A) < \infty$, so $\beta(Ax, Ay), \beta(Ay, Ax) < \infty$ for all $x, y \in \mathbb{R}^n_+ \setminus \{0\}$. Note that $\Delta(A) < \infty$ if and only if A is strictly positive (i.e. $A_{ij} > 0$ for all i, j), so in particular Ax has strictly positive entries for all $x \in \mathbb{R}^n_+$, which implies that $w^\intercal Ax > 0$ for all $x, w \in \mathbb{R}^n_+$. Using (2.23) in the second and fourth equalities below, we get

$$\begin{aligned} e^{\Delta(A)} &= \sup_{x,y\in\mathbb{R}^n_+\setminus\{0\}} \sup_{w,z\in\mathbb{R}^n_+\setminus\{0\}} \left\{ \frac{w^\intercal Ay}{w^\intercal Ax} \frac{z^\intercal Ax}{z^\intercal Ay} \right\} = \sup_{x,y\in\mathbb{R}^n_+\setminus\{0\}} \max_{i,j} \left\{ \frac{e_i^\intercal Ay}{e_i^\intercal Ax} \frac{e_j^\intercal Ax}{e_j^\intercal Ay} \right\} \\ &= \max_{i,j} \sup_{x,y\in\mathbb{R}^n_+\setminus\{0\}} \left\{ \frac{y^\intercal A^\intercal e_i}{y^\intercal A^\intercal e_j} \frac{x^\intercal A^\intercal e_j}{x^\intercal A^\intercal e_i} \right\} = \max_{i,j} \max_{k,l} \left\{ \frac{e_k^\intercal A^\intercal e_i}{e_k^\intercal A^\intercal e_j} \frac{e_l^\intercal A^\intercal e_j}{e_l^\intercal A^\intercal e_i} \right\} \\ &= \max_{i,j,k,l} \frac{A_{ik}A_{jl}}{A_{jk}A_{il}}, \end{aligned}$$

so finally Birkhoff's contraction coefficient is given by

$$\tau(A) = \tanh\left(\frac{\Delta(A)}{4}\right) = \frac{e^{\frac{\Delta(A)}{2}} - 1}{e^{\frac{\Delta(A)}{2}} + 1} = \frac{1 - \sqrt{\phi(A)}}{1 + \sqrt{\phi(A)}}, \quad \text{with } \phi(A) = \min_{i,j,k,l} \frac{A_{ik}A_{jl}}{A_{jk}A_{il}}. \tag{2.26}$$

We can check that if A is not strictly positive, so $\Delta(A) = \infty$, the formula above still holds with the convention that $\tanh(\infty) := 1$. □

In fact, we obtain a similar representation of Birkhoff's contraction coefficient for a class of transition kernels on more general spaces. Let E be a Polish space with Borel σ-algebra $\mathcal{B}(E)$, and consider a positive kernel K on $\mathcal{B}(E) \times E$. Then there exists an associated positive linear operator (again denoted by K) such that $K : \mathcal{M}_+(E) \to \mathcal{M}_+(E)$ and

$$K\mu(\,\mathrm{d}a) = \int_E K(\,\mathrm{d}a, x)\,\mathrm{d}\mu(x). \tag{2.27}$$

In the statement below, we restict our attention to kernels that have a density with respect to a reference measure $\rho \in \mathcal{M}_+(E)$. In other words, let $K(\,\mathrm{d}a, x) = \kappa(a, x)\,\mathrm{d}\rho(a)$ for some positive function $\kappa : \mathrm{Supp}(\rho) \times E \to \mathbb{R}_+$. Then (2.27) reduces to

$$K\mu(A) = \int_{a\in A} \int_{x\in E} \kappa(a, x)\,\mathrm{d}\mu(x)\,\mathrm{d}\rho(a), \quad \forall A \in \mathcal{B}(E), \tag{2.28}$$

(where we have swapped the order of integration using Tonelli's theorem). We then have the following infinite dimensional counterpart to Proposition 2.4.1.

Proposition 2.4.2. *Let E be a Polish space with Borel σ-algebra $\mathcal{B}(E)$ and reference measure $\rho \in \mathcal{M}_+(E)$ with support $\mathrm{Supp}(\rho) \subset E$. Consider a kernel operator $K : \mathcal{M}_+(E) \to \mathcal{M}_+(E)$ of the form (2.28) defined by a density*

$$\frac{\mathrm{d}(K\mu)}{\mathrm{d}\rho}(a) := \int_E \kappa(a,x)\,\mathrm{d}\mu(x),$$

for $\kappa : \mathrm{Supp}(\rho) \times E \to (0,\infty)$. Assume κ is a bounded continuous function. Then the Birkhoff coefficient of K is given by

$$\tau(K) = \frac{1-\sqrt{\phi(K)}}{1+\sqrt{\phi(K)}}, \quad \textit{with} \quad \phi(K) = \inf_{\substack{a,b \in \mathrm{Supp}(\rho) \\ x,y \in E}} \left\{ \frac{\kappa(a,x)\kappa(b,y)}{\kappa(a,y)\kappa(b,x)} \right\}. \tag{2.29}$$

Proof. From the structure of the operator, we know that the Radon–Nikodym derivative of $K\mu$ and $K\nu$ (for $\nu \neq 0$) is given by

$$\frac{\mathrm{d}(K\mu)}{\mathrm{d}(K\nu)}(a) = \frac{\int_E \kappa(a,x)\,\mathrm{d}\mu(x)}{\int_E \kappa(a,x)\,\mathrm{d}\nu(x)}, \quad \forall a \in \mathrm{Supp}(\rho).$$

As κ is continuous and bounded, by dominated convergence we have that $a \mapsto \frac{\mathrm{d}(K\mu)}{\mathrm{d}(K\nu)}(a)$ is continuous, and so the ρ-essential supremum of $\frac{\mathrm{d}(K\mu)}{\mathrm{d}(K\nu)}(a)$ is equal to its (pointwise) supremum on $\mathrm{Supp}(\rho)$. Using the definition of $\mathcal{T}$ and (2.17), we know that

$$\begin{aligned}
\mathcal{T}(K\mu, K\nu) &= \tanh\left(\frac{1}{4}\log\Big(\sup_a \left\{\frac{\int_E \kappa(a,x)\,\mathrm{d}\mu(x)}{\int_E \kappa(a,x)\,\mathrm{d}\nu(x)}\right\} \sup_b \left\{\frac{\int_E \kappa(b,y)\,\mathrm{d}\nu(y)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)}\right\}\Big)\right) \\
&= \sup_{a,b}\left\{\tanh\left(\frac{1}{4}\log\Big(\frac{\int_E \kappa(a,x)\,\mathrm{d}\mu(x)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)}\frac{\int_E \kappa(b,y)\,\mathrm{d}\nu(y)}{\int_E \kappa(a,x)\,\mathrm{d}\nu(x)}\Big)\right)\right\}.
\end{aligned}$$

From the definition of τ in Theorem 2.2.6 and monotonicity of tanh, we know that

$$\begin{aligned}
\tau(K) &= \sup_{\mu,\nu \in \mathcal{M}_+(E)} \mathcal{T}(K\mu, K\nu) \\
&= \sup_{a,b}\left\{\tanh\left(\frac{1}{4}\log\Big(\sup_{\mu \in \mathcal{M}_+(E)} \left\{\frac{\int_E \kappa(a,x)\,\mathrm{d}\mu(x)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)}\right\} \sup_{\nu \in \mathcal{M}_+(E)} \left\{\frac{\int_E \kappa(b,y)\,\mathrm{d}\nu(y)}{\int_E \kappa(a,x)\,\mathrm{d}\nu(x)}\right\}\Big)\right)\right\}.
\end{aligned} \tag{2.30}$$

In order to compute the inner suprema, we observe that for all $\mu \in \mathcal{M}_+(E)$ and $a, b \in E$,

$$\frac{\int_E \kappa(a,x)\,\mathrm{d}\mu(x)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)} = \int_E \frac{\kappa(b,x)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)}\frac{\kappa(a,x)}{\kappa(b,x)}\,\mathrm{d}\mu(x) = \int_E \frac{\kappa(a,x)}{\kappa(b,x)}\,\mathrm{d}\tilde{\mu}_b(x)$$

where $\tilde{\mu}_b \in \mathcal{P}(E)$ is defined by $\frac{\mathrm{d}\tilde{\mu}_b}{\mathrm{d}\mu}(x) = \frac{\kappa(b,x)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)}$. Therefore, as κ is continuous,

$$\sup_{\mu \in \mathcal{M}_+(E)} \frac{\int_E \kappa(a,x)\,\mathrm{d}\mu(x)}{\int_E \kappa(b,y)\,\mathrm{d}\mu(y)} = \sup_{\tilde{\mu} \in \mathcal{P}(E)} \int_E \frac{\kappa(a,x)}{\kappa(b,x)}\,\mathrm{d}\tilde{\mu}(x) = \sup_{x \in E} \frac{\kappa(a,x)}{\kappa(b,x)}.$$

Substituting in (2.30) yields

$$\tau(K) = \tanh\Big(\frac{\Delta(K)}{4}\Big), \quad \text{with} \quad \Delta(K) = \sup_{\substack{a,b \in \mathrm{Supp}(\rho) \\ x,y \in E}} \Big\{ \log\Big(\frac{\kappa(a,x)\kappa(b,y)}{\kappa(a,y)\kappa(b,x)}\Big)\Big\},$$

and essentially the same calculation as (2.26) gives the form (2.29). □

2.4.1 Hexagonal polytopes in Hilbert projective geometry

We now introduce a change of coordinates from the interior of $\mathcal{S}^n$ to $\mathbb{R}^n$ which allows us to build an understanding of the geometry of the probability simplex as a metric space equipped with the Hilbert distance $\mathcal{H}$.

Notation. For simplicity, let $\mathbf{N} := \{0, 1, \dots, n\}$ throughout this section.

Let $\mathring{\mathcal{S}}^n$ be the interior of the n-dimensional probability simplex $\mathcal{S}^n \subset \mathbb{R}^{n+1}$. Following Amari [2,3], we map any discrete distribution $\mu \in \mathring{\mathcal{S}}^n$ to its natural parameters $\theta \in \mathbb{R}^n$. In other words, for all $k = 0, \dots, n$, let $\theta_k : \mathring{\mathcal{S}}^n \to \mathbb{R}^n$ such that

$$\theta_k^i(\mu) = \log \frac{\mu^i}{\mu^k}, \quad \forall i \in \mathbf{N} \setminus \{k\}. \tag{2.31}$$

Since we exclude the k^{th} component, $\theta_k(\mu) = \{\theta_k^i(\mu)\}_{i \neq k}$ is an n-dimensional vector. The inverse mapping $\theta_k^{-1} : \mathbb{R}^n \to \mathring{\mathcal{S}}^n$ is given by

$$\mu^i(\theta_k) = \frac{e^{\theta_k^i}}{\sum_{i=0}^n e^{\theta_k^i}}, \quad \forall i \in \mathbf{N} \quad \text{(where for notational simplicity } \theta_k^k \equiv 0). \tag{2.32}$$

Then θ_k is a diffeomorphism $\mathring{\mathcal{S}}^n \to \mathbb{R}^n$, and a global chart for $\mathring{\mathcal{S}}^n$. Note that we have $n+1$ choices for k, so in fact we have a family of $n+1$ coordinate transformations.

We note that we have the equivalence

$$\begin{aligned} \mathcal{H}(\mu,\nu) &= \max_i \log \frac{\mu^i}{\nu^i} - \min_i \log \frac{\mu^i}{\nu^i} = \max_{i,k} \Big\{ \log \frac{\mu^i}{\nu^i} - \log \frac{\mu^k}{\nu^k} \Big\} \\ &= \max_{i,k} \big\{ \theta_k^i(\mu) - \theta_k^i(\nu) \big\} = \max_k \|\theta_k(\mu) - \theta_k(\nu)\|_{\ell^\infty}, \end{aligned} \tag{2.33}$$

where by ℓ^∞ we denote the standard supremum norm between vectors.

Remark 2.10. It is informative to compare this with Proposition 2.3.6. We fix $k = 0$ for simplicity, and ρ as the counting measure on E. Writing $\mathbf{1} \in \mathbb{R}^n$ for the vector of ones, we then have

$$\begin{aligned} \theta_0(\mu) &= \big(\log(\mu^1/\mu^0), \log(\mu^2/\mu^0), ..., \log(\mu^n/\mu^0) \big) \\ &= \big(\log(\mu^1), \log(\mu^2), ..., \log(\mu^n) \big) - \log(\mu^0)\mathbf{1}. \end{aligned}$$

On the other hand, in the notation of Proposition 2.3.6, we have the equivalence class

$$\theta(\mu) = \Big\{ \big(\log(\mu^0), \log(\mu^1), ..., \log(\mu^n) \big) + c\mathbf{1};\ c \in \mathbb{R} \Big\} \in \Theta_\rho \cong \mathbb{R}^{n+1} \big/ \sim_{\text{const}} .$$

Of course, we can identify $\theta_0(\mu)$ with $\big(0, \theta_0(\mu)\big) \in \theta(\mu)$. Therefore, we see that our θ_0-coordinates (2.31) simply choose the representative element in $\theta(\mu)$ with $c = -\log(\mu^0)$, or equivalently, the (unique) element with 0 in the first entry. Consequently, Proposition 2.3.6 gives the representation of the metric (in θ_0-coordinates)

$$\mathcal{H}(\mu, \nu) = \Big(\max_{i \neq 0} \big\{ \theta_0^i(\mu) - \theta_0^i(\nu) \big\} \Big)^+ + \Big(\min_{i \neq 0} \big\{ \theta_0^i(\mu) - \theta_0^i(\nu) \big\} \Big)^-,$$

with $x^+ = \max\{0, x\}$ and $x^- = \max\{0, -x\}$. As this representation shows the Hilbert metric is given by a norm in θ_k-coordinates (for any k), we know that translation in θ_k-coordinates will not change the size or shape of a ball.

The θ_k-coordinates allow us to investigate in detail the shape of Hilbert balls in $\mathcal{S}^n$. The main idea is as follows: let $\mu, \nu \in \mathring{\mathcal{S}}^n$, so μ and ν are equivalent, and for all $k \in \mathbf{N}$ consider the transformations $\mu \mapsto \theta_k(\mu)$ and $\nu \mapsto \theta_k(\nu)$. Let $\mathcal{H}(\mu, \nu) = R < \infty$. Fixing ν and $k = 0$, we prove that the $\mathcal{H}$-ball of radius R around $\theta_0(\nu)$ is a convex polytope $\mathcal{C} \subset \mathbb{R}^n$. Mapping $\mathcal{C}$ to the simplex $\mathcal{S}^n$, we find that the image of $\mathcal{C}$ through the inverse transformation θ_0^{-1} is also a convex polytope.

We start with the following lemma.

Lemma 2.4.3. *Consider two probability measures $\mu, \nu \in \mathring{\mathcal{S}}^n$ such that $\mathcal{H}(\mu, \nu) = R > 0$. Let $\theta_0(\mu), \theta_0(\nu) \in \mathbb{R}^n$ be their natural parameters under the mapping θ_0 given by (2.31). Then $\theta_0(\mu)$ belongs to the boundary $\partial\mathcal{C}$ of an n-dimensional convex polytope $\mathcal{C} \subset \mathbb{R}^n$ centred at $\theta_0(\nu)$, with $2(2^n - 1)$ vertices at the points*

$$v_{\mathcal{I}}^+ := \theta_0(\nu) + R \sum_{i \in \mathcal{I}} e_i, \quad v_{\mathcal{I}}^- := \theta_0(\nu) - R \sum_{i \in \mathcal{I}} e_i, \tag{2.34}$$

where $\{e_i\}_{i=1}^n$ denote the basis vectors of $\mathbb{R}^n$ and $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$, $\mathcal{I} \neq \emptyset$.

Proof. The first thing we do, to simplify our calculations, is to translate $\theta_0(\nu) \in \mathbb{R}^n$ to the origin $\mathbf{0}$. Now let $\theta_k^i := \log(\mu^i / \mu^k) \in \mathbb{R}$ for all $i, k = 0, \ldots, n$. We fix the coordinate system in $\mathbb{R}^n$ to be given by $(x^1, \ldots, x^n) \equiv (\theta_0^1, \ldots, \theta_0^n)$, so that the basis vectors e_i are the unit vectors in the θ_0^i-direction. We look for a representation of the $\mathcal{H}$-ball of radius R around the origin in this coordinate system.

By (2.33), clearly $|\theta_k^i| \leq R$ for all pairs $(i, k) \in \mathbf{N} \times \mathbf{N}$. We consider all these inequalities, noting that, since $|\theta_k^i| = |\theta_i^k|$ by properties of log, we can avoid needless repetitions by restricting our consideration to all pairs of indices $(i, k) \in I^n :=$

$\{(i,k) \in \mathbf{N} \times \mathbf{N} : i > k\}$. Then we have in total $n(n+1)/2$ unique inequalities. Recalling that $\theta^i_k = \theta^i_0 - \theta^k_0$, for $(i,k) \in I^n$ we define the $(n-1)$-dimensional hyperplanes

$$\hbar^{\pm}_{i,k} := \begin{cases} \{(x^1, \ldots, x^n) \in \mathbb{R}^n : x^i - x^k = \pm R\} & k \neq 0, \\ \{(x^1, \ldots, x^n) \in \mathbb{R}^n : x^i = \pm R\} & k = 0, \end{cases} \tag{2.35}$$

and denote by $\wp^+_{i,k} := \{x \in \mathbb{R}^n : x^i - x^k \leq R\}$ the half-spaces bounded by $\hbar^+_{i,k}$, and similarly by $\wp^-_{i,k} := \{x \in \mathbb{R}^n : x^i - x^k \geq -R\}$ those bounded by $\hbar^-_{i,k}$ (and equivalently when $k = 0$).

We let $\mathcal{C}^n = \bigcap_{(i,k) \in I^n} \wp^+_{i,k} \cap \wp^-_{i,k}$. By standard results in n-dimensional geometry, $\mathcal{C}^n \subset \mathbb{R}^n$ is a polyhedron, since it is the intersection of a finite number of closed half-spaces. We claim $\mathcal{C}^n$ is bounded.

Note that the intersection $\mathcal{C}^n_0 = \bigcap_{i=1}^n \wp^+_{i,0} \cap \wp^-_{i,0}$ is the n-cube with side-length $2R$ centred at $\mathbf{0}$ with 2^n vertices at all possible positive/negative combinations of the coordinates $(\pm R, \ldots, \pm R)$. Then

$$\mathcal{C}^n = \bigcap_{\substack{(i,k) \in I^n \\ k \neq 0}} \wp^+_{i,k} \cap \wp^-_{i,k} \cap \mathcal{C}^n_0, \tag{2.36}$$

and the intersection of a hypercube with closed half-spaces is bounded, so $\mathcal{C}^n$ is a bounded polyhedron, and therefore a convex polytope. (Note that $\mathcal{C}^n$ is non-empty, since one can easily check that $\mathbf{0} \in \mathcal{C}^n$.)

We now would like to find the vertices of $\mathcal{C}^n$. We look for all the points in $\mathbb{R}^n$ where exactly n of the hyperplanes (2.35) intersect uniquely.

Consider a linear system S given by n equations from (2.35). We say indices i, j are *linked* if an equation of the form $x^i - x^j = \pm R$ appears in the system S, and extend this definition by transitivity to partition the indices appearing in S into *linked classes*. We say an index i is a *base case* if $x_i = \pm R$ appears in S.

1. Consider a class not containing a base case. Then we can add a constant $r \in \mathbb{R}$ to each component x^i in the class without altering the equations of the form $x^i - x^j = \pm R$. Therefore the subsystem of S containing all the equations for this class cannot have a unique solution, so S cannot give a vertex.

2. For any class containing a base case, let's say x^i, note that $(x^i - x^j)/R \in \mathbb{Z}$ for any x^j linked to x^i. By transitivity and additive closure of $\mathbb{Z}$, we observe that all indices in a class containing a base case must have $x^j/R \in \mathbb{Z}$.

By combining the above observations, all vertices of $\mathcal{C}^n$ must have coordinates that are integer multiples of R, with at least one coordinate given by a base-case, i.e. any

point $v \in \mathbb{R}^n$ that solves uniquely S must be of the form $(m_1R, \ldots, m_nR)$ for $m_i \in \mathbb{Z}$ with $-n \leq m_i \leq n$ for all $i = 1, \ldots, n$, and at least one $m_i \in \{\pm 1\}$.

Now, by (2.36) we must have that all the vertices of $\mathcal{C}^n$ belong to $\mathcal{C}_0^n$. Thus, any point $v \in \mathbb{R}^n$ that uniquely solves S and is potentially a vertex of $\mathcal{C}^n$ must be of the form $(m_1R, \ldots, m_nR)$ with $m_i \in \{-1, 0, 1\}$ for all $i = 1, \ldots, n$. Moreover, assume that $m_i = 1$ and $m_k = -1$ for $i > k$. Then $m_iR - m_kR = 2R \geq R$, so $v \notin p_{i,k}^+$ and $v \notin \mathcal{C}^n$. Similarly, if $m_i = -1$ and $m_k = 1$ for $i > k$, then $m_iR - m_kR \leq -2R$ so $v \notin p_{i,k}^-$ and $v \notin \mathcal{C}^n$.

Thus we must have that any vertex of $\mathcal{C}^n$ is of the form $(m_1R, \ldots, m_nR)$ with either $m_i \in \{-1, 0\}$ for all $i = 1, \ldots, n$ or $m_i \in \{0, 1\}$ for all $i = 1, \ldots, n$. Conversely, consider any point $v \in \mathbb{R}^n$ of this form. It is easy to construct a system S for a choice of n equations in $\{x^i - x^k = \pm R\} \cup \{x^j = \pm R\}$ such that v solves S. Then the points $(m_1R, \ldots, m_nR) \in \mathbb{R}^n$ with either $m_i \in \{-1, 0\}$ for all $i = 1, \ldots, n$ or $m_i \in \{0, 1\}$ for all $i = 1, \ldots, n$ (and not all m_i identically 0) are in fact all the vertices of $\mathcal{C}^n$. Letting $\mathcal{C} = \mathcal{C}^n + \theta_0(\nu)$, we are done. □

Remark 2.11. As can be deduced by (2.33), the $\mathcal{H}$-ball is, in a way, nothing but the intersection of $n+1$ skewed ℓ_∞-balls, or, geometrically speaking, the intersection of $n+1$ skewed hypercubes. Recall the notation from our proof above. Define the following intersections

$$\mathcal{C}_k^n := \Big[\bigcap_{j=k+1}^{n} p_{j,k}^+ \cap p_{j,k}^- \Big] \cap \Big[\bigcap_{j=0}^{k-1} p_{k,j}^+ \cap p_{k,j}^- \Big], \quad \forall k = 1, \ldots, n.$$

Then for each $k = 1, \ldots, n$, $\mathcal{C}_k^n$ is the image of the hypercube centred at $\theta_k(\nu)$ with side-length $2R$ under the linear transformation $\theta_k \mapsto \theta_0$.

We now map our convex polytope from $\mathbb{R}^n$ to S^n through the inverse transformation θ_0^{-1} given by (2.32). Note that if θ_0^{-1} were an affine transformation, then Lemma 2.4.4 stated below would be trivially true. However, as θ_0^{-1} is not affine, a bit more work is required to prove that convexity, linearity of the boundary, and intersections are preserved. We give a depiction of this result in Figure 2.1.

Lemma 2.4.4. *The $\mathcal{H}$-ball of radius R around $\theta_0(\nu)$, given by the convex polytope $\mathcal{C}$ of Lemma 4.2, maps to a convex polytope $\mathcal{D}$ centred at ν in $\mathring{\mathcal{S}}^n$ under the inverse transformation $\theta_0^{-1} : \mathbb{R}^n \to \mathring{\mathcal{S}}^n$. The vertices of $\mathcal{D}$ are given by the images of the vertices of $\mathcal{C}$ under the same transformation.*

Proof. We use the same notation as in Lemma 2.4.3. Recall that we centred our $\mathcal{H}$-ball at $\hat{\theta}_0 = \theta_0(\nu)$, the image of ν under the mapping $\theta_0 : \mathcal{S}^n \to \mathbb{R}^n$. Similarly, we will fix the centre of $\mathcal{D}$, the representation of the $\mathcal{H}$-ball in $\mathcal{S}^n$, at ν, and proceed as if ν were known.

We start by noting that the bounds $|\theta_k^i - \hat{\theta}_k^i| \leq R$, which correspond to the linear constraints (2.35), are equivalent to linear constraints in $\mathcal{S}^n$. Recalling the notation of the proof of Lemma 2.4.3, consider the half-spaces $\mathfrak{p}_{i,k}^+$ under the transformation $\theta_0^{-1} : \mathbb{R}^n \to \mathcal{S}^n$. We compute

$$\begin{aligned}\theta_0^{-1}(\mathfrak{p}_{i,k}^+) &= \{\theta_0^{-1}(\theta_0) \in \mathcal{S}^n \,:\, \theta_0^i - \theta_0^k \leq \hat{\theta}_0^i - \hat{\theta}_0^k + R\} \\ &= \Big\{\mu \in \mathcal{S}^n \,:\, \log\frac{\mu^i}{\mu^k} \leq \log\frac{\nu^i}{\nu^k} + R\Big\} = \Big\{\mu \in \mathcal{S}^n \,:\, \mu^i - \mu^k\frac{\nu^i}{\nu^k}e^R \leq 0\Big\},\end{aligned}$$

where we have used bijectivity of θ_0^{-1} and the fact that exp is increasing. Similarly,

$$\theta_0^{-1}(\mathfrak{p}_{i,k}^-) = \Big\{\mu \in \mathcal{S}^n \,:\, \mu^i - \mu^k\frac{\nu^i}{\nu^k}e^{-R} \geq 0\Big\}.$$

Note that $\theta_0^{-1}(\mathfrak{p}_{i,k}^+)$ and $\theta_0^{-1}(\mathfrak{p}_{i,k}^-)$ are $(n-1)$-dimensional flat subspaces of $\mathcal{S}^n$. In particular, since $\mathcal{S}^n$ is a subset of an n-dimensional affine space $A \cong \mathbb{R}^n$, we see that $\theta_0^{-1}(\mathfrak{p}_{i,k}^+)$ and $\theta_0^{-1}(\mathfrak{p}_{i,k}^-)$ are closed half-spaces of A, bounded by the hyperplanes $\mathfrak{l}_{i,k}^+ := \{\mu \in A \,:\, \mu^i = \mu^k \frac{\nu^i}{\nu^k} e^R\}$ and $\mathfrak{l}_{i,k}^- := \{\mu \in A \,:\, \mu^i = \mu^k \frac{\nu^i}{\nu^k} e^{-R}\}$, which are the images (extended to A) of respectively $\mathfrak{h}_{i,k}^+$ and $\mathfrak{h}_{i,k}^-$ under θ_0^{-1}. Then, recalling that $I^n := \{(i,k) \in \mathbf{N} \times \mathbf{N} \,:\, i > k\}$, the intersection

$$\mathcal{D} := \bigcap_{(i,k)\in I^n} \theta_0^{-1}(\mathfrak{p}_{i,k}^+) \cap \theta_0^{-1}(\mathfrak{p}_{i,k}^-)$$

is a convex polyhedron in A, and in particular, since θ_0^{-1} is a bijection from $\mathbb{R}^n$ into $\mathring{\mathcal{S}}^n$,

$$\theta_0^{-1}(\mathcal{C}) = \theta_0^{-1}\Bigg(\bigcap_{(i,k)\in I^n} \mathfrak{p}_{i,k}^+ \cap \mathfrak{p}_{i,k}^-\Bigg) = \mathcal{D} \subset \mathcal{S}^n.$$

Finally, both boundedness of $\mathcal{D}$ in $\mathcal{S}^n$ (in the sense that $\mathcal{D}$ is bounded away from $\partial\mathcal{S}^n$), so that $\mathcal{D}$ is a convex polytope in $\mathcal{S}^n$, and the fact that vertices are preserved under the mapping follow easily from θ_0 being an homeomorphism between $\mathring{\mathcal{S}}^n$ and $\mathbb{R}^n$. □

Remark 2.12. Using Figure 2.1, we can build an intuition of how the transformation θ_0^{-1} deforms $\mathcal{C}$ by considering what happens to the parallel pairs of hyperplanes $\mathfrak{h}_{i,k}^+$ and $\mathfrak{h}_{i,k}^-$ when mapped into $\mathcal{S}^n$. For each pair $(i,k) \in \mathbf{N} \times \mathbf{N}$, $\mathfrak{l}_{i,k}^+ = \theta_0^{-1}(\mathfrak{h}_{i,k}^+)$ and

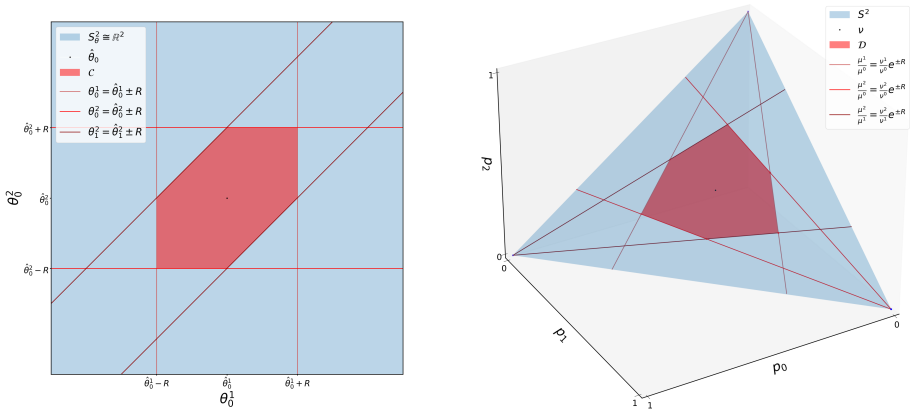

Figure 2.1: On the left: representation in θ_0-coordinates of a 2-dimensional $\mathcal{H}$-ball $\mathcal{C}$ of radius R around $\hat{\theta}_0 = \theta_0(\nu) = (0,0)$. On the right, the image of $\mathcal{C}$ under θ_0^{-1}, which gives the $\mathcal{H}$-ball $\mathcal{D}$ around $\nu = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ as a hexagonal polygon in the simplex $\mathcal{S}^2$.

$\ell_{i,k}^- = \theta_0^{-1}(\hbar_{i,k}^-)$ are not parallel in $\mathcal{S}^n$, but meet at the $(n-2)$-face of the simplex given by $f_{i,k} = \{\mu \in \mathcal{S}^n : \mu^i = 0, \mu^k = 0\}$. In other words, the 'point at infinity' at which $\hbar_{i,k}^+$ and $\hbar_{i,k}^-$ meet in $\mathbb{R}^n$ is mapped to the boundary of the simplex, and in particular to $f_{i,k}$, under θ_0^{-1}.

For example, in dimension 2, $f_{i,k}$ are vertices of $\mathcal{S}^2$: when mapping $\mathcal{C} \in \mathbb{R}^2$ to $\mathcal{S}^2$, we can think of squeezing together the ∞-extremities of each pair of parallel lines $\hbar_{i,k}^{\pm}$ (for $(i,k) \in \{(1,0),(2,0),(2,1)\}$) so that they meet at an angle of $\alpha = a(e^R - e^{-R})/(1+a^2)$, where $a = \nu^i/\nu^k$. Then we place the intersection point at the vertex $f_{i,k}$ of $\mathcal{S}^2$, so that, intuitively, the strip of plane between $\hbar_{i,k}^+$ and $\hbar_{i,k}^-$ is mapped to a slice of $\mathcal{S}^2$ of width α bounded by $l_{i,k}^+$ and $l_{i,k}^-$. Then it is easy to visualize how the straight lines that compose the boundary of $\mathcal{C}$ are mapped to straight lines, and how intersections are preserved, making $\theta_0^{-1}(\mathcal{C})$ into a polytope as well. However, these lines (and those parallel to them) are in fact the only straight lines in θ_0-coordinates that map to straight lines in $\mathcal{S}^2$ (as expected, since θ_0 and its inverse are nonlinear). We illustrate this in Figure 2.2 below.

Remark 2.13. The regularity of the $\mathcal{H}$-balls, when represented in θ_0-coordinates, has other surprising consequences[1]—for example, Hilbert balls of constant radius naturally tile the space, as illustrated in Figure 2.3.

[1]In a more artistic vein, Figure 2.3 also illustrates that the map $\theta_0 : S^n \to \mathbb{R}^n$, for $n = 2$, corresponds to the classical transformation between parallel oblique perspective (θ_0-coordinates) and three-point perspective (by viewing S^n with its vertices at the three vanishing points), linking back well beyond Birkhoff (1957) and Hilbert (1895), at least as far as the work of Jean Pelerin (Viator) in *De Artificiali Perspectiva* (1505).

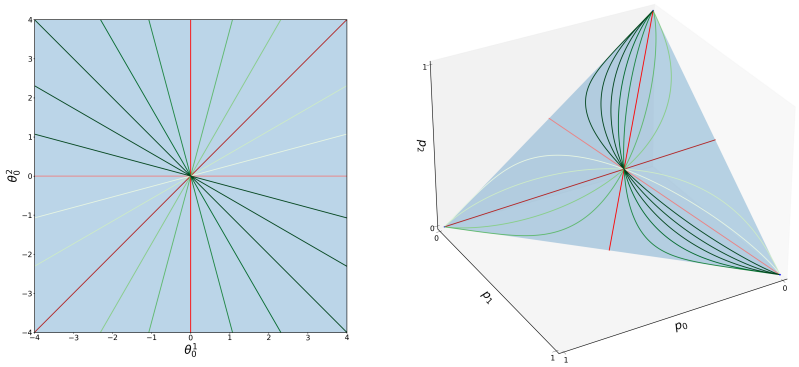

Figure 2.2: Straight lines through the origin in θ_0-coordinates on the left, and their images in $\mathcal{S}^2$ under the inverse mapping θ_0^{-1} on the right. In red the lines parallel to the axes and the diagonal in $\mathbb{R}^2$, which remain straight in $\mathcal{S}^2$.

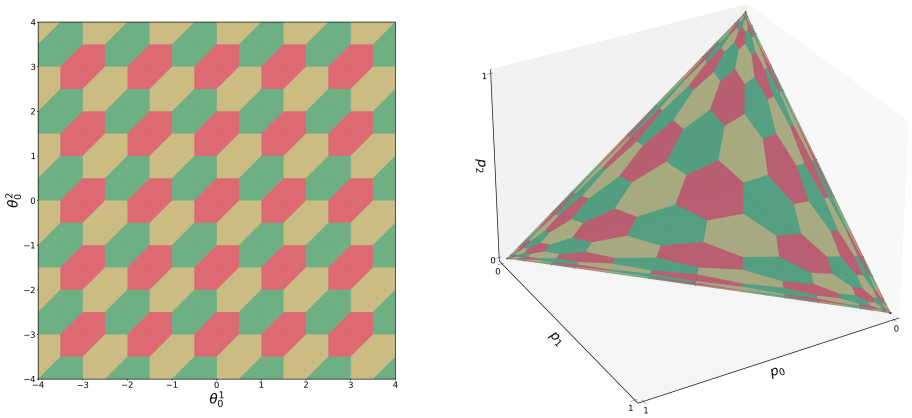

Figure 2.3: Tiling of the 2-dimensional probability simplex $\mathcal{S}^2$ with Hilbert balls of radius 0.5, starting from the ball around the center $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ (corresponding to the origin in the θ_0-coordinates on the left).

2.5 Metric-comparisons for probability measures: TV and $\mathcal{H}$

We now exploit the geometric intuition we gathered in the previous section to derive a bound, sharper than (2.18), for the total variation norm with respect to the Hilbert projective metric. We start by working with discrete probabilities in the simplex $\mathcal{S}^n$ and then extend our result to probability measures on a general measurable space. Perhaps unsurprisingly, the distance $\mathcal{T}(\mu, \nu) = \tanh(\mathcal{H}(\mu, \nu)/4)$, which we defined in

Section 2.2.1, plays a role once more in the computations below.

2.5.1 Probabilities on finite state-space

First of all, recall that for a measurable space $(E, \mathcal{F})$ the total variation distance (2.11) between two probability measures $\mu, \nu \in \mathcal{P}(E)$ is equivalent to

$$\|\mu - \nu\|_{\mathsf{TV}} = 2 \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|, \tag{2.37}$$

which, in the case of $E \cong \{0, \cdots n\}$ and $\mu, \nu \in \mathcal{S}^n$, reduces to

$$\|\mu - \nu\|_{\mathsf{TV}} = \sum_{i=0}^{n} |\mu^i - \nu^i| = \|\mu - \nu\|_{\ell^1}. \tag{2.38}$$

Remark 2.14. Note that the factor of 2 in (2.37) is usually dropped, in which case (2.38) would be stated as $\|\mu - \nu\|_{\mathsf{TV}} = \frac{1}{2}\|\mu - \nu\|_{\ell^1}$. We keep the factor of 2 in analogy with Atar and Zeitouni [8].

Theorem 2.5.1. *Given two probability measures $\mu, \nu \in \mathcal{S}^n$, we have that*

$$\|\mu - \nu\|_{\mathsf{TV}} = \|\mu - \nu\|_{\ell^1} \leq 2 \tanh \frac{\mathcal{H}(\mu, \nu)}{4}. \tag{2.39}$$

Equivalently, we have

$$\sup_{A \subseteq \{0,\ldots,n\}} \sum_{i \in A} |\mu^i - \nu^i| \leq \mathcal{T}(\mu, \nu). \tag{2.40}$$

From Lemma 2.4.4 we know that if $\mu, \nu \in \mathring{\mathcal{S}}^n$ with $\mathcal{H}(\mu, \nu) = R < \infty$, then μ belongs to the boundary of a convex polytope $\mathcal{D} \in \mathcal{S}^n$, centred at ν and with vertices $\{\theta_0^{-1}(v_{\mathcal{I}}^+), \theta_0^{-1}(v_{\mathcal{I}}^-) : \mathcal{I} \subseteq \{1, \ldots, n\}, \mathcal{I} \neq \emptyset\}$. Finding an upper bound for $\|\mu - \nu\|_{\ell^1}$ is now a simple convex optimization problem: we know the ℓ^1-distance between ν and μ is maximized when μ is at one of the vertices of $\mathcal{D}$, so we compute the ℓ^1-distance between ν and each of these vertices, and then maximize over the choice of vertex.

Lemma 2.5.2. *Assume $\nu \in \mathring{\mathcal{S}}^n$ is known. If $\mathcal{H}(\mu, \nu) = R$, the ℓ^1-distance between μ and ν is bounded by*

$$\|\mu - \nu\|_{\ell^1} \leq 2 \max_{\substack{\mathcal{I} \subseteq \{1,\ldots,n\} \\ \mathcal{I} \neq \emptyset}} \left\{ \frac{\mathbf{S}_{\mathcal{I}}(1 - \mathbf{S}_{\mathcal{I}})(e^R - 1)}{1 + \mathbf{S}_{\mathcal{I}}(e^R - 1)} \;\vee\; \frac{\mathbf{S}_{\mathcal{I}}(1 - \mathbf{S}_{\mathcal{I}})(1 - e^{-R})}{1 + \mathbf{S}_{\mathcal{I}}(e^{-R} - 1)} \right\}, \tag{2.41}$$

where $\mathbf{S}_{\mathcal{I}} := \sum_{i \in \mathcal{I}} \nu^i$ for any subset $\mathcal{I}$ (not $\emptyset$) of the indices.

Proof. Recall (2.34). Consider a vertex $\mathbf{v}_{\mathcal{I}+} = \theta_0^{-1}(v_{\mathcal{I}}^+)$ of $\mathcal{D} \in \mathcal{S}^n$, and let $\mathbf{N}_0 = \mathbf{N} \setminus \{0\}$. Compute

$$\begin{aligned}
\|\nu - \mathbf{v}_{\mathcal{I}+}\|_{\ell^1} &= \sum_{i=0}^{n} |\nu^i - \mathbf{v}^i_{\mathcal{I}+}| \\
&= \left| \nu^0 - \frac{1}{1 + \sum_{i\in\mathcal{I}} \exp\{\theta_0^i(\nu) + R\} + \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \exp\{\theta_0^i(\nu)\}} \right| \\
&\quad + \sum_{k\in\mathcal{I}} \left| \nu^k - \frac{\exp\{\theta_0^k(\nu) + R\}}{1 + \sum_{i\in\mathcal{I}} \exp\{\theta_0^i(\nu) + R\} + \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \exp\{\theta_0^i(\nu)\}} \right| \\
&\quad + \sum_{k\in\mathbf{N}_0\setminus\mathcal{I}} \left| \nu^k - \frac{\exp\{\theta_0^k(\nu)\}}{1 + \sum_{i\in\mathcal{I}} \exp\{\theta_0^i(\nu) + R\} + \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \exp\{\theta_0^i(\nu)\}} \right| \\
&= \left| \nu^0 - \frac{1}{1 + e^R \sum_{i\in\mathcal{I}} \frac{\nu^i}{\nu^0} + \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \frac{\nu^i}{\nu^0}} \right| \\
&\quad + \sum_{k\in\mathcal{I}} \left| \nu^k - \frac{e^R \frac{\nu^k}{\nu^0}}{1 + e^R \sum_{i\in\mathcal{I}} \frac{\nu^i}{\nu^0} + \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \frac{\nu^i}{\nu^0}} \right| \\
&\quad + \sum_{k\in\mathbf{N}_0\setminus\mathcal{I}} \left| \nu^k - \frac{\frac{\nu^k}{\nu^0}}{1 + e^R \sum_{i\in\mathcal{I}} \frac{\nu^i}{\nu^0} + \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \frac{\nu^i}{\nu^0}} \right|.
\end{aligned}$$

Letting $\mathbf{S}_{\mathcal{I}} := \sum_{i\in\mathcal{I}} \nu^i$ and $\breve{\mathbf{S}}_{\mathcal{I}} := \sum_{i\in\mathbf{N}_0\setminus\mathcal{I}} \nu^i$, so that $\mathbf{S}_{\mathcal{I}} + \breve{\mathbf{S}}_{\mathcal{I}} + \nu^0 = 1$, some algebra yields

$$\|\nu - \mathbf{v}_{\mathcal{I}+}\|_{\ell^1} = 2(e^R - 1)\frac{\mathbf{S}_{\mathcal{I}}(1 - \mathbf{S}_{\mathcal{I}})}{1 + \mathbf{S}_{\mathcal{I}}\, e^R - \mathbf{S}_{\mathcal{I}}} =: g_R^+(\mathbf{S}_{\mathcal{I}})$$

Equivalently, if we let $\mathbf{v}_{\mathcal{I}-} = \theta_0^{-1}(v_{\mathcal{I}}^-)$, we obtain

$$\|\nu - \mathbf{v}_{\mathcal{I}-}\|_{\ell^1} = 2(1 - e^{-R})\frac{\mathbf{S}_{\mathcal{I}}(1 - \mathbf{S}_{\mathcal{I}})}{1 + \mathbf{S}_{\mathcal{I}}\, e^{-R} - \mathbf{S}_{\mathcal{I}}} =: g_R^-(\mathbf{S}_{\mathcal{I}})$$

Then the ℓ^1-distance between μ and ν is bounded by the maximum between $\|\nu - \mathbf{v}_{\mathcal{I}+}\|_{\ell^1}$ and $\|\nu - \mathbf{v}_{\mathcal{I}-}\|_{\ell^1}$ over all choices of vertices, which yields the lemma. □

Proof of Theorem 2.5.1. Let $\mu, \nu \in \mathring{\mathcal{S}}^n$ be such that $\mathcal{H}(\mu, \nu) = R$. Recall the notation from the proof of Lemma 2.5.2. Note that $g_R^+(x)$ and $g_R^-(x)$ for $x \in [0, 1]$ are symmetric around $x = \frac{1}{2}$. By standard calculus, we find that the maximum of g_R^+ is attained at $x_+^* = \frac{1}{1+e^{R/2}}$; while g_R^- is maximized at $x_-^* = 1 - x_+^* = \frac{e^{R/2}}{1+e^{R/2}}$. Evaluating g_R^+ and g_R^-

at their respective maximizers gives the upper bound

$$\begin{aligned}\|\nu - \mathbf{v}\|_{\ell^1} &\leq \max_{\substack{\mathcal{I}\subseteq\{1,\dots,n\}\\ \mathcal{I}\neq\emptyset}} \{\|\nu - \mathbf{v}_{\mathcal{I}+}\|_{L^1} \vee \|\nu - \mathbf{v}_{\mathcal{I}-}\|_{L^1}\} \\ &\leq \max_{\nu\in\mathring{\mathcal{S}}^n} \max_{\substack{\mathcal{I}\subseteq\{1,\dots,n\}\\ \mathcal{I}\neq\emptyset}} \{\|\nu - \mathbf{v}_{\mathcal{I}+}\|_{L^1} \vee \|\nu - \mathbf{v}_{\mathcal{I}-}\|_{L^1}\} \\ &\leq \max_{\mathbf{S}_{\mathcal{I}}\in[0,1]} \{g_R^+(\mathbf{S}_{\mathcal{I}}) \vee g_R^-(\mathbf{S}_{\mathcal{I}})\} \\ &\leq 2\tanh\frac{R}{4}.\end{aligned}$$

Finally, note that the statement is trivial if $\mu \nsim \nu$. Moreover, if $\mu \sim \nu$ and $\mu, \nu \in \partial\mathcal{S}^n$, then they must belong to the same $(n-d)$-face of $\mathcal{S}^n$, which is also a probability simplex $\mathcal{S}^d$ with $1 \leq d < n$. In particular, $\mu, \nu \in \mathring{\mathcal{S}}^d$, which is the same as the case we considered originally, so we are done. □

Remark 2.15. For low dimensions, Lemma 2.5.2 provides a way to compute a bound for $\|\mu - \nu\|_{\ell^1}$ which is tighter than Theorem 2.5.1. While this still holds true in higher dimensions, the improvement gained by computing explicitly the bound (2.41) instead of using (2.39) can be negligible, since it is likely that at least one of the possible combinations for $\mathbf{S}_{\mathcal{I}}$ comes very close to the maximizer.

2.5.2 Probabilities on a general measurable space

We now use the discrete result of Theorem 2.5.1 to give bounds for general probability measures.

Corollary 2.5.2.1. *Let $(E, \mathcal{F})$ be a measurable space. Consider $\mu, \nu \in \mathcal{P}(E)$. We have*

$$\frac{1}{2}\|\mu - \nu\|_{\mathsf{TV}} = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)| \leq \mathcal{T}(\mu, \nu).$$

Proof. Note that if $\mu \nsim \nu$, then the statement follows trivially, so assume that $\mu \sim \nu$. By definition of the total variation distance (2.37), for all $n \in \mathbb{N}$ there exists a set $A_n \in \mathcal{F}$ such that

$$\frac{1}{2}\|\mu - \nu\|_{\mathsf{TV}} - \frac{1}{n} \leq |\mu(A_n) - \nu(A_n)| \leq \frac{1}{2}\|\mu - \nu\|_{\mathsf{TV}}. \tag{2.42}$$

Let $\mathcal{F}_n$ be the σ-algebra generated by A_n, i.e. $\mathcal{F}_n = \{A_n, A_n^c, E, \emptyset\}$, and let π_n be the partition of E given by $\pi_n = \{A_n, A_n^c\}$. Consider the probability measures μ_n, ν_n on the space $(E, \mathcal{F}_n)$ given by

$$\mu_n = \mu|_{\pi_n}, \qquad \nu_n = \nu|_{\pi_n}.$$

Then μ_n and ν_n are probabilities on the finite state space $\{A_n, A_n^c\}$, and in particular $\mu_n, \nu_n \in \mathcal{S}^1$. Therefore it holds that

$$\begin{aligned}\sup_{A\in\mathcal{F}_n} |\mu_n(A) - \nu_n(A)| &= |\mu_n(A_n) - \nu_n(A_n)| \\ &\leq \tanh \frac{\mathcal{H}(\mu_n, \nu_n)}{4} = \tanh \frac{\left|\log \frac{\mu_n(A_n)}{\nu_n(A_n)} - \log \frac{\mu_n(A_n^c)}{\nu_n(A_n^c)}\right|}{4} \\ &\leq \tanh \frac{\sup_{A,B\in\mathcal{F}} \left(\log \frac{\mu(A)}{\nu(A)} - \log \frac{\mu(B)}{\nu(B)}\right)}{4} = \tanh \frac{\mathcal{H}(\mu, \nu)}{4}.\end{aligned}$$

As for the left-hand side, we have that

$$\sup_{A\in\mathcal{F}_n} |\mu_n(A) - \nu_n(A)| = |\mu_n(A_n) - \nu_n(A_n)| = |\mu(A_n) - \nu(A_n)| \leq \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|.$$

But by (2.42) we also have that

$$\sup_{A\in\mathcal{F}_n} |\mu_n(A) - \nu_n(A)| = |\mu_n(A_n) - \nu_n(A_n)| = |\mu(A_n) - \nu(A_n)| \geq \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)| - \frac{1}{n}.$$

Therefore, putting together the two above inequalities, we arrive at the final expression

$$\lim_{n\to\infty} \sup_{A\in\mathcal{F}_n} |\mu_n(A) - \nu_n(A)| = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|,$$

and the result follows. □

Remark 2.16. Note that the bounds in Theorem 2.5.1 and Corollary 2.5.2.1 can be attained, and therefore are sharp. In particular, $2\mathcal{T}$ is nothing but the maximum TV (or ℓ^1) norm between two probability measures which are a fixed $\mathcal{H}$-distance apart. Theorem 2.2.6 then tells us that this quantity contracts under (positive) linear transformations.

Remark 2.17. For $\mu, \nu \in \mathcal{P}(E)$, we can also find an upper bound for $\mathcal{T}(\mu, \nu)$ in terms of TV. For all $A \in \mathcal{F}$, we have

$$\tanh\left(\frac{1}{2}\log\frac{\mu(A)}{\nu(A)}\right) = \frac{\mu(A) - \nu(A)}{\mu(A) + \nu(A)},$$

therefore

$$\mathcal{T}(\mu, \nu) \leq \tanh\left(\frac{1}{2}\log\Big(\sup_{A\in\mathcal{F}}\frac{\mu(A)}{\nu(A)} \vee \sup_{B\in\mathcal{F}}\frac{\mu(B)}{\nu(B)}\Big)\right) \leq \frac{\sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|}{2\big(\inf_{B\in\mathcal{F}}\mu(B) \wedge \inf_{B\in\mathcal{F}}\nu(B)\big)}. \tag{2.43}$$

However, we note that the denominator of the right-hand side may become arbitrarily small, which is unsurprising given $\mathcal{T}$ is a stronger metric than TV on S^n.

Chapter 3

Exponential contraction estimates for the Wonham filter

This chapter is concerned with the study of the stability of the Wonham filter. In it, we establish the fundamental contraction estimates that allow us to move on to the question of robustness in Chapter 4, and ultimately to compute quantitatively meaningful error bounds for approximate filters.

3.1 Discussion of known results

Filtering stability has been an active field of study since the 1990's. A key paper in the literature is [71], in which Ocone and Pardoux establish a relationship between the stability of the Kalman filter and detectability/stabilizability of the signal-observation linear control system. Their arguments for stability in the nonlinear setting, however, rely on a result by Kunita [59], which was later found to contain a mistake (see Baxendale, Chigansky and Liptser [12, Section 2] for a detailed explanation and a counterexample, and Budhiraja [20] for an analysis of its relevance in the context of nonlinear filtering stability). The gap in Kunita's proof was addressed by Van Handel, who established the necessary conditions for the stability of the nonlinear filter in different settings (for ergodic signals in discrete and continuous time in [85], non-ergodic signal with compact state-space in [84], and with Polish state-space in [86]). More recently, in Kim, Mehta and Meyn [56] and Kim and Mehta [55], stability of the Wonham filter is shown to be equivalent to stabilizability of a dual control problem, in an extension of [71] to the nonlinear case. We refer the interested reader to Chigansky [24] for an extensive review of nonlinear filtering stability results (in discrete time with finite state-space) and to [33, Part 3] for a broad collection of

survey papers. Of particular relevance to our setting, Chigansky, Liptser and Van Handel [25] gives an accessible introduction to the stability results of [84–86].

While the above results guarantee stability of the filter in the strongest possible generality (and under the weakest possible assumptions), their qualitative nature makes them unsuitable for understanding general approximation errors. On the other hand, if one is willing to impose relatively strong ergodicity assumptions on the signal process, there are explicit decay rates available in the literature, at least for the particular case of the Wonham filter. Delyon and Zeitouni [35] introduced the study of the top Lyapunov exponent for the Wonham filter, and proved that it is negative under certain conditions on the model parameters. This method was expanded by Atar and Zeitouni [8, 9], who, under a fairly strong mixing assumption for the signal, compute an explicit exponential decay rate for the stability error. Applying the techniques of [8], Baxendale, Chigansky and Liptser weakened the ergodicity assumptions slightly by proving a.s. negativity of the decay rate if all the states of X communicate [12, Theorem 4.1] (although we lose an explicit rate). Finally, by working with the *smoother process* (as described in e.g. Liptser and Shiryayev [66, Theorem 9.5]), they provide an explicit exponential rate of decay for a mixing signal in terms of its ergodic distribution [12, Theorem 4.2], and a *non-asymptotic* exponential bound for the stability error [12, Theorem 4.3], with the same decay rate as [8, 9].

As far as we are aware, the bound in [12, Theorem 4.3] is the only non-asymptotic bound available in the literature for the stability error of the Wonham filter in continuous time. The prefactor to the exponential decay term is proportional to the dimension of the Wonham SDE and the Radon–Nikodym derivatives of the true and the 'wrong' initial distribution, and it is far too large for the bound to be useful from a quantitative point of view. Van Handel improves it significantly (although the result still remains far from a contraction), and the best estimate for the prefactor is found by combining [26, Proposition 3.5] and [83, Corollary 2.3.2]. This stability result is central in the robustness analysis for the Wonham filter carried out in [26]. On the other hand, the robustness results for the nonlinear filter in discrete time [21, 62, 63] that we mentioned previously build on the work on stability by Atar and Zeitouni (in [8, 9] the analysis is carried out for both discrete and continuous time settings).

The fundamental contribution of [8, 9] is to introduce the use of the Hilbert projective distance as a metric on the space of probability measures to carry out stability estimates for the nonlinear filter. As we have seen in Chapter 2 a key advantage of using the Hilbert metric is that positive linear operators contract under this distance. Recalling that the generator of a discrete-time Markov chain is a stochastic matrix,

Birkhoff's and Seneta's works make the stability results for discrete-time nonlinear filtering intuitively straightforward.

Atar and Zeitouni provide asymptotic rates for the decay of the stability error of the filter, for both the discrete and continuous time case. Building on these ideas, and on Seneta's work, Le Gland and Mevel [61, 62], and then Le Gland and Oudjane [63] proved non-asymptotic and non-logarithmic stability bounds for the discrete time setting, conditional on a strong mixing assumption for the signal process. In [63], they are also able to tackle the issue of robustness of the nonlinear filter (in discrete time) and in particular they study the global error of interacting particle approximations to the filtering process. Our results in this chapter and in Chapter 4 follow roughly along the same lines, although in the continuous time setting. Moreover, our approach is fundamentally different from that in [8, 9, 61–63]; the only common aspect is the use of the Hilbert metric in the stability analysis.

3.2 Filtering set-up and a key result

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a filtration $\{\mathcal{F}_t, t \geq 0\}$ satisfying the usual conditions. Consider an $\{\mathcal{F}_t\}$-adapted continuous-time, time-homogeneous Markov chain $X = (X_t)_{t\geq 0}$ with finite state-space $\mathbb{S} = \{a_0, \ldots, a_n\}$, and associated transition intensity matrix $Q = (q_{ij}) \in \mathbb{R}^{(n+1)\times(n+1)}$. We let $\mathcal{M}^+(\mathbb{S})$ and $\mathcal{P}(\mathbb{S})$ denote respectively the non-negative measures and the probability measures on $\mathbb{S}$. Let the initial distribution of X be given by $\mu^i = \mathbb{P}(X_0 = a_i)$.

Recall that the Q-matrix is defined as the matrix of transition rates such that its entries for each row sum to 0, and its off-diagonal entries are non-negative, i.e. $\sum_j q_{ij} = 0$ for all i, and $q_{ij} \geq 0$ for all $i, j \leq n+1$, $i \neq j$, and

$$M_t^{\varphi} = \varphi(X_t) - \varphi(X_0) - \int_0^t Q\varphi(X_s)\,\mathrm{d}s, \quad t \geq 0$$

is an $\{\mathcal{F}_t\}$-adapted, right-continuous martingale for all bounded functions $\varphi : \mathbb{S} \to \mathbb{R}$, with $Q\varphi(a_i) = \sum_{j=0}^n q_{ij}\varphi(a_j)$ for all $a_i \in \mathbb{S}$.

Let $h = (h_i)_{i=1}^d : \mathbb{S} \to \mathbb{R}^d$ be a bounded function and $\sigma \neq 0$. Suppose W is a standard $\{\mathcal{F}_t\}$-adapted d-dimensional Brownian motion independent of X, and let $Y = (Y_t)_{t\geq 0}$ be the process satisfying the SDE

$$Y_t = Y_0 + \int_0^t h(X_s)\,\mathrm{d}s + \sigma W_t. \tag{3.1}$$

Let $\{\mathcal{Y}_t\}_{t\geq 0}$ be the (completed) natural filtration generated by the observation process Y. This describes the information available from observing Y in the time-interval $[0,t]$.

By common practice, we identify the state-space $\mathbb{S}$ with $\{e_0,\dots,e_n\}$, the standard basis for $\mathbb{R}^{n+1}$. Denote by $\pi_t = \mathbf{E}\left[X_t|\mathcal{Y}_t\right]$ the conditional expectation of X given $\mathcal{Y}_t$. In other words, by abuse of notation, $\pi_t^i = \mathbb{P}(X_t = a_i|\mathcal{Y}_t)$.

The process π_t satisfies the Wonham form of the *Kushner–Stratonovich equation* (see e.g. [10, Eq. 3.53]):

$$\mathrm{d}\pi_t = Q^\top \pi_t\,\mathrm{d}t + \frac{1}{\sigma^2}\sum_{k=1}^{d}\left(H^k - \pi_t^\top h_k\,\mathbb{I}_{n+1}\right)\pi_t\left(\mathrm{d}Y_t^k - \pi_t^\top h_k\,\mathrm{d}t\right), \quad \pi_0 = \mu, \tag{3.2}$$

where, for $k = 1,\dots,d$, $H^k = \mathrm{diag}\left(h_k(a_i)\right)$ is an $(n+1)\times(n+1)$-dimensional diagonal matrix and $\mathbb{I}_{n+1}$ is the identity matrix. Note that (3.2) is initialized at $\mu = \mathrm{law}(X_0) = \mathbf{E}\left[X_0\right]$.

The probabilities π_t for $t\geq 0$ are $(n+1)$-dimensional (column) vectors, so (3.2) is a $(n+1)$-dimensional nonlinear SDE. In fact, since the components π_t^i must sum to 1 for all $t\geq 0$, the SDE (3.2) describes a flow on the n-dimensional probability simplex $\mathcal{S}^n$, where

$$\mathcal{S}^n = \left\{x\in\mathbb{R}^{n+1} \,:\, \sum_i x_i = 1,\, x_i\geq 0\right\}.$$

We write $\mathring{\mathcal{S}}^n$ for the interior of the simplex, that is, $x\in\mathring{\mathcal{S}}^n$ if $x\in\mathcal{S}^n$ and $x^i>0$ for all i.

Our choice of metric on $\mathcal{S}^n$ for the stability analysis of (3.2) is the Hilbert projective distance $\mathcal{H}$, which we recall is given by

$$\mathcal{H}(\mu,\nu) = \begin{cases} \log\left(\dfrac{\max_{j:\nu^j>0}\frac{\mu^j}{\nu^j}}{\min_{i:\nu^i>0}\frac{\mu^i}{\nu^i}}\right), & \mu\sim\nu,\\ \infty, & \mu\nsim\nu, \end{cases} \tag{3.3}$$

for $\mu,\nu\in\mathcal{S}^n$ expressed as non-negative vectors in $\mathbb{R}^{n+1}$.

In this chapter and the next, we make the following assumptions on the nonlinear filtering system described above

(A1) *X is a time-homogeneous continuous-time Markov chain on $n+1$ states.*

(A2) *$h = (h_i)_{i=1}^d$ is bounded for all i.*

(A3) *$\sigma = 1$ and $d = 1$ in (3.1) and (3.2).*

The final assumption only serves the purpose of simplifying notation – all our results are easily extendable to the case of multi-dimensional Y and invertible $\sigma \in \mathbb{R}^{d\times d}$. Similarly, we could easily allow for time-dependence in σ and h, as long as the first is bounded away from zero, and the second stays bounded for all t, and for time inhomogeneity in the Markov chain dynamics of X.

Notation. Given that we take the observations Y to be one-dimensional, the sensor function $h : \mathbb{S} \to \mathbb{R}$ can be seen as a vector $h \in \mathbb{R}^{n+1}$ with entries $h^i = h(a_i)$ for $i = 0, \dots, n$. From now on we will employ this notation. We also denote by $H = \operatorname{diag}(h)$ the diagonal matrix with entries $(H)_{ii} = h^i$. In general, we will always denote the components of vectors (or vector-valued processes) with superscripts. We denote by $\mathbf{N}$ the set of natural numbers $\{0, \dots, n\}$. Sometimes we will write $\mathrm{d}A_t \leq \mathrm{d}\tilde{A}_t$ for two Lebesgue–Stieltjes measures A_t and $\tilde{A}_t$ on $[0, \infty)$, by which we mean $\int_s^t \mathrm{d}A_r \leq \int_s^t \mathrm{d}\tilde{A}_r$ for all $0 \leq s < t < \infty$.

For reference, we rewrite here equation (3.2) for the Wonham filter given the above assumptions and notation

$$\mathrm{d}\pi_t = Q^\top \pi_t \,\mathrm{d}t + \left(H - \pi_t^\top h \,\mathbb{I}_{n+1}\right) \pi_t \left(\mathrm{d}Y_t - \pi_t^\top h \,\mathrm{d}t\right), \quad \pi_0 = \mu. \tag{3.4}$$

We consider the long time behaviour of the error between π_t and $\tilde{\pi}_t$, where $\tilde{\pi}_t$ is the filter initialized with the 'wrong' initial data $\tilde{\pi}_0 = \nu \neq \mu$ but the same dynamics as π. The evolution equation for $\tilde{\pi}_t$ is given by

$$\mathrm{d}\tilde{\pi}_t = Q^\top \tilde{\pi}_t \,\mathrm{d}t + \left(H - \tilde{\pi}_t^\top h \,\mathbb{I}_{n+1}\right) \tilde{\pi}_t \left(\mathrm{d}Y_t - \tilde{\pi}_t^\top h \,\mathrm{d}t\right), \quad \tilde{\pi}_0 = \nu. \tag{3.5}$$

Our key result is the following pathwise estimate on the stability of the filter.

Theorem 3.2.1 (Contraction rate of $\mathcal{H}(\pi_t, \tilde{\pi}_t)$)**.** *Let π_t be the solution to* (3.4) *and $\tilde{\pi}_t$ the solution to* (3.5)*. Suppose $q_{ij} > 0$ for all $i \neq j$. Then for all $t < \infty$,*

$$\tanh\left(\frac{\mathcal{H}(\pi_t, \tilde{\pi}_t)}{4}\right) \leq \tanh\left(\frac{\mathcal{H}(\mu, \nu)}{4}\right) e^{-\lambda t},$$

where $\lambda = 2 \min_{i\neq j} \sqrt{q_{ij} q_{ji}}$. In particular,

$$\mathcal{H}(\pi_t, \tilde{\pi}_t) \leq \mathcal{H}(\mu, \nu) e^{-\lambda t}.$$

Unsurprisingly, our contraction rate is the same as the asymptotic rate in [9], and the non-asymptotic rate in [12, Theorem 4.3], and shares the issue of only being (strictly) positive if all the off-diagonal entries of Q are (strictly) positive. This is a very strong mixing assumption on X; however, it seems necessary to be able to

compute an explicit contraction rate, and in fact a similar assumption is made in [63] in the discrete-time setting (see [63, Definition 3.2]).

Given these stability estimates, in Chapter 4 we will be able to proceed to the next challenge of understanding the error of approximate filters, and then apply these estimates to projection filters on the simplex in Chapter 5.

3.3 Contraction rates in the Hilbert projective metric

The rest of this chapter is dedicated to proving Theorem 3.2.1 and a few more results related to the stability of the nonlinear filter with respect to its initial conditions.

We start by reintroducing the family of coordinate transformations (2.31) from $\mathring{\mathcal{S}}^n$ to $\mathbb{R}^n$ that map a discrete probability distribution to its natural parameters. We derive the evolution equation for the Wonham filter in these new parametrizations, and then consider the difference between the natural parameters of the Wonham filter initialized at $\mu = \text{law}(X_0)$ and those of the Wonham filter 'wrongly' initialized at $\nu \neq \mu$. By relating the ℓ_∞ norm of the difference, maximized over parametrizations, to the Hilbert projective metric, we are able to compute explicitly an exponential contraction rate in the Hilbert metric for the Wonham filter. Up until the proof of Theorem 3.2.1, we will regularly make the extra assumption that $\pi_0 = \mu$ and $\tilde{\pi}_0 = \nu$ belong to the interior of the simplex.

Remark 3.1. For the entirety of this chapter, $(\pi_t)_{t\geq 0}$ represents the Wonham filter initialized at $\pi_0 = \mu$, and $(\tilde{\pi}_t)_{t\geq 0}$ the Wonham filter initialized at $\tilde{\pi}_0 = \nu$.

3.3.1 Coordinate transformations

Recall the coordinate transformation (2.31) from Section 2.4.1 which sends a probability distribution to what, in statistics, are called the *natural (or canonical) parameters*. For all $k \in \mathbf{N}$, we have the diffeomorphism $\theta_k : \mathring{\mathcal{S}}^n \ni p \mapsto \theta_k \in \mathbb{R}^n$ given by

$$\theta_k^i = \log \frac{p^i}{p^k}, \quad \forall i \in \mathbf{N}, \tag{3.6}$$

and its inverse map θ_k^{-1} is

$$p^i = \frac{\exp \theta_k^i}{1 + \sum_{j\neq k} \exp \theta_k^j}, \quad \forall i \in \mathbf{N}. \tag{3.7}$$

We remark that $\theta_k^k = 0$ can be ignored as an entry of the vector θ_k (and it can be 'skipped'), so that indeed $\theta_k \in \mathbb{R}^{k-1} \times \{0\} \times \mathbb{R}^{n-k} \cong \mathbb{R}^n$.

Natural parameters are the usual choice of parametrization for an exponential family of distributions, which have probability densities that can be written in general form as

$$p(x,\theta) = \exp\{\theta \cdot c(x) + k(x) - \psi(\theta)\}, \tag{3.8}$$

where $\theta \in \mathbb{R}^n$ is the n-dimensional vector of natural parameters, $c(x)$ is the vector of *sufficient statistics* of the distribution (and its n components are linearly independent), $k(x)$ is a function of x and $\psi(\theta)$ is the log partition function. A change of measure from $\mathrm{d}x$ to $\mathrm{d}\upsilon(x) = \exp\{k(x)\}\,\mathrm{d}x$ allows us to ignore $k(x)$, as long as $p(x,\theta)$ is understood as a density with respect to the measure $\mathrm{d}\upsilon(x)$ instead. We will assume $k(x) = 0$ for simplicity. Note that $\mathring{\mathcal{S}}^n$ is an n-dimensional exponential family, and we can write a discrete distribution $p \in \mathring{\mathcal{S}}^n$ in the form (3.8) by fixing $k \in \mathbf{N}$ and choosing $c^i(x) = \delta_{a_i}(x)$ (for $a_i \in \mathbb{S}$).

Our choice of studying the filtering equations in the coordinate system θ of natural parameters is motivated by Amari's theory of information geometry [2, 3]. If $\mu, \nu \in \mathring{\mathcal{S}}^n$, then the filtering process π_t lives in $\mathring{\mathcal{S}}^n$, which, in the language of information geometry, is an n-dimensional statistical manifold, with θ (and its dual affine, the *expectation parameter* η) as a global chart. The Riemannian metric for $\mathring{\mathcal{S}}^n$ is the Fisher Information, which infinitesimally agrees with the KL-divergence. In this thesis, we will not make much use of the differential geometrical structures for $\mathring{\mathcal{S}}^n$ developed by Amari, as our choice of distance between probability vectors is the Hilbert metric, which does not allow for a smooth geometry. However, it will still be convenient to work in the global coordinate system given by the θ-parametrization.

We now would like to apply the coordinate transformation (3.6) to $(\pi_t)_{t\geq 0}$ and $(\tilde{\pi}_t)_{t\geq 0}$ and derive evolution equations for the parameters $\theta_k(\pi_t)$ and $\theta_k(\tilde{\pi}_t)$. For all $k \in \mathbf{N}$, for notational simplicity define

$$\theta_k(t) := \theta_k(\pi_t), \quad \tilde{\theta}_k(t) := \theta_k(\tilde{\pi}_t),$$

so that, component-wise, we have

$$\theta_k^i(t) := \log \frac{\pi_t^i}{\pi_t^k}, \quad \tilde{\theta}_k^i(t) := \log \frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}, \quad \forall (i,k) \in \mathbf{N} \times \mathbf{N}.$$

The following lemma guarantees that these processes are almost surely well-defined for all $t < \infty$. For its proof we refer to [26].

Lemma 3.3.1 (Lemma 2.1 in [26])**.** *Denote by $\pi_{s,t}(\mu)$ the solution at time $t \geq 0$ to* (3.4) *initialized at time $s \leq t$ with $\pi_s = \mu$. Then*

$$\mathbb{P}\big(\pi_{s,t}(\mu) \in \mathring{\mathcal{S}}^n \text{ for all } \mu \in \mathring{\mathcal{S}}^n \text{ and all } 0 \leq s \leq t < \infty\big) = 1.$$

Corollary 3.3.1.1. *Assume $\mu, \nu \in \mathring{\mathcal{S}}^n$. We have that, almost surely,*

$$\mathcal{H}(\pi_t, \tilde{\pi}_t),\ |\theta_k^i(t)|,\ |\tilde{\theta}_k^i(t)| < \infty, \quad \forall\, (i,k) \in \mathbf{N} \times \mathbf{N},$$

for all times $0 \leq t < \infty$.

The proof of Lemma 3.3.1 also directly yields the following alternative result.

Lemma 3.3.2. *Denote by $\pi_t(\mu)$ the solution at time $t \geq 0$ to (3.4) initialized at time $0 \leq t$ with $\pi_0 = \mu$. Suppose $q_{ij} > 0$ for all $i \neq j$. Then*

$$\mathbb{P}\big(\pi_t(\mu) \in \mathring{\mathcal{S}}^n \text{ for all } \mu \in \mathcal{S}^n \text{ and all } 0 < t < \infty\big) = 1.$$

We now proceed to study the dynamics of the natural parameters $\theta_k^i(t)$ and $\tilde{\theta}_k^i(t)$. For all pairs of indices $(i,k) \in \mathbf{N} \times \mathbf{N}$, define the difference process

$$\Delta_{ik}(t) := (\theta_k^i(t) - \tilde{\theta}_k^i(t))_{t \geq 0}, \tag{3.9}$$

where $\Delta_{ii} = 0$ for all $i \in \mathbf{N}$.

We start with the following proposition.

Proposition 3.3.3. *Assume $\mu, \nu \in \mathring{\mathcal{S}}^n$. For all $0 \leq t < \infty$ and all pairs of indices $(i,k) \in \mathbf{N} \times \mathbf{N}$, the process $\Delta_{ik}(t)$ is C^1 in time and has the dynamics*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(t) &= -\sum_{\substack{j=0 \\ j \neq k}}^{n} q_{jk}\big(e^{\theta_k^j} - e^{\tilde{\theta}_k^j}\big) + \sum_{\substack{j=0 \\ j \neq i}}^{n} q_{ji}\big(e^{\theta_i^j} - e^{\tilde{\theta}_i^j}\big), \\ \Delta_{ik}(0) &= \log\frac{\mu^i}{\mu^k} - \log\frac{\nu^i}{\nu^k}. \end{aligned} \tag{3.10}$$

Proof. For $i = k$ the process Δ_{kk} is identically 0, so the statement holds trivially. Assume $i \neq k$. Consider $\theta_k^i(t) = \log(\pi_t^i/\pi_t^k)$ for $i \neq k$. We apply Itô's formula and obtain that, for any choice of $k \in \mathbf{N}$, and $i \neq k$, we have

$$\begin{aligned} \mathrm{d}\log\frac{\pi^i}{\pi^k}(t) = &-\sum_{\substack{j=0 \\ j \neq k}}^{n} q_{jk}\frac{\pi_t^j}{\pi_t^k}\,\mathrm{d}t + \sum_{\substack{j=0 \\ j \neq i}}^{n} q_{ji}\frac{\pi_t^j}{\pi_t^i}\,\mathrm{d}t + (q_{ii} - q_{kk})\,\mathrm{d}t + (h^i - h^k)\,\mathrm{d}B_t \\ &+ \frac{1}{2}\Big((h^k)^2 - (h^i)^2 + 2(h^i - h^k)\pi_t^\top h\Big)\,\mathrm{d}t, \\ \log\frac{\pi^i}{\pi^k}(0) = &\log\frac{\mu^i}{\mu^k}, \end{aligned} \tag{3.11}$$

where for readability we have introduced the innovation process $B_t = Y_t - \int_0^t \pi_s^\top h \, \mathrm{d}s$, which is a $\{\mathcal{Y}_t\}$-adapted Brownian motion (see e.g. [10, Proposition 2.30]). Similarly,

$$\begin{aligned}
\mathrm{d}\log\frac{\tilde{\pi}^i}{\tilde{\pi}^k}(t) &= -\sum_{\substack{j=0\\ j\neq k}}^{n} q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\,\mathrm{d}t + \sum_{\substack{j=0\\ j\neq i}}^{n} q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\,\mathrm{d}t + (q_{ii}-q_{kk})\,\mathrm{d}t + (h^i - h^k)\,\mathrm{d}B_t \\
&\quad + \frac{1}{2}\Big((h^k)^2 - (h^i)^2 + 2(h^i-h^k)\tilde{\pi}_t^\top h\Big)\,\mathrm{d}t + (h^i-h^k)\big(\pi_t^\top h - \tilde{\pi}_t^\top h\big)\,\mathrm{d}t, \\
\log\frac{\tilde{\pi}^i}{\tilde{\pi}^k}(0) &= \log\frac{\nu^i}{\nu^k}.
\end{aligned}$$

Subtracting the two equations, we see that the difference has absolutely continuous dynamics

$$\begin{aligned}
\mathrm{d}\bigg(\log\frac{\pi^i}{\pi^k}(t) - \log\frac{\tilde{\pi}^i}{\tilde{\pi}^k}(t)\bigg) &= -\sum_{\substack{j=0\\ j\neq k}}^{n} q_{jk}\bigg(\frac{\pi_t^j}{\pi_t^k} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\bigg)\,\mathrm{d}t + \sum_{\substack{j=0\\ j\neq i}}^{n} q_{ji}\bigg(\frac{\pi_t^j}{\pi_t^i} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\bigg)\,\mathrm{d}t, \\
\log\frac{\pi^i}{\pi^k}(0) - \log\frac{\tilde{\pi}^i}{\tilde{\pi}^k}(0) &= \log\frac{\mu^i}{\mu^k} - \log\frac{\nu^i}{\nu^k}.
\end{aligned} \tag{3.12}$$

Noting that the right-hand side of the above equation is continuous in time (since π_t and $\tilde{\pi}_t$ are both continuous), we have that the derivative of Δ_{ik} exists and is continuous for every $t \geq 0$, and (3.10) follows. □

3.3.2 The Hilbert error

Recalling the convenient equality (2.33), we now observe that

$$\mathcal{H}(\pi_t, \tilde{\pi}_t) = \Delta_\infty(t), \tag{3.13}$$

where we have defined the the maximal process

$$\Delta_\infty(t) := \max_{k\in\mathbf{N}} \Big\|\theta_k(t) - \tilde{\theta}_k(t)\Big\|_{\ell^\infty} = \max_{(i,k)\in\mathbf{N}\times\mathbf{N}} \Delta_{ik}(t).$$

We want to study the evolution in time of the stochastic process $\Delta_\infty(t)$. We here adapt some arguments from [12], since it turns out that our difference processes Δ_{ik} of Proposition 3.3.3 have dynamics somewhat similar to the equations of the smoother process considered in [12, Section 5.2, Eq. 5.6 & Eq. 5.7].

We shall be needing the next lemma in what follows.

Lemma 3.3.4 (Theorem A.6.3 in Dupuis and Ellis [38])**.** *Let $g : [0,1] \to \mathbb{R}$ be an absolutely continuous function. Then for every real number $r \in \mathbb{R}$, the set $\{t : g(t) = r, \dot{g}(t) \neq 0\}$ has Lebesgue measure 0.*

Unless specified otherwise, when we say that an adapted stochastic process $Z(t,\omega)$ is absolutely continuous or has absolutely continuous paths (a.s.), we mean not only that it can be written as $\mathrm{d}Z(t,\omega) = g(t,\omega)\,\mathrm{d}t$ with $g \in L^1([0,t])$, for all $t > 0$ (a.s.), but also that the weak derivative $g(t,\omega)$ is jointly measurable and adapted to the underlying filtration. The next lemma confirms that this is the case for the process $\Delta_\infty(t,\omega)$.

Lemma 3.3.5. *Assume $\mu, \nu \in \mathring{\mathcal{S}}^n$. The stochastic process $(t,\omega) \mapsto \Delta_\infty(t,\omega)$ has absolutely continuous paths (in particular, it is predictable).*

Proof. Fix an arbitrary $k \in \mathbf{N}$. Start by considering the processes $\Delta^\star_{i,k}(t) = \Delta_{0k} \vee \Delta_{1k} \vee \cdots \vee \Delta_{ik}$ for $i \in \mathbf{N}$. We proceed by induction to prove absolute continuity of $\Delta^\star_{n,k}(t) = \max_{i\in\mathbf{N}} \Delta_{ik}$. Trivially, $\Delta^\star_{0,k}(t) = \Delta_{0k}(t)$ is absolutely continuous, since it is either constant 0 by definition (if $k = 0$), or is absolutely continuous by Proposition 3.3.3 (if $k \neq 0$). Consider the case $i = 1$, with $\Delta^\star_{1,k}(t) = \Delta_{0k}(t) \vee \Delta_{1k}(t)$. Recall that $a \vee b = \frac{1}{2}(a + b + |a - b|)$. Then

$$\Delta^\star_{1,k}(t) = \frac{1}{2}\big(\Delta_{0k}(t) + \Delta_{1k}(t) + |\Delta_{0k}(t) - \Delta_{1k}(t)|\big).$$

By Proposition 3.3.3 we have that $\Delta_{0k}(t)$ and $\Delta_{1k}(t)$ are C^1 in time, and $\mathcal{F}_t$-measurable in ω. By the chain rule for weakly differentiable functions, if $F(t)$ is absolutely continuous with weak derivative $f(t)$, then

$$\mathrm{d}|F(t)| = \operatorname{sign}(F(t))f(t)\,\mathrm{d}t. \tag{3.14}$$

Thus we have that $|\Delta_{0k}(t) - \Delta_{1k}(t)|$ is absolutely continuous in time (for each ω), and it is clear from the form of (3.14) that the weak derivative is jointly measurable in (t,ω) and $\mathcal{F}_t$-adapted. Hence the same is true for $\Delta^\star_{1,k}(t)$.

Now noting that $\Delta^\star_{i,k}(t) = \Delta^\star_{i-1,k}(t) \vee \Delta_{ik}(t)$ for all $2 \leq i \leq n$, as before we can write

$$\Delta^\star_{i,k}(t) = \frac{1}{2}\big(\Delta^\star_{i-1,k}(t) + \Delta_{ik}(t) + |\Delta^\star_{i-1,k}(t) - \Delta_{ik}(t)|\big),$$

and by induction it follows that $\Delta^\star_{n,k}(t) = \max_{i\in\mathbf{N}} \Delta_{ik}$ has absolutely continuous paths.

Since the argument above is independent of our choice of k, we have that $\Delta^\star_{n,k}(t)$ is absolutely continuous for all $k \in \mathbf{N}$. Now all we have to do is take the maximum of $\Delta^\star_{n,k}(t)$ over all $k \in \mathbf{N}$ and prove it is also absolutely continuous. Consider the processes $\Delta^\star_k(t) = \Delta^\star_{n,0} \vee \Delta^\star_{n,1} \vee \cdots \vee \Delta^\star_{n,k}$ for $k \in \mathbf{N}$. Proceeding by induction exactly as above, by exploiting the absolute continuity of the processes $\Delta^\star_{n,k}$, we finally obtain that the process $\Delta^\star_n(t) = \max_{k\in\mathbf{N}} \Delta^\star_{n,k}$ is measurable in (t,ω) and absolutely continuous in time. Noting that $\Delta^\star_n(t) = \Delta_\infty(t)$, we are done. □

Lemma 3.3.6. *Assume $\mu, \nu \in \mathring{\mathcal{S}}^n$. There exists a $\{\mathcal{Y}_t\}$-predictable selection of indices $(t,\omega) \mapsto (i^\star(t,\omega), k^\star(t,\omega))$ such that*

$$\Delta_\infty(t,\omega) = \Delta_{i^\star(t,\omega)k^\star(t,\omega)}(t,\omega) \quad \textit{for all } t,\omega.$$

Moreover, the dynamics of $\Delta_\infty(t,\omega)$ are given by

$$\begin{aligned} \mathrm{d}\Delta_\infty(t) &= \sum_{i\in\mathbf{N}}\sum_{k\in\mathbf{N}} \mathbf{1}_{\{(i^\star,k^\star)(t)=(i,k)\}} \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(t)\,\mathrm{d}t, \\ \Delta_\infty(0) &= \log\frac{\mu^{i^\star(0)}}{\mu^{k^\star(0)}} - \log\frac{\nu^{i^\star(0)}}{\nu^{k^\star(0)}}. \end{aligned} \tag{3.15}$$

Proof. Consider the measurable space $(M,\mathcal{M})$, where $M = ([0,\infty)\times\Omega)$ and $\mathcal{M}$ is the $\{\mathcal{Y}_t\}$-predictable σ-algebra. Let $U = \mathbf{N}\times\mathbf{N}$ be endowed with the discrete topology. Consider the function $f : M\times U \to \mathbb{R}$ such that $f((t,\omega),(i,k)) = \Delta_{ik}(t,\omega)$. Note that $z(\cdot,(i,k)) = \Delta_{ik}(\cdot)$ is $\mathcal{M}$-measurable for all $(i,k)\in U$ by Proposition 3.3.3. Moreover, $z((t,\omega),\cdot) = \Delta_\cdot(t,\omega)$ is continuous as a function $U\to\mathbb{R}$ (because it is defined on the discrete space $U = \mathbf{N}\times\mathbf{N}$). The function $\Delta_\infty : M\to\mathbb{R}$ is $\mathcal{M}$-measurable by Lemma 3.3.5. Since $\Delta_\infty = \max_{(i,k)\in\mathbf{N}\times\mathbf{N}}\Delta_{ik}$, we must have that the image of Δ_∞ is contained in the image of f. In other words, we have

$$\Delta_\infty(t,\omega) \in f((t,\omega),U) \quad \forall (t,\omega)\in M.$$

Then by Filippov's implicit function lemma (see e.g. [28, Theorem A.10.2]) there exists an $\mathcal{M}$-measurable (i.e. $\{\mathcal{Y}_t\}$-predictable) map $u : M\to U$ that maps $(t,\omega)\mapsto (i^\star(t,\omega),k^\star(t,\omega))$ such that

$$\Delta_\infty(t,\omega) = f\big((t,\omega),u(t,\omega)\big) = f\big((t,\omega),(i^\star(t,\omega),k^\star(t,\omega))\big) = \Delta_{i^\star(t,\omega)k^\star(t,\omega)}(t,\omega).$$

To prove the second part of the Lemma, recall that by Lemma 3.3.5 we have that $\Delta_\infty(t,\omega)$ is absolutely continuous. Then $\mathrm{d}\Delta_\infty(t) = g(t)\,\mathrm{d}t$ for some density $g(t)$ such that $\int_0^t |g(s)|\,\mathrm{d}s < \infty$ a.s. for each $t\geq 0$. Since $\sum_{i\in\mathbf{N}}\sum_{k\in\mathbf{N}}\mathbf{1}_{\{(i^\star,k^\star)(t)=(i,k)\}} = 1$, we can write

$$\Delta_\infty(t) = \Delta_\infty(0) + \int_0^t \sum_{i\in\mathbf{N}}\sum_{k\in\mathbf{N}} \mathbf{1}_{\{(i^\star,k^\star)(t)=(i,k)\}} g(s)\,\mathrm{d}s.$$

So, if we can show that for any $(i,k)\in\mathbf{N}\times\mathbf{N}$ and any $t>0$ we have

$$\int_0^t \mathbf{1}_{\{(i^\star,k^\star)(t)=(i,k)\}}\Big|g(s) - \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(s)\Big|\,\mathrm{d}s = 0 \quad \text{a.s.}, \tag{3.16}$$

we are done. Rewriting the left-hand side of the above, we have

$$\begin{aligned}
0 &\leq \int_0^t \mathbf{1}_{\{(i^\star,k^\star)(t)=(i,k)\}} \Big| g(s) - \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(s) \Big| \,\mathrm{d}s \\
&\leq \int_0^t \mathbf{1}_{\{\Delta_\infty(s)-\Delta_{ik}(s)=0\}} \Big| g(s) - \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(s) \Big| \,\mathrm{d}s \\
&= \int_0^t \mathbf{1}_{\left\{\Delta_\infty(s)-\Delta_{ik}(s)=0,\, g(s)-\frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(s)\neq 0\right\}} \Big| g(s) - \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(s) \Big| \,\mathrm{d}s,
\end{aligned}$$

and since the set $\left\{s \,:\, \Delta_\infty(s) - \Delta_{ik}(s) = 0,\ g(s) - \frac{\mathrm{d}}{\mathrm{d}t}\Delta_{ik}(s) \neq 0\right\}$ has measure 0 by Lemma 3.3.4, we see (3.16) holds and the proof is complete. □

The equality (3.13) gives us the chance to spell out the following lemmata, which will be useful later.

Lemma 3.3.7. *Assume $\mu, \nu \in \mathring{\mathcal{S}}^n$. For all $t < \infty$, the indices $i^\star(t,\omega)$ and $k^\star(t,\omega)$ respectively maximize and minimize the quantity $\frac{\pi_t^j}{\tilde{\pi}_t^j}$ over $j \in \mathbf{N}$. Moreover, we have that $\frac{\pi_t^{i^\star}}{\tilde{\pi}_t^{i^\star}} =: M_t \geq 1$ and $\frac{\pi_t^{k^\star}}{\tilde{\pi}_t^{k^\star}} =: \frac{1}{m_t} \leq 1$ for all $t < \infty$.*

Proof. Fix $(t,\omega) \in [0,\infty) \times \Omega$. Recalling Lemma 3.3.6, we see from the definition of Δ_{ik} and (3.13) that

$$\begin{aligned}
\log \frac{\pi_t^{i^\star(t,\omega)}}{\tilde{\pi}_t^{i^\star(t,\omega)}}(\omega) - \log \frac{\pi_t^{k^\star(t,\omega)}}{\tilde{\pi}_t^{k^\star(t,\omega)}}(\omega) &= \Delta_{i^\star(t,\omega)k^\star(t,\omega)}(t,\omega) \\
&= \Delta_\infty(t,\omega) = \log \max_{i\in\mathbf{N}} \frac{\pi_t^i}{\tilde{\pi}_t^i}(\omega) - \log \min_{k\in\mathbf{N}} \frac{\pi_t^k}{\tilde{\pi}_t^k}(\omega),
\end{aligned}$$

so the first part of the lemma follows. For the second part, assume for contradiction that there exists $\omega \in \Omega$ such that $M_t(\omega) < 1$. Then, for all $j \in \mathbf{N}$

$$\frac{\pi_t^j}{\tilde{\pi}_t^j}(\omega) \leq M_t(\omega) < 1 \Longrightarrow \pi_t^j < \tilde{\pi}_t^j,$$

which implies that $\sum_j \pi_t^j < 1$, and contradicts the fact that π_t is a probability distribution. The argument for $1/m_t(\omega)$ is analogous. □

Lemma 3.3.8. *Assume $\mu, \nu \in \mathring{\mathcal{S}}^n$. For all $i,k \in \mathbf{N} \times \mathbf{N}$, define $T_{ik}(t) := \frac{\pi_t^i}{\pi_t^k} - \frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}$. For all $t < \infty$, we have that $T_{ji^\star}(t) \leq 0$ and $T_{jk^\star}(t) \geq 0$, where $i^\star = i^\star(t,\omega)$ and $k^\star = k^\star(t,\omega)$ are the maximizing/minimizing indices from Lemma 3.3.7.*

Proof. Trivially, $T_{i^\star i^\star}(t) = T_{k^\star k^\star}(t) = 0$. Now consider $T_{jk^\star}(t)$ for $j \neq k^\star$. Note that

$$T_{jk^\star}(t) = \frac{\pi_t^j}{\pi_t^{k^\star}} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^{k^\star}} = \left(\frac{\pi_t^j}{\tilde{\pi}_t^j} - \frac{\pi_t^{k^\star}}{\tilde{\pi}_t^{k^\star}}\right)\frac{\tilde{\pi}_t^j}{\pi_t^{k^\star}}, \quad \forall j \in \mathbf{N}, j \neq k^\star.$$

By Lemma 3.3.7, $k^\star$ minimizes $\frac{\pi_t^j}{\tilde{\pi}_t^j}$ over $j \in \mathbf{N}$, so we have that $\frac{\pi_t^j}{\tilde{\pi}_t^j} - \frac{\pi_t^{k^\star}}{\tilde{\pi}_t^{k^\star}} \geq 0$ for all $j \neq k^\star$. Moreover, $\tilde{\pi}_t^j / \pi_t^{k^\star} > 0$ as well, since π_t and $\tilde{\pi}_t$ have positive entries for $t < \infty$. We conclude that $T_{jk^\star}(t) \geq 0$ for all $t < \infty$.

For the case of $T_{ji^\star}(t)$ we argue in the same way by noting that

$$T_{ji^\star}(t) = \frac{\pi_t^j}{\pi_t^{i^\star}} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^{i^\star}} = -\left(\frac{\pi_t^{i^\star}}{\tilde{\pi}_t^{i^\star}} - \frac{\pi_t^j}{\tilde{\pi}_t^j}\right)\frac{\tilde{\pi}_t^j}{\pi_t^{i^\star}}, \quad \forall j \in \mathbf{N}, j \neq i^\star,$$

and using that $i^\star$ maximizes $\frac{\pi_t^j}{\tilde{\pi}_t^j}$. □

Finally, we will need the following continuity result for the Hilbert metric.

Lemma 3.3.9. *For any $(\mu, \nu) \in \mathcal{S}^n \times \mathcal{S}^n$, it holds that*

(i) $\liminf_{m\to\infty} \mathcal{H}(\mu_m, \nu_m) \geq \mathcal{H}(\mu, \nu)$*, for all sequences $(\mu_m, \nu_m) \to (\mu, \nu)$ converging in the Euclidean metric;*

(ii) there exists a sequence $(\mu_m, \nu_m) \in \mathring{\mathcal{S}}^n \times \mathring{\mathcal{S}}^n$ such that $\lim_{m\to\infty} \mathcal{H}(\mu_m, \nu_m) = \mathcal{H}(\mu, \nu)$, and $(\mu_m, \nu_m) \to (\mu, \nu)$ in the Euclidean metric;

(iii) if $\mu, \nu \in \mathring{\mathcal{S}}^n$, then $\lim_{m\to\infty} \mathcal{H}(\mu_m, \nu_m) = \mathcal{H}(\mu, \nu)$, for all sequences $(\mu_m, \nu_m) \to (\mu, \nu)$ converging in the Euclidean metric.

Proof. If $(\mu, \nu) \in \mathring{\mathcal{S}}^n \times \mathring{\mathcal{S}}^n$, the result is immediate from the definition of the Hilbert metric and continuity of division and logarithms, establishing *(iii)*. Consider now a pair of sequences $\{\mu_m\}, \{\nu_m\} \in \mathring{\mathcal{S}}^n$ convergent in the Euclidean metric, with respective limits $\mu, \nu \in \mathcal{S}^n$. Suppose first that $\mu \nsim \nu$, then it is easy to verify that either $\max_i\{\mu_m^i/\nu_m^i\} \to \infty$ or $\min_i\{\mu_m^i/\nu_m^i\} \to 0$, hence $\mathcal{H}(\mu_m, \nu_m) \to \infty = \mathcal{H}(\mu, \nu)$.
Suppose instead that $\mu \sim \nu$, let $J = \{j : \mu^j = \nu^j = 0\}$. If

$$\limsup_{m\to\infty} \max_{j\in J}\{\mu_m^j/\nu_m^j\} \leq \max_{i:\nu^i>0}\{\mu^i/\nu^i\} \quad \text{and} \quad \liminf_{m\to\infty} \min_{j\in J}\{\mu_m^j/\nu_m^j\} \geq \min_{i:\nu^i>0}\{\mu^i/\nu^i\},$$

then a direct calculation shows $\mathcal{H}(\mu_m, \nu_m) \to \mathcal{H}(\mu, \nu)$. Since we can always choose $\{\mu_m\}$ and $\{\nu_m\}$ satisfying the two above inequalities, this proves *(ii)*.
If a given sequence (μ_m, ν_m) does not satisfy the inequalities above, then take any subsequence, still indexed by m, such that $\mathcal{H}(\mu_m, \nu_m)$ converges in $[0, \infty]$, and such that at least one of the above inequalities is violated for every term in the subsequence; in particular, suppose that for some $\epsilon > 0$, for all m,

$$\max_{j\in J}\{\mu_m^j/\nu_m^j\} > \max_{i:\nu^i>0}\{\mu^i/\nu^i\} + \epsilon.$$

Then

$$\lim_{m\to\infty} \mathcal{H}(\mu_m,\nu_m) \geq \log\left(\frac{\max_{i:\nu^i>0}\{\mu^i/\nu^i\}+\epsilon}{\min_{i:\nu^i>0}\{\mu^i/\nu^i\}}\right) > \mathcal{H}(\mu,\nu).$$

For any subsequence with $\min_{j\in J}\{\mu_m^j/\nu_m^j\} < \min_{i:\nu^i>0}\{\mu^i/\nu^i\}-\epsilon$, a similar argument holds. Therefore, we conclude $\liminf_{m\to\infty}\mathcal{H}(\mu_m,\nu_m)\geq\mathcal{H}(\mu,\nu)$, which is *(i)*. □

3.3.3 Proof of Theorem 3.2.1

We are now ready to prove Theorem 3.2.1.

Proof of Theorem 3.2.1. Let us start by considering (3.15), and assuming $\mu,\nu\in\mathring{\mathcal{S}}^n$. Writing it out in full we have

$$\begin{aligned}\mathrm{d}\Delta_\infty(t) &= \sum_{i\in\mathbf{N}}\sum_{k\in\mathbf{N}} \mathbf{1}_{\{(i^\star,k^\star)(t)=(i,k)\}}\frac{\mathrm{d}}{\mathrm{d}t}\left(\log\frac{\pi_t^i}{\pi_t^k}-\log\frac{\tilde\pi_t^i}{\tilde\pi_t^k}\right)\mathrm{d}t\\ &= \left[-\sum_{\substack{j=0\\ j\neq k^\star}}^n q_{jk^\star}\left(\frac{\pi_t^j}{\pi_t^{k^\star}}-\frac{\tilde\pi_t^j}{\tilde\pi_t^{k^\star}}\right)+\sum_{\substack{j=0\\ j\neq i^\star}}^n q_{ji^\star}\left(\frac{\pi_t^j}{\pi_t^{i^\star}}-\frac{\tilde\pi_t^j}{\tilde\pi_t^{i^\star}}\right)\right]\mathrm{d}t, \end{aligned}\tag{3.17}$$

where we have dropped the (t,ω)-dependence of $(i^\star,k^\star)$ for readability.

Rewriting (3.17) in the notation of Lemma 3.3.8, we have

$$\begin{aligned}\mathrm{d}\Delta_\infty(t) &= -\left[\sum_{\substack{j=0\\ j\neq k^\star}}^n q_{jk^\star}T_{jk^\star}(t)-\sum_{\substack{j=0\\ j\neq i^\star}}^n q_{ji^\star}T_{ji^\star}(t)\right]\mathrm{d}t\\ &= -\big(q_{i^\star k^\star}T_{i^\star k^\star}(t)-q_{k^\star i^\star}T_{k^\star i^\star}(t)\big)\,\mathrm{d}t-\sum_{\substack{j=0\\ j\neq i^\star,k^\star}}^n\left(q_{jk^\star}T_{jk^\star}(t)-q_{ji^\star}T_{ji^\star}(t)\right)\mathrm{d}t.\end{aligned}\tag{3.18}$$

and the right-hand side is non-positive, since the off-diagonal entries of the Q-matrix are non-negative by definition, and the differences $q_{i^\star k^\star}T_{i^\star k^\star}(t)-q_{k^\star i^\star}T_{k^\star i^\star}(t)$ and $q_{jk^\star}T_{jk^\star}(t)-q_{ji^\star}T_{ji^\star}(t)$ for $j\neq i^\star,k^\star$ are also all non-negative by Lemma 3.3.8. We look for an upper bound on the Stieltjes measure $\mathrm{d}\Delta_\infty(t)$ on $[0,\infty)$. Dropping non-positive terms in the sum, we simplify to

$$\begin{aligned}\mathrm{d}\Delta_\infty(t) &\leq -\big(q_{i^\star k^\star}T_{i^\star k^\star}(t)-q_{k^\star i^\star}T_{k^\star i^\star}(t)\big)\,\mathrm{d}t\\ &= -\left[q_{i^\star k^\star}\left(\frac{\pi_t^{i^\star}}{\pi_t^{k^\star}}-\frac{\tilde\pi_t^{i^\star}}{\tilde\pi_t^{k^\star}}\right)+q_{k^\star i^\star}\left(\frac{\tilde\pi_t^{k^\star}}{\tilde\pi_t^{i^\star}}-\frac{\pi_t^{k^\star}}{\pi_t^{i^\star}}\right)\right]\mathrm{d}t\\ &= -\left[q_{i^\star k^\star}\left(M_t-\frac{1}{m_t}\right)\frac{\tilde\pi_t^{i^\star}}{\pi_t^{k^\star}}+q_{k^\star i^\star}\left(m_t-\frac{1}{M_t}\right)\frac{\pi_t^{k^\star}}{\tilde\pi_t^{i^\star}}\right]\mathrm{d}t\\ &\leq -2\sqrt{q_{i^\star k^\star}q_{k^\star i^\star}}\left(\frac{M_tm_t-1}{\sqrt{M_tm_t}}\right)\mathrm{d}t,\end{aligned}$$

where $M_t := \frac{\pi_t^{i^\star}}{\tilde{\pi}_t^{i^\star}}$ and $m_t := \frac{\tilde{\pi}_t^{k^\star}}{\pi_t^{k^\star}}$ as in Lemma 3.3.7, and we have made use of the inequality $a + b \geq 2\sqrt{ab}$ for $a, b \geq 0$. Recall that $\Delta_\infty(t) = \Delta_{i^\star k^\star}(t) = \log(M_t m_t) \geq 0$, since $M_t m_t \geq 1$ by Lemma 3.3.7. Then we can rewrite the inequality above as

$$\mathrm{d}\Delta_\infty(t) \leq -4\sqrt{q_{i^\star k^\star} q_{k^\star i^\star}} \sinh\left(\frac{\Delta_\infty(t)}{2}\right) \mathrm{d}t$$

Now, if $\Delta_\infty(t) = 0$, then the theorem holds trivially, so we can assume $\Delta_\infty(t) > 0$. Since $\sinh(x) > 0$ for $x > 0$, we can divide both sides by $\sinh\big(\Delta_\infty(t)/2\big)$. Integrating over $[s, t]$ yields

$$\log\tanh\left(\frac{\Delta_\infty(t)}{4}\right) - \log\tanh\left(\frac{\Delta_\infty(s)}{4}\right) \leq -\lambda(t - s),$$

where we have defined $\lambda := 2\min_{i \neq j} \sqrt{q_{ij} q_{ji}}$, and it follows that

$$\tanh\left(\frac{\Delta_\infty(t)}{4}\right) \leq \tanh\left(\frac{\Delta_\infty(0)}{4}\right) e^{-\lambda t}.$$

Concavity and monotonicity of $\tanh(x)$ for $x \geq 0$ imply that $\tanh(x)e^{-\lambda t} \leq \tanh(xe^{-\lambda t})$ and hence

$$\Delta_\infty(t) \leq \Delta_\infty(0) e^{-\lambda t},$$

and, recalling (3.13), we are done.

Finally, we lift the assumption that $\mu, \nu \in \mathring{\mathcal{S}}^n$. From Lemma 3.3.9, we can choose sequences $\mu_m, \nu_m \in \mathring{\mathcal{S}}^n$ such that $\mathcal{H}(\mu_m, \nu_m) \to \mathcal{H}(\mu, \nu)$, and $(\mu_m, \nu_m) \to (\mu, \nu)$ in the Euclidean norm. Consider the corresponding filters $\pi_t(\mu_m), \tilde{\pi}_t(\nu_m)$ (the solutions to (3.4) and (3.5) initialized at μ_m and ν_m respectively). As the Kushner–Stratonovich equations are Lipschitz on $\mathcal{S}^n$, by standard stability results for SDEs (e.g. [28, Theorem 16.4.3]), we know that $\pi_t(\mu_m) \to \pi_t$ and $\tilde{\pi}_t(\nu_m) \to \tilde{\pi}_t$ in probability as $m \to \infty$.

We know from Lemma 3.3.2 that $\pi_t, \tilde{\pi}_t \in \mathring{\mathcal{S}}^n$. Using the continuity given in Lemma 3.3.9, and applying the result above, we know that

$$\begin{aligned}\tanh\Big(\frac{\mathcal{H}(\pi_t, \tilde{\pi}_t)}{4}\Big) &= \lim_{m\to\infty}\Big[\tanh\Big(\frac{\mathcal{H}(\pi_t(\mu_m), \tilde{\pi}_t(\nu_m))}{4}\Big)\Big] \\ &\leq \lim_{m\to\infty}\Big[\tanh\Big(\frac{\mathcal{H}(\mu_m, \nu_m)}{4}\Big)\Big] e^{-\lambda t} = \tanh\Big(\frac{\mathcal{H}(\mu, \nu)}{4}\Big) e^{-\lambda t}\end{aligned}$$

as desired. Monotonicity and concavity of tanh again complete the argument. □

Remark 3.2. Previously, the best non-asymptotic stability estimate for the continuous-time finite state-space nonlinear filter was given by

$$\|\pi_t - \tilde{\pi}_t\|_{\ell^1} \leq \Big(2 \wedge \frac{\|\mu - \nu\|_{\ell^1}}{\min_k\{\mu^k, \nu^k\}}\Big) e^{-\lambda t} \tag{3.19}$$

(where λ is the same rate as in Theorem 3.2.1), which is obtained by combining [26, Proposition 3.5] (which relies on [12, Lemma 5.7]) and [83, Corollary 2.3.2]. Combining Theorem 3.2.1 with Theorem 2.5.1, we obtain the tighter bound

$$\|\pi_t - \tilde{\pi}_t\|_{\ell^1} \leq 2\tanh\Big(\frac{\mathcal{H}(\mu,\nu)}{4}\Big)e^{-\lambda t}. \tag{3.20}$$

3.3.4 On the optimality of the contraction rate

The deterministic contraction rate $\lambda = 2\min_{i\neq j}\sqrt{q_{ij}q_{ji}}$ that we just proved is sharp for the case $\pi_t, \tilde{\pi}_t \in \mathcal{S}^1$ uniformly in $\mu, \nu \in \mathring{\mathcal{S}}^1$, in the sense that if we have $\rho \in \mathbb{R}$ s.t. $\Delta_\infty(t) \leq \Delta_\infty(s)e^{-\rho(t-s)}$ *a.s.* for all $s < t$, we know $\rho \leq \lambda$. In this basic case, the maximum process is simply given by $\Delta_\infty = |\theta_0^1 - \tilde{\theta}_0^1|$. Consider the specific situation when Q is symmetric, so that the diagonal entries are given by $q_{00} = q_{11} = -q$ and the off-diagonal entries by $q_{01} = q_{10} = q$. Then $\lambda = 2q$. We compute

$$\begin{aligned}
\mathrm{d}|\theta_0^1(t) - \tilde{\theta}_0^1(t)| &= \operatorname{sign}\big(\theta_0^1(t) - \tilde{\theta}_0^1(t)\big)\Big[-q\big(e^{\theta_0^1(t)} - e^{\tilde{\theta}_0^1(t)} + e^{-\tilde{\theta}_0^1(t)} - e^{-\theta_0^1(t)}\big)\Big]\,\mathrm{d}t \\
&= \operatorname{sign}\big(\theta_0^1(t) - \tilde{\theta}_0^1(t)\big)\Big[-2q\big[(\theta_0^1(t) - \tilde{\theta}_0^1(t)) + \frac{1}{3!}((\theta_0^1(t))^3 - (\tilde{\theta}_0^1(t))^3) + \dots\big]\Big]\,\mathrm{d}t \\
&= -2q|\theta_0^1(t) - \tilde{\theta}_0^1(t)|\Big(1 + \frac{1}{3!}\big((\theta_0^1(t))^2 + (\tilde{\theta}_0^1(t))^2 + \theta_0^1(t)\tilde{\theta}_0^1(t)\big) + \dots\Big)\,\mathrm{d}t \\
&=: -2q|\theta_0^1(t) - \tilde{\theta}_0^1(t)|R_t\,\mathrm{d}t,
\end{aligned}$$

from which we deduce

$$|\theta_0^1(t) - \tilde{\theta}_0^1(t)| = e^{-2q\int_s^t R_u\,\mathrm{d}u}|\theta_0^1(s) - \tilde{\theta}_0^1(s)|, \quad \text{for all } 0 \leq s \leq t. \tag{3.21}$$

Note that R_t is close to 1 iff $\theta_0^1(t) \approx \tilde{\theta}_0^1(t) \approx 0$, which happens if π_t and $\tilde{\pi}_t$ are near the centre of the simplex, i.e. $\pi_t \approx \tilde{\pi}_t \approx \big(\frac{1}{2}, \frac{1}{2}\big)$. Let $\tau < \infty$ be a time such that $R_\tau \approx 1$. Note that such τ exists with positive probability, since the Brownian dynamics (under a change of measure) for $\theta_0^1(t)$ ensure that $\theta_0^1(t)$ must visit 0 infinitely many times; then by Theorem 3.2.1, for all large enough τ, there exists $\varepsilon \ll 1$ such that $|\tilde{\theta}_0^1(\tau) - \theta_0^1(\tau)| \leq \varepsilon$.

Since R_t is continuous in time, for all $\delta > 0$ there exists t_δ such that $R_s \in (1, 1+\delta]$ for all $s \in [\tau, \tau + t_\delta]$. Then by (3.21) for all $s \in [\tau, \tau + t_\delta]$,

$$|\theta_0^1(\tau + \tau_\delta) - \tilde{\theta}_0^1(\tau + \tau_\delta)| \geq e^{-2q(1+\delta)t_\delta}\Big|\theta_0^1(\tau) - \tilde{\theta}_0^1(\tau)\Big|.$$

Since δ was arbitrary, we see that the bound $\lambda = 2q$ is achieved. We illustrate this in a simulated example in Figure 3.1(left).

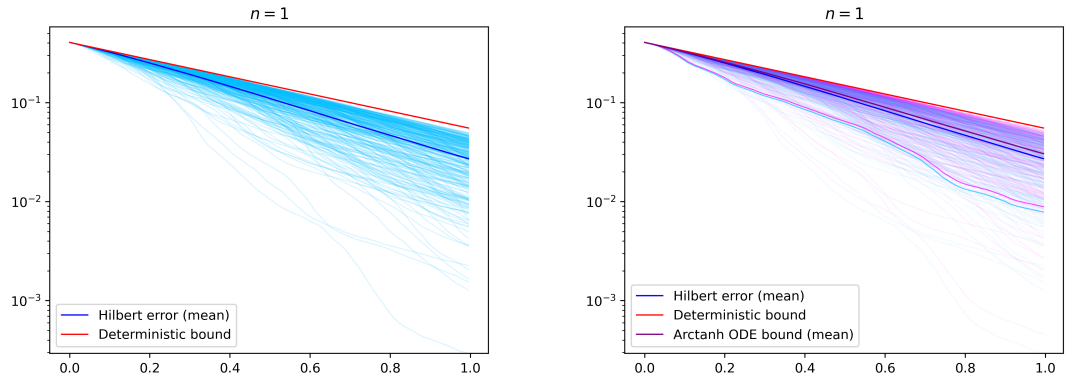

Figure 3.1: On the left, in log scale, we plot 300 realizations of the Hilbert projective error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$ (in light blue), where $\pi_t, \tilde{\pi}_t \in \mathring{\mathcal{S}}^1 \subset \mathbb{R}^2$ are solutions to (3.4) initialized at $\mu, \nu \in \mathring{\mathcal{S}}^1$ respectively. In blue we plot the sample mean, and in red we have the deterministic pathwise bound $\mathcal{H}(\mu, \nu)e^{-\lambda t}$ from Theorem 3.2.1. We see that the bound is attained. For this simulation, the signal X is a 2-state Markov chain with symmetric rate matrix Q and jump rate $q = 1$, initialized at its invariant distribution $\mu = \left(\frac{1}{2}, \frac{1}{2}\right)$. This gives $\pi_0 = \mu$ as the initial condition for the Wonham filter. The 'wrong' filter $\tilde{\pi}_t$ is initialized at $\nu = \left(\frac{2}{5}, \frac{3}{5}\right)$. Fixing these parameters, and setting the sensor function to be $h = (-1, 1)$, we simulate 300 paths for the signal and the observation processes, compute the two filters for these paths, and plot their Hilbert error. On the right, we plot once more the realizations of the Hilbert error (very light blue), and add to the same plot the pathwise ODE bounds (fuchsia) given by solving numerically, for each realization of $\tilde{\pi}_t$, the ODE (3.23) from Proposition 3.3.11. In purple we plot the mean of the ODE bounds; out of 300 pairs of Hilbert error and ODE bound, we highlight one at random.

On the other hand, as the dimension of the state-space increases, numerical experiments suggest that λ becomes less optimal. Recalling the notation from the previous subsection and looking back at our proof, it is easy to pinpoint the cause of this sub-optimality to having discarded the negative sum of terms of the form $q_{jk^\star}T_{jk^\star} - q_{ji^\star}T_{ji^\star}$ on the right-hand side of equation (3.18). In particular, there are $n-2$ such negative terms, and they can take values in $(0, \infty)$, which suggests that, as n increases, the derivative of Δ_∞ becomes more negative (and potentially quite substantially so), and Δ_∞ should in fact tend to 0 faster than our bound indicates.

Unfortunately, we have not been able to find a uniform bound from below of the form $K_q\Delta_\infty$ for $\sum_j (q_{jk^\star}T_{jk^\star} - q_{ji^\star}T_{ji^\star})$, where K_q is some constant depending only on Q. However, assuming one can observe the path of the wrongly initialized filter $\tilde{\pi}_t$, we can provide a sharper, $\tilde{\pi}_t$-dependent bound for the decay rate. This allows one to dynamically compute a more refined estimate of the stability error.

We will need the following classical result, which we include for completeness.

Lemma 3.3.10 (Comparison principle)**.** *Let $X_t \in \mathbb{R}$ be an absolutely continuous process such that its almost everywhere derivative satisfies*

$$\mathrm{d}X_t \leq \alpha(t, X_t)\,\mathrm{d}t, \qquad X_0 = x_0,$$

on $[0, \infty)$, where $x \mapsto \alpha(t, x)$ is locally Lipschitz continuous. Let u_t be the unique solution (up to its first explosion time $T > 0$) to the ODE

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} = \alpha(t, u_t), \qquad u_0 = x_0. \tag{3.22}$$

Then $X_t \leq u_t$ for all $t < T$.

Proof. First of all, recall that standard results in ODE theory (see e.g. [81, Theorem 2.5]) give that (3.22) has a unique solution u_t up to its first explosion time $T > 0$. For $t < T$, consider $H_t = X_t - u_t$. Note that $H_0 = 0$, and that H_t is absolutely continuous with a.e. derivative satisfying

$$\frac{\mathrm{d}H_t}{\mathrm{d}t} \leq \alpha(t, X_t) - \alpha(t, u_t).$$

Assume for a contradiction that there exists $\tau < T$ s.t. $H_\tau > 0$. By continuity of H_t, there exists $t_0 \in [0, \tau)$ such that $H_{t_0} = 0$ and $H_s \geq 0$ for all $s \in [t_0, \tau]$. Moreover, by continuity of X_t and u_t, there exists $R \in \mathbb{R}$ such that $X_s, u_s \in (-R, R)$ for all $s \in [t_0, \tau]$. Then

$$\begin{aligned} H_\tau = \int_{t_0}^{\tau} \frac{\mathrm{d}H_s}{\mathrm{d}s}\,\mathrm{d}s &\leq \int_{t_0}^{\tau} \big(\alpha(s, X_s) - \alpha(s, u_s)\big)\,\mathrm{d}s \\ &\leq \int_{t_0}^{\tau} C_R(s)|X_s - u_s|\,\mathrm{d}s = \int_{t_0}^{\tau} C_R(s) H_s\,\mathrm{d}s, \end{aligned}$$

where $C_R(t) \geq 0$ is the Lipschitz constant of $\alpha(t, x_t)$ for $x_t \in (-R, R)$, and we have used that $H_s = X_s - u_s > 0$ on $[t_0, \tau]$ by assumption. Then Grönwall's inequality yields that $H_\tau \leq 0$, which is a contradiction. Therefore $X_t \leq u_t$ for all $t < T$. □

Proposition 3.3.11. *Suppose $\mu^i, \nu^i > 0$ for all $i \in \mathbf{N}$. For all $t \geq 0$, let $u_t \in (0, 1)$ be the unique solution to the ODE with random coefficients given by*

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} = -\tilde{\lambda}^\star(t, u_t)u_t, \qquad u_0 = \tanh\left(\frac{\mathcal{H}(\mu, \nu)}{4}\right), \tag{3.23}$$

where

$$\tilde{\lambda}^\star(t,u_t)=\min_{i\neq k}\left\{\left(q_{ik}\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}+\sum_{\substack{j\neq i,k,\\ j\notin\tilde{\mathcal{J}}_k^i(t,u_t)}}q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\frac{1+u_t}{1-u_t}+\left(q_{ki}\frac{\tilde{\pi}_t^k}{\tilde{\pi}_t^i}+\sum_{\substack{j\neq i,k,\\ j\in\tilde{\mathcal{J}}_k^i(t,u_t)}}q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\right)\frac{1-u_t}{1+u_t}\right\}, \tag{3.24}$$

and

$$\tilde{\mathcal{J}}_k^i(t,u_t):=\left\{j\in\mathbf{N}\,:\,\frac{q_{jk}}{\tilde{\pi}_t^k}\geq\frac{q_{ji}}{\tilde{\pi}_t^i}\Big(\frac{1-u_t}{1+u_t}\Big)^2\right\}.$$

Then for all $t<\infty$,

$$\tanh\left(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\right)\leq u_t. \tag{3.25}$$

In particular, $\tilde{\lambda}^\star(t,u_t)\geq\tilde{\lambda}_t^\star$, *where*

$$\begin{aligned}\tilde{\lambda}_t^\star&:=2\min_{i\neq k}\left\{\min_{S\subseteq\mathbf{N}}\sqrt{q_{ik}q_{ki}+\sum_{j\in S,\,j\neq i,k}q_{ji}q_{ik}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}+\sum_{l\notin S,\,l\neq i,k}q_{lk}q_{ki}\frac{\tilde{\pi}_t^l}{\tilde{\pi}_t^i}+\sum_{\substack{j\in S,\\ j\neq i,k}}\sum_{\substack{l\notin S,\\ l\neq i,k}}q_{ji}q_{lk}\frac{\tilde{\pi}_t^j\tilde{\pi}_t^l}{\tilde{\pi}_t^i\tilde{\pi}_t^k}}\right\},\\ &\geq 2\min_{i\neq k}\left\{\sqrt{q_{ik}q_{ki}+\sum_{j\neq i,k}\min\left\{\frac{q_{ji}q_{ik}}{\tilde{\pi}_t^k},\frac{q_{jk}q_{ki}}{\tilde{\pi}_t^i}\right\}\tilde{\pi}_t^j}\right\}=:\tilde{\lambda}_t,\end{aligned} \tag{3.26}$$

which gives that for all $t<\infty$,

$$\tanh\left(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\right)\leq\tanh\left(\frac{\mathcal{H}(\mu,\nu)}{4}\right)e^{-\int_0^t\tilde{\lambda}_s^\star\,\mathrm{d}s}. \tag{3.27}$$

Proof. Recall the notation from the proof of Theorem 3.2.1. Consider (3.18) and apply the chain-rule to derive the dynamics of $\tanh(\Delta_\infty(t)/4)$, to yield

$$\begin{aligned}\mathrm{d}\tanh\left(\frac{\Delta_\infty(t)}{4}\right)=&-\frac{1}{4}\cosh^{-2}\left(\frac{\Delta_\infty(t)}{4}\right)\big(q_{i^\star k^\star}T_{i^\star k^\star}(t)-q_{k^\star i^\star}T_{k^\star i^\star}(t)\big)\,\mathrm{d}t\\ &-\frac{1}{4}\cosh^{-2}\left(\frac{\Delta_\infty(t)}{4}\right)\sum_{\substack{j=0\\ j\neq i^\star,k^\star}}^n\Big(q_{jk^\star}T_{jk^\star}(t)-q_{ji^\star}T_{ji^\star}(t)\Big)\,\mathrm{d}t.\end{aligned} \tag{3.28}$$

We consider the terms on the right-hand side of the above equation one by one. Start from $T_{i^\star k^\star}(t)$ and notice that

$$\begin{aligned}T_{i^\star k^\star}(t)&=e^{\tilde{\theta}_{k^\star}^{i^\star}(t)}\big(e^{\theta_{k^\star}^{i^\star}(t)-\tilde{\theta}_{k^\star}^{i^\star}(t)}-1\big)=e^{\tilde{\theta}_{k^\star}^{i^\star}(t)}\big(e^{\Delta_\infty(t)}-1\big)\\ &=e^{\tilde{\theta}_{k^\star}^{i^\star}(t)+\frac{\Delta_\infty(t)}{2}}\big(e^{\frac{\Delta_\infty(t)}{2}}-e^{-\frac{\Delta_\infty(t)}{2}}\big)=2e^{\tilde{\theta}_{k^\star}^{i^\star}(t)+\frac{\Delta_\infty(t)}{2}}\sinh\left(\frac{\Delta_\infty(t)}{2}\right).\end{aligned}$$

Recalling the identity $\sinh(2x) = 2\sinh(x)\cosh(x)$, we have that

$$\frac{T_{i^\star k^\star}(t)}{\cosh^2\left(\frac{\Delta_\infty(t)}{4}\right)} = 4e^{\tilde{\theta}^{i^\star}_{k^\star}(t)+\frac{\Delta_\infty(t)}{2}}\tanh\left(\frac{\Delta_\infty(t)}{4}\right) = 4\frac{\tilde{\pi}^{i^\star}_t}{\tilde{\pi}^{k^\star}_t}e^{\frac{\Delta_\infty(t)}{2}}\tanh\left(\frac{\Delta_\infty(t)}{4}\right).$$

Similarly,

$$-\frac{T_{k^\star i^\star}(t)}{\cosh^2\left(\frac{\Delta_\infty(t)}{4}\right)} = 4e^{-\tilde{\theta}^{i^\star}_{k^\star}(t)-\frac{\Delta_\infty(t)}{2}}\tanh\left(\frac{\Delta_\infty(t)}{4}\right) = 4\frac{\tilde{\pi}^{k^\star}_t}{\tilde{\pi}^{i^\star}_t}e^{-\frac{\Delta_\infty(t)}{2}}\tanh\left(\frac{\Delta_\infty(t)}{4}\right).$$

Now, for $j \neq i^\star, k^\star$, consider $q_{jk^\star}T_{jk^\star}(t) - q_{ji^\star}T_{ji^\star}(t)$. Note that

$$\Delta_{jk^\star}(t) = \theta^j_{k^\star}(t) - \tilde{\theta}^j_{k^\star}(t) = \left(\theta^{i^\star}_{k^\star}(t) - \tilde{\theta}^{i^\star}_{k^\star}(t)\right) + \left(\theta^j_{i^\star}(t) - \tilde{\theta}^j_{i^\star}(t)\right) = \Delta_\infty(t) + \Delta_{ji^\star}.$$

By Lemma 3.3.8, and recalling that log is increasing, we have that $\Delta_{jk^\star}(t) \geq 0$ and $\Delta_{ji^\star} \leq 0$. Moreover, by definition of $\Delta_\infty(t)$, we have that $\Delta_{jk^\star}(t) \leq \Delta_\infty(t)$ and $|\Delta_{ji^\star}(t)| \leq \Delta_\infty(t)$. Then we can write

$$\begin{aligned} q_{jk^\star}T_{jk^\star}(t) - q_{ji^\star}T_{ji^\star}(t) &= q_{jk^\star}e^{\tilde{\theta}^j_{k^\star}}(e^{\Delta_{jk^\star}(t)} - 1) + q_{ji^\star}e^{\tilde{\theta}^j_{i^\star}}(1 - e^{\Delta_{jk^\star}(t)-\Delta_\infty(t)}) \\ &= \tilde{\pi}^j_t\left[\frac{q_{jk^\star}}{\tilde{\pi}^{k^\star}_t}(e^{\Delta_{jk^\star}(t)} - 1) + \frac{q_{ji^\star}}{\tilde{\pi}^{i^\star}_t}(1 - e^{\Delta_{jk^\star}(t)-\Delta_\infty(t)})\right], \end{aligned}$$

and in particular if $\frac{q_{jk^\star}}{\tilde{\pi}^{k^\star}_t} \geq \frac{q_{ji^\star}}{\tilde{\pi}^{i^\star}_t}e^{-\Delta_\infty(t)}$, then $q_{jk^\star}T_{jk^\star}(t) - q_{ji^\star}T_{ji^\star}(t)$ is increasing in $\Delta_{jk^\star}(t)$; otherwise it is decreasing. Therefore

$$\frac{q_{jk^\star}}{\tilde{\pi}^{k^\star}_t} \geq \frac{q_{ji^\star}}{\tilde{\pi}^{i^\star}_t}e^{-\Delta_\infty(t)} \Longrightarrow \min_{0\leq\Delta_{jk^\star}(t)\leq\Delta_\infty(t)}\left\{q_{jk^\star}T_{jk^\star}(t) - q_{ji^\star}T_{ji^\star}(t)\right\} = q_{ji^\star}\frac{\tilde{\pi}^j_t}{\tilde{\pi}^{i^\star}_t}(1 - e^{-\Delta_\infty(t)}),$$

$$\frac{q_{jk^\star}}{\tilde{\pi}^{k^\star}_t} < \frac{q_{ji^\star}}{\tilde{\pi}^{i^\star}_t}e^{-\Delta_\infty(t)} \Longrightarrow \min_{0\leq\Delta_{jk^\star}(t)\leq\Delta_\infty(t)}\left\{q_{jk^\star}T_{jk^\star}(t) - q_{ji^\star}T_{ji^\star}(t)\right\} = q_{jk^\star}\frac{\tilde{\pi}^j_t}{\tilde{\pi}^{k^\star}_t}(e^{\Delta_\infty(t)} - 1).$$

For all $t < \infty$, and all $i, k \in \mathbf{N} \times \mathbf{N}$, let

$$\tilde{J}^i_k(t, \Delta_\infty(t)) := \left\{j \in \mathbf{N} \,:\, \frac{q_{jk}}{\tilde{\pi}^k_t} \geq \frac{q_{ji}}{\tilde{\pi}^i_t}e^{-\Delta_\infty(t)} \text{ and } j \neq i, k\right\},$$

and ${}^{\mathsf{c}}\tilde{J}^i_k(t, \Delta_\infty(t)) := \mathbf{N} \setminus \left(\tilde{J}^i_k(t, \Delta_\infty(t)) \cup \{i, k\}\right)$. Putting all the above estimates together, we can bound (in the sense of Lebesgue–Stieltjes measures) the right-hand side of (3.28) as

$$\begin{aligned} \mathrm{d}\tanh\left(\frac{\Delta_\infty(t)}{4}\right) &\leq -\left[q_{i^\star k^\star}\frac{\tilde{\pi}^{i^\star}_t}{\tilde{\pi}^{k^\star}_t}e^{\frac{\Delta_\infty(t)}{2}} + q_{k^\star i^\star}\frac{\tilde{\pi}^{k^\star}_t}{\tilde{\pi}^{i^\star}_t}e^{-\frac{\Delta_\infty(t)}{2}}\right]\tanh\left(\frac{\Delta_\infty(t)}{4}\right)\mathrm{d}t \\ &\quad - \left[\sum_{j\in\tilde{J}^{i^\star}_{k^\star}(t,\Delta_\infty(t))} q_{ji^\star}\frac{\tilde{\pi}^j_t}{\tilde{\pi}^{i^\star}_t}e^{-\frac{\Delta_\infty(t)}{2}} + \sum_{j\in{}^{\mathsf{c}}\tilde{J}^i_k(t,\Delta_\infty(t))} q_{jk^\star}\frac{\tilde{\pi}^j_t}{\tilde{\pi}^{k^\star}_t}e^{\frac{\Delta_\infty(t)}{2}}\right]\tanh\left(\frac{\Delta_\infty(t)}{4}\right)\mathrm{d}t \\ &\leq -\tilde{\lambda}(t, \Delta_\infty(t))\tanh\left(\frac{\Delta_\infty(t)}{4}\right)\mathrm{d}t, \end{aligned} \tag{3.29}$$

where we have defined

$$\tilde{\lambda}(t,\Delta_\infty(t)) := \min_{i\neq k}\Bigg\{ q_{ik}\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}e^{\frac{\Delta_\infty(t)}{2}} + q_{ki}\frac{\tilde{\pi}_t^k}{\tilde{\pi}_t^i}e^{-\frac{\Delta_\infty(t)}{2}} + \sum_{j\in\tilde{J}_k^i(t,\Delta_\infty(t))} q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}e^{-\frac{\Delta_\infty(t)}{2}} + \sum_{j\in{}^{\mathsf{c}}\tilde{J}_k^i(t,\Delta_\infty(t))} q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}e^{\frac{\Delta_\infty(t)}{2}}\Bigg\}.$$

Using the inequality $a+b\geq 2\sqrt{ab}$ for $a,b\geq 0$, we have, $\forall (i,k)\in\mathbf{N}\times\mathbf{N}$, and $\forall t<\infty$,

$$\begin{aligned}
&q_{ik}\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}e^{\frac{\Delta_\infty(t)}{2}} + q_{ki}\frac{\tilde{\pi}_t^k}{\tilde{\pi}_t^i}e^{-\frac{\Delta_\infty(t)}{2}} + \sum_{j\in\tilde{J}_k^i(t,\Delta_\infty(t))} q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}e^{-\frac{\Delta_\infty(t)}{2}} + \sum_{j\in{}^{\mathsf{c}}\tilde{J}_k^i(t,\Delta_\infty(t))} q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}e^{\frac{\Delta_\infty(t)}{2}} \\
&\geq 2\left[\frac{e^{\frac{\Delta_\infty(t)}{2}}}{\tilde{\pi}_t^k}\left(q_{ik}\tilde{\pi}_t^i + \sum_{j\in{}^{\mathsf{c}}\tilde{J}_k^i(t,\Delta_\infty(t))} q_{jk}\tilde{\pi}_t^j\right)\right]^{\frac{1}{2}}\left[\frac{e^{\frac{-\Delta_\infty(t)}{2}}}{\tilde{\pi}_t^i}\left(q_{ki}\tilde{\pi}_t^k + \sum_{j\in\tilde{J}_k^i(t,\Delta_\infty(t))} q_{ji}\tilde{\pi}_t^j\right)\right]^{\frac{1}{2}} \\
&\geq 2\left(q_{ik}q_{ki} + \sum_{j\in\tilde{J}_k^i(t,\Delta_\infty(t))} q_{ik}q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k} + \sum_{l\in{}^{\mathsf{c}}\tilde{J}_k^i(t,\Delta_\infty(t))} q_{ki}q_{lk}\frac{\tilde{\pi}_t^l}{\tilde{\pi}_t^i} + \sum_{\substack{j\in\tilde{J}_k^i(t,\Delta_\infty(t)),\\ l\in{}^{\mathsf{c}}\tilde{J}_k^j(t,\Delta_\infty(t))}} q_{ji}q_{lk}\frac{\tilde{\pi}_t^j\tilde{\pi}_t^l}{\tilde{\pi}_t^i\tilde{\pi}_t^k}\right)^{\frac{1}{2}},
\end{aligned}$$

which yields

$$\tilde{\lambda}(t,\Delta_\infty(t)) \geq 2\min_{i\neq k}\min_{S\subseteq\mathbf{N}}\left(q_{ik}q_{ki} + \sum_{\substack{j\in S,\\ j\neq i,k}} q_{ji}q_{ik}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k} + \sum_{\substack{l\notin S,\\ l\neq i,k}} q_{lk}q_{ki}\frac{\tilde{\pi}_t^l}{\tilde{\pi}_t^i} + \sum_{\substack{j\in S,\\ j\neq i,k}}\sum_{\substack{l\notin S,\\ l\neq i,k}} q_{ji}q_{lk}\frac{\tilde{\pi}_t^j\tilde{\pi}_t^l}{\tilde{\pi}_t^i\tilde{\pi}_t^k}\right)^{\frac{1}{2}}. \tag{3.30}$$

Then, bounding the right-hand side of (3.29) and applying a Grönwall's argument (for absolutely continuous processes) yields (3.27). The inequality (3.26) follows immediately by minimizing further the right-hand side of (3.30), and in particular

$$\tilde{\lambda}(\tilde{\pi}_t,\Delta_\infty(t)) \geq \tilde{\lambda}_t := 2\min_{i\neq k}\left\{\sqrt{q_{ik}q_{ki} + \sum_{j\neq i,k}\min\left\{\frac{q_{ji}q_{ik}}{\tilde{\pi}_t^k}, \frac{q_{jk}q_{ki}}{\tilde{\pi}_t^i}\right\}\tilde{\pi}_t^j}\right\} > 0, \tag{3.31}$$

since by assumption Q has strictly positive non-diagonal entries and $\tilde{\pi}_t\in\mathring{\mathcal{S}}^n$ for all $t<\infty$ by Lemma 3.3.1.

Now let $X_t := \tanh(\Delta_\infty(t)/4)$. Then we can rewrite (3.29) as

$$\mathrm{d}X_t \leq -\tilde{\lambda}^\star(t,X_t)X_t\,\mathrm{d}t,$$

where

$$\tilde{\lambda}^\star(t, X_t) := \tilde{\lambda}(t, 4\,\mathrm{arctanh}(X_t))$$
$$= \min_{i\neq k}\left\{\left(q_{ik}\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k} + \sum_{j\in{}^{\mathsf{c}}\tilde{\mathcal{J}}_k^i(t,X_t)} q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\frac{1+X_t}{1-X_t} + \left(q_{ki}\frac{\tilde{\pi}_t^k}{\tilde{\pi}_t^i} + \sum_{j\in\tilde{\mathcal{J}}_k^i(t,X_t)} q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\frac{1-X_t}{1+X_t}\right\},$$

and $\tilde{\mathcal{J}}_k^i(t, X_t) := \tilde{J}_k^i(t, 4\,\mathrm{arctanh}(X_t))$ and ${}^{\mathsf{c}}\tilde{\mathcal{J}}_k^i(t, X_t) = \mathbf{N} \setminus \big(\tilde{\mathcal{J}}_k^i(t, X_t) \cup \{i, k\}\big)$. We want to compare X_t to the solution to (3.23). Next, we argue that (3.23) has a well-defined solution for all $t < \infty$. Then the proposition follows from a direct application of Lemma 3.3.10.

Let u_t be a solution to (3.23). Note that $\mathcal{H}(\mu, \nu) \in [0, \infty)$ since $\mu, \nu \in \mathring{\mathcal{S}}^n$ by assumption, so $u_0 \in [0, 1)$. If $\mathcal{H}(\mu, \nu) = 0$, then $u_0 = 0$, and $u_t = 0$ for all $t \geq 0$, so the proposition holds trivially. So from now on, assume $u_0 \in (0, 1)$. Now, the coefficient $\tilde{\lambda}^\star(\tilde{\pi}_t, u_t)$ depends on the process $\tilde{\pi}_t$, which is fixed ω-by-ω. Observe that $\tilde{\lambda}^\star(\tilde{\pi}_t, u_t)$ blows up when $u_t \uparrow 1$. Since $u_0 < 1$, the explosion time T such that $u_{T-} = 1$ is strictly positive. Recall that by Lemma (3.3.1) $\tilde{\pi}_t \in \mathring{\mathcal{S}}^n$ for all $t < \infty$. Then on the interval $[0, T)$, $x \mapsto \tilde{\lambda}^\star(\tilde{\pi}_t, x)x$ is locally Lipschitz continuous (with Lipschitz constant dependent on t, ω and x) and standard results in ODE theory (see e.g. [81, Theorem 2.5]) give that (3.23) has a unique solution u_t in $[0, T)$. On the other hand, $\tilde{\lambda}^\star(\tilde{\pi}_t, u_t) \geq \tilde{\lambda}_t \geq 2\min_{i\neq k}\sqrt{q_{ik}q_{ki}}$ is strictly positive for $u_t \in (-1, 1)$, so $-\tilde{\lambda}^\star(\tilde{\pi}_t, u_t)u_t$ is strictly negative for $u_t \in (0, 1)$. Then $u_t \leq u_0 < 1$ for all $t \geq 0$, so in fact the explosion time $T = \infty$ and (3.23) has a unique solution for all $t \geq 0$. Moreover, $-\tilde{\lambda}^\star(\tilde{\pi}_t, u_t)u_t$ tends to 0 as u_t approaches 0, hence it readily follows that $u_t \in (0, u_0]$ for all $t \geq 0$. □

By symmetry, the bounds of Proposition 3.3.11 can also be expressed in terms of the true filter π_t.

Corollary 3.3.11.1. *Suppose $\mu^i, \nu^i > 0$ for all $i \in \mathbf{N}$, and π_t is observed. For all $t \geq 0$, let $u_t \in (0, 1)$ be the unique solution to the ODE with random coefficients given by*

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} = -\lambda^\star(t, u_t)u_t, \qquad u_0 = \tanh\left(\frac{\mathcal{H}(\mu, \nu)}{4}\right), \tag{3.32}$$

where

$$\lambda^\star(t, u_t) = \min_{i\neq k}\left\{\left(q_{ik}\frac{\pi_t^i}{\pi_t^k} + \sum_{\substack{j\neq i,k,\\ j\in\mathcal{J}_k^i(t,u_t)}} q_{jk}\frac{\pi_t^j}{\pi_t^k}\right)\frac{1-u_t}{1+u_t} + \left(q_{ki}\frac{\pi_t^k}{\pi_t^i} + \sum_{\substack{j\neq i,k,\\ j\in\mathcal{J}_k^i(t,u_t)}} q_{ji}\frac{\pi_t^j}{\pi_t^i}\right)\frac{1+u_t}{1-u_t}\right\}, \tag{3.33}$$

and $\mathcal{J}_k^i(t,u_t) := \left\{ j \in \mathbf{N} \,:\, \frac{q_{jk}}{\pi_t^k} \leq \frac{q_{ji}}{\pi_t^i}\left(\frac{1+u_t}{1-u_t}\right)^2 \right\}$. *Then for all* $t < \infty$,

$$\tanh\left(\frac{\mathcal{H}(\pi_t, \tilde{\pi}_t)}{4}\right) \leq u_t.$$

In particular, for all $t < \infty$, $\tanh\left(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\right) \leq \tanh\left(\frac{\mathcal{H}(\mu,\nu)}{4}\right) e^{-\int_0^t \lambda_s^\star \,\mathrm{d}s}$, *where* $\lambda_t^\star \geq \lambda_t$, *with* $\lambda_t^\star$ *and* λ_t *defined equivalently to* $\tilde{\lambda}_t^\star$ *and* $\tilde{\lambda}_t$ *in* (3.26), *with* π_t *in place of* $\tilde{\pi}_t$.

Proof. Similar to the proof of Proposition 3.3.11. □

In (3.26) we state a lower bound $\tilde{\lambda}_t$ for $\tilde{\lambda}_t^\star$ because, numerically, finding $\tilde{\lambda}_t^\star$ by minimizing over all possible subsets of $\mathbf{N}$ for each t can be costly, especially in high dimensions. On the other hand, $\tilde{\lambda}_t$ is easy to compute. We illustrate the performance of the bounds from Proposition 3.3.11 in Figure 3.2: we plot both the pathwise bound $\mathcal{H}(\mu,\nu)e^{-2\int_0^t \tilde{\lambda}_s \,\mathrm{d}s}$, which follows directly from (3.27), using the rate $\tilde{\lambda}_t$, and the ODE bound given by $4\,\mathrm{arctanh}(u_t)$, where u_t is the numerical solution to (3.23), and compare them with the deterministic rate from Theorem 3.2.1.

As we can see from the plots in Figure 3.2, even when using the pathwise contraction rate or the ODE bound from Proposition 3.3.11, our bounds are not tight in dimension $n \geq 2$. This affects our simulations for the error bounds in Section 4.4 as well. Since further algebraic manipulations in the spirit of what we have attempted so far do not seem likely to yield a better bound, one could think of improving our estimates by looking instead at the rate of decay of the expected Hilbert error, which from our simulations seems very well behaved, or even at the expected contraction rate. To proceed in either of these directions, one would need to find a way to estimate the expectation of the indicators of the argmax and argmin of the ratios between the components of π_t and $\tilde{\pi}_t$.

Our numerical experiments also suggest, at least for the examples we consider, that there is a concentration of measure phenomenon occurring in high-dimensional examples, where a much faster convergence rate than we have established will hold with overwhelming probability. We leave the study of this problem open for future research.

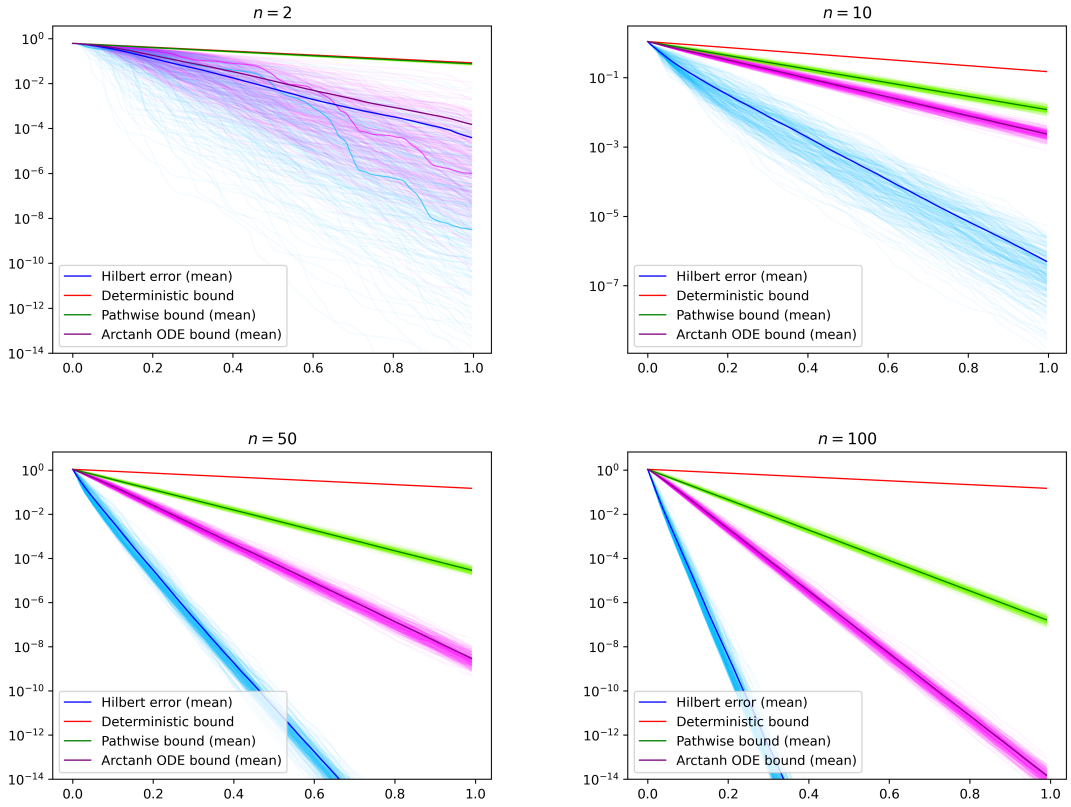

Figure 3.2: For dimensions $n = 2, 20, 50, 100$, and $\pi_t, \tilde{\pi}_t \in \mathring{\mathcal{S}}^n$, initialized at $\mu, \nu \in \mathring{\mathcal{S}}^n$ respectively, we plot 300 realizations of the Hilbert projective error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$ (light blue), of the pathwise bound $\mathcal{H}(\mu, \nu)e^{-\int_0^t \tilde{\lambda}_s \, \mathrm{d}s}$ (light green), and of the ODE bound from Proposition 3.3.11 (fuchsia), all in log scale, for $t \in [0, 1]$. In blue, darker green and purple we have the respective sample means. In red we plot the deterministic bound $\mathcal{H}(\mu, \nu)e^{-\lambda t}$ from Theorem 3.2.1. In the case of $n = 2$, we highlight one realization of the Hilbert error (selected randomly), and its corresponding ODE bound. For this simulation, we keep the structure of the rate matrix fixed across all dimensions, so that the deterministic rate λ is also fixed, and it is equal to 2 throughout. We see that as the dimension increases, the deterministic contraction rate becomes less and less optimal, and that the improvement gained by computing the pathwise rate or the ODE bound instead is significant (although the bound still remains far from sharp). For further details about this simulation, see Appendix A.2.

Chapter 4

Robustness and error bounds

In this chapter we build on the contraction results of Chapter 3 to investigate the behaviour of the error when approximate filters, rather than the optimal filter, are employed. The error bounds are stated for a general approximate filter, but are most useful when the coefficient of the stochastic term of the approximate filter matches that of the Wonham SDE (3.4).

4.1 Main results

Consider a general approximate filter of the form

$$\mathrm{d}\tilde{\pi}_t = \tilde{f}_t \,\mathrm{d}t + \tilde{g}_t \,\mathrm{d}Y_t, \quad \tilde{\pi}_0 = \nu, \tag{4.1}$$

where $\tilde{f}_t, \tilde{g}_t$ are $\mathbb{R}^{n+1}$-valued $\{\mathcal{Y}_t\}$-predictable process and $\tilde{\pi}_t \in \mathring{\mathcal{S}}^n$ for all t, and we refer to Section 4.3 for the necessary assumptions on $\tilde{f}_t$, $\tilde{g}_t$ and $\tilde{\pi}_t$.

Theorem 4.1.1 (Bounds for the expected Hilbert error)**.** *Let π_t be the solution to* (3.4) *and $\tilde{\pi}_t$ the solution to* (4.1)*. Suppose $\mu, \nu \in \mathring{\mathcal{S}}^n$ and $q_{ij} > 0$ for all $i \neq j$. Assuming sufficient integrability in* (4.1) *(see Assumption (A4)), for all $t < \infty$, we have that*

$$\begin{aligned}\mathbf{E}\left[\mathcal{H}(\pi_t, \tilde{\pi}_t)\right] \leq \mathcal{H}(\mu,\nu)e^{-\lambda t} &+ \int_0^t e^{-\lambda(t-s)}\mathbf{E}\left[\max_{i,k}\left\{\mathcal{E}_s^{1,i} - \mathcal{E}_s^{1,k} - \frac{1}{2}\big(\mathcal{E}_s^{2,i} - \mathcal{E}_s^{2,k}\big)\right\}\right]\mathrm{d}s \\ &+ \max_j |h^j| \int_0^t e^{-\lambda(t-s)}\mathbf{E}\left[\max_{i,k}\left\{\mathcal{E}_s^{3,i} - \mathcal{E}_s^{3,k}\right\}\right]\mathrm{d}s \\ &+ \frac{1}{4}\sum_{(i,k)}\sum_{(j,l)\neq(i,k)} \mathbf{E}\left[\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}L_s^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))\right],\end{aligned}$$

where $\lambda = 2\min_{i\neq k}\sqrt{q_{ik}q_{ki}}$ *is the deterministic contraction rate from Theorem 3.2.1. For* $j \in \mathbf{N}$ *the error terms are given by*

$$\mathcal{E}_t^{1,j} = \left(\sum_{m=0}^{n} q_{mj}\frac{\tilde{\pi}_t^m}{\tilde{\pi}_t^j}\right) - \frac{\tilde{f}_t^j}{\tilde{\pi}_t^j}, \quad \mathcal{E}_t^{2,j} = (h^j)^2 - \frac{(\tilde{g}_t^j)^2}{(\tilde{\pi}_t^j)^2}, \quad \mathcal{E}_t^{3,j} = h^j - \frac{\tilde{g}_t^j}{\tilde{\pi}_t^j}, \tag{4.2}$$

and the processes $(\Delta_{ik}(t))_{t\geq 0}$ *for* $(i,k) \in \mathbf{N}\times\mathbf{N}$ *are defined as* $\Delta_{ik}(t) = \log\frac{\pi_t^i}{\pi_t^k} - \log\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}$. *For all* (i,k), $(j,l) \in \mathbf{N}\times\mathbf{N}$, $L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))$ *denotes the local time at* 0 *of the difference process* $(\Delta_{ik} - \Delta_{jl})$.

Assuming there is no error in the stochastic terms when comparing (4.1) and (3.4), a stronger result is possible.

Theorem 4.1.2 (Pathwise decay rate for the Hilbert error)**.** *Under the same assumptions as in Theorem 4.1.1, suppose that the error terms* $\mathcal{E}_t^{3,i}$ *defined in* (4.2) *vanish for all* $i \in \mathbf{N}$ *and all* $t \geq 0$, *and* $\tilde{\pi}_t$ *is observable. Let* $u_t \in (0,1)$ *be the unique solution to the ODE with random coefficients given by*

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} = -\tilde{\lambda}^\star(t,u_t)u_t + \frac{1}{4}\max_{i,k}\left\{\mathcal{E}_s^{1,i} - \mathcal{E}_s^{1,k}\right\}(1-u_t^2), \qquad u_0 = \tanh\left(\frac{\mathcal{H}(\mu,\nu)}{4}\right), \tag{4.3}$$

where

$$\tilde{\lambda}^\star(t,u_t) = \min_{i\neq k}\left\{\left(q_{ik}\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k} + \sum_{\substack{j\neq i,k,\\ j\notin\tilde{\mathcal{J}}_k^i(t,u_t)}} q_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\frac{1+u_t}{1-u_t} + \left(q_{ki}\frac{\tilde{\pi}_t^k}{\tilde{\pi}_t^i} + \sum_{\substack{j\neq i,k,\\ j\in\tilde{\mathcal{J}}_k^i(t,u_t)}} q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\right)\frac{1-u_t}{1+u_t}\right\},$$

and $\tilde{\mathcal{J}}_k^i(t,u_t) := \left\{j \in \mathbf{N} : \frac{q_{jk}}{\tilde{\pi}_t^k} \geq \frac{q_{ji}}{\tilde{\pi}_t^i}\left(\frac{1-u_t}{1+u_t}\right)^2\right\}$. *Then for all* $t < \infty$,

$$\tanh\left(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\right) \leq u_t.$$

In particular, $\tilde{\lambda}^\star(t,u_t) \geq \tilde{\lambda}_t^\star$, *where*

$$\begin{aligned}\tilde{\lambda}_t^\star &:= 2\min_{i\neq k}\left\{\min_{S\subseteq\mathbf{N}}\sqrt{q_{ik}q_{ki} + \sum_{j\in S,\, j\neq i,k} q_{ik}q_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k} + \sum_{l\notin S,\, l\neq i,k} q_{ki}q_{lk}\frac{\tilde{\pi}_t^l}{\tilde{\pi}_t^i} + \sum_{\substack{j\in S,\\ j\neq i,k}}\sum_{\substack{l\notin S,\\ l\neq i,k}} q_{ji}q_{lk}\frac{\tilde{\pi}_t^j\tilde{\pi}_t^l}{\tilde{\pi}_t^i\tilde{\pi}_t^k}}\right\},\\ &\geq 2\min_{i\neq k}\left\{\sqrt{q_{ik}q_{ki} + \sum_{j\neq i,k}\min\left\{\frac{q_{ji}q_{ik}}{\tilde{\pi}_t^k}, \frac{q_{jk}q_{ki}}{\tilde{\pi}_t^i}\right\}\tilde{\pi}_t^j}\right\} =: \tilde{\lambda}_t,\end{aligned} \tag{4.4}$$

which gives that for all $t < \infty$, *we have the two bounds*

$$\tanh\left(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\right) \leq \tanh\left(\frac{\mathcal{H}(\mu,\nu)}{4}\right)e^{-\int_0^t\tilde{\lambda}_s^\star\,\mathrm{d}s} + \frac{1}{4}\int_0^t e^{-\int_s^t\tilde{\lambda}_r^\star\,\mathrm{d}r}\max_{i,k}\left\{\mathcal{E}_s^{1,i} - \mathcal{E}_s^{1,k}\right\}\mathrm{d}s, \tag{4.5}$$

and

$$\mathcal{H}(\pi_t, \tilde{\pi}_t) \leq \mathcal{H}(\mu, \nu) e^{-\int_0^t \tilde{\lambda}_s^\star \, \mathrm{d}s} + \int_0^t e^{-\int_s^t \tilde{\lambda}_r^\star \, \mathrm{d}r} \max_{i,k} \left\{ \mathcal{E}_s^{1,i} - \mathcal{E}_s^{1,k} \right\} \mathrm{d}s. \tag{4.6}$$

Remark 4.1. Theorem 4.1.2 suggests an approach to constructing approximate filters with relatively small error. If h is known, and we can choose $\tilde{g}_t = H\tilde{\pi}_t$ such that the error terms $\mathcal{E}^{3,i}$ (and consequently $\mathcal{E}^{2,i}$) vanish for all $i \in \mathbf{N}$, then the errors due to the stochastic term vanish, and the local time terms as well. From a numerical perspective, this is equivalent to killing the infinitesimal errors of order $\sqrt{\mathrm{d}t}$; this is natural when looking for an approximate solution to the Wonham SDE.

Remark 4.2. As in the proof of Theorem 3.2.1, by using Lemma 3.3.9, it is possible to lift the assumption that $\mu, \nu \in \mathring{\mathcal{S}}^n$ in Theorem 4.1.2, provided one takes sufficient care in constructing the solution to the ODE (4.3).

The following corollary provides some exchangeability between π_t and $\tilde{\pi}_t$ when computing the decay rates.

Corollary 4.1.2.1. *Assume the Wonham filter π_t is observable. Theorem 4.1.2 holds equivalently if one substitutes $\lambda^\star(t, u_t)$ for $\tilde{\lambda}^\star(t, u_t)$ in* (4.3) *and $\lambda_t^\star$ for $\tilde{\lambda}_t^\star$ in* (4.5) *and* (4.6)*, where*

$$\lambda^\star(t, u_t) = \min_{i \neq k} \left\{ \left(q_{ik} \frac{\pi_t^i}{\pi_t^k} + \sum_{\substack{j \neq i,k, \\ j \in \mathcal{J}_k^i(t,u_t)}} q_{jk} \frac{\pi_t^j}{\pi_t^k} \right) \frac{1-u_t}{1+u_t} + \left(q_{ki} \frac{\pi_t^k}{\pi_t^i} + \sum_{\substack{j \neq i,k, \\ j \in \mathcal{J}_k^i(t,u_t)}} q_{ji} \frac{\pi_t^j}{\pi_t^i} \right) \frac{1+u_t}{1-u_t} \right\},$$

and $\mathcal{J}_k^i(t, u_t) := \left\{ j \in \mathbf{N} \,:\, \frac{q_{jk}}{\pi_t^k} \leq \frac{q_{ji}}{\pi_t^i} \left(\frac{1+u_t}{1-u_t} \right)^2 \right\}$*, and $\lambda_t^\star$ is defined equivalently to* (4.4) *with π_t in place of $\tilde{\pi}_t$.*

4.2 Continuity of the Wonham filter with respect to the model parameters

In this section, we recover a version of Chigansky and Van Handel's results on robustness of the Wonham filter with respect to the model parameters (see [26]). Note that our approach is entirely different from [26], and we obtain robustness in terms of the Hilbert error instead of the ℓ^1-norm. Since the Hilbert metric is stronger than ℓ^1 (see Lemma 1 in [8]), the error estimates we obtain here are tighter than those in [26].

Consider an approximate Wonham filter with incorrect model parameters

$$\mathrm{d}\tilde{\pi}_t = \tilde{Q}^T \tilde{\pi}_t \, \mathrm{d}t + \left(\tilde{H} - \tilde{\pi}_t^\top \tilde{h} \, \mathbb{I}_{n+1} \right) \tilde{\pi}_t \left(\mathrm{d}Y_t - \tilde{\pi}_t^\top \tilde{h} \, \mathrm{d}t \right), \quad \tilde{\pi}_0 = \nu, \tag{4.7}$$

where $\tilde{Q} = (\tilde{q}_{ij})$ and $\tilde{h}$ are respectively a transition rate matrix and a bounded sensor function different from Q and h, and $\tilde{H} = \mathrm{diag}(\tilde{h})$ the diagonal matrix with entries $(\tilde{H})_{ii} = \tilde{h}^i$. We are interested in the Hilbert error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$.

Notation. Compared to Chapter 3, $(\tilde{\pi}_t)_{t\geq 0}$ now denotes the solution to (4.7), while $(\pi_t)_{t\geq 0}$ is still the solution to (3.4).

Remark 4.3. Once more, we recall that all our results can be extended to the case of multidimensional observations ($d \neq 1$) and invertible σ. In particular, note that the arguments we present here allow easily for $\sigma \neq 1$ in (3.1) and (3.4) (which would correspond to invertible $\sigma \neq \mathbb{I}_d$ in higher dimensions). This would add another 'misspecified' parameter $\tilde{\sigma}$ to (4.7). In the proofs we present below, σ and $\tilde{\sigma}$ can be directly incorporated into respectively h and $\tilde{h}$ in the equations (3.4) and (4.7) for the 'right' and 'wrong' Wonham filter. This case cannot be treated in [26] since the arguments therein require knowledge of the quadratic variation of the observations Y (see also [26, Remark 4.2]).

The rest of this section is devoted to proving the following theorem.

Theorem 4.2.1 (Model robustness)**.** *Let π_t be the solution to* (3.4) *and $\tilde{\pi}_t$ the solution to* (4.7)*. Assume $\mu^i, \nu^i > 0$ for all $i \in \mathbf{N}$ and $q_{ij}, \tilde{q}_{ij} > 0$ for all $i \neq j$. For all $t < \infty$,*

$$\mathbf{E}\left[\mathcal{H}(\pi_t, \tilde{\pi}_t)\right] \leq \mathcal{H}(\mu, \nu)e^{-\lambda t} + K_q \int_0^t e^{-\lambda(t-s)} \mathbf{E}\left[\frac{1}{\min_k \tilde{\pi}_s^k}\right] \mathrm{d}s + K_h \int_0^t e^{-\lambda(t-s)}\, \mathrm{d}s + \frac{1}{4} \sum_{(i,k)} \sum_{(j,l)\neq(i,k)} \mathbf{E}\left[\int_0^t e^{-\lambda(t-s)}\, \mathrm{d}L_s^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))\right],$$

where $\lambda = 2\min_{i\neq j} \sqrt{q_{ij}q_{ji}}$, $K_q = 2\max_{j,k} |\tilde{q}_{jk} - q_{jk}|$ and $K_h = 2\max_j |h^j| \max_i |h^i - \tilde{h}^i| + \max_i |(h^i)^2 - (\tilde{h}^i)^2|$, and $L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))$ denotes the local time at 0 of the process $(\Delta_{ik}(t) - \Delta_{jl}(t))_{t\geq 0}$, where Δ_{ik} evolves according to (4.8) *below for all $(i,k) \in \mathbf{N} \times \mathbf{N}$.*

Moreover, for all $t < \infty$, we have that the local time terms disappear as $\tilde{h} \to h$ for $\tilde{h}$ in a compact set around h, in the sense that there exists a constant $\tilde{C} < \infty$ such that

$$\lim_{\tilde{h}\to h} \mathbf{E}\left[\mathcal{H}(\pi_t, \tilde{\pi}_t)\right] \leq \mathcal{H}(\mu, \nu)e^{-\lambda t} + K_q \tilde{C}(1 - e^{-\lambda t}).$$

Even in the case where $\tilde{h}$ remains fixed, this result gives us good control over $\mathcal{H}(\pi_t, \tilde{\pi}_t)$, as shown by the next Proposition.

Proposition 4.2.2. *The error terms in Theorem 4.2.1 stay finite as $t \to \infty$. Specifically,*

$$\sup_{t\geq 0} \mathbf{E}\left[\int_0^t e^{-\lambda(t-s)}\, \mathrm{d}L_s^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))\right] < \infty, \quad \forall (i,k),(j,l) \in \mathbf{N}\times\mathbf{N}, (i,k)\neq(j,l).$$

Remark 4.4. We have stated Theorem 4.2.1 above with the decay rate λ a function of Q, the 'true' dynamics of the Markov chain, and the (first) error term a function of the 'misspecified' process $\tilde{\pi}_t$. However, nothing in our proof prevents us from doing the opposite, if we so wish: the theorem still holds if we replace λ with $\tilde{\lambda} = 2\min_{i\neq j}\sqrt{\tilde{q}_{ij}\tilde{q}_{ji}}$, and $\tilde{\pi}_t$ with π_t. Note that the error term due to the misspecification of h (and the local time terms) stay the same.

We now set out to prove these results. As in section 3.3, we start by transforming $\pi_t, \tilde{\pi}_t$ into $\theta_t, \tilde{\theta}_t$ and derive the dynamics of $\Delta_{ik}(t) = \theta_k^i(t) - \tilde{\theta}_k^i(t) = \log\frac{\pi^i}{\pi^k}(t) - \log\frac{\tilde{\pi}^i}{\tilde{\pi}^k}(t)$ for $(i,k)\in\mathbf{N}\times\mathbf{N}$ (recalling that $\Delta_{ii}=0$). Note that θ_k^i, $\tilde{\theta}_k^i$ and $\mathcal{H}(\pi_t,\tilde{\pi}_t)$ are all a.s. well-defined for finite $t\geq 0$, (since Lemma 3.3.1 holds equivalently when the dynamics of the filter $\tilde{\pi}$ are given by parameters $\tilde{Q}$ and $\tilde{h}$).

Applying Itô's formula to (3.4) and (4.7), we derive the dynamics of the difference process $\Delta_{ik}(t)$ as

$$\begin{aligned}
\mathrm{d}\Delta_{ik}(t) = &-\sum_{\substack{j=0\\ j\neq k}}^{n}\left(q_{jk}\frac{\pi_t^j}{\pi_t^k} - \tilde{q}_{jk}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\mathrm{d}t + \sum_{\substack{j=0\\ j\neq i}}^{n}\left(q_{ji}\frac{\pi_t^j}{\pi_t^i} - \tilde{q}_{ji}\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\right)\mathrm{d}t \\
&+ (q_{ii} - \tilde{q}_{ii} - q_{kk} + \tilde{q}_{kk})\,\mathrm{d}t + (h^i - \tilde{h}^i - h^k + \tilde{h}^k)(\,\mathrm{d}B_t + \pi_t^\top h\,\mathrm{d}t) \\
&+ \frac{1}{2}\Big((h^k)^2 - (\tilde{h}^k)^2 - (h^i)^2 + (\tilde{h}^i)^2\Big)\,\mathrm{d}t, \\
\Delta_{ik}(0) = &\log\frac{\mu^i}{\mu^j} - \log\frac{\nu^i}{\nu^j},
\end{aligned} \tag{4.8}$$

where we have again introduced the innovation process $B_t = Y_t - \int_0^t \pi_s^\top h\,\mathrm{d}s$.

We note straight away that if $\tilde{h} = h$, the stochastic term disappears, as well as all the drift terms involving h and $\tilde{h}$. So, in this simple case, we recover once again C^1 dynamics for $\Delta_{ik}(t)$ and arguments similar to the ones in Section 3.3 apply, to yield the following estimate for all $t<\infty$

$$\begin{aligned}
\mathcal{H}(\pi_t,\tilde{\pi}_t) &\leq \mathcal{H}(\mu,\nu)e^{-\int_0^t\tilde{\lambda}_s\,\mathrm{d}s} + \int_0^t e^{-\int_s^t\tilde{\lambda}_r\,\mathrm{d}r}\max_{i,k}\left\{\frac{(\delta Q^\top\tilde{\pi}_s)^i}{\tilde{\pi}_s^i} - \frac{(\delta Q^\top\tilde{\pi}_s)^k}{\tilde{\pi}_s^k}\right\}\mathrm{d}s \\
&\leq e^{-\lambda t}\mathcal{H}(\mu,\nu) + 2\max_{i,k}|\tilde{q}_{ik} - q_{ik}|\int_0^t e^{-\lambda(t-s)}\frac{1}{\min_j\tilde{\pi}^j(s)}\,\mathrm{d}s \\
&\leq e^{-\lambda t}\mathcal{H}(\mu,\nu) + \frac{2}{\lambda}\max_{i,k}|\tilde{q}_{ik} - q_{ik}|\,\frac{1}{\min_{j\in\mathbf{N},\,s\in[0,t]}\tilde{\pi}^j(s)}(1-e^{-\lambda t}),
\end{aligned} \tag{4.9}$$

where $\delta Q := Q - \tilde{Q}$ and $\min_{j \in \mathbf{N},\, s \in [0,t]} \tilde{\pi}^j(s) \neq 0$ almost surely by Lemma 3.3.1, and $\tilde{\lambda}_t$ and λ are respectively the pathwise contraction rate from (3.26) in Proposition 3.3.11 and the deterministic rate of Theorem 3.2.1. Tighter bounds for this error, in the spirit of Proposition 3.3.11, are possible, but we will state them later in Section 4.3, when we treat the error of general approximate filters (of which a filter with misspecified model parameters is a specific example).

In the case $\tilde{h} \neq h$, the strategy of proof developed in Section 3.3 cannot be applied directly. It is not unlikely that, by carefully modifying the arguments to account for the stochastic terms, for example by iterated application of Tanaka's formula, one could derive dynamics for $\Delta_\infty(t) = \max_{(i,k) \in \mathbf{N} \times \mathbf{N}} \Delta_{ik}(t)$ similar to (3.15). Here, however, we present a different strategy. We start by introducing a few definitions.

Recall the following smooth approximations of the maximum and the argmax.

Definition 4.2.3. Let $\alpha \in (0, \infty)$ and let $\mathbf{X}_t = \{X_t^0, \ldots, X_t^n\}$ be a family of real-valued, continuous random variables. Define the *LogSumExp* function $LSE_\alpha(\mathbf{X}_\cdot)(t)$ as

$$LSE_\alpha(\mathbf{X}_\cdot)(t) = \frac{1}{\alpha} \log \sum_i e^{\alpha X_t^i}. \tag{4.10}$$

For $\mathbf{g}_t = \{g_t^i(x)\}_{i=0}^n$ a family of real-valued functions, define the *SoftArgMax* (also known as *SoftMax*) function $S_\alpha^{arg}(\mathbf{X}_\cdot, \mathbf{g}_\cdot)(t)$ as

$$S_\alpha^{arg}(\mathbf{X}_\cdot, \mathbf{g}_\cdot)(t) = \frac{\sum_i g_t^i(X_t^i) e^{\alpha X_t^i}}{\sum_k e^{\alpha X_t^k}}. \tag{4.11}$$

Note that, for each ω, we have pointwise convergence of $LSE_\alpha(\mathbf{X}_\cdot)$ to $\max_i X^i$ and of $S_\alpha^{arg}(\mathbf{X}_\cdot, \mathbf{g}_\cdot(\mathbf{X}_\cdot))(t)$ to $\frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} g_t^j(X_t^j)$ where $\mathcal{I} = \operatorname{argmax}_i X^i$ as $\alpha \to \infty$ for every t (see Appendix A.1).

We will also need the definition of the *local time* L_a^t of a continuous semimartingale in the arguments that follow. In particular, we will only be concerned with local times of continuous semimartingales whose finite variation part is absolutely continuous (e.g. Itô processes). In this case we can take the following definition for the local time of a semimartingale Z with absolutely continuous finite variation part (adapting from Revuz and Yor [74, Chapter 6, Corollary 1.9] and noting that if the finite variation part of Z is absolutely continuous, then the proof of [74, Chapter 6, Theorem 1.7] yields that L_t^a has a a.s. bicontinuous modification in a and t).

Definition 4.2.4. Let $Z_t = M_t + A_t$ be a real valued continuous semimartingale, where M is a local martingale and A is an absolutely continuous finite variation

process. We take the *local time* of Z at $a \in \mathbb{R}$, at time t, to be the process L_a^t continuous in $t \in \mathbb{R}^+$ and a given by

$$L_t^a = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_0^t \mathbb{1}_{(a-\varepsilon, a+\varepsilon)}(Z_s)\, \mathrm{d}\langle Z \rangle_s \quad a.s.$$

4.2.1 Proof of Theorem 4.2.1

Our strategy for the proof of Theorem 4.2.1 is as follows. We use (4.10) to define a smooth approximation of the maximal process Δ_∞, and we derive its dynamics through Itô's formula. Then a bit of care is required when taking the limit as $\alpha \to \infty$, as some of the integrands converge to Dirac masses when the maximal process is attained at multiple indices at the same time. In Appendix A.1 we show how to deal with these terms, which determine the emergence of local times in our estimates.

Proof of Theorem 4.2.1. Consider the family of processes $\mathbf{\Delta}_t = \{\Delta_{ik}(t)\}_{i \in \mathbf{N}, k \in \mathbf{N}}$ with evolution equations given by (4.8). We apply Itô's Lemma to derive the dynamics of $LSE_\alpha(\mathbf{\Delta}_\cdot)(t)$

$$\begin{aligned}
\mathrm{d}LSE_\alpha(\mathbf{\Delta}_\cdot)(t) &= \sum_{(j,l)} \frac{e^{\alpha \Delta_{jl}(t)}}{\sum_{(i,k)} e^{\alpha \Delta_{ik}(t)}}\, \mathrm{d}\Delta_{jl}(t) \\
&+ \frac{1}{2} \sum_{(j,l)} \sum_{(u,v)} \alpha e^{\alpha \Delta_{jl}(t)} \left(\frac{\mathbb{1}_{\{(u,v)=(j,l)\}}}{\sum_{(i,k)} e^{\alpha \Delta_{ik}(t)}} - \frac{e^{\alpha \Delta_{uv}(t)}}{(\sum_{(i,k)} e^{\alpha \Delta_{ik}(t)})^2} \right) \mathrm{d}\langle \Delta_{jl}(\cdot), \Delta_{uv}(\cdot) \rangle_t,
\end{aligned}$$

where all the summations happen over the set of double indices $\mathbf{N} \times \mathbf{N}$.

Recalling our notation $T_{ik}(t) := \frac{\pi_t^i}{\pi_t^k} - \frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}$ from Section 3.3, we add and subtract terms appropriately to the dynamics of Δ_{ik} to yield

$$\begin{aligned}
\mathrm{d}\Delta_{ik}(t) &= \left(-\sum_{\substack{j=0 \\ j \neq k}}^n q_{jk} T_{jk}(t) + \sum_{\substack{j=0 \\ j \neq i}}^n q_{ji} T_{ji}(t) \right) \mathrm{d}t + \left(\frac{(\delta Q^\top \tilde{\pi}_t)^i}{\tilde{\pi}_t^i} - \frac{(\delta Q^\top \tilde{\pi}_t)^k}{\tilde{\pi}_t^k} \right) \mathrm{d}t \\
&\quad + (h^i - \tilde{h}^i - h^k + \tilde{h}^k)\, \mathrm{d}B_t + \big(h^i - \tilde{h}^i - h^k + \tilde{h}^k\big) \pi_t^\top h\, \mathrm{d}t \\
&\quad + \frac{1}{2} \Big((h^k)^2 - (\tilde{h}^k)^2 - (h^i)^2 + (\tilde{h}^i)^2 \Big)\, \mathrm{d}t \\
&=: C_{ik}^{q,1}(t)\, \mathrm{d}t + C_{ik}^{q,2}(t)\, \mathrm{d}t + C_{ik}^{h,1}\, \mathrm{d}B_t + C_{ik}^{h,1} \pi_t^\top h\, \mathrm{d}t + \frac{1}{2} C_{ik}^{h,2}\, \mathrm{d}t,
\end{aligned}$$

where again $\delta Q = Q - \tilde{Q}$.

Then we have, for all $s \leq t < \infty$

$$\begin{aligned}
LSE_\alpha(\boldsymbol{\Delta}_\cdot)(t) = LSE_\alpha(\boldsymbol{\Delta}_\cdot)(s) \\
&+ \int_s^t \sum_{(j,l)} \frac{e^{\alpha\Delta_{jl}(r)}}{\sum_{(i,k)} e^{\alpha\Delta_{ik}(r)}} \left(C_{jl}^{q,1}(r) + C_{jl}^{q,2}(r) + C_{jl}^{h,1}\, \pi_r^\top h + \frac{1}{2} C_{jl}^{h,2} \right) \mathrm{d}r \\
&+ \int_s^t \sum_{(j,l)} \frac{e^{\alpha\Delta_{jl}(r)}}{\sum_{(i,k)} e^{\alpha\Delta_{ik}(r)}} C_{jl}^{h,1}\, \mathrm{d}B_r \\
&+ \frac{1}{2} \int_s^t \sum_{(j,l)} \sum_{(u,v)\neq(j,l)} \frac{\alpha e^{\alpha(\Delta_{jl}(r)+\Delta_{uv}(r))}}{(\sum_{(i,k)} e^{\alpha\Delta_{ik}(r)})^2} \left((C_{jl}^{h,1})^2 - C_{jl}^{h,1} C_{uv}^{h,1} \right) \mathrm{d}r \\
&=: LSE_\alpha(\boldsymbol{\Delta}_\cdot)(s) + I_1 + I_2 + \frac{1}{2} I_3.
\end{aligned}$$

We want to take the limit, on both sides, as $\alpha \to \infty$.

Denote

$$\Delta_\infty(t) = \max_{(j,l)\in\mathbf{N}\times\mathbf{N}} \Delta_{jl}(t),$$

and we immediately have that $LSE_\alpha(\boldsymbol{\Delta}_\cdot)(t)$ converges to $\Delta_\infty(t)$ as $\alpha \to \infty$.

We let $\lambda = 2 \min_{i\neq k} \sqrt{q_{ik}\, q_{ki}}$, and for $j \in \mathbf{N}$ define the error terms

$$\mathcal{E}_t^{1,j} = \frac{(\delta Q^\top \tilde{\pi}_s)^j}{\tilde{\pi}_s^j}, \quad \mathcal{E}^{2,j} = (h^j)^2 - (\tilde{h}^j)^2, \quad \mathcal{E}^{3,j} = h^j - \tilde{h}^j.$$

For each time t, define by $\mathcal{I}_t \subset \mathbf{N}\times\mathbf{N}$ the argmax of $\boldsymbol{\Delta}_t$, i.e. the set of double indices (i,k) such that $\Delta_{ik}(t) = \Delta_\infty(t)$ for all $(i,k) \in \mathcal{I}_t$. Let $|\mathcal{I}_t|$ denote the size of $\mathcal{I}_t$. Let us consider I_1, I_2 and I_3 one at a time.

Start with I_1. We recognize as integrands $S_\alpha^{arg}(\boldsymbol{\Delta}_\cdot, \mathbf{C}^{q,1}(\cdot))(r)$, $S_\alpha^{arg}(\boldsymbol{\Delta}_\cdot, \mathbf{C}^{q,2}(\cdot))(r)$, and $S_\alpha^{arg}(\boldsymbol{\Delta}_\cdot, \mathbf{C}^{h,1})(r)\, \pi_r^\top h$, as well as $S_\alpha^{arg}(\boldsymbol{\Delta}_\cdot, \mathbf{C}^{h,2})(r)$. The first two terms are bounded respectively by $\max_{i,j\in\mathbf{N}\times\mathbf{N}} C_{ij}^{q,1}$ and $\max_{i,j\in\mathbf{N}\times\mathbf{N}} C_{ij}^{q,2}$, which are continuous in time and therefore integrable on $[0,t]$. The second two respectively by $\max_k |h^k| \max_{i,j\in\mathbf{N}\times\mathbf{N}} C^{h,1} < \infty$ and $\max_{i,j\in\mathbf{N}\times\mathbf{N}} C^{h,2} < \infty$ which are bounded by assumptions on h, and therefore integrable. Then we can apply the dominated convergence theorem to bring the limit inside the integral and Lemma A.1.2 yields, for

all $s \leq t$,

$$\begin{aligned}
\lim_{\alpha\to\infty} I_1 &= \int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \Big(C_{ik}^{q,1}(r) + C_{ik}^{q,2}(r) + C_{ik}^{h,1}\, \pi_r^\top h + \frac{1}{2} C_{ik}^{h,3} \Big)\, \mathrm{d}r \\
&\leq -\int_s^t \frac{4}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \sqrt{q_{ik}\, q_{ki}}\, \sinh\left(\frac{\Delta_{ik}(r)}{2} \right) \mathrm{d}r \\
&\quad + \int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \left(\frac{(\delta Q^\top \tilde{\pi}_r)^i}{\tilde{\pi}_r^i} - \frac{(\delta Q^\top \tilde{\pi}_r)^k}{\tilde{\pi}_r^k} \right) \mathrm{d}r \\
&\quad + \max_j |h^j| \int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \big(h^i - \tilde{h}^i - h^k + \tilde{h}^k \big)\, \mathrm{d}r \\
&\quad + \int_s^t \frac{1}{2|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \Big((h^k)^2 - (\tilde{h}^k)^2 - (h^i)^2 + (\tilde{h}^i)^2 \Big)\, \mathrm{d}r \\
&\leq -\lambda \int_s^t \Delta_\infty(r)\, \mathrm{d}r + \int_s^t \max_{i,k} \Big\{ \mathcal{E}_r^{1,i} - \mathcal{E}_r^{1,k} + \frac{1}{2} \big(\mathcal{E}^{2,k} - \mathcal{E}^{2,i} \big) \Big\}\, \mathrm{d}r \\
&\quad + \max_j |h^j| \max_{i,k} \big\{ \mathcal{E}^{3,i} - \mathcal{E}^{3,k} \big\} (t-s),
\end{aligned}$$

where we have bounded $C_{ik}^{q,1}(r)$ as in the proof of Theorem 3.2.1, and noted that $2\sinh(x/2) \geq x$ for $x \geq 0$.

Similarly, we can swap limit and integration when dealing with I_2 by dominated convergence for stochastic integrals, and we get

$$\lim_{\alpha\to\infty} I_2 = \int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} (h^i - \tilde{h}^i - h^k + \tilde{h}^k)\, \mathrm{d}B_r.$$

Finally, recalling (4.8) and noting that the processes $\{\Delta_{ik}\}$ are continuous semi-martingales of the form (A.1) considered in Appendix A.1, Proposition A.1.6 applies and we have

$$\lim_{\alpha\to\infty} I_3 \leq \frac{1}{2} \sum_{(i,k)} \sum_{(j,l)\neq(i,k)} \Big(L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot)) - L_s^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot)) \Big) \quad \text{a.s.},$$

where $L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))$ denotes the local time at 0, at time t, of the difference process $(\Delta_{ik}(r) - \Delta_{jl}(r))_{r\geq 0}$.

Putting all these estimates together, we have that, for all $s \leq t$,

$$\begin{aligned}
\Delta_\infty(t) &\leq \Delta_\infty(s) - \lambda \int_s^t \Delta_\infty(r)\, \mathrm{d}r + \int_s^t \max_{i,k} \Big\{ \mathcal{E}_r^{1,i} - \mathcal{E}_r^{1,k} + \frac{1}{2} \big(\mathcal{E}^{2,k} - \mathcal{E}^{2,i} \big) \Big\}\, \mathrm{d}r \\
&\quad + \max_j |h^j| \max_{i,k} \big\{ \mathcal{E}^{3,i} - \mathcal{E}^{3,k} \big\} (t-s) + \int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} (h^i - \tilde{h}^i - h^k + \tilde{h}^k)\, \mathrm{d}B_r \\
&\quad + \frac{1}{4} \sum_{(i,k)} \sum_{(j,l)\neq(i,k)} \Big(L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot)) - L_s^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot)) \Big).
\end{aligned}$$

Taking expectation with respect to the reference measure $\mathbb{P}$ on both sides, the stochastic integral vanishes since the integrand is bounded, so we have

$$\begin{aligned}\mathrm{d}\mathbf{E}\left[\Delta_\infty(t)\right] \leq -\lambda\mathbf{E}\left[\Delta_\infty(t)\right]\,\mathrm{d}t + \mathbf{E}\left[\max_{i,k}\left\{\mathcal{E}_t^{1,i} - \mathcal{E}_t^{1,k} + \frac{1}{2}\big(\mathcal{E}^{2,k} - \mathcal{E}^{2,i}\big)\right\}\right]\mathrm{d}t \\ + \max_j |h^j| \max_{i,k}\left\{\mathcal{E}^{3,i} - \mathcal{E}^{3,k}\right\}\mathrm{d}t + \frac{1}{4}\sum_{(i,k)}\sum_{(j,l)\neq(i,k)} \mathrm{d}\mathbf{E}\left[L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))\right],\end{aligned}$$

where the left-hand side and the last term on the right-hand side are to be understood as Lebesgue–Stieltjes measures. Using the product rule to find the dynamics of $e^{\lambda t}\mathbf{E}\left[\Delta_\infty(t)\right]$, and integrating both sides of the resulting differential inequality yields that, for all $s \leq t < \infty$,

$$\begin{aligned}\mathbf{E}\left[\Delta_\infty(t)\right] \leq \mathbf{E}\left[\Delta_\infty(s)\right]e^{-\lambda(t-s)} + \int_s^t e^{-\lambda(t-r)}\mathbf{E}\left[\max_{i,k}\left\{\mathcal{E}_r^{1,i} - \mathcal{E}_r^{1,k} + \frac{1}{2}\big(\mathcal{E}^{2,k} - \mathcal{E}^{2,i}\big)\right\}\right]\mathrm{d}r \\ + \max_j |h^j| \max_{i,k}\left\{\mathcal{E}^{3,i} - \mathcal{E}^{3,k}\right\}\int_s^t e^{-\lambda(t-r)}\,\mathrm{d}r \\ + \frac{1}{4}\sum_{(i,k)}\sum_{(j,l)\neq(i,k)}\int_s^t e^{-\lambda(t-r)}\,\mathrm{d}\mathbf{E}\left[L_r^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))\right].\end{aligned}$$

Bounding the error terms, since the local time is of finite variation (hence also its expectation), we can apply integration by parts for Stieltjes integrals twice and use Fubini–Tonelli on the last term of the right-hand side to obtain

$$\begin{aligned}\mathbf{E}\left[\Delta_\infty(t)\right] \leq \mathbf{E}\left[\Delta_\infty(s)\right]e^{-\lambda(t-s)} + 2\max_{i,k}|\tilde{q}_{ik} - q_{ik}|\int_s^t e^{-\lambda(t-r)}\mathbf{E}\left[\frac{1}{\min_j \tilde{\pi}_r^j}\right]\mathrm{d}r \\ + \left(2\max_j |h^j| \max_i |h^i - \tilde{h}^i| + \max_i |(h^i)^2 - (\tilde{h}^i)^2|\right)\int_s^t e^{-\lambda(t-r)}\,\mathrm{d}r \\ + \frac{1}{4}\sum_{(i,k)}\sum_{(j,l)\neq(i,k)}\mathbf{E}\left[\int_s^t e^{-\lambda(t-r)}\,\mathrm{d}L_r^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))\right], \qquad (4.12)\end{aligned}$$

for all $s \leq t < \infty$, which is what we set out to prove.

We now move on to the second part of the theorem. First of all, analogously to [26, Lemma 3.6], one can get an explicit bound on $\mathbf{E}[(\min_j \tilde{\pi}_t^j)^{-1}]$ which depends continuously on the parameters $(\nu, \tilde{Q}, \tilde{h})$ for $\nu \in \mathring{\mathcal{S}}^n$. In particular, we have

$$\mathbf{E}\left[\frac{1}{\min_j \tilde{\pi}_t^j}\right] \leq \max_j\left\{\frac{1}{\nu^j}\exp\left\{-\tilde{q}_{jj}t + \max_k(\tilde{h}^j - \tilde{h}^k)^2 t\right\}\right\}$$

for all $t < \infty$, and the first integral on the right-hand side of (4.12) is controlled as we take the limit as $\tilde{h} \to h$, for $\tilde{h}$ in a compact set around h. The second term clearly vanishes as $\tilde{h} \to h$.

Next, we move to the integrals against the local times. Let $\tilde{\pi}_t(v)$ denote the unique solution to (4.7) with $v \in \mathbb{R}^{n+1}$ in place of $\tilde{h}$. Note that the drift of each process $\Delta_{ik} - \Delta_{jl}$, for $(i,k),(j,l) \in \mathbf{N}\times\mathbf{N}$ and $(i,k) \neq (j,l)$, is then given by $b_t^{ik,jl}(\tilde{h})$, where

$$\begin{aligned} b_t^{ik,jl}(v) := &- \sum_{r=0}^{n} \left(q_{rk} \frac{\pi_t^r}{\pi_t^k} - \tilde{q}_{rk} \frac{\tilde{\pi}_t^r(v)}{\tilde{\pi}_t^k(v)} \right) + \sum_{r=0}^{n} \left(q_{ri} \frac{\pi_t^r}{\pi_t^i} - \tilde{q}_{ri} \frac{\tilde{\pi}_t^r(v)}{\tilde{\pi}_t^i(v)} \right) \\ &+ \sum_{r=0}^{n} \left(q_{rl} \frac{\pi_t^r}{\pi_t^l} - \tilde{q}_{rl} \frac{\tilde{\pi}_t^r(v)}{\tilde{\pi}_t^l(v)} \right) - \sum_{r=0}^{n} \left(q_{rj} \frac{\pi_t^r}{\pi_t^j} - \tilde{q}_{rj} \frac{\tilde{\pi}_t^r(v)}{\tilde{\pi}_t^j(v)} \right) \\ &+ (h^i - v^i - h^k + v^k - h^j + v^j + h^l - v^l) \pi_t^\top h \\ &+ \frac{1}{2} \Big((h^k)^2 - (v^k)^2 - (h^i)^2 + (v^i)^2 - (h^l)^2 + (v^l)^2 + (h^j)^2 - (v^j)^2 \Big). \end{aligned} \tag{4.13}$$

Consider the difference of $b^{ik,jl}(\tilde{h})$ and $b^{ik,jl}(h)$ on $[0,t]$. Using that $\tilde{\pi}(h)$ and $\tilde{\pi}(\tilde{h})$ live in the simplex, we get

$$\begin{aligned} &\mathbf{E}\left[\sup_{s\le t} \left| b_s^{ik,jl}(h) - b_s^{ik,jl}(\tilde{h}) \right| \right]^2 \\ &\le \sum_{u\in\{i,k,j,l\}} \sum_{r\neq u} \tilde{q}_{ru} \mathbf{E}\left[\sup_{s\le t} \left(\frac{1}{\tilde{\pi}_s^u(h)\tilde{\pi}_s^u(\tilde{h})} \right)^2 \right] \mathbf{E}\left[\sup_{s\le t} \left(\left| \tilde{\pi}_s^u(\tilde{h}) - \tilde{\pi}_s^u(h) \right| + \left| \tilde{\pi}_s^r(\tilde{h}) - \tilde{\pi}_s^r(h) \right| \right)^2 \right]. \end{aligned}$$

For all $u \in \mathbf{N}$, a minor extension of [26, Lemma 3.6] gives that the first expectation is controlled uniformly in $\tilde{h}$, for $\tilde{h}$ belonging to a compact set around h. Since $\tilde{\pi}_t$ lives in the simplex, the SDE (4.7) has Lipschitz coefficients, and we can apply standard stability arguments (such as [28, Theorem 16.4.3]) to see that the second expectation tends to 0 as $\tilde{h} \to h$. Hence we have ucp convergence $b^{ik,jl}(\tilde{h}) \to b^{ik,jl}(h)$ on $[0,t]$ as $\tilde{h} \to h$.

Now fix an arbitrary sequence $\{\tilde{h}_n\}_{n\in\mathbb{N}}$ such that $\tilde{h}_n \to h$. By the above, we can take a subsequence $\{\tilde{h}_{n_r}\}_{r\in\mathbb{N}}$ such that $b_s^{ik,jl}(\tilde{h}_{n_r})$ converges uniformly to $b_s^{ik,jl}(h)$ on $[0,t]$ a.s. From now on, when we write $\tilde{h} \to h$, we mean the limit along this subsequence. Denote by $(\Delta_{ik} - \Delta_{jl})_t^\star$ the limit of $(\Delta_{ik} - \Delta_{jl})_t$ as $\tilde{h} \to h$. Using this uniform convergence, we get that, a.s., for all $s \in [0,t]$,

$$\begin{aligned} (\Delta_{ik} - \Delta_{jl})_s^\star &= (\Delta_{ik} - \Delta_{jl})_0 + \lim_{\tilde{h}\to h} \int_0^s b_r^{ik,jl}(\tilde{h})\, \mathrm{d}r \\ &\quad + \lim_{\tilde{h}\to h} (h^i - \tilde{h}^i - h^k + \tilde{h}^k - h^j + \tilde{h}^j + h^l - \tilde{h}^l) B_s \\ &= (\Delta_{ik} - \Delta_{jl})_0 + \int_0^s b_r^{ik,jl}(h)\, \mathrm{d}r, \end{aligned}$$

and $(\Delta_{ik} - \Delta_{jl})_s^\star$ is absolutely continuous with derivative $b_s^{ik,jl}(h)$.

Now, by Tanaka's formula we have that

$$L_t^0(\Delta_{ik}-\Delta_{jl}) = |(\Delta_{ik}-\Delta_{jl})_t| - |(\Delta_{ik}-\Delta_{jl})_0| + \int_0^t \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s)\,\mathrm{d}(\Delta_{ik}-\Delta_{jl})_s,$$

with the convention $\operatorname{sign}(0) = -1$. Taking the limit as $\tilde{h} \to h$ on both sides of the equation above, the stochastic integral vanishes, and applying dominated convergence to the integral involving $b^{ik,jl}$, we have

$$\lim_{\tilde{h}\to h} L_t^0(\Delta_{ik}-\Delta_{jl}) = |(\Delta_{ik}-\Delta_{jl})_t^\star| - |(\Delta_{ik}-\Delta_{jl})_0| + \int_0^t \lim_{\tilde{h}\to h} \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s) b_s^{ik,jl}(\tilde{h})\,\mathrm{d}s. \tag{4.14}$$

Consider the limit inside the integral. Note that for all $s \le t$ such that $b_s^{ik,jl}(h) \neq 0$ and $(\Delta_{ik}-\Delta_{jl})_s^\star \neq 0$, we have a.s.

$$\lim_{\tilde{h}\to h} \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s) b_s^{ik,jl}(\tilde{h}) = \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s^\star) b_s^{ik,jl}(h).$$

Now consider $s \le t$ such that $b_s^{ik,jl}(h) = 0$. Then we have

$$\lim_{\tilde{h}\to h} \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s) b_s^{ik,jl}(\tilde{h}) = 0 = \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s^\star) b_s^{ik,jl}(h),$$

for all such s. Finally, consider times $s \le t$ such that $b_s^{ik,jl}(h) \neq 0$ but $(\Delta_{ik}-\Delta_{jl})_s^\star = 0$. Then potentially we have $\operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s) b_s^{ik,jl}(\tilde{h}) \nrightarrow \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s^\star) b_s^{ik,jl}(h)$ as $\tilde{h}$ goes to h. However, by Lemma 3.3.4, the set

$$\left\{ s \,:\, (\Delta_{ik}-\Delta_{jl})_s^\star = 0, \frac{\mathrm{d}}{\mathrm{d}s}(\Delta_{ik}-\Delta_{jl})_s^\star = b_s^{ik,jl}(h) \neq 0 \right\}$$

has Lebesgue measure zero. So finally we can conclude that a.s.

$$\lim_{\tilde{h}\to h} \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s) b_s^{ik,jl} = \operatorname{sign}((\Delta_{ik}-\Delta_{jl})_s^\star) b_s^{ik,jl}(h), \quad \text{for a.a. } s \le t,$$

and hence, by absolute continuity of $(\Delta_{ik}-\Delta_{jl})_t^\star$, the right-hand side of (4.14) is 0. Thus we have proven a.s. convergence of $L_t^0(\Delta_{ik}-\Delta_{jl}) \to 0$ as $\tilde{h} \to h$ along the subsequence $\{\tilde{h}_{n_r}\}_{r\in\mathbb{N}}$, which implies convergence in probability along the same subsequence. On the other hand, the original sequence $\{\tilde{h}_n\}_{n\in\mathbb{N}}$ was arbitrary, so we can repeat the argument above along any sequence and always find a subsequence along which $L_t^0(\Delta_{ik}-\Delta_{jl})$ converges to 0 in probability. It follows that $L_t^0(\Delta_{ik}-\Delta_{jl})$ vanishes in probability as $\tilde{h} \to h$. By Tanaka's formula, we can also check, similarly to how the ucp convergence was deduced, that $\mathbf{E}[L_t^0(\Delta_{ik}-\Delta_{jl})^2]$ is bounded uniformly in $\tilde{h}$, for $\tilde{h}$ in a compact set around h, and thus Vitali's convergence theorem gives $\mathbf{E}[L_t^0(\Delta_{ik}-\Delta_{jl})] \to 0$ as $\tilde{h} \to h$. This yields the theorem. □

Proof of Proposition 4.2.2. We focus on the local time terms, since by similar arguments to [26, Proposition 3.7], we immediately have that $\sup_{t>0} \mathbf{E}[(\min_k \tilde{\pi}_t^k)^{-1}] < \infty$. Let $(i,k),(j,l) \in \mathbf{N} \times \mathbf{N}$, with $(i,k) \neq (j,l)$. Recall that by Tanaka's formula we can write the local time at 0 of $X_t := (\Delta_{ik} - \Delta_{jl})_t$ as

$$L_t^0(X) = |X_t| - |X_0| + \int_0^t \operatorname{sign}(X_s)\,\mathrm{d}X_s.$$

Then we have

$$\begin{aligned}
\mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}L_s^0(X_\cdot)\Big] &= \mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}|X_s|\Big] + \mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\operatorname{sign}(X_s)\,\mathrm{d}X_s\Big] \\
&= \mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}|X_s|\Big] + \mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\operatorname{sign}(X_s) b_s^{ik,jl}(\tilde{h})\,\mathrm{d}s\Big] \\
&\leq \mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}|X_s|\Big] + \sup_{s\leq t}\mathbf{E}\Big[\big|b_s^{ik,jl}(\tilde{h})\big|\Big]\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}s,
\end{aligned} \tag{4.15}$$

where $b_s^{ik,jl}(\tilde{h})$ is the drift of X_t, defined in (4.13). Since

$$|b_t^{ik,jl}| \leq K_q \frac{1}{\min_i \pi_t^i} + K_{\tilde{q}} \frac{1}{\min_i \tilde{\pi}_t^i} + K_h,$$

where K_q, $K_{\tilde{q}}$ and K_h are constants only depending on Q, $\tilde{Q}$, h and $\tilde{h}$, it follows that the second term in (4.15) is finite as we take the supremum over all $t > 0$. As for the first term, integrating by parts twice we have that

$$\mathbf{E}\Big[\int_0^t e^{-\lambda(t-s)}\,\mathrm{d}|X_s|\Big] = \mathbf{E}\big[|X_t|\big] - |X_0|e^{-\lambda t} - \lambda \mathbf{E}\Big[\int_0^t |X_s| e^{-\lambda(t-s)}\,\mathrm{d}s\Big],$$

and since $X_t = \Delta_{ik} - \Delta_{jl}$, $\Delta_{ik} = \log \pi_i/\pi_k - \log \tilde{\pi}_i/\tilde{\pi}_k$ and $|\log(x)| \leq 1/x$, we can again bound the right-hand side by multiples of $\mathbf{E}\big[1/\min_i \pi_t^i\big]$ and $\mathbf{E}\big[1/\min_i \tilde{\pi}_t^i\big]$, which remain finite as we take a supremum over $t > 0$. □

4.3 Error bounds for an approximate filter

The approach we presented in the previous section allows for a more general result. We can proceed exactly as before to compute the error of a general approximate filter, rather than simply the filter with modified Q and h (and σ, if we allow for $\sigma \neq 1$). The discussion in this section will yield the proofs of Theorem 4.1.1 and Theorem 4.1.2.

Consider a general approximate filtering model given by (4.1), i.e.

$$\mathrm{d}\tilde{\pi}_t = \tilde{f}_t\,\mathrm{d}t + \tilde{g}_t\,\mathrm{d}Y_t, \quad \tilde{\pi}_0 = \nu, \tag{4.16}$$

where $\tilde{f}_t, \tilde{g}_t$ are $\mathbb{R}^{n+1}$-valued predictable processes. We will also need the following assumption:

(A4) *With probability 1, $\tilde{\pi}_t \in \mathring{\mathcal{S}}^n$ for all $t < \infty$. Moreover, $\tilde{f}_t$ and $\tilde{g}_t$ are locally bounded and satisfy the integrability condition*

$$\mathbf{E}\left[\int_0^t \max_i \frac{|\tilde{f}_s^i|}{\pi_s^i}\,\mathrm{d}s + \left(\int_0^t \max_i \left(\frac{\tilde{g}_s^i}{\pi_s^i}\right)^2 \mathrm{d}s\right)^{1/2}\right] < \infty$$

for all $t < \infty$.

Note that the Wonham filter SDE (3.4), or the Wonham filter with misspecified model parameters given by (4.7), immediately satisfy Assumption (A4) by Lemma 3.3.1 and (a simple extension of) [26, Lemma 3.6].

We start by proving an intermediate result.

Proposition 4.3.1 (Dynamics of the Hilbert error of an approximate filter)**.** *Let π_t be the solution to* (3.4) *and $\tilde{\pi}_t$ the solution to* (4.16)*. Suppose $\mu^i, \nu^i > 0 \quad \forall i$ and $q_{ij} > 0$ for all $i \neq j$. Under Assumption (A4), for all $s \leq t < \infty$, we have*

$$\begin{aligned}\mathcal{H}(\pi_t, \tilde{\pi}_t) \leq \mathcal{H}(\pi_s, \tilde{\pi}_s) &- 2\int_s^t \kappa_r \sinh\left(\frac{\mathcal{H}(\pi_r, \tilde{\pi}_r)}{2}\right)\mathrm{d}r \qquad (4.17)\\ &+ \int_s^t \max_{i,k}\left\{\mathcal{E}_r^{1,i} - \mathcal{E}_r^{1,k} + \frac{1}{2}\left(\mathcal{E}_r^{2,k} - \mathcal{E}_r^{2,i}\right)\right\}\mathrm{d}r\\ &+ \max_j |h^j| \int_s^t \max_{i,k}\left\{\mathcal{E}_r^{3,i} - \mathcal{E}_r^{3,k}\right\}\mathrm{d}r + \int_s^t \frac{1}{|\mathcal{I}_r|}\sum_{(i,k)\in\mathcal{I}_r}\left(\mathcal{E}_r^{3,i} - \mathcal{E}_r^{3,k}\right)\mathrm{d}B_r\\ &+ \frac{1}{4}\sum_{(i,k)}\sum_{(j,l)\neq(i,k)}\int_s^t \mathrm{d}L_r^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot)), \qquad (4.18)\end{aligned}$$

where $B_t = Y_t - \int_0^t \pi_s^\top h\,\mathrm{d}s$ is the innovation process. For $j \in \mathbf{N}$ the error terms are given by

$$\mathcal{E}_t^{1,j} = \left(\sum_{m=0}^n q_{mj}\frac{\tilde{\pi}_t^m}{\tilde{\pi}_t^j}\right) - \frac{\tilde{f}_t^j}{\tilde{\pi}_t^j}, \quad \mathcal{E}_t^{2,j} = (h^j)^2 - \frac{(\tilde{g}_t^j)^2}{(\tilde{\pi}_t^j)^2}, \quad \mathcal{E}_t^{3,j} = h^j - \frac{\tilde{g}_t^j}{\tilde{\pi}_t^j},$$

and the processes $(\Delta_{ik}(t))_{t\geq 0}$ for $(i,k) \in \mathbf{N}\times\mathbf{N}$ are defined as $\Delta_{ik}(t) = \log\frac{\pi_t^i}{\pi_t^k} - \log\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}$. The set $\mathcal{I}_t = \{(i,k) : \Delta_{ik}(t) = \mathcal{H}(\pi_t, \tilde{\pi}_t)\}$ is the argmax of these processes for all $t < \infty$, and $L_t^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot))$ denotes the local time at 0 of the difference process $(\Delta_{ik} - \Delta_{jl})$ for all (i,k), $(j,l) \in \mathbf{N}\times\mathbf{N}$. The decay coefficient $\kappa_t > 0$ can be taken to

be any of

$$\kappa_t = \begin{cases} \lambda, \\ \tilde{\lambda}^\star\Big(t, \tanh\Big(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\Big)\Big), \\ \lambda^\star\Big(t, \tanh\Big(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\Big)\Big), \end{cases} \tag{4.19}$$

where λ is the deterministic rate from Theorem 3.2.1, and $\tilde{\lambda}^\star$ and $\lambda^\star$ are the functions defined in Proposition 3.3.11 and Corollary 3.3.11.1 (in (3.24) *and* (3.33) *respectively).*

Proof. Assumption (A4) allows us to move our analysis from the simplex to $\mathbb{R}^n$ by defining the usual transformations $\theta_k^i : (0,1)^{\times 2} \to \mathbb{R}$. The dynamics of $\tilde{\theta}_k^i = \log \frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}$ are given by

$$\mathrm{d}\log\frac{\tilde{\pi}^i}{\tilde{\pi}^k}(t) = \frac{1}{\tilde{\pi}_t^i}\big(\tilde{f}_t^i\,\mathrm{d}t + \tilde{g}_t^i\,\mathrm{d}Y_t\big) - \frac{1}{\tilde{\pi}_t^k}\big(\tilde{f}_t^k\,\mathrm{d}t + \tilde{g}_t^k\,\mathrm{d}Y_t\big) + \frac{1}{2}\left(\left(\frac{\tilde{g}_t^k}{\tilde{\pi}_t^k}\right)^2 - \left(\frac{\tilde{g}_t^i}{\tilde{\pi}_t^i}\right)^2\right)\mathrm{d}t,$$

so that, letting $\theta_k^i(t) = \log \frac{\pi_t^i}{\pi_t^k}$ and $\Delta_{ik}(t) = \theta_k^i(t) - \tilde{\theta}_k^i(t)$, defining the innovation process $B_t = Y_t - \int_0^t \pi_s^\top h\,\mathrm{d}s$, and recalling (3.11) for the dynamics of $\theta_k^i(t)$, we have

$$\begin{aligned} \mathrm{d}\Delta_{ik}(t) = &\left[\frac{1}{\pi_t^i}\Big(\sum_{j=0}^n q_{ji}\pi_t^j\Big) - \frac{\tilde{f}_t^i}{\tilde{\pi}_t^i}\right]\mathrm{d}t + \left[\frac{\tilde{f}_t^k}{\tilde{\pi}_t^k} - \frac{1}{\pi_t^k}\Big(\sum_{j=0}^n q_{jk}\pi_t^j\Big)\right]\mathrm{d}t \\ &+ \frac{1}{2}\left((h^k)^2 - (h^i)^2 + \left(\frac{\tilde{g}_t^i}{\tilde{\pi}_t^i}\right)^2 - \left(\frac{\tilde{g}_t^k}{\tilde{\pi}_t^k}\right)^2\right)\mathrm{d}t \\ &+ \Big(h^i - h^k - \frac{\tilde{g}_t^i}{\tilde{\pi}_t^i} + \frac{\tilde{g}_t^k}{\tilde{\pi}_t^k}\Big)\big(\,\mathrm{d}B_t + \pi_t^\top h\,\mathrm{d}t\big). \end{aligned} \tag{4.20}$$

This equation might seem a bit daunting at first, but it can be treated exactly as we did in the case of misspecified Q and h. Adding and subtracting terms as appropriate, we can rewrite (4.20) as

$$\begin{aligned} \mathrm{d}\Delta_{ik}(t) = &\left(\sum_{\substack{j=0\\ j\neq i}}^n q_{ji}\left(\frac{\pi_t^j}{\pi_t^i} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\right) - \sum_{\substack{j=0\\ j\neq k}}^n q_{jk}\left(\frac{\pi_t^j}{\pi_t^k} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\right)\mathrm{d}t + \big(\mathcal{E}_t^{1,i} - \mathcal{E}_t^{1,k}\big)\,\mathrm{d}t \\ &+ \frac{1}{2}\big(\mathcal{E}_t^{2,k} - \mathcal{E}_t^{2,i}\big)\,\mathrm{d}t + \big(\mathcal{E}_t^{3,i} - \mathcal{E}_t^{3,k}\big)\big(\,\mathrm{d}B_t + \pi_t^\top h\,\mathrm{d}t\big), \end{aligned} \tag{4.21}$$

where for all $j \in \mathbf{N}$ we have defined the error terms

$$\mathcal{E}_t^{1,j} = \left(\sum_{m=0}^n q_{mj}\frac{\tilde{\pi}_t^m}{\tilde{\pi}_t^j}\right) - \frac{\tilde{f}_t^j}{\tilde{\pi}_t^j}, \qquad \mathcal{E}_t^{2,j} = (h^j)^2 - \left(\frac{\tilde{g}_t^j}{\tilde{\pi}_t^j}\right)^2, \qquad \mathcal{E}_t^{3,j} = h^j - \frac{\tilde{g}_t^j}{\tilde{\pi}_t^j}. \tag{4.22}$$

Now we proceed as in the proof of Theorem 4.2.1 by letting $\mathbf{\Delta}_t = \{\Delta_{ik}(t)\}_{i\in\mathbf{N},k\in\mathbf{N}}$ be the family of processes with evolution equations given by (4.21), defining the process $LSE_\alpha(\Delta_\cdot)(t)$ and its dynamics, and finally taking $\alpha \to \infty$ to yield our error estimates.

Letting once more $T_{ik}(t) = \frac{\pi_t^i}{\pi_t^k} - \frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k}$ for all $(i,k) \in \mathbf{N}\times\mathbf{N}$, we have, for all $s \le t < \infty$,

$$
\begin{aligned}
LSE_\alpha(\mathbf{\Delta}_\cdot)(t) &= LSE_\alpha(\mathbf{\Delta}_\cdot)(s) + \int_s^t S_\alpha^{arg}\Big(\mathbf{\Delta}_\cdot, \sum_{j\neq i}^n q_{ji}T_{ji}(\cdot) - \sum_{j\neq k}^n q_{jk}T_{jk}(\cdot)\Big)(r)\,\mathrm{d}r \\
&+ \int_s^t S_\alpha^{arg}\big(\mathbf{\Delta}_\cdot, \mathcal{E}_\cdot^{1,i} - \mathcal{E}_\cdot^{1,k}\big)(r)\,\mathrm{d}r + \frac{1}{2}\int_s^t S_\alpha^{arg}\big(\mathbf{\Delta}_\cdot, \mathcal{E}_\cdot^{2,k} - \mathcal{E}_\cdot^{2,i}\big)(r)\,\mathrm{d}r \\
&+ \int_s^t S_\alpha^{arg}\big(\mathbf{\Delta}_\cdot, \mathcal{E}_\cdot^{3,i} - \mathcal{E}_\cdot^{3,k}\big)(r)(\,\mathrm{d}B_r + \pi_s^\top h\,\mathrm{d}r) \\
&+ \frac{1}{2}\int_s^t \sum_{(j,l)}\sum_{(u,v)\neq(j,l)} \frac{\alpha e^{\alpha(\Delta_{jl}(r)+\Delta_{uv}(r))}}{(\sum_{(i,k)} e^{\alpha\Delta_{ik}(r)})^2}\left(\big(\mathcal{E}_r^{3,j} - \mathcal{E}_r^{3,l}\big)^2 - \big(\mathcal{E}_r^{3,j} - \mathcal{E}_r^{3,l}\big)\big(\mathcal{E}_r^{3,u} - \mathcal{E}_r^{3,v}\big)\right)\mathrm{d}r.
\end{aligned}
$$

Note that by Lemma 3.3.1 and Assumption (A4), the first four integrands on the right-hand side have enough regularity to apply dominated convergence for Lebesgue or stochastic integrals when taking the limit as $\alpha \to \infty$. For the final term, we invoke once more Proposition A.1.6, which is justified by Assumption (A4), to bound the integral in terms of the local times of the difference processes $\Delta_{ik} - \Delta_{jl}$ as we let $\alpha \to \infty$. This yields, for all $s \le t < \infty$,

$$
\begin{aligned}
\Delta_\infty(t) \le \Delta_\infty(s) &+ \int_s^t \frac{1}{|\mathcal{I}_r|}\sum_{(i,k)\in\mathcal{I}_r}\left(\sum_{j\neq i}^n q_{ji}T_{ji}(r) - \sum_{j\neq k}^n q_{jk}T_{jk}(r)\right)\mathrm{d}r \\
&+ \int_s^t \max_{i,k}\left\{\mathcal{E}_r^{1,i} - \mathcal{E}_r^{1,k} + \frac{1}{2}\big(\mathcal{E}_r^{2,k} - \mathcal{E}_r^{2,i}\big)\right\}\mathrm{d}r + \max_j |h^j| \int_s^t \max_{i,k}\left\{\mathcal{E}_r^{3,i} - \mathcal{E}_r^{3,k}\right\}\mathrm{d}r \\
&+ \int_s^t \frac{1}{|\mathcal{I}_r|}\sum_{(i,k)\in\mathcal{I}_r}\big(\mathcal{E}_r^{3,i} - \mathcal{E}_r^{3,k}\big)\,\mathrm{d}B_r + \frac{1}{4}\sum_{(i,k)}\sum_{(j,l)\neq(i,k)}\int_s^t \mathrm{d}L_r^0(\Delta_{ik}(\cdot) - \Delta_{jl}(\cdot)),
\end{aligned}
\tag{4.23}
$$

where for all $r \in [s,t]$, we have defined $\mathcal{I}_r = \{(i,k) : \Delta_{ik}(r) = \Delta_\infty(r)\} \subset \mathbf{N}\times\mathbf{N}$ to be the argmax of $\mathbf{\Delta}_r$, and let $|\mathcal{I}_r|$ denote its size.

Finally, consider the drift terms in the first integral on the right-hand side. By Lemma 3.3.7, for all $r \in [s,t]$ and all $(i,k) \in \mathcal{I}_r$ we have that

$$
\Delta_{ik}(r) = \Delta_\infty(r) = \mathcal{H}(\pi_r, \tilde{\pi}_r) = \log\max_j \frac{\pi_r^j}{\tilde{\pi}_r^j} - \log\min_j \frac{\pi_r^j}{\tilde{\pi}_r^j} =: \log M_r - \log\frac{1}{m_r},
$$

where $M_r \ge 1$ and $1/m_r \le 1$ are respectively the pointwise maximum and minimum ratio between the components of π_r and $\tilde{\pi}_r$. By Lemma 3.3.8 we have that, for all

$r \in [s,t]$ and for all $(i,k) \in \mathcal{I}_r$, $T_{ji}(r) \leq 0$ and $T_{jk}(r) \geq 0$, for all $j \in \mathbf{N}$. Then the first integral on the right-hand side is negative and in particular

$$\int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \left(\sum_{j\neq i}^n q_{ji} T_{ji}(r) - \sum_{j\neq k}^n q_{jk} T_{jk}(r) \right) \mathrm{d}r \\ \leq - \int_s^t \min_{(i,k)\in\mathcal{I}_r} \left(\sum_{j\neq k}^n q_{jk} T_{jk}(r) - \sum_{j\neq i}^n q_{ji} T_{ji}(r) \right) \mathrm{d}r.$$

Now we can minimize the integrand with algebraic calculations as in the proof of Theorem 3.2.1, or Proposition 3.3.11, or Corollary 3.3.11.1, which yields

$$\int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{(i,k)\in\mathcal{I}_r} \left(\sum_{j\neq i}^n q_{ji} T_{ji}(r) - \sum_{j\neq k}^n q_{jk} T_{jk}(r) \right) \mathrm{d}r \leq -2 \int_s^t \kappa_r \sinh\left(\frac{\Delta_\infty(r)}{2} \right) \mathrm{d}r,$$

where the decay rate κ_r can be chosen to be any of the rates λ from Theorem 3.2.1, $\tilde{\lambda}^\star(t, \tanh(\Delta_\infty(t)/4))$ from (3.24) in Proposition 3.3.11, or $\lambda^\star(t, \tanh(\Delta_\infty(t)/4))$ from (3.33) in Corollary 3.3.11.1. □

Theorem 4.1.1 and Theorem 4.1.2 now follow easily from the above proposition.

Proof of Theorem 4.1.1. Start from (4.18). For all $t < \infty$, we bound κ_t from below by the deterministic rate $\lambda = 2 \min_{i\neq k} \sqrt{q_{ik} q_{ki}}$. Moreover, recall that $2\sinh(x/2) \geq x$ for $x \geq 0$. Substitute both these bound in the first integral in the right-hand side of (4.18). We take expectation and note that the stochastic integral vanishes, since it is a martingale (as the integrand is locally L^2-integrable by assumption (A4)). A modification of the standard Grönwall argument to deal with Lebesgue–Stieltjes measures (as in the proof of Theorem 4.2.1) concludes the proof. □

Proof of Theorem 4.1.2. If $\mathcal{E}_t^{3,i} = 0$ for all $i \in \mathbf{N}$ and $t < \infty$, then $\mathcal{E}_t^{2,i} = 0$ as well. Then (4.21) reduces to

$$\mathrm{d}\Delta_{ik}(t) = \left(\sum_{\substack{j=0 \\ j\neq i}}^n q_{ji} \left(\frac{\pi_t^j}{\pi_t^i} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i} \right) - \sum_{\substack{j=0 \\ j\neq k}}^n q_{jk} \left(\frac{\pi_t^j}{\pi_t^k} - \frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k} \right) \right) \mathrm{d}t + \left(\mathcal{E}_t^{1,i} - \mathcal{E}_t^{1,k} \right) \mathrm{d}t,$$

for all $(i,k) \in \mathbf{N} \times \mathbf{N}$, so we recover C^1 dynamics for the difference processes $\Delta_{ik}(t)$. A C^1 process does not generate local time, so (4.18) simplifies to

$$\Delta_\infty(t) \leq \Delta_\infty(s) - 2 \int_s^t \kappa_r \sinh\left(\frac{\Delta_\infty(r)}{2} \right) \mathrm{d}r + \int_s^t \max_{i,k} \left\{ \mathcal{E}_r^{1,i} - \mathcal{E}_r^{1,k} \right\} \mathrm{d}r,$$

for all $s \leq t$, where $\kappa_t > 0$ is the coefficient given by any of the rates in (4.19). Now the second part of the theorem follows easily, by first bounding κ_r from below by a positive (measurable) process γ_t given by one of

$$\gamma_t = \begin{cases} \lambda, \\ \tilde{\lambda}^\star_t \leq \tilde{\lambda}^\star\Big(t, \tanh\Big(\frac{\mathcal{H}(\pi_t, \tilde{\pi}_t)}{4}\Big)\Big), \\ \lambda^\star_t \leq \lambda^\star\Big(t, \tanh\Big(\frac{\mathcal{H}(\pi_t, \tilde{\pi}_t)}{4}\Big)\Big), \end{cases}$$

where $\tilde{\lambda}^\star_t$ and $\lambda^\star_t$ are as in Proposition 3.3.11 and Corollary 3.3.11.1. Then, recalling once more that $2\sinh(x/2) \geq x$ for $x \geq 0$, the usual Grönwall argument yields

$$\Delta_\infty(t) \leq \Delta_\infty(0) e^{-\int_0^t \gamma_s \,\mathrm{d}s} + \int_0^t e^{-\int_s^t \gamma_r \,\mathrm{d}r} \max_{i,k} \left\{\mathcal{E}^{1,i}_s - \mathcal{E}^{1,k}_s\right\} \mathrm{d}s,$$

which is (4.6) for $\gamma_t = \tilde{\lambda}^\star_t$.

We now look for a tighter bound. Consider the process $X_t = \tanh\big(\Delta_\infty(t)/4\big)$. Applying the chain rule we have

$$\begin{aligned} \mathrm{d}X_t &= \frac{1}{4}\cosh^{-2}\left(\frac{\Delta_\infty(t)}{4}\right) \mathrm{d}\Delta_\infty(t) \\ &\leq -\kappa_t \frac{\sinh\left(\frac{\Delta_\infty(t)}{2}\right)}{2\cosh^2\left(\frac{\Delta_\infty(t)}{4}\right)} \mathrm{d}t + \frac{1}{4}\frac{\max_{i,k}\left\{\mathcal{E}^{1,i}_r - \mathcal{E}^{1,k}_r\right\}}{\cosh^2\left(\frac{\Delta_\infty(t)}{4}\right)} \\ &\leq -\kappa_t X_t \,\mathrm{d}t + \frac{1}{2}\max_{i,k}\left\{\mathcal{E}^{1,i}_t - \mathcal{E}^{1,k}_t\right\} \frac{X_t}{\sinh\big(2\operatorname{arctanh}(X_t)\big)} \,\mathrm{d}t, \end{aligned}$$

where we have used the identity $\sinh(2x) = 2\sinh(x)\cosh(x)$. Moreover, since $\sinh(2\operatorname{arctanh}(x)) = \frac{2x}{1-x^2}$, we can rewrite the above as

$$\mathrm{d}X_t \leq \alpha(t, X_t)\,\mathrm{d}t, \quad \text{where} \quad \alpha(t, X_t) = -\tilde{\lambda}^\star(t, X_t) X_t + \frac{1}{4}\max_{i,k}\left\{\mathcal{E}^{1,i}_t - \mathcal{E}^{1,k}_t\right\}\big(1 - X_t^2\big),$$

where we have substituted $\tilde{\lambda}^\star(t, X_t)$ for κ_t for clarity in the exposition below (but the arguments are analogous whether κ_t is the deterministic rate λ from Theorem 3.2.1, the pathwise rate $\tilde{\lambda}^\star_t$ from Proposition 3.3.11, or the coefficient $\lambda^\star(t, X_t)$ or the pathwise rate $\lambda^\star_t$ from Corollary 3.3.11.1). Bounding $\tilde{\lambda}^\star(t, X_t)$ from below by $\tilde{\lambda}^\star_t$, and $(1 - X_t^2)$ from above by 1, another application of Grönwall yields (4.5).

We recall (3.24) for the definition of $\tilde{\lambda}^\star$. Note that the mapping $x \mapsto \alpha(t, x)$ is locally Lipschitz continuous (with Lipschitz constant dependent on ω, t and x), since $x \mapsto \tilde{\lambda}^\star(t, x)x$ is locally Lipschitz continuous and $\max_{i,k}\left\{\mathcal{E}^{1,i}_t - \mathcal{E}^{1,k}_t\right\}$ is locally

bounded by Lemma 3.3.1 and Assumption (A4). Now let u_t be the solution to the ODE with random coefficients given by

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} = \alpha(t, u_t), \qquad u_0 = X_0 = \tanh\left(\frac{\Delta_\infty(0)}{4}\right), \tag{4.24}$$

where α, or, specifically, $\tilde{\lambda}^\star$ and $\max_{i,k}\left\{\mathcal{E}_t^{1,i} - \mathcal{E}_t^{1,k}\right\}$ depend on the process $\tilde{\pi}_t$, which is fixed for each ω. Recall that, since $\mu, \nu \in \mathring{\mathcal{S}}^n$ by assumption, $\mathcal{H}(\mu, \nu) < \infty$, and therefore $u_0 \in (0, 1)$. Since the right-hand side is locally Lipschitz, (4.24) has a unique solution u_t up to its first explosion time $T > 0$ (again, see e.g. [81, Theorem 2.5]). Now, if $T < \infty$, then T is the first time such that $u_{T-} = 1$. By continuity of u_t, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for $s \in (T - \delta, T)$, we have $u_s \in (1 - \varepsilon, 1)$. Then

$$\begin{aligned} 1 &= u_{T-\delta} + \int_{T-\delta}^{T} \alpha(s, u_s)\,\mathrm{d}s \\ &\leq u_{T-\delta} - \frac{\delta(2-\varepsilon)(1-\varepsilon)}{\varepsilon} \inf_{s\in[T-\delta,T]} \min_{i\neq k}\left\{ q_{ik}\frac{\tilde{\pi}_s^i}{\tilde{\pi}_s^k}\right\} + \frac{\delta\varepsilon(2-\varepsilon)}{4} \sup_{s\in[T-\delta,T]} \max_{i,k}\left\{\mathcal{E}_s^{1,i} - \mathcal{E}_s^{1,k}\right\}, \end{aligned}$$

using (strict) positivity of $\lambda^\star$ and $\max_{i,k}\left\{\mathcal{E}_t^{1,i} - \mathcal{E}_t^{1,k}\right\}$, and that $\frac{1+u_t}{1-u_t} \geq \frac{(2-\varepsilon)(1-\varepsilon)}{\varepsilon}$ and $(1 - u_s^2) \leq \varepsilon(2 - \varepsilon)$ for $s \in (T - \delta, T)$. Since, for small enough ε, the negative term dominates the positive term, this implies $1 < u_{T-\delta}$, which is strictly less than 1, and therefore a contradiction. Then $T = \infty$ and (4.24) has a unique solution for all $t \geq 0$.

A similar argument proves that $u_t > 0$ for all $t \geq 0$, and finally Lemma 3.3.10 yields the theorem. □

Proof of Corollary 4.1.2.1. Analogous to the proof of Theorem 4.1.2. □

4.4 A numerical example

We conclude this chapter by testing our bounds in a couple of simulations.

We consider a Wonham filter with approximate model parameters, whose dynamics are given by (4.7). We assume h to be known, so $\tilde{h} = h$. We approximate Q by applying a non-negative factorization algorithm: we subtract the diagonal from Q, approximate the resulting positive matrix using the `NMF` class from the python package `sklearn.decomposition`, and reconstruct the diagonal to ensure all the rows sum to 0 to yield $\tilde{Q}$. In this setting, there is no error due to the misspecification of h, so we do not have to worry with estimating the local time terms. We consider the error bounds given in Theorem 4.1.2.

In the figure below we compare this approximate filter with the Wonham filter for a 3-state and a 6-state Markov chain. We take the Q matrices to be given by

$$Q = \begin{pmatrix} -3 & 1 & 2 \\ 1 & -3 & 2 \\ 1.5 & 1.5 & -3 \end{pmatrix}, \qquad Q = \begin{pmatrix} -9 & 3 & 1 & 1.5 & 2.5 & 1 \\ 1 & -7.5 & 1 & 2 & 2.3 & 1.2 \\ 3 & 2 & -8 & 1 & 1 & 1 \\ 2 & 1.3 & 1 & -6 & 0.7 & 1 \\ 1.1 & 1 & 0.9 & 3 & -9 & 3 \\ 1 & 1 & 3 & 2 & 2.5 & -9.5 \end{pmatrix}, \tag{4.25}$$

respectively, and the sensor functions h to be

$$h = (-1, 0, 1)\,, \qquad h = (-3, -2, -1, 1, 2, 3)\,. \tag{4.26}$$

For the 3-state Markov chain, we take the initial law of the signal X to be given by its ergodic distribution, i.e. $\text{law}(X_0) = \big(0.3, 0.3, 0.4\big)$. This is also the initial condition for the Wonham filter π_t. The approximate rate matrix $\tilde{Q}$ for the approximate filter $\tilde{\pi}_t$ is obtained using a 2-channel NMF approximation of Q. We take the initial condition for $\tilde{\pi}_t$ to be $\tilde{\pi}_0 = \big(0.2, 0.2, 0.6\big)$. In the 6-state case, we start the signal X quite close to the boundary of $\mathcal{S}^5$, with its initial law given by $\mu = \big(0.5, 0.04, 0.09, 0.2, 0.04, 0.13\big)$, which is also the initial condition for π_t. For the rate matrix $\tilde{Q}$ for $\tilde{\pi}_t$ we use a 4-channel NMF approximation of Q. We start $\tilde{\pi}_t$ also relatively close to the boundary of $\mathcal{S}^5$, but near a different edge from μ, and take $\tilde{\pi}_0 = \big(0.25, 0.1, 0.06, 0.07, 0.22, 0.3\big)$.

For transparency, we write here the matrices $\tilde{Q}$ (rounded to the second significant digit) resulting from the NMF approximation in each case:

$$\tilde{Q} \approx \begin{pmatrix} -2.5 & 0.5 & 2 \\ 0.5 & -2.5 & 2 \\ 1.5 & 1.5 & -3 \end{pmatrix}, \quad \tilde{Q} \approx \begin{pmatrix} -9 & 3.04 & 1.04 & 1.54 & 2.43 & 0.95 \\ 0.94 & -7.25 & 1.70 & 2.02 & 1.58 & 1.01 \\ 2.92 & 2.05 & -7.8 & 0.69 & 0.88 & 1.26 \\ 2.11 & 1.24 & 0.52 & -5.34 & 0.84 & 0.62 \\ 1.13 & 0.86 & 0.63 & 3.02 & -8.68 & 3.04 \\ 1.02 & 0.77 & 2.64 & 1.92 & 2.92 & -9.28 \end{pmatrix}.$$

In Figure 4.1, on the left, both for the 3-state and the 6-state nonlinear filter, we plot 100 realizations of the Hilbert error between π_t and $\tilde{\pi}_t$ (and their sample mean) in blue, and of the error bounds from Theorem 4.1.2 (and their sample means). Since these bounds are path-by-path, each realization of the error between π_t and $\tilde{\pi}_t$ has three corresponding error bounds: in fuchsia we plot $4\,\text{arctanh}(u_t)$, where u_t is the (numerical) solution to the ODE (4.3); in green we plot the bound (4.6) where the decay rate is given by $\tilde{\lambda}_t$ as defined in (4.4); in red we plot again the bound (4.6), but using the deterministic decay rate λ from Theorem 3.2.1 instead. The error terms

(4.2) are evaluated pathwise at each time-step. In the pictures on the right, for the same simulations, we plot 100 realizations of $\tanh(\mathcal{H}(\pi_t, \tilde{\pi}_t)/4)$ (blue), and of the numerical solution u_t to (4.3) (fuchsia).

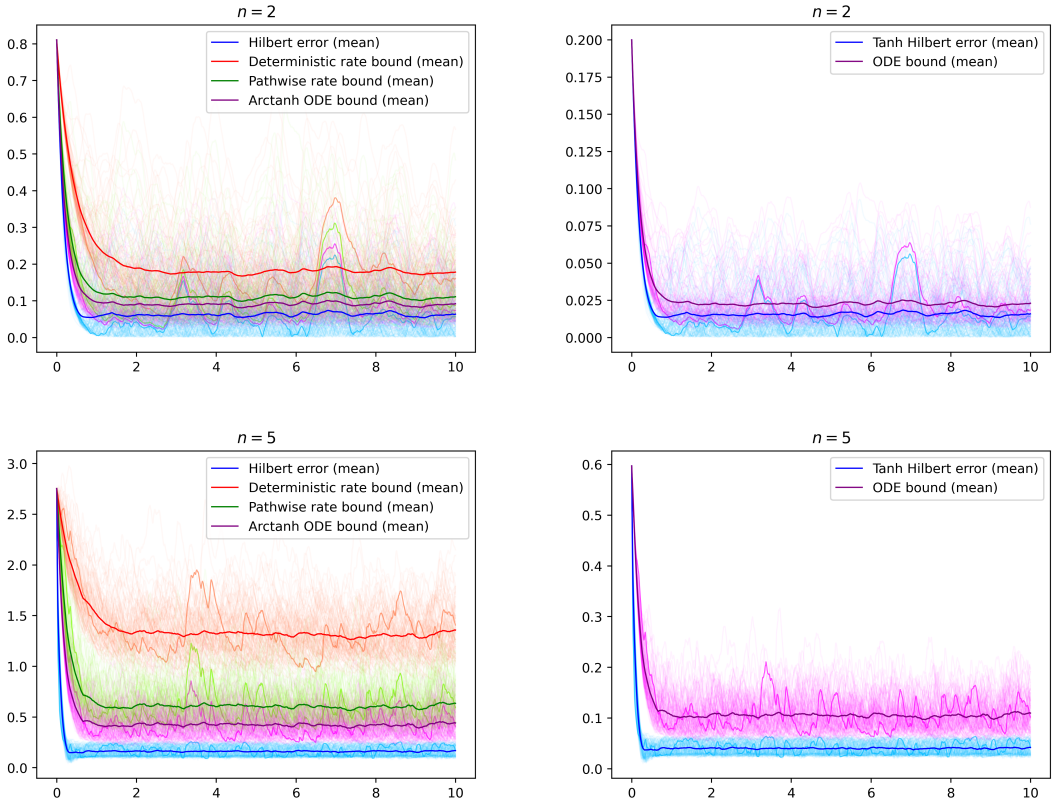

Figure 4.1: For dimensions $n = 2, 5$, we test our error bounds from Theorem 4.1.2 against the actual Hilbert error between the Wonham filter and an approximate filter. On the right we plot 100 realizations of the Hilbert projective error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$ (faded, light blue), of the ODE bound given by $4 \operatorname{arctanh}(u_t)$, where u_t solves (4.3) (faded, fuchsia), of the pathwise bound (4.6) with pathwise decay rate $\tilde{\lambda}_t$ (faded, light green), and of the pathwise bound with deterministic decay rate λ (faded, orange), for $t \in [0, 10]$. We highlight one sample path of the Hilbert error at random, together with its three corresponding pathwise bounds. In blue, purple, green and red we plot the sample means of the errors and of the three bounds. On the right, for the same simulations, we plot 100 realizations of the quantity $\tanh(\mathcal{H}(\pi_t, \tilde{\pi}_t)/4)$ together with the ODE bound u_t, and the sample means of both. Again, we highlight at random one realization of the tanh error and its corresponding ODE bound.

In the 3-state case, where the filter lives in $\mathcal{S}^2$, we can see that our estimates for the error are very close to its actual value (the ODE bounds given by the solution to (4.3) in particular). In the 6-state case, with a 5-dimensional filter, our error bounds are less sharp. In fact, it is safe to assume that our error bounds get progressively

worse as we increase the dimension of the state-space.

Why is this the case? As we already mentioned in Section 3.3.4, the main issue with our error bounds is the contraction rate. Our numerical experiments for the stability estimates (see Figure 3.2) show that the error contracts at a much faster rate than what we can prove. This makes sense, since by minimizing over all entries of Q, the decay rates in Theorem 3.2.1 and Proposition 3.3.11 give pathwise bounds for 'worst case'-type of scenarios. A similar argument applies to our treatment of the approximation-error terms (4.2): to ensure our bounds hold, we need to maximize over all possible indices, and this implies that we are more and more likely to overestimate the errors as the dimension increases.

There are a couple of directions that one could pursue at this point, to tighten our error estimates. The first would be to try to exploit some averaging over the indices, instead of simply minimizing/maximizing over them, to yield tighter decay rates/error terms. This could potentially be achieved if one looked for a bound in expectation instead of pathwise. This problem seems difficult, however, as it involves estimating the expectation of the argmin and argmax of the log differences between the ratios of π_t and $\tilde{\pi}_t$. On the other hand, if, instead of proceeding analytically, one were able to estimate quantities numerically, it should be relatively easy to obtain good numerical estimates for the decay rate of the stability error in high dimensions, as we can see from the plots in Figure 3.2. The estimated rate can then be substituted into (4.6) to yield tighter bounds for the error of approximate filters in high dimensions (which should hold with high probability). The issue of overestimating the error terms remains, but, since they are dominated by the negative exponentials, tightening the decay rate would yield a significant overall improvement.

Numerical estimation of the decay rate opens up other possibilities as well. The arguments we developed in this chapter work when the signal is given by any ergodic time-continuous Markov chain – the strict positivity of the off-diagonal entries of Q is only required to guarantee that the decay rates we derive are nonzero. In other words, the stability error of the Wonham filter decays as long as the signal is ergodic (as discussed by [12]) even when the Q-matrix is sparse. By discretization, the nonlinear filter on a compact state space, given by the solution to the Kushner–Stratonovich SPDE, is often approximated by a Wonham filter on a high number of states. However, the diffusion operator corresponds to a very sparse transition matrix. Given a numerical estimate for the decay rate of the Hilbert error of the discretized diffusion operator, one could then use our error estimates to understand the error of approximate filters in infinite dimensions.

Chapter 5

The projection filter in finite dimensions

In this last chapter we finally consider the question of how to define an 'optimal' projection filter for the continuous-time, finite-dimensional filtering problem introduced in Chapter 3. We use the error bounds we developed in Chapter 4 both to gain an intuition of how to achieve this, and to evaluate the error of our low dimensional approximation compared to the Wonham filter.

5.1 Introduction

The projection filter was introduced by Brigo, Hanzon and Le Gland in [19]. The filtering setting in [19] is different from the one we are considering here: the signal X is taken to be a diffusion process on $\mathbb{R}$, and the optimal filter is then the (infinite dimensional) solution to the Kushner–Stratonovich SPDE (see e.g. [10, Thm. 3.30]). However, the methodology presented in [19] can easily be adapted to fit our context as well. We give a short overview of the main ideas in [19] and subsequent related works [4,6], both to provide some background to those unfamiliar with the projection filter, and to motivate the direction of our own research.

In [19], the authors start with the assumption that the nonlinear filter, which, in the setting under their consideration, is a probability measure on $\mathbb{R}$, has a density with respect to the Lebesgue measure. They endow the space of probability measures with the Hellinger distance, which induces an L^2 structure on the square-root of the corresponding probability densities. Then, they consider a subspace $M \subset \sqrt{P}$, where $\sqrt{P} = \{\sqrt{p(x)} \,:\, p(x) \geq 0, \|p\|_{L^1} = 1\}$ is the space of root-densitites, such that $p(x) \in M$ has a finite dimensional parametrization $\xi \in \mathbb{R}^n$. In other words, they choose M to be a parametric family of (square-root) densities $M = \{\sqrt{p(x,\xi)} \in \sqrt{P} \,:\, \xi \in \mathbb{R}^n\}$.

In the language of Amari's theory of information geometry [2, 3], M is a statistical manifold, with Riemannian metric given by the Fisher information. By choosing M to be an exponential family, a parametrization is immediately provided by the natural parameters, so $\xi = \theta$. At this point, proceeding very much in the spirit of Amari, the authors of [19] identify the tangent space $T_{p(x,\theta)}M$ with

$$\operatorname{span}\Big(\big\{\partial_1\sqrt{p(x,\theta)},\cdots,\partial_n\sqrt{p(x,\theta)}\big\}\Big),$$

where ∂_i denotes the partial derivative with respect to θ^i for $i = 1,\ldots,n$, and $x \in \mathbb{R}, \theta \in \mathbb{R}^n$. Note that the tangent vectors in T_pM are in fact random variables (see [2, Sec. 2.2]). Then, the L^2 inner product on the square-rooted probabilities becomes a stochastic operation on T_pM, specifically

$$\left\langle \frac{\partial\sqrt{p(\cdot,\theta)}}{\partial\theta^i}, \frac{\partial\,\sqrt{p(\cdot,\theta)}}{\partial\theta^j} \right\rangle_{L^2} = \frac{1}{4}\mathbf{E}_\mu\left[\partial_i \log p(\cdot,\theta)\partial_j \log p(\cdot,\theta)\right],$$

where the expectation is taken with respect to μ, the probability measure associated to $p(x,\theta) \in M$. On the right-hand side we recognize (within a factor of $1/4$) the variance of the score function, which is the Fisher information metric. Adopting the Fisher information as a Riemannian metric for M, in [19] Brigo, Hanzon and Le Gland define the orthogonal projection operator $T_{p(x,\theta)}\sqrt{P} \to T_{p(x,\theta)}M$, which, when applied to the flow operators of the Kushner–Stratonovich SPDE, constrains the flow to the tangent bundle of M. Therefore the solution of this modified equation, which they name the *projection filter*, remains on M, thus simplifying the SPDE into an n-dimensional SDE for the parameter θ, and, crucially, reduces the filtering problem from infinite to finite dimension.

At this point, one might wonder how good an approximation the projection filter actually provides. Or, turning the problem around, whether it is possible to define an 'optimal' projection filter with small (or even minimal) error when compared to the optimal filter. In particular, there are a couple of questions that arise naturally.

1. How does one select the statistical family M?

2. What is the impact of the choice of ambient space and geometric structure (the square-root densities with the L^2 inner product in the case of [19]) on the projection filter?

3. Is it possible to bound the error of the projection filter (as an approximation for the true filter) without solving the original filtering equations?

To attempt to answer either of these questions with some degree of completion, one first needs a good understanding of the error between the optimal filter and the projection filter. In the setting considered in [19], this amounts to obtaining 'sharp enough' estimates to quantify the error between the solution to the Kushner–Stratonovich SPDE and the projection filter. As we have already stressed many times before, satisfactory quantitative estimates of this type, necessarily related to the study of the stability and robustness of the Kushner–Stratonovich SPDE, are not available in the literature, and are hard to find. This is why here we focus on the simpler case of finite state-space nonlinear filtering, where we hope that the estimates we obtained in Chapter 4 can help us shed some light on the answers to these questions.

It should be noted that Brigo and collaborators are certainly aware of the challenges that arise when attempting to construct an optimal projection filter. The selection of the statistical family M is a topic already explored in [19, Sec. 6], where, by considering the residual of the projection operator, Brigo, Hanzon and Le Gland show how to choose an exponential family which kills part of the projection error. The lack of understanding of the cumulative error between the projection filter and the optimal filter makes this manifold selection process somewhat heuristic, although numerical experiments presented in [18] show remarkably good results for the case of the cubic filter (which is known to be infinite-dimensional, see [44]). We will see in Section 5.5 that the manifold selection approach we suggest for the finite state-space nonlinear filtering setting shares similarities with [19].

The issue of the choice of geometric structure on the ambient space has also been partly explored. In [19], the choice to work in the space $\sqrt{P}$ of square-root probability densities endowed with the L^2 inner product seems quite arbitrary. In [4], Armstrong and Brigo consider instead $P = \{p(x) : p(x) \geq 0, \|p\|_{L^1} = 1\} \cup L^2$ as their ambient space, endowed with the L^2 inner product. This results in a different metric on $T_{p(x,\theta)}M$ (which we call the L^2-metric for lack of a better name), and therefore a different orthogonal projection $T_{p(x,\theta)}P \to T_{p(x,\theta)}M$. They also take M to be a mixture family of probability distributions (instead of an exponential family). These choices lead to the definition of a new projection filter, which numerically is found to perform similarly to (and, in one case, slightly better than) the exponential projection filter. While these combinations (i.e. Fisher information metric with projection onto exponential families, and L^2-metric onto mixture families) might be computationally convenient, neither of them is, a priori, justified. One could just as easily mix up the choices of metric and family of distribution (e.g. projecting using the Fisher information onto a mixture family), and define yet another projection filter. In Amari's

work, for example, the metric of choice is the Fisher information for both exponential and mixture families. And of course, one could work directly with the space of probability measures instead of the space of densities, and choose a geometric structure appropriate to this more general context. In our simpler finite dimensional setting we still have a choice of metric on $\mathcal{S}^n$,

Finally, somewhat related to these geometric musings, one more question that we might ask in our pursuit of the optimal projection filter is:

4. How do we define our projection operator so that the infinitesimal error is minimized?

This is the topic of [6, 7] by Armstrong, Brigo and Rossi Ferrucci, which builds on [5] and the classical theory of stochastic calculus on manifolds (see e.g. [39, 47]) to define and compare three different ways to project SDEs from $\mathbb{R}^n$ to a submanifold $M \cong \mathbb{R}^d$, with $d \ll n$. Throughout the chapter we will consider these three types of projection and keep track of the error estimates for each of them. However, the choice of projection (in the sense of *how to project an SDE*, not in the sense of what *geometric projection* to employ!) will turn out to be somewhat irrelevant in our work, for two reasons. The first one is that the optimality criteria developed in [6] are not applicable to our problem, and the way we measure the error between the optimal filter and the projection filter. The second reason is that, once we are able to select the "right" submanifold for the projection filter, the three projections defined in [6] are all equivalent!

5.2 Hidden Markov models and SDEs on the probability simplex

For the purpose of keeping the exposition self-contained, we start by recalling our filtering setting. Compared to Chapter 3 and Chapter 4, we now consider time-dependent coefficients, since this will allow for interesting observations when it comes to the projection filter. Note that the stability and robustness results of Chapters 3 and 4 apply to this time-dependent setting as well.

Assume X is a continuous-time Markov chain taking values in the state-space given by the standard basis $\mathbb{S} = \{e_0, \dots, e_n\}$ of $\mathbb{R}^{n+1}$. For $t \geq 0$, denote by $Q_t = (q_{ij}(t)) \in \mathbb{R}^{(n+1)\times(n+1)}$ the transition rate matrix of X, so that $X_t - X_0 - \int_0^t Q_s^\top X_s \, \mathrm{d}s$ is a right-continuous martingale. Let the initial distribution of X be given by $\mu = \mathbf{E}[X_0]$, where μ is an element of the n-dimensional probability simplex $\mathcal{S}^n$. Suppose W is a

standard d-dimensional Brownian motion independent of X at time t, and let Y be the $\mathbb{R}^d$-valued process satisfying the SDE

$$\mathrm{d}Y_t^i = h_i(t)^\top X_s \,\mathrm{d}s + \sigma \,\mathrm{d}W_t^i, \quad Y_0^i = 0, \quad \text{for all } i = 1, \ldots, d,$$

where $h_i(t) \in \mathbb{R}^{n+1}$ for $i = 1, \ldots, d$ and $t \geq 0$, and $\sigma \neq 0$. Assume that $h_i(t)$ is bounded for all $i = 1, \ldots, d$ and $t \geq 0$. Let $\{\mathcal{Y}_t\}_{t \geq 0}$ be the (completed) natural filtration generated by the observation process Y. We consider the problem of estimating the state of X given $\mathcal{Y}_t$. Denote the posterior distribution of X at time t by $\pi_t = \mathbf{E}\left[X_t | \mathcal{Y}_t\right]$. The $\mathcal{S}^n$-valued process $\pi = (\pi_t)_{t \geq 0}$ satisfies the Wonham SDE:

$$\mathrm{d}\pi_t = Q_t^\top \pi_t \,\mathrm{d}t + \frac{1}{\sigma^2} \sum_{k=1}^{d} \left(H_k(t) - \pi_t^\top h_k(t) \,\mathbb{I}_{n+1}\right) \pi_t \left(\mathrm{d}Y_t^k - \pi_t^\top h_k(t) \,\mathrm{d}t\right), \quad \pi_0 = \mu, \tag{5.1}$$

where, for $k = 1, \ldots, d$ and $t \geq 0$, $H_k(t)$ is the $(n+1) \times (n+1)$-dimensional diagonal matrix $\mathrm{diag}\left(h_k(t)\right)$ and $\mathbb{I}_{n+1}$ is the identity matrix. Note that (5.1) is initialized at $\mu = \mathbf{E}\left[X_0\right]$. We call π, the solution to the SDE (5.1), the *Wonham filter*: π_t is the optimal estimate for X_t given all the information collected by observing the process Y over the time interval $[0, t]$. In the next section it will be useful to have expressed (5.1) in Stratonovich form, instead of Itô. The Stratonovich dynamics of (5.1) are

$$\begin{aligned} \mathrm{d}\pi_t = Q_t^\top \pi_t \,\mathrm{d}t + \frac{1}{2\sigma^2} \sum_{k=1}^{d} \left(h_k(t)^\top H_k(t) \pi_t \,\mathbb{I}_{n+1} - H_k(t)^2\right) \pi_t \,\mathrm{d}t \\ + \frac{1}{\sigma^2} \sum_{k=1}^{d} \left(H_k(t) - \pi_t^\top h_k(t) \,\mathbb{I}_{n+1}\right) \pi_t \circ \mathrm{d}Y_t^k, \qquad \pi_0 = \mu. \end{aligned} \tag{5.2}$$

(A5) *For the sake of clarity in the exposition below, we assume $\sigma = 1$, although our results remain valid for $\sigma \neq 1$ or even invertible and time-dependent $\sigma_t \in \mathbb{R}^{d \times d}$, as long as it is bounded away from 0.*

Now that our set-up is clear, let us state our goal: for $m \ll n$, we want to find an m-dimensional SDE of the form

$$\mathrm{d}\xi_t = A_t(\xi_t) \,\mathrm{d}t + B_t(\xi_t) \,\mathrm{d}Y_t \tag{5.3}$$

together with a map $\mathbb{R}^m \ni \xi_t \mapsto \tilde{\pi}_t \in \mathcal{S}^n$ such that $\tilde{\pi}_t$ is a good approximation of the Wonham filter π_t, the solution to (5.1). We will reduce the dimensionality of (5.1) using geometric projections, in the spirit of the *projection filter*. Thus, the first thing to do is understand how to project the flow of the Wonham SDE (5.1) onto a chosen m-dimensional submanifold M.

5.2.1 Projecting the Wonham SDE

The Wonham filter π lives in the probability simplex $\mathcal{S}^n$, so the SDE (5.1) describes a flow on $\mathcal{S}^n$. To define a projection filter for the problem under consideration we need a subset $M \subset \mathcal{S}^n$, such that $M \cong \mathbb{R}^m$ with $m \ll n$, and a way to project the Wonham SDE (5.1) onto M. In this section we consider three different ways to define the projection of (5.1) onto a given submanifold $M \subset \mathcal{S}^n$, following the work in [6]. Note that the exposition in [6] is tailored to SDEs living in $\mathbb{R}^n$ projected onto a m-dimensional submanifold of $\mathbb{R}^n$, and consequently it is mainly concerned with the standard Euclidean metric as the choice of Riemannian metric tensor on $\mathbb{R}^n$. Since we can embed $\mathcal{S}^n \hookrightarrow \mathbb{R}^{n+1}$ and $M \hookrightarrow \mathbb{R}^{n+1}$, we see that [6] is closely related to our setting. We keep our exposition here as simple as possible, without losing sight of our objective, which is ultimately to use this theory to define a projection filter. For further details, we refer to [6] and to Rossi Ferrucci's PhD thesis [75, Chapter 1][1], a rewriting of the differential geometric content of [6] which provides an excellent introduction to the topic of SDEs on manifolds embedded in $\mathbb{R}^n$.

Before we start to describe the three different ways to define a 'projected Wonham SDE', there is a technicality, which was not an issue in [6,75], that we have deal with: the boundary of $\mathcal{S}^n$. The probability simplex $\mathcal{S}^n$ is an n-dimensional submanifold of $\mathbb{R}^{n+1}$ with boundary. This could make the analysis a little bit more complicated, because one needs to make sure the projection operator is well-defined at boundary points. We will start by making an assumption so that we can get rid of problems at the boundary.

(A6) *The Wonham SDE* (5.1) *is initialized in the interior of* $\mathcal{S}^n$*, i.e.* $\pi_0 = \mu = \mathbf{E}[X_0] \in \mathring{\mathcal{S}}^n$.

By [26, Lemma 2.1], we then have that $\pi_t \in \mathring{\mathcal{S}}^n$ almost surely for all $t < \infty$, so we can work with the interior of the simplex $\mathring{\mathcal{S}}^n$ and project (5.1) on a submanifold $M \cong \mathbb{R}^m$, with $M \subset \mathring{\mathcal{S}}^n$. Note that we do not have to worry about local coordinates: by embedding $\mathcal{S}^n \hookrightarrow \mathbb{R}^{n+1}$, we can take $p = (p^0, \ldots, p^n) \in \mathbb{R}^{n+1}$ as the global coordinate system for $\mathcal{S}^n$. Moreover, the SDE (5.1) is driven by the observation process Y, which is an $\mathbb{R}^d$-valued semi-martingale. Thus we can also dispense with having to use local coordinates for the ambient space of Y (which would be needed if Y were manifold-valued instead).

[1] A version of this chapter has recently been published in [7].

Notation. For any manifold M and $p \in M$, we denote by T_xM the tangent space of M at p. By TM we denote the tangent bundle of M and by $\Gamma(TM)$ the set of tangent vector fields along M. If M is embedded in another manifold, we let $N_pM = (T_pM)^\perp$ denote the normal space of M at p, NM the normal bundle, and $\Gamma(TN)$ the normal vector fields along M. By $g = g_{ij}$ we denote the (Riemannian) metric tensor, and by $\langle \cdot, \cdot \rangle_{g(p)} : T_pM \times T_pM \to \mathbb{R}$ the inner product associated to g at $p \in M$.

Let M be a smooth m-dimensional submanifold of $\mathring{\mathcal{S}}^n$. In particular, we take M to be a parametrized statistical family embedded in $\mathring{\mathcal{S}}^n$, i.e.

$$M = \left\{ p \in \mathring{\mathcal{S}}^n \,:\, p = F(\xi) = \big(p^0(\xi), \dots, p^n(\xi)\big) \text{ for } \xi \in \mathbb{R}^m \right\}, \tag{5.4}$$

where M is smooth as long as $p^0(\xi), \dots, p^n(\xi)$ are smooth. Note that M is defined globally, and ξ is a global coordinate system for M. We will see in Section 5.5 that we can take M to vary with time, but for the sake of clarity we consider M fixed, for now. The map $F : M \to \mathring{\mathcal{S}}^n$ is an immersion. It is an embedding if F is a homemorphism onto its image in the subspace topology.

Now we equip $\mathring{\mathcal{S}}^n$ with a Riemannian metric g (for example, we might take g to be the Fisher Information metric, or the Euclidean metric induced by the embedding $\mathring{\mathcal{S}}^n \hookrightarrow \mathbb{R}^{n+1}$, or anything else that we might think suitable). The submanifold M inherits the Riemannian structure of $\mathring{\mathcal{S}}^n$ induced by F, and its metric tensor is F^*g (the pullback of g through F, or the restriction of g to TM). Canonically, the tangent space of M at a point $p \in M$ is given by $T_pM \cong \operatorname{span}(\{\partial_i|_p\})$ where ∂_i denotes the derivative of F with respect to ξ^i for $i = 1, \dots, m$.

At each point $p \in M$, the tangent space of $\mathring{\mathcal{S}}^n$ at p decomposes into the direct sum $T_p\mathring{\mathcal{S}}^n = T_pM \oplus N_pM$. In particular, we can define the smooth bundle homomorphisms

$$\Pi^\top : T\mathring{\mathcal{S}}^n|_M \to TM, \quad \Pi^\perp : T\mathring{\mathcal{S}}^n|_M \to NM,$$

called the tangential and normal projections, which for each $p \in M$ restrict to orthogonal projections from $T_p\mathring{\mathcal{S}}^n$ to T_pM and N_pM respectively. Since the tangent vectors $\partial_i|_p \in T_pM$ are not necessarily orthogonal, we can either apply the Gram–Schmidt algorithm to obtain an orthonormal basis, or equivalently we can define the orthogonal projection, for $p \in M$, as

$$\begin{aligned} \mathfrak{p}_p \,:\, & T_p\mathring{\mathcal{S}}^n \to T_pM, \\ & w \mapsto \sum_{i=1}^m \sum_{j=1}^m g^{-1}_{ij}(p) \langle w, \partial_j|_p \rangle_{g(p)} \partial_i|_p, \end{aligned} \tag{5.5}$$

where $g^{-1} = (g^{-1}_{ij})$ is the inverse of the matrix $g = (g_{ij}) = \big(\langle \partial_i|_p, \partial_j|_p \rangle_{g(p)}\big)$.

Remark 5.1. If we take the metric g on $\mathring{\mathcal{S}}^n$ to be the canonical Euclidean metric (induced from $\mathbb{R}^{n+1}$), for $p \in M$ the matrix $g(p)$ defined above is given by $g(p) = JF(p)^\top JF(p)$, where $JF \in \mathbb{R}^{(n+1)\times m}$ is the Jacobian matrix of the the immersion $F : M \to \mathring{\mathcal{S}}^n$ (i.e. the matrix with columns given by the basis vectors of T_pM). If g is not Euclidean, we can still express $g(p)$ in matrix form as $JF(p)^\top G(p) JF(p)$ for some positive semi-definite matrix $G(p) \in \mathbb{R}^{(n+1)\times(n+1)}$ which defines the inner product on $T_p\mathring{\mathcal{S}}^n$, so that in particular the orthogonal projection (5.5) can be written in matrix form as

$$w \mapsto JF(p)(JF(p)^\top G(p) JF(p))^{-1} JF(p)^\top G(p) w. \tag{5.6}$$

Compare this expression with [75, Eq. 1.56] (or [7, Eq. 45]), where the Riemannian metric is taken to be Euclidean (and instead of an immersion, the authors define the submanifold M through a submersion).

Given the projection $\mathfrak{p}_p$, let us now investigate how to project the Wonham SDE (5.1) on M. There are three classical ways to define an SDE on a manifold: *Stratonovich*, *Schwartz–Meyer* and *Itô* . Each of these comes with a natural way to project the SDE on a submanifold: respectively the *Stratonovich projection*, the *Itô-jet projection* and the *Itô-vector projection* (where we stick with the nomenclature chosen in [6]). We consider each of these in order, starting from the Stratonovich case, which is the most straightforward.

Consider the Wonham SDE in Stratonovich form (5.2). For simplicity we rewrite it as

$$\mathrm{d}\pi_t = b_0(\pi_t, t)\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d} c_k(\pi_t, t)\,\mathrm{d}t + \sum_{k=1}^{d} \sigma_k(\pi_t, t) \circ \mathrm{d}Y_t^k, \quad \pi_0 = \mu, \tag{5.7}$$

where $b_0(p,t) = Q_t^\top p$, $c_k(p,t) = (h_k(t)^\top H_k(t) p\, \mathbb{I}_{n+1} - H_k(t)^2)p$ and $\sigma_k(p,t) = (H_k(t) - p^\top h_k(t)\, \mathbb{I}_{n+1})p$ for $k = 1, \dots, d$ and $p \in \mathring{\mathcal{S}}^n$. Note that the coefficients b_0, c_k and σ_k are independent of Y. They are smooth linear maps from $\mathbb{R}_{\geq 0}$ to $T\mathring{\mathcal{S}}^n$: letting $f = b_0, c_k$ or σ_k (for $k = 1, \dots, d$), f corresponds to a smooth map $(\mathring{\mathcal{S}}^n, \mathbb{R}_{\geq 0}) \ni (p,t) \to f(p,t) \in \mathrm{Hom}(\mathbb{R}_{\geq 0}, T_p\mathring{\mathcal{S}}^n)$. In other words, the coefficients of the Stratonovich SDE (5.7) are elements of $\Gamma(\mathrm{Hom}(\mathbb{R}_{\geq 0}, T\mathring{\mathcal{S}}^n)) \cong C^\infty(\mathbb{R}_{\geq 0}, \Gamma(T\mathring{\mathcal{S}}^n))$.

The *Stratonovich projection* of (5.2) is given by applying the orthogonal projection (5.5) to the coefficients of (5.7). For $p \in M$ and $X \in T_p\mathring{\mathcal{S}}^n$, we denote by $\overline{X} = \mathfrak{p}_p(X)$ the orthogonal projection of X onto T_pM. The resulting projected Wonham SDE on M, in Stratonovich form, is given by

$$\mathrm{d}\tilde{\pi}_t = \overline{b}_0(\tilde{\pi}_t, t)\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d} \overline{c}_k(\tilde{\pi}_t, t)\,\mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k(\tilde{\pi}_t, t) \circ \mathrm{d}Y_t^k, \quad \tilde{\pi}_0 \in M. \tag{5.8}$$

Note that unless $\pi_0 = \mu \in M$, we need to choose a suitable initial condition $\tilde{\pi}_0$: for example, $\tilde{\pi}_0$ can be taken to be the point on M with the shortest geodesic path to μ, i.e. $\tilde{\pi}_0 = \operatorname{argmin}\{d_g(\mu, p) : p \in M\}$ (although this might be hard to compute). Recall that Stratonovich SDEs started from an embedded submanifold M in $\mathring{\mathcal{S}}^n$ will stay on M as long as their flow is mapped to T_pM for all $p \in M$ (see e.g. [47, Prop. 1.2.8]), which is exactly what we are ensuring by applying the orthogonal projection (5.5) to the coefficients of (5.7).

For comparison with the other projections, we transform (5.8) from Stratonovich to Itô form. For each component $\overline{\sigma}_k^i(\tilde{\pi}_t, t)$ of $\sigma_k(\tilde{\pi}_t, t) \in T_{\tilde{\pi}_t}M$, we compute

$$\mathrm{d}\langle \overline{\sigma}_k^i(\tilde{\pi}_\cdot, \cdot), Y_\cdot^k \rangle_t = \sum_{j=0}^{n} \frac{\partial \overline{\sigma}_k^i(\tilde{\pi}_t, t)}{\partial p^j} \overline{\sigma}_k^j(\tilde{\pi}_t, t)\, \mathrm{d}\langle Y_\cdot^k \rangle_t,$$

which yields the Itô form of the Stratonovich projection

$$\begin{aligned} \mathrm{d}\tilde{\pi}_t &= \overline{b}_0\, \mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\left[\overline{c}_k + \sum_{j=0}^{n} \frac{\partial \overline{\sigma}_k}{\partial p^j}\overline{\sigma}_k^j\right] \mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k\, \mathrm{d}Y_t^k, \\ &= \overline{b}_0\, \mathrm{d}t - \sum_{k=1}^{d} \overline{b}_k\, \mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\left[\sum_{j=0}^{n} \frac{\partial \overline{\sigma}_k}{\partial p^j}\overline{\sigma}_k^j - \mathfrak{p}_{\tilde{\pi}_t}\Big(\frac{\partial \sigma_k}{\partial p^j}\Big)\sigma_k^j\right] \mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k\, \mathrm{d}Y_t^k, \\ \tilde{\pi}_0 &\in M, \end{aligned} \tag{5.9}$$

where it is implied that all the terms are evaluated at $(\tilde{\pi}_t, t)$, and for $k = 1, \dots, d$ we have defined $b_k(p,t) = p^\top h_k(t)(H_k(t) - p^\top h_k(t)\, \mathbb{I}_{n+1})p$, so we have that $c_k(p,t) = -b_k(p,t) - \frac{1}{2}\sum_{j=0}^{n} \frac{\partial \sigma_k(p,t)}{\partial p^j}\sigma_k^j(p,t)$. For reference, we write here the Wonham SDE (5.1) using this simplified notation:

$$\mathrm{d}\pi_t = b_0(\pi_t, t)\, \mathrm{d}t - \sum_{k=1}^{d} b_k(\pi_t, t)\, \mathrm{d}t + \sum_{k=1}^{d} \sigma_k(\pi_t, t)\, \mathrm{d}Y_t^k, \quad \pi_0 = \mu. \tag{5.10}$$

Now, before we proceed with the *Itô-vector* and the *Itô-jet* projections, we need to be a bit more precise when it comes to the immersion of M in $\mathring{\mathcal{S}}^n$. In particular, to define the next two projections, we need a better understanding of the relationship between the geometry of the ambient manifold $\mathring{\mathcal{S}}^n$ and the submanifold M. The first step in this direction would be to compare the extrinsic Levi–Civita connection of $\mathring{\mathcal{S}}^n$ with the intrinsic connection of M: a measure of this difference is given by the second fundamental form of M (see e.g. [64, Chapter 8]). However, to carry out such analysis precisely would require the introduction of many elements of differential geometry, which is beyond the scope of this thesis. To keep the exposition fluid, we give here a more intuitive approach, following [75, Chapter 1].

Let U be a tubular neighbourhood of M in $\mathring{\mathcal{S}}^n$. For $p \in U$, consider the smooth Riemannian submersion $\mathfrak{p} : U \to M$ given by

$$\mathfrak{p}(p) = \operatorname{argmin}\big\{d_g(p,q) \,:\, q \in M\big\}, \tag{5.11}$$

where d_g is the Riemannian (geodesic) distance of $p \in U$ from $q \in M$. For $q \in M$, each fiber $\mathring{\mathcal{S}}^n_q = \mathfrak{p}^{-1}(q)$ is an embedded smooth submanifold of $\mathring{\mathcal{S}}^n$. Then at each point $p \in \mathring{\mathcal{S}}^n$ we can decompose the tangent space $T_p\mathring{\mathcal{S}}^n$ into the direct sum of the vertical tangent space at p, given by $V_p = T_p\big(\mathring{\mathcal{S}}^n_{\mathfrak{p}(p)}\big)$, and its orthogonal (horizontal) component $(V_p)^\perp$. In particular, the Riemannian submersion $\mathfrak{p}$ gives a linear isometry from $(V_p)^\perp$ to $T_{\mathfrak{p}(p)}M$, and consequently also from V_p to $N_{\mathfrak{p}(p)}M$, so it automatically induces the decomposition of any vector field along $\mathring{\mathcal{S}}^n$ into a tangential and a normal component along M (for more details, see [64, Chapter 2]). In other words, locally on the tangent bundle of M the Riemannian submersion $\mathfrak{p}$ agrees with the orthogonal projection (5.5). As readers familiar with differential geometry might imagine, differentiating the map $\mathfrak{p}$ provides an easy alternative (at least formally) to working directly with connections.

Given the tangential othogonal projection (5.5), we can define the normal orthogonal projection as

$$\mathfrak{p}_p^\perp \,:\, T_p\mathring{\mathcal{S}}^n \to N_pM, \quad \mathfrak{p}_p^\perp(w) = (\mathbb{I}\mathrm{d} - \mathfrak{p}_p)(w). \tag{5.12}$$

For $p \in M$ and $X \in T_p\mathring{\mathcal{S}}^n$, denote by $\breve{X} = \mathfrak{p}_p^\perp(X)$ the orthogonal projection of X onto N_pM.

Now let us consider the Itô SDE (5.10) on $\mathring{\mathcal{S}}^n$. In general, the theory of Itô calculus on manifolds is more complicated than that of Stratonovich calculus, due to the second order chain rule for the Itô integral (which reduces to the standard Itô formula in $\mathbb{R}^n$). Thus, it makes sense that to define an Itô SDE on a general manifold M it is necessary to look at differential operators of order higher than 1. Generally an Itô SDE on a manifold M, driven by a manifold-valued semimartingale, will depend explicitely on the connections on M and on the semimartingale-manifold. In the case of the *Schwartz–Meyer* formulation of an SDE on M, the matter of higher order terms is taken care of by considering the second order tangent bundle of M, $\mathbb{T}M$, which consists of second order differential operators (without a constant term). Then a Schwartz–Meyer equation on M driven by an $\mathbb{R}^d$-valued semimartingale Y has diffusion coefficients which are elements of $C^\infty(\mathbb{R}_{\geq 0}, \Gamma(T\mathring{\mathcal{S}}^n))$, but drift terms which are instead in $C^\infty(\mathbb{R}_{\geq 0}, \Gamma(\mathbb{T}\mathring{\mathcal{S}}^n))$. For the (better known) *Itô* formulation, the Christoffel symbols which describe the metric connection should appear in any description of the

SDE in local coordinates, although the diffusion and drift coefficients remain elements of $C^\infty(\mathbb{R}_{\geq 0}, \Gamma(T\mathring{\mathcal{S}}^n))$.

This being said, our Wonham SDE (5.10) is quite simple. The Christoffel symbols have vanished—this is because $\mathring{\mathcal{S}}^n$ is a flat manifold in $\mathbb{R}^{n+1}$. Projecting the SDE onto M is simply a matter of finding the right correction term for the drift so that the projected SDE does not leave M. To do so, one goes back to Stratonovich form and sets the condition that the Stratonovich drift and diffusion coefficients belong to TM. The procedure for this is explained in [75]. Here, we limit ourselves to state the final results. The *Itô-jet projection* of (5.10) is given by

$$\mathrm{d}\tilde{\pi}_t = \overline{b}_0 \,\mathrm{d}t - \sum_{k=1}^{d} \overline{b}_k \,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j} \sigma_k^i \sigma_k^j \,\mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k \,\mathrm{d}Y_t^k, \quad \tilde{\pi}_0 \in M. \tag{5.13}$$

The *Itô-vector projection* of (5.10) is given by

$$\mathrm{d}\tilde{\pi}_t = \overline{b}_0 \,\mathrm{d}t - \sum_{k=1}^{d} \overline{b}_k \,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j} \overline{\sigma}_k^i \overline{\sigma}_k^j \,\mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k \,\mathrm{d}Y_t^k, \quad \tilde{\pi}_0 \in M. \tag{5.14}$$

Both in (5.13) and (5.14) evaluation of all the terms at $(\tilde{\pi}_t, t)$ is implied. Note that the 'drift correction' $\frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}(\tilde{\pi}_t)}{\partial p^i \partial p^j} \overline{\sigma}_k^i \overline{\sigma}_k^j$ appears in all three types of projection (5.9), (5.13) and (5.14). To see this, use [75, Eq.1.57, Eq.1.63] to decompose the drift of (5.9) into elements of TM and of NM. This results in the following formulation for the *Stratonovich projection* (in Itô form)

$$\begin{aligned} \mathrm{d}\tilde{\pi}_t = \overline{b}_0 \,\mathrm{d}t - \sum_{k=1}^{d} \overline{b}_k \,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\Bigg[\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j} \overline{\sigma}_k^i \breve{\sigma}_k^j - \sum_{j=0}^{n} \mathfrak{p}_{\tilde{\pi}_t}\Big(\frac{\partial \sigma_k}{\partial p^j}\Big)\breve{\sigma}_k^j\Bigg]\,\mathrm{d}t \\ + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j} \overline{\sigma}_k^i \overline{\sigma}_k^j \,\mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k \,\mathrm{d}Y_t^k, \qquad \tilde{\pi}_0 \in M, \end{aligned} \tag{5.15}$$

while for the *Itô-jet projection*, decomposing $\sigma_k = \overline{\sigma}_k + \breve{\sigma}_k$ for all $k = 1, \ldots, d$, we have

$$\begin{aligned} &\mathrm{d}\tilde{\pi}_t = \overline{b}_0 \,\mathrm{d}t - \sum_{k=1}^{d} \overline{b}_k \,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n}\left[2\frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j} \overline{\sigma}_k^i \breve{\sigma}_k^j + \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j} \overline{\sigma}_k^i \overline{\sigma}_k^j\right]\mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k \,\mathrm{d}Y_t^k, \\ &\tilde{\pi}_0 \in M. \end{aligned} \tag{5.16}$$

5.3 Error bounds for the projection filter

The work we did in the previous section has given us three possible candidates for an SDE of the form (5.3). The Stratonovich, Itô-jet and Itô-vector projections given by

(5.15) (5.16) and (5.14) respectively are nothing but projection filters, living in the parametrized statistical family M, fixed by the expression (5.4). Assume $F(\xi)$, the parametrization of M, is bijective and invertible over its domain (so that essentially $F : M \to \mathring{\mathcal{S}}^n$ is an embedding). We can express the three projections (5.15) (5.16) and (5.14) as m-dimensional SDEs for the parameter $\xi \in \mathbb{R}^m$ of M in the required form (5.3). For example, consider the Stratonovich projection (5.8) (in Stratonovich form to make the calculations easier). Since the solution $\tilde{\pi}_t \in M$ for all $t \geq 0$, we can write $\tilde{\pi}_t = F(\xi_t)$ for $\xi_t \in \mathbb{R}^m$ the parameter of the distribution evolving in time. Then, applying the chain rule and writing down explicitly the orthogonal projection $\mathfrak{p}_{\tilde{\pi}_t}$ defined in (5.5), (5.8) can be expressed as

$$\begin{aligned}
\mathrm{d}F(\xi_t) &= \sum_{i=1}^{m} \partial_i|_{F(\xi_t)} \circ \mathrm{d}\xi_t^i \\
&= \overline{b}_0(F(\xi_t), t)\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d} \overline{c}_k(F(\xi_t), t)\,\mathrm{d}t + \sum_{k=1}^{d} \overline{\sigma}_k(F(\xi_t), t) \circ \mathrm{d}Y_t^k, \\
&= \sum_{i=1}^{m} \left[\sum_{j=1}^{m} g_{ij}^{-1}(F(\xi_t)) \langle b_0, \partial_j|_{F(\xi_t)} \rangle_{g(F(\xi_t))} \right] \partial_i|_{F(\xi_t)}\,\mathrm{d}t \\
&\quad + \sum_{i=1}^{m} \left[\frac{1}{2} \sum_{k=1}^{d} \sum_{j=1}^{m} g_{ij}^{-1}(F(\xi_t)) \langle c_k, \partial_j|_{F(\xi_t)} \rangle_{g(F(\xi_t))} \right] \partial_i|_{F(\xi_t)}\,\mathrm{d}t \\
&\quad + \sum_{i=1}^{m} \left[\sum_{k=1}^{d} \sum_{j=1}^{m} g_{ij}^{-1}(F(\xi_t)) \langle \sigma_k, \partial_j|_{F(\xi_t)} \rangle_{g(F(\xi_t))} \right] \partial_i|_{F(\xi_t)} \circ \mathrm{d}Y_t^k, \\
\xi_0 &= F^{-1}(\tilde{\pi}_0), \quad \text{with } \tilde{\pi}_0 \in M,
\end{aligned}$$

where $\{\partial_i|_{F(\xi_t)}\}_i$ are the tangent vectors $\partial_{\xi^i} F(\xi)$ evaluated at points $F(\xi_t)$ along the submanifold M, and the coefficients b_0, c_k and σ_k are implicitely intended to be evaluated at $(F(\xi_t), t)$. Then, simplifying, the above reduces to the following m-dimensional SDE for $\xi_t \in \mathbb{R}^m$

$$\begin{aligned}
\mathrm{d}\xi_t &= g^{-1}(F(\xi_t))\mathfrak{P}_{\mathrm{vec}}(b_0, F(\xi_t))\,\mathrm{d}t \\
&\quad + g^{-1}(F(\xi_t)) \sum_{k=1}^{d} \left[\frac{1}{2}\mathfrak{P}_{\mathrm{vec}}(c_k, F(\xi_t))\,\mathrm{d}t + \mathfrak{P}_{\mathrm{vec}}(\sigma_k, F(\xi_t)) \circ \mathrm{d}Y_t^k \right], \quad \xi_0 = F^{-1}(\tilde{\pi}_0),
\end{aligned}$$

where, for $p \in M$ and $w \in T_p\mathring{\mathcal{S}}^n$, we denote by $\mathfrak{P}_{\mathrm{vec}}(w, p)$ the m-dimensional vector with each entry the component of w with respect to each basis vector of T_pM, i.e.

$$\mathfrak{P}_{\mathrm{vec}}(w, p) = \begin{bmatrix} \langle w, \partial_1|_{F(\xi_t)} \rangle_{g(F(\xi_t))} \\ \vdots \\ \langle w, \partial_m|_{F(\xi_t)} \rangle_{g(F(\xi_t))} \end{bmatrix}.$$

Remark 5.2. If g is the standard Euclidean metric, for $p \in M$ we can write $g^{-1}(p) = (JF(p)^\top JF(p))^{-1}$ and for $w \in T_p\mathring{\mathcal{S}}^n$ we have $\mathfrak{P}_{\mathrm{vec}}(w,p) = JF(p)^\top w$.

The calculations to obtain SDEs for the parameters $\xi_t \in \mathbb{R}^m$ in the cases of the Itô-jet and Itô-vector projection are analogous (although tedious, since we have to take care of the second order terms that appear in Itô's formula), and we omit them.

In Section 5.2.1 we essentially gave a recipe to construct three types of projection filters given a submanifold $M \in \mathring{\mathcal{S}}^n$. To proceed any further with our objective of defining a projection filter which gives a 'good approximation' of the optimal filter, we need some understanding of the error between the Wonham filter and these projection filters. We apply Theorem 4.1.1, which provides a first answer in this direction. We recall that the error analysis in Chapter 4 was carried out through the use of the Hilbert projective metric $\mathcal{H}$, given by (3.3).

Notation. Let $\mathbf{N}$ denote the set $\{0, \dots, n\}$.

For clarity, we quickly recall here some notation from the previous section. Consider a parametrized family of discrete probability distributions $M \subset \mathring{\mathcal{S}}^n$ given by

$$M = \left\{ p \in \mathring{\mathcal{S}}^n \,:\, p = F(\xi) = \big(p^0(\xi), \dots, p^n(\xi)\big) \ \text{ for } \xi \in \mathbb{R}^m \right\},$$

where $F : M \to \mathring{\mathcal{S}}^n$ is an embedding. Equip $\mathring{\mathcal{S}}^n$ with a Riemannian metric g and define the orthogonal projection $\mathfrak{p}_p : T_p\mathring{\mathcal{S}}^n \to T_pM$ for $p \in M$ and the Riemannian submersion $\mathfrak{p} : \mathring{\mathcal{S}}^n \to M$ as in (5.5) and (5.11). For $p \in \mathcal{S}^n$ and $t \geq 0$, denote the coefficients of the Wonham SDE (5.1) by

$$\begin{aligned} b_0(p,t) &= Q_t^\top p, \\ b_k(p,t) &= p^\top h_k(t)(H_k(t) - p^\top h_k(t)\, \mathbb{I}_{n+1})p, \quad \text{for } k = 1, \dots, d, \\ \sigma_k(p,t) &= (H_k(t) - p^\top h_k(t)\, \mathbb{I}_{n+1})p, \quad \text{for } k = 1, \dots, d. \end{aligned} \tag{5.17}$$

For $p \in M$, let the tangential orthogonal projections of the above coefficients be denoted by $\overline{b}_0(p,t) = \mathfrak{p}_p(b_0(p,t))$, and similarly for $\overline{b}_k(p,t)$ and $\overline{\sigma}_k(p,t)$; denote the normal projection of σ_k by $\breve{\sigma}_k(p,t) = \mathfrak{p}_p^\perp(\sigma_k(p,t)) = (\mathbb{I}\mathrm{d} - \mathfrak{p}_p)(\sigma_k(p,t))$.

We have the following result.

Theorem 5.3.1 (Expected Hilbert error bounds for the projection filter)**.** *Let π_t be the solution to* (5.1) *and let ${}^{\text{Strat}}\tilde{\pi}_t$, ${}^{\text{Itô-j}}\tilde{\pi}_t$ and ${}^{\text{Itô-v}}\tilde{\pi}_t$ be the solutions respectively to the Stratonovich projection* (5.15)*, the Itô-jet projection* (5.16) *and the Itô-vector projection* (5.14) *of* (5.1) *on M. Let $\pi_0 = \mu \in \mathring{\mathcal{S}}^n$ and ${}^{\mathrm{p}}\tilde{\pi}_0 = \nu \in M$ for* $\mathrm{p} \in$

$\{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, *and assume* $q_{ij}(t) > 0$ *for all* $i \neq j$ *and all* $t \geq 0$. *For all* $t < \infty$, *and* $\mathrm{p} \in \{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, *we have that*

$$\begin{aligned}\mathbf{E}\left[\mathcal{H}(\pi_t, {}^{\mathrm{p}}\tilde{\pi}_t)\right] \leq \mathcal{H}(\mu,\nu)e^{-\int_0^t \lambda_s \,\mathrm{d}s} \\ + \int_0^t e^{-\int_s^t \lambda_r \,\mathrm{d}r}\mathbf{E}\Big[\max_{i,j}\Big\{{}^{\mathrm{p}}\mathcal{E}_s^{1,i} - {}^{\mathrm{p}}\mathcal{E}_s^{1,j} - \frac{1}{2}\sum_{k=1}^{d}\big({}^{\mathrm{p}}\mathcal{E}_s^{2,k,i} - {}^{\mathrm{p}}\mathcal{E}_s^{2,k,j}\big)\Big\}\Big]\,\mathrm{d}s \\ + \sum_{k=1}^{d}\int_0^t \max_l |h_k^l(s)| e^{-\int_s^t \lambda_r \,\mathrm{d}r}\mathbf{E}\Big[\max_{i,j}\big\{{}^{\mathrm{p}}\mathcal{E}_s^{3,k,i} - {}^{\mathrm{p}}\mathcal{E}_s^{3,k,j}\big\}\Big]\,\mathrm{d}s \\ + \frac{1}{4}\sum_{(i,j)}\sum_{(u,v)\neq(i,j)}\mathbf{E}\left[\int_0^t e^{-\int_s^t \lambda_r \,\mathrm{d}r}\,\mathrm{d}L_s^0({}^{\mathrm{p}}\Delta_{ij}(\cdot) - {}^{\mathrm{p}}\Delta_{uv}(\cdot))\right],\end{aligned} \tag{5.18}$$

where

- $\lambda_t = 2\min_{i\neq j}\sqrt{q_{ij}(t)q_{ji}(t)}$ *is the time-dependent deterministic contraction rate;*
- *for* $(i,j) \in \mathbf{N}\times\mathbf{N}$, *the processes* $({}^{\mathrm{p}}\Delta_{ij}(t))_{t\geq 0}$ *are defined as* ${}^{\mathrm{p}}\Delta_{ij}(t) = \log\frac{\pi_t^i}{\pi_t^j} - \log\frac{{}^{\mathrm{p}}\tilde{\pi}_t^i}{{}^{\mathrm{p}}\tilde{\pi}_t^j}$;
- $L_t^0({}^{\mathrm{p}}\Delta_{ij}(\cdot) - {}^{\mathrm{p}}\Delta_{uv}(\cdot))$ *denotes the local time at* 0 *of the difference processes* $({}^{\mathrm{p}}\Delta_{ij} - {}^{\mathrm{p}}\Delta_{uv})$, *for all* (i,j), $(u,v) \in \mathbf{N}\times\mathbf{N}$;
- *the error terms* ${}^{\mathrm{p}}\mathcal{E}_t^{1,j}$, ${}^{\mathrm{p}}\mathcal{E}_t^{2,k,j}$, ${}^{\mathrm{p}}\mathcal{E}_t^{3,k,j}$ *are given by*

$$\begin{aligned}{}^{\mathrm{p}}\mathcal{E}_t^{1,j} &= \left(\sum_{r=0}^{n} q_{rj}(t)\frac{{}^{\mathrm{p}}\tilde{\pi}_t^r}{{}^{\mathrm{p}}\tilde{\pi}_t^j}\right) - \frac{{}^{\mathrm{p}}B^j\big({}^{\mathrm{p}}\tilde{\pi}_t, t\big)}{{}^{\mathrm{p}}\tilde{\pi}_t^j}, \\ {}^{\mathrm{p}}\mathcal{E}_t^{2,k,j} &= h_k^j(t)^2 - \frac{\overline{\sigma}_k^j\big({}^{\mathrm{p}}\tilde{\pi}_t, t\big)^2}{\big({}^{\mathrm{p}}\tilde{\pi}_t^j\big)^2}, \quad \textit{for } k = 1,\ldots,d, \\ {}^{\mathrm{p}}\mathcal{E}_t^{3,k,j} &= h_k^j(t) - \frac{\overline{\sigma}_k^j\big({}^{\mathrm{p}}\tilde{\pi}_t, t\big)}{{}^{\mathrm{p}}\tilde{\pi}_t^j}, \quad \textit{for } k = 1,\ldots,d,\end{aligned} \tag{5.19}$$

where the terms ${}^{\mathrm{p}}B$ *are given by the drifts of* (5.15), (5.16) *and* (5.14), *i.e.*

$$
\begin{aligned}
{}^{\text{Strat}}B(p,t) &= \overline{b}_0 - \sum_{k=1}^{d} \overline{b}_k \\
&\quad + \frac{1}{2}\sum_{k=1}^{d}\left[\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j}\overline{\sigma}_k^i \breve{\sigma}_k^j - \sum_{j=0}^{n} \mathfrak{p}_{\tilde{\pi}_t}\Big(\frac{\partial \sigma_k}{\partial p^j}\Big)\breve{\sigma}_k^j + \sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j}\overline{\sigma}_k^i \overline{\sigma}_k^j\right],
\end{aligned} \tag{5.20}
$$

$$
{}^{\text{Itô-j}}B(p,t) = \overline{b}_0 - \sum_{k=1}^{d} \overline{b}_k + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n}\left[2\frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j}\overline{\sigma}_k^i \breve{\sigma}_k^j + \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j}\overline{\sigma}_k^i \overline{\sigma}_k^j\right], \tag{5.21}
$$

$$
{}^{\text{Itô-v}}B(p,t) = \overline{b}_0 - \sum_{k=1}^{d} \overline{b}_k + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n} \frac{\partial^2 \mathfrak{p}}{\partial p^i \partial p^j}\overline{\sigma}_k^i \overline{\sigma}_k^j, \tag{5.22}
$$

and evaluation of all terms at $(p,t) \in M \times \mathbb{R}_{\geq 0}$ *is implied.*

Proof. This is a direct application of Theorem 4.1.1. For $\mathrm{p} \in \{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, note that ${}^{\mathrm{p}}\tilde{\pi}_t \in M \subset \mathring{\mathcal{S}}^n$ by construction, and that the drifts ${}^{\text{Strat}}B(p,t)$ and diffusion coefficients $\overline{\sigma}_k$ for $k = 1, \dots d$ inherit local-boundedness and integrability from the coefficients of (5.1). Then (A4) is satisfied and we can indeed apply Theorem 4.1.1. The generalizations of the results in Theorem 4.1.1 to time-dependent transition matrix Q_t and sensor function $h_t = (h_k(t))$ and multi-dimensional Y_t require some work but are straightforward. □

The error bounds of Theorem 5.3.1 are mainly useful from a qualitative point of view: we see that if the transition matrix Q_t has positive off-diagonal entries for all $t \geq 0$, then the expected Hilbert error of the projection filter does not accumulate, and in fact it stays finite (bounded) as $t \to \infty$ as long as the infinitesimal error terms (5.19) are finite (bounded) for all $t \geq 0$ (to prove this, in light of the local time terms, one would need arguments similar to those in Section 4.2—and Proposition 4.2.2 in particular). Clearly, for a fixed manifold M, the error terms in (5.18) are minimized if the orthogonal projection (5.5) and the submersion (5.11) minimize the Hilbert distance, locally on TM and in a tubular neighbourhood U of M. If it were possible, the obvious choice would be to take $\mathcal{H}$ as a Riemannian metric on $\mathring{\mathcal{S}}^n$. However, the geometry induced by $\mathcal{H}$ on $\mathring{\mathcal{S}}^n$ is not Riemannian (as we saw in Chapter 2), so this is clearly not a viable option. One possible avenue to explore is approximate the $\mathcal{H}$ metric by a Riemannian metric on $\mathring{\mathcal{S}}^n$: we will suggest a way to do this in Section 5.6.

Meanwhile, looking at the error terms (5.19) and at the drifts (5.20), it is also not immediately clear if one out of the Stratonovich, Itô-jet and Itô-vector projections

is preferable to the others. In [6, 7] a case is made for the optimality of the Itô-jet and Itô-vector projections. Provided that ${}^{\mathrm{proj}}\tilde{\pi}_0 = \pi_0 \in M$, the Itô-jet projection minimizes the first and second order coefficients in the Taylor expansion around $t = 0$ of the error $\mathbf{E}\left[d_g\big({}^{\mathrm{proj}}\tilde{\pi}_t, \mathfrak{p}(\pi_t)\big)^2\right]$ (where d_g is the Riemannian distance under the metric g). The Itô-vector projection satisfies a similar optimality criteria by minimizing the first/second order terms of the strong/weak Taylor expansion of the error $\mathbf{E}\left[||\psi({}^{\mathrm{p}}\tilde{\pi}_t), \psi(\pi_t)||_{\ell^2}\right]$, where ψ is any normal chart for $\mathring{\mathcal{S}}^n$ centred at $\tilde{\pi}_0 \in M$(see [7, Thm. 4.3, Thm. 4.8, Rmk. 5.6]). However, in our case we are not interested in either of these errors, but wish to minimize the expected infinitesimal Hilbert errors appearing on the right-hand side of (5.18) instead. Therefore which projection among Stratonovich, Itô-jet and Itô-vector performs better in our setting remains an open question.

Overall, it looks like working with the error bounds from Theorem 5.3.1 might not be as useful as we could have hoped, in terms of making progress towards the definition of an 'optimal' projection filter. The error terms (5.19) are hard to analyze, and therefore hard to minimize. The error bound (5.18) does not shed much light directly on the type of projection to employ, nor on the choice of Riemannian metric for $\mathring{\mathcal{S}}^n$, nor on the choice of submanifold M. Intuitively, we know that all of the above should be selected so that the infinitesimal Hilbert error between the dynamics of ${}^{\mathrm{p}}\tilde{\pi}$ and those of π is minimized, but Theorem 5.3.1 does not offer much insight on how to accomplish this. From Theorem 4.1.2 we know that tighter, pathwise error estimates of the error between the Wonham filter π and an approximate filter (such as the projection filter ${}^{\mathrm{p}}\tilde{\pi}$) are possible if the error terms ${}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t$ from (5.19) vanish for all $i \in \mathbf{N}$ and $k = 1, \ldots, d$ (and consequently ${}^{\mathrm{p}}\mathcal{E}^{2,k,j}_t$ vanish as well). Finding a way to force these error terms to vanish is going to be our goal in the next sections, although we will need to slightly alter our approach to achieve it.

Remark 5.3. (We refer to Theorem 5.3.1 above for the notation.) Another issue of the bounds of Theorem 5.3.1 is the presence of local times in our estimates. Unless the processes ${}^{\mathrm{p}}\Delta_{ij} - {}^{\mathrm{p}}\Delta_{uv}$ have a density (w.r.t. Lebesgue), these terms are very hard to estimate. However, we know heuristically that the local time of the process ${}^{\mathrm{p}}\Delta_{ij} - {}^{\mathrm{p}}\Delta_{uv}(t)$ is 'proportional' to its quadratic variation (see e.g. [74, Chapter VI,Cor. 1.9], which is often taken as the definition of the local time). The quadratic variation of ${}^{\mathrm{p}}\Delta_{ij} - {}^{\mathrm{p}}\Delta_{uv}(t)$ is given by

$$\sum_{k=1}^{d}\Big({}^{\mathrm{p}}\mathcal{E}^{3,k,i}_t - {}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t - {}^{\mathrm{p}}\mathcal{E}^{3,k,u}_t + {}^{\mathrm{p}}\mathcal{E}^{3,k,v}_t\Big)^2,$$

so minimizing the error terms ${}^{\mathrm{p}}\mathcal{E}_t^{3,k,i}$ in (5.19) for all $i \in \mathbf{N}$ and $k = 1, \ldots, d$ would potentially minimize the error due to the local times as well.

Remark 5.4. As we mentioned earlier, in [4, 19] the authors work with the space of probability distributions on $\mathbb{R}$, not with the simplex $\mathcal{S}^n$. However, the methodology that we implemented in Section 5.2.1 to derive the projection filter is exactly the one proposed in these papers. In fact, our manifold M is a statistical family of discrete distributions, which means that it is both an exponential and a mixture family. Then, endowing $\mathring{\mathcal{S}}^n$ with the Fisher Information or the Euclidean metric (both of which make $\mathring{\mathcal{S}}^n$ into a Riemannian manifold), and defining the inner product on T_pM to be the one induced by either of these metrics, yields respectively the Hellinger projection filter of [19] or the mixture projection filter of [4]. Finally, we note that the fact that the error terms of Theorem 5.3.1 are hard to quantify does not necessarily signify that the projection filters defined in Section 5.2.1 perform badly as approximations of the Wonham filter. As long as the infinitesimal Hilbert errors between the dynamics of the Wonham filter and those of the projection filter are small enough at each point of the submanifold M, the total error should remain small as well.

5.4 Projection filters in the space of natural parameters

It is quite curious to note that, for all that the Wonham SDE (5.1) is an equation on a manifold, we have never really needed to work in local coordinates, since the SDE is perfectly well-defined globally in the coordinate-system of $\mathbb{R}^{n+1}$. Now, it would probably not be particularly useful to move from a global coordinate system to a local one. However, not counting $p \in \mathring{\mathcal{S}}^n \subset \mathbb{R}^{n+1}$, there are two global charts for $\mathring{\mathcal{S}}^n$ that are natural when working in a statistical setting. These are given by the *natural parameters*, denoted by θ (which we have employed many times throughout this thesis) and the *expectation parameters*, denoted by η, by viewing discrete probabilities in the simplex respectively as an exponential and a mixture statistical family. These two parametrizations, and an interesting duality relationship between them, are the starting point of information geometry [1, 3]. These two charts are given by

$$\mathring{\mathcal{S}}^n \ni p \mapsto \theta \in \mathbb{R}^n \,:\, \theta^i = \log \frac{p^i}{p^0}, \quad \text{and} \quad \mathring{\mathcal{S}}^n \ni p \to \eta \in (0,1)^{\times n} \,:\, \eta^i = p^i, \tag{5.23}$$

for $i = 1, \ldots, n$, and their respective inverses

$$p^0 = \frac{1}{1 + \sum_k e^{\theta^k}},\ p^i = \frac{e^{\theta^i}}{1 + \sum_k e^{\theta^k}}, \quad \text{and} \quad p^0 = 1 - \sum_k \eta^k,\ p^i = \eta^i. \tag{5.24}$$

Notation. For $p \in \mathring{\mathcal{S}}^n$, we might write $\theta(p)$ or $\eta(p)$ for the representation of p in θ- or η-coordinates, i.e. the image of p under the diffeomorphisms in (5.23). Similarly, by $p(\theta)$ or $p(\eta)$ we denote the mapping of $\theta \in \mathbb{R}^n$ or $\eta \in (0,1)^{\times n}$ back to $\mathring{\mathcal{S}}^n$.

Now, we do not expect the η-coordinates to offer any significant advantage compared to the standard coordinates $p \in \mathring{\mathcal{S}}^n$, since the Wonham SDE (5.1) expressed in η is essentially the same as when expressed in p, removing the first component. On the other hand, as we have seen already in Chapters 3 and 4, the θ-coordinates can be very effective when working with the Wonham filter. In particular, we attribute some of the difficulties we encounter in the analysis of the error of the projection filters in Theorem 5.3.1 to the nonlinearities in the stochastic term of the Wonham SDE (5.1). Moving to the θ-coordinates gets rid of this problem. Applying Itô's formula (or from (3.11)) we have that the Wonham SDE (5.1) in the natural parameters, componentwise, is given by

$$\begin{aligned} \mathrm{d}\theta_t^i &= \sum_{j=0}^n \left[q_{ji}(t)\frac{\pi_t^j}{\pi_t^i} - q_{j0}(t)\frac{\pi_t^j}{\pi_t^0} \right] \mathrm{d}t + \sum_{k=1}^d \Big[(h_k^i(t) - h_k^0(t))\,\mathrm{d}Y_t^k + \frac{1}{2}\big(h_k^0(t)^2 - h_k^i(t)^2\big)\,\mathrm{d}t \Big], \\ \theta_0^i &= \log\frac{\mu^i}{\mu^0}, \end{aligned} \tag{5.25}$$

for $i = 1, \dots, n$ and $\theta_t^i = \log \frac{\pi_t^i}{\pi_t^0}$. In vector form, the Wonham SDE for $\theta_t \in \mathbb{R}^n$ is

$$\mathrm{d}\theta(t) = AD(\pi_t)Q_t^\top \pi_t\,\mathrm{d}t + \sum_{k=1}^d \Big[\mathbb{h}_k(t)\,\mathrm{d}Y_t^k - \frac{1}{2}\mathbb{H}_k(t)\,\mathrm{d}t \Big], \quad \theta_0 = \theta(\mu), \tag{5.26}$$

where

$$D(\pi_t) = \operatorname{diag}\left(\left\{ \frac{1}{\pi_t^i} \right\}_{i=0}^n \right) \in \mathbb{R}^{(n+1)\times(n+1)}, \quad A = \begin{pmatrix} -1 & \\ \vdots & \mathrm{Id}_{n\times n} \\ -1 & \end{pmatrix} \in \mathbb{R}^{n\times(n+1)},$$

and for $k = 1, \dots d$,

$$\mathbb{h}_k(t) = \begin{pmatrix} h_k^1(t) - h_k^0(t) \\ \vdots \\ h_k^n(t) - h_k^0(0) \end{pmatrix} \in \mathbb{R}^n, \quad \mathbb{H}_k(t) = \begin{pmatrix} h_k^1(t)^2 - h_k^0(t)^2 \\ \vdots \\ h_k^n(t)^2 - h_k^0(t)^2 \end{pmatrix} \in \mathbb{R}^n. \tag{5.27}$$

Remark 5.5. Since the diffusion coefficients $\mathbb{h}_k(t)$ are deterministic for all $k = 1, \dots, d$, the Itô and Stratonovich integrals agree, so (5.26) does not change when written in Stratonovich form.

Notation. Let the first drift term of (5.26) be denoted by $a(\theta, t) := AD(p(\theta))Q_t^\top p(\theta)$ (where $p(\theta)$ is the image of $\theta \in \mathbb{R}^n$ in $\mathring{\mathcal{S}}^n$).

We now proceed to define a projection filter in the space Θ of the natural parameters. The methodology is exactly the same as in Section 5.2.1, although we consider $\mathbb{R}^n$, instead of $\mathring{\mathcal{S}}^n$, as our ambient space now, and instead of distributions $p \in \mathring{\mathcal{S}}^n$ we will work with their natural parameters $\theta \in \mathbb{R}^n$. We find that our calculations are much simplified.

Let $m \ll n$ and let $M \subset \mathbb{R}^n$ be an m-dimensional submanifold embedded in $\mathbb{R}^n$ of the form

$$M = \{\theta(\xi) \in \mathbb{R}^n \,:\, \theta(\xi) = F(\xi) = (\theta^1(\xi), \dots, \theta^n(\xi)),\, \xi \in \mathbb{R}^d\}, \tag{5.28}$$

where $F : M \to \mathring{\mathcal{S}}^n$ is an embedding. Equip $\mathbb{R}^n$ with a Riemannian metric g and induce the same metric on M. Define the Riemannian submersion $\mathfrak{q} : U \to M$, where U is a tubular neighbourhood of M, by $\mathfrak{q}(\theta) = \operatorname{argmin}\{d_g(\theta, x) \,:\, x \in M\}$, which reduces to the orthogonal projection on the tangent spaces $\mathfrak{q}_\theta : T_\theta\mathbb{R}^n \cong \mathbb{R}^n \to T_\theta M \cong \operatorname{span}(\{\partial_i|_\theta\})$ given by, for $\theta \in M$,

$$w \mapsto \sum_{i=1}^m \sum_{j=1}^m g^{-1}_{ij}(\theta)\langle w, \partial_j|_\theta\rangle_{g(\theta)}\partial_i|_\theta. \tag{5.29}$$

Equivalently to Section 5.2.1, we define the Stratonovich, Itô-jet and Itô-vector projections of (5.26) on M. For $\theta \in M$ and $X \in T_\theta\mathbb{R}^n$, let $\overline{X} = \mathfrak{q}_\theta(X) \in T_\theta M$, and $\breve{X} = \mathfrak{q}^\perp_\theta(X) = (\mathbb{I}\mathrm{d} - \mathfrak{q}_\theta)(X) \in N_\theta M$.

The *Stratonovich projection* (in Stratonovich form) of (5.26) is given by

$$\mathrm{d}\tilde{\theta}_t = \overline{a}(\tilde{\theta}_t, t)\,\mathrm{d}t - \frac{1}{2}\sum_{k=1}^d \overline{\mathbb{H}}_k(\tilde{\theta}_t, t)\,\mathrm{d}t + \sum_{k=1}^d \overline{\mathbb{h}}_k(\tilde{\theta}_t, t) \circ \mathrm{d}Y^k_t, \quad \tilde{\theta}_0 \in M, \tag{5.30}$$

where we have stressed dependence of $\overline{\mathbb{h}}_k$ and $\overline{\mathbb{H}}_k$ on both t and θ since the projection operator (5.29) depends on θ (although the vectors $\mathbb{h}_k$ and $\mathbb{H}_k$ do not). From now on, unless specifically stated, evaluation of the projected SDE's coefficients at $(\tilde{\theta}_t, t)$ is implied. Note that as in Section 5.2.1, if the initial conditions of (5.26) $\theta(\mu) \notin M$, we need to fix a criterion to choose the initial conditions of (5.30) (e.g. choose $\tilde{\theta}_0 = \operatorname{argmin}\{d_g(\theta, \theta(\mu)) \,:\, \theta \in M\}$). We now transform (5.30) back into Itô form, which results in

$$\mathrm{d}\tilde{\theta}_t = \overline{a}\,\mathrm{d}t - \frac{1}{2}\sum_{k=1}^d \overline{\mathbb{H}}_k\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^d\sum_{j=1}^n \frac{\partial\overline{\mathbb{h}}_k}{\partial\theta^j}\overline{\mathbb{h}}^j_k\,\mathrm{d}t + \sum_{k=1}^d \overline{\mathbb{h}}_k\,\mathrm{d}Y^k_t, \quad \tilde{\theta}_0 \in M. \tag{5.31}$$

The *Itô-jet projection* of (5.26) onto M is the SDE

$$\mathrm{d}\tilde{\theta}_t = \overline{a}\,\mathrm{d}t - \frac{1}{2}\sum_{k=1}^d \overline{\mathbb{H}}_k\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^d\sum_{i,j=1}^n \frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\mathbb{h}^i_k\mathbb{h}^j_k\,\mathrm{d}t + \sum_{k=1}^d \overline{\mathbb{h}}_k\,\mathrm{d}Y^k_t, \quad \tilde{\theta}_0 \in M, \tag{5.32}$$

while the *Itô-vector projection* is given by

$$\mathrm{d}\tilde{\theta}_t = \overline{a}\,\mathrm{d}t - \frac{1}{2}\sum_{k=1}^{d}\overline{\mathbb{H}}_k\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=1}^{n}\frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\overline{\mathbb{h}}_k^i\overline{\mathbb{h}}_k^j\,\mathrm{d}t + \sum_{k=1}^{d}\overline{\mathbb{h}}_k\,\mathrm{d}Y_t^k, \quad \tilde{\theta}_0 \in M. \tag{5.33}$$

For comparison, we rewrite here the drifts of (5.31), (5.32) and (5.33), using once more [75, Eq.1.57, Eq.1.63] to decompose the drift of (5.31). Recall that we can write the orthogonal projection (5.29) as a an $n \times n$ matrix acting on $w \in T_\theta\mathbb{R}^n$ analgously to (5.6). Denote this matrix by $\mathfrak{L}(\theta)$, so that for $\theta \in M$ and $w \in T_\theta M$ we can write $\mathfrak{q}_\theta(w) = \mathfrak{L}(\theta)w$. Consider the last term of the drift of the Stratonovich projection (5.31). Componentwise, we compute

$$\begin{aligned}\sum_{j=1}^{n}\frac{\partial\overline{\mathbb{h}}_k^i}{\partial\theta^j}\overline{\mathbb{h}}_k^j\,\mathrm{d}t &= \sum_{j=1}^{n}\frac{\partial(\mathfrak{L}\,\mathbb{h}_k)^i}{\partial\theta_j}\overline{\mathbb{h}}_k^j\,\mathrm{d}t = \sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\partial\mathfrak{L}_{il}}{\partial\theta_j}\mathbb{h}_k^l\overline{\mathbb{h}}_k^j\,\mathrm{d}t \\ &= \sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\partial\mathfrak{L}_{il}}{\partial\theta_j}\overline{\mathbb{h}}_k^j(\overline{\mathbb{h}}_k^l + \breve{\mathbb{h}}_k^l)\,\mathrm{d}t = \sum_{u=1}^{n}\sum_{v=1}^{n}\frac{\partial^2\mathfrak{q}^i}{\partial\theta^u\theta^v}\overline{\mathbb{h}}_k^j(\overline{\mathbb{h}}_k^u + \breve{\mathbb{h}}_k^v).\end{aligned}$$

Then finally, for $\mathrm{p} \in \{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, $t \geq 0$ and $\theta \in M$, we write the drifts of the three projected SDEs as

$$^{\mathrm{p}}D(\theta,t) = -\frac{1}{2}\sum_{k=1}^{d}\overline{\mathbb{H}}_k + {}^{\mathrm{p}}C(\theta,t), \tag{5.34}$$

where

$$\begin{aligned}{}^{\text{Strat}}C(\theta,t) &= \overline{a} + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n}\left[\frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\overline{\mathbb{h}}_k^i\breve{\mathbb{h}}_k^j + \frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\overline{\mathbb{h}}_k^i\overline{\mathbb{h}}_k^j\right], \\ {}^{\text{Itô-j}}C(\theta,t) &= \overline{a} + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=0}^{n}\left[2\frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\overline{\mathbb{h}}_k^i\breve{\mathbb{h}}_k^j + \frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\overline{\mathbb{h}}_k^i\overline{\mathbb{h}}_k^j\right], \\ {}^{\text{Itô-v}}C(\theta,t) &= \overline{a} + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=1}^{n}\frac{\partial^2\mathfrak{q}}{\partial\theta^i\partial\theta^j}\overline{\mathbb{h}}_k^i\overline{\mathbb{h}}_k^j.\end{aligned} \tag{5.35}$$

Note in particular that the the drift of the Stratonovich projection and that of the Itô-jet projection only differ by a factor of 2 (compare with (5.20)). This is due to the fact that the vectors $\mathbb{h}_k(t)$ in (5.26) are independent of θ_t for all $k = 1, \ldots, d$, while in Section 5.2.1 we were working with the Wonham SDE (5.1), where the diffusion coefficients depend on π_t.

Just as in the case of the projected SDEs in Section 5.2.1, the Stratonovich (5.31), Itô-jet (5.32) and Itô vector projections (5.33) of (5.26) on $M \subset \mathbb{R}^n$ are examples of

projection filters, and can be expressed as m-dimensional SDEs for the parameter of M. To distinguish these new projection filters from those we defined in Section 5.2.1, we refer to (5.31), (5.32) and (5.33) as *θ-projection filters*, to emphasize that the projections are defined in the space of the natural parameters θ. We state the equivalent result to Theorem 5.3.1.

Theorem 5.4.1 (Expected Hilbert error bounds for the θ-projection filter)**.** *Let π_t be the solution to* (5.1) *and let ${}^{\text{Strat}}\tilde{\theta}_t$, ${}^{\text{Itô-j}}\tilde{\theta}_t$ and ${}^{\text{Itô-v}}\tilde{\theta}_t$ be the solutions respectively to the Stratonovich projection* (5.31)*, the Itô-jet projection* (5.32) *and the Itô-vector projection* (5.26) *of the SDE* (5.26) *on the submanifold $M \subset \mathbb{R}^n$ defined by* (5.28)*. Assume $q_{ij}(t) > 0$ for all $i \neq j$ and all $t \geq 0$, and $\pi_0 = \mu \in \mathring{\mathcal{S}}^n$. For $\mathrm{p} \in \{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, let ${}^{\mathrm{p}}\tilde{\pi}_t = p\big({}^{\mathrm{p}}\tilde{\theta}_t\big) \in \mathring{\mathcal{S}}^n$ be the image of ${}^{\mathrm{p}}\tilde{\theta}_t$ in $\mathring{\mathcal{S}}^n$ under the inverse of the θ-chart given by* (5.24) *and let ${}^{\mathrm{p}}\tilde{\pi}_0 = \nu \in p(M)$ (where $p(M)$ is the image of $M \subset \mathbb{R}^n$ in $\mathring{\mathcal{S}}^n$). Then the bounds* (5.18) *hold, with the error terms given by, for $\mathrm{p} \in \{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, $t \geq 0$, and for $j \in \mathbf{N}$,*

$$\begin{aligned}
{}^{\mathrm{p}}\mathcal{E}_t^{1,j} &= \left(\sum_{r=0}^{n} q_{rj}(t)\frac{{}^{\mathrm{p}}\tilde{\pi}_t^r}{{}^{\mathrm{p}}\tilde{\pi}_t^j}\right) - \widehat{\delta}_{j0}\,{}^{\mathrm{p}}C^j\big({}^{\mathrm{p}}\tilde{\theta}_t, t\big), \\
{}^{\mathrm{p}}\mathcal{E}_t^{2,k,j} &= \widehat{\delta}_{j0}\Big[\mathbb{H}_k^j(t) - \overline{\mathbb{H}}_k^j\big({}^{\mathrm{p}}\tilde{\theta}_t, t\big)\Big], \quad \text{for } k = 1, \ldots, d, \\
{}^{\mathrm{p}}\mathcal{E}_t^{3,k,j} &= \widehat{\delta}_{j0}\Big[\mathbb{h}_k^j(t) - \overline{\mathbb{h}}_k^j\big({}^{\mathrm{p}}\tilde{\theta}_t, t\big)\Big], \quad \text{for } k = 1, \ldots, d,
\end{aligned} \tag{5.36}$$

where we have defined $\widehat{\delta}_{ij} := 1 - \delta_{ij}$ (with δ_{ij} the Kronecker delta), and the drift terms ${}^{\mathrm{p}}C$ are given by (5.35)*.*

Proof. This result is again a consequence of Theorem 4.1.1. One way to prove it is to apply Itô's formula to derive the SDEs for ${}^{\mathrm{p}}\tilde{\pi}_t = p\big({}^{\mathrm{p}}\tilde{\theta}_t\big)$ for $\mathrm{p} \in \{\text{Strat}, \text{Itô-j}, \text{Itô-v}\}$, where $p : \mathbb{R}^n \to \mathring{\mathcal{S}}^n$ is the inverse of the θ-chart from (5.24). Then a direct application of Theorem 4.1.1 yields the result. To avoid these tedious calculations, we present a simpler alternative proof.

Let $\theta_t \in \mathbb{R}^n$ be the solution to (5.26) and $\tilde{\theta}_t \in \mathbb{R}^n$ the solution to any of the equations (5.31), (5.32) or (5.33). Let $\pi_t = p(\theta_t)$ (so π_t is effectively the solution to (5.1)) and $\tilde{\pi}_t = p(\tilde{\theta}_t)$. Consider the processes $\Delta_{ij}(t) = \log\frac{\pi_t^i}{\pi_t^j} - \log\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^j}$ for $i, j \in \mathbf{N}$. Since $\tilde{\pi}_t$ and $\tilde{\theta}_t$ are connected by diffeomorphism, we can write Δ_{ij} in terms of θ_t and $\tilde{\theta}_t$ as $\Delta_{ij} = \theta_t^i - \theta_t^j - \tilde{\theta}_t^i + \tilde{\theta}^j$ for $i, j = 1, \ldots, n$, and $\Delta_{i0} = \theta_t^i - \tilde{\theta}_t^i$, and $\Delta_{0i} = -\Delta_{i0}$ for $i = 1, \ldots, n$. Writing (5.26) and (any of) (5.31), (5.32), (5.33) (for p = Strat, Itô-j and Itô-v respectively) componentwise, we get that the SDEs for the Δ_{ij} processes

are given by

$$\begin{aligned}
\mathrm{d}\Delta_{ij}(t) &= \sum_{r=0}^{n}\left[q_{ri}(t)\frac{\pi_t^r}{\pi_t^i} - q_{rj}(t)\frac{\pi_t^r}{\pi_t^j}\right]\mathrm{d}t \\
&\quad + \sum_{k=1}^{d}\left[\big(\mathbb{h}_k^i(t) - \mathbb{h}_k^j(t)\big)\,\mathrm{d}Y_t^k - \frac{1}{2}\big(\mathbb{H}_k^i(t) - \mathbb{H}_k^j(t)\big)\,\mathrm{d}t\right] \\
&\quad - \Big({}^{\mathrm{p}}C^i(\tilde{\theta}_t,t) - {}^{\mathrm{p}}C^j(\tilde{\theta}_t,t)\Big)\,\mathrm{d}t - \sum_{k=1}^{d}\Big(\overline{\mathbb{h}}_k^i(\tilde{\theta}_t,t) - \overline{\mathbb{h}}_k^j(\tilde{\theta}_t,t)\Big)\,\mathrm{d}Y_t^k \\
&\quad + \frac{1}{2}\sum_{k=1}^{d}\Big(\overline{\mathbb{H}}_k^i(\tilde{\theta}_t,t) - \overline{\mathbb{H}}_k^j(\tilde{\theta}_t,t)\Big)\,\mathrm{d}t, \\
\Delta_{ij}(0) &= \log\frac{\mu^i}{\mu^j} - \log\frac{\nu^i}{\nu^j},
\end{aligned}$$

for $i,j = 1,\dots,n$, and similarly

$$\begin{aligned}
\mathrm{d}\Delta_{i0}(t) &= \sum_{r=0}^{n}\left[q_{ri}(t)\frac{\pi_t^r}{\pi_t^i} - q_{r0}(t)\frac{\pi_t^r}{\pi_t^0}\right]\mathrm{d}t + \sum_{k=1}^{d}\left[\mathbb{h}_k^i(t)\,\mathrm{d}Y_t^k - \frac{1}{2}\mathbb{H}_k^i(t)\,\mathrm{d}t\right] \\
&\quad - {}^{\mathrm{p}}C^i(\tilde{\theta}_t,t)\,\mathrm{d}t - \sum_{k=1}^{d}\left[\overline{\mathbb{h}}_k^i(\tilde{\theta}_t,t)\,\mathrm{d}Y_t^k - \frac{1}{2}\overline{\mathbb{H}}_k^i(\tilde{\theta}_t,t)\,\mathrm{d}t\right], \\
\Delta_{ij}(0) &= \log\frac{\mu^i}{\mu^0} - \log\frac{\nu^i}{\nu^0}.
\end{aligned}$$

Putting these two equations together and adding and subtracting terms appropriately, we can write, for all $i,j \in \mathbf{N}$,

$$\begin{aligned}
\mathrm{d}\Delta_{ij}(t) &= \sum_{r=0}^{n}\left[q_{ri}(t)\Big(\frac{\pi_t^r}{\pi_t^i} - \frac{\tilde{\pi}_t^r}{\tilde{\pi}_t^i}\Big) - q_{rj}(t)\Big(\frac{\pi_t^r}{\pi_t^j} - \frac{\tilde{\pi}_t^r}{\tilde{\pi}_t^j}\Big)\right]\mathrm{d}t \\
&\quad + \left[\Big(\sum_{r=0}^{n} q_{ri}(t)\frac{\tilde{\pi}_t^r}{\tilde{\pi}_t^i} - \sum_{r=0}^{n} q_{rj}(t)\frac{\tilde{\pi}_t^r}{\tilde{\pi}_t^j}\Big) - \Big(\widehat{\delta}_{i0}\,{}^{\mathrm{p}}C^i\big(\tilde{\theta}_t,t\big) - \widehat{\delta}_{j0}\,{}^{\mathrm{p}}C^j\big(\tilde{\theta}_t,t\big)\Big)\right]\mathrm{d}t \\
&\quad - \frac{1}{2}\sum_{k=1}^{d}\left[\widehat{\delta}_{i0}\Big(\mathbb{H}_k^i(t) - \overline{\mathbb{H}}_k^i\big(\tilde{\theta}_t,t\big)\Big) - \widehat{\delta}_{j0}\Big(\mathbb{H}_k^j(t) - \overline{\mathbb{H}}_k^j\big(\tilde{\theta}_t,t\big)\Big)\right]\mathrm{d}t \\
&\quad + \sum_{k=1}^{d}\left[\widehat{\delta}_{i0}\Big(\mathbb{h}_k^i(t) - \overline{\mathbb{h}}_k^i\big(\tilde{\theta}_t,t\big)\Big) - \widehat{\delta}_{j0}\Big(\mathbb{h}_k^j(t) - \overline{\mathbb{h}}_k^j\big(\tilde{\theta}_t,t\big)\Big)\right]\mathrm{d}Y_t, \\
\Delta_{ij}(0) &= \log\frac{\mu^i}{\mu^j} - \log\frac{\nu^i}{\nu^j},
\end{aligned} \tag{5.37}$$

where we have defined $\widehat{\delta}_{ij} := 1 - \delta_{ij}$ (where δ_{ij} is the Kronecker delta). On the right-hand side of the above we recognize our error terms (5.36). Then we can proceed exactly as in the proof of Theorem 4.1.1, by recalling the equality $\mathcal{H}(\pi_t,\tilde{\pi}_t) =$

$\max_{ij} \Delta_{ij}(t)$ from (3.13), approximating this process using a smooth function and taking the limit to yield the desired result. □

Remark 5.6. The error terms ${}^{\mathrm{p}}\mathcal{E}^{2,k,j}_t$ and ${}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t$ from (5.36) can equivalently be expressed as

$$ {}^{\mathrm{p}}\mathcal{E}^{2,k,j}_t = h^j_k(t)^2 - \widehat{\delta}_{j0}\overline{\mathbb{H}}^j_k\big({}^{\mathrm{p}}\tilde{\theta}_t, t\big) \quad \text{and} \quad {}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t = h^j_k(t) - \widehat{\delta}_{j0}\overline{\mathbb{h}}^j_k\big({}^{\mathrm{p}}\tilde{\theta}_t, t\big). $$

5.5 The primary natural submanifold of the Wonham filter

As we mentioned in Section 5.2.1, our goal in defining the θ-projection filter was to find a simple way to understand, and subsequently eliminate, the error terms ${}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t$ from (5.19) in Theorem 5.3.1, for all $k = 1, \dots, d$, $j \in \mathbf{N}$ and $t \geq 0$ (and for $\mathrm{p} \in \{\mathrm{Strat}, \mathrm{Itô\text{-}j}, \mathrm{Itô\text{-}v}\}$). When dealing with projection filters defined directly on a submanifold of the simplex (such as those of Section 5.2.1), this seemed a daunting task. If we now look instead at the error terms (5.36) in Theorem 5.4.1, we see that the errors of the θ-projection filter are much easier to work with. In fact, it is straightforward to notice that if the vectors $\mathbb{h}_k(t)$ are invariant under the orthogonal projection (5.29), the errors ${}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t$ from (5.36) in Theorem 5.4.1 vanish for all $k = 1, \dots, d$ and all $i \in \mathbf{N}$. In particular, this means that the stochastic terms in (5.37) disappear, which in turn implies that we can find tighter, pathwise error estimates for the Hilbert error of the θ-projection filter in the spirit of Theorem 4.1.2. We proceed to implement all these ideas, starting from the definition of the *primary natural submanifold of the Wonham filter*, along which the vectors $\mathbb{h}_k(t)$ are invariant.

Remark 5.7. Note that in Theorem 5.3.1 if the error terms ${}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t$ from (5.19) vanish, then so do the errors ${}^{\mathrm{p}}\mathcal{E}^{2,k,j}_t$ (for $k = 1, \dots, d$ and $j \in \mathbf{N}$). This is not the case for Theorem 5.4.1: even if ${}^{\mathrm{p}}\mathcal{E}^{3,k,j}_t$ from (5.36) vanish, there is no guarantee that the same should happen to ${}^{\mathrm{p}}\mathcal{E}^{2,k,j}_t$. If one is curious about the lack of symmetry, recall that to compute the error bounds in Theorem 5.3.1 we essentially have to go through the proof of Theorem 4.1.1, which relies on transforming both the SDE for the Wonham filter and that for the projection filter into θ-coordinates. So in the case of a projection filter of the type constructed in Section 5.2.1, defined directly by projecting on a submanifold of $\mathring{\mathcal{S}}^n$, we first define the SDE (in $\mathcal{S}^n$-coordinates) and then change coordinates to θ to obtain Theorem 5.3.1, while in the case of the θ-projection filters we first move to the θ-coordinate system and then project. Since projections and θ-transformation do not commute, some differences are to be expected.

Definition 5.5.1 (Primary natural submanifold)**.** Consider the vectors $\mathbb{h}_k(t) \in \mathbb{R}^n$ given in (5.27), for $k = 1, \dots, n$. Assume that we can find a set B (such that $|B| = \hat{d}$ constant, for simplicity) of time-dependendent, C^2, linearly independent vectors $\{v_k(t)\}_k$ such that $\text{span}\left(\{\mathbb{h}_k(t)\}_k\right) \subseteq \text{span}\left(\{v_k(t)\}_k\right)$. Then $\hat{d} \leq d$. Fix a point $\widehat{\theta} \in \mathbb{R}^n$. We define the *primary natural submanifold through* $\widehat{\theta}$ *of the Wonham filter* to be the statistical family given by

$$\gamma_t = \left\{ p(\xi, \widehat{\theta}) \in \mathring{\mathcal{S}}^n \,:\, p(\xi, \widehat{\theta}) = \exp\left\{ \widehat{\theta} + \sum_{k=1}^{\hat{d}} v_k(t)\xi^k - \mathbf{1}\log(\psi_t(\xi, \widehat{\theta})) \right\} \text{for } \xi \in \mathbb{R}^{\hat{d}} \right\}, \tag{5.38}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector of 1's, and $\psi_t(\xi, \widehat{\theta}) = 1 + \sum_{i=1}^n \exp\left\{\widehat{\theta^i} + \sum_{k=1}^{\hat{d}} v_k^i(t)\xi^k\right\}$.

At this point we should make a few remarks.

First of all, at each time $t \geq 0$ the dimension of γ_t is at most d, which is the dimension of the observation process Y. It is exactly d if the sensor function $h(t) = (h_k(t))_k$ is 'distinct enough' for each k: clearly, if there exists i, j such that $h_i(t)$ is a multiple of $h_j(t)$, we gain no more information from observing both Y^i and Y^j than we would from observing only one of them. Thus, for the sake of keeping the dimension of γ_t as low as possible, we make a selection of the vectors $\mathbb{h}_k(t)$ so that the set B in Definition 5.5.1 is maximally informative, but as small as possible. Related to this, we note that we could potentially allow the dimension of B to change over time: for example, if two vectors $\mathbb{h}_i(t)$ and $\mathbb{h}_j(t)$ are initially linearly independent but after some time become linearly dependent, we might want to remove one of them from B (and vice versa, they might start out dependent but become independent as time passes, in which case we might want to add to B the vector we are not yet keeping track of). As long as these changes happen smoothly, having a time-varying dimension for γ_t should not constitute a problem, although for simplicity we will not treat this case here.

Finally, we should spend a few words on $\widehat{\theta} \in \mathbb{R}^n$, since its role in Definition 5.5.1 is (purposefully) vague. As we will shortly see, the submanifold γ_t is flat when mapped to $\mathbb{R}^n$ using the θ-chart from (5.23). So for example, when γ_t is 1-dimensional, it is a line in $\mathbb{R}^n$ (at each time $t \geq 0$). The slope of this line is fixed by the vectors in B, so that, in a way, the vectors $\mathbb{h}_k$ are parallel to it (i.e. the orthogonal projection of $\mathbb{h}_k$ on γ_t is the identity). However, this criterion does not help us decide which, of the infinite number of lines in $\mathbb{R}^n$ with the correct slope (the same slope as γ_t), we should select. The point $\widehat{\theta} \in \mathbb{R}^n$ serves the purpose of fixing this selection. Although neither of them might be optimal, we suggest two possible choices for $\widehat{\theta}$.

Remark 5.8. Two sensible choices for the point $\widehat{\theta} \in \mathbb{R}^n$ in Definition 5.5.1 are

(i) $\widehat{\theta} = \theta(\mu)$, where $\mu \in \mathring{\mathcal{S}}^n$ is the initial condition for the Wonham filter SDE (5.1);

(ii) $\widehat{\theta} = \theta(p_{\mathrm{inv}})$, where $p_{\mathrm{inv}} \in \mathring{\mathcal{S}}^n$ is the underlying Markov chain X (assuming it is stationary).

Let us now move on to the definition of the the projection filter onto the primary submanifold γ_t, which we call the *γ-projection filter*.

First of all, let us embed γ_t in $\mathbb{R}^n$ through the usual θ-transformation. This yields

$$\widetilde{\gamma}_t := \theta(\gamma_t) = \left\{ \theta(\xi, \widehat{\theta}) \in \mathbb{R}^n \,:\, \theta(\xi, \widehat{\theta}) = \widehat{\theta} + \sum_{k=1}^{\hat{d}} v_k(t) \xi^k \text{ for } \xi \in \mathbb{R}^d \right\}, \tag{5.39}$$

which is a linear subspace of $\mathbb{R}^n$. At a point $\theta \in \widetilde{\gamma}_t$, the tangent space of $\widetilde{\gamma}_t$ is given by

$$T_\theta \widetilde{\gamma}_t \cong \operatorname{span}\left(\{v_k(t)\}_{k=1}^{\hat{d}}\right) = \operatorname{span}\left(\{\mathbb{h}_k(t)\}_{k=1}^{d}\right).$$

Equip $\mathbb{R}^n$ with a Riemannian metric g, and let $\mathfrak{q}_t : U_t \to \mathbb{R}^n$ (where U_t is a tubular neighbourhood of $\widetilde{\gamma}_t$) be defined as the Riemannian submersion $\mathfrak{q}_t(\theta) = \operatorname{argmin}\{d_g(\theta, y) : y \in \gamma_t\}$ at each time $t \geq 0$. Let the time-dependent orthogonal projection $\mathfrak{q}_{t,\theta} : T_\theta \mathbb{R}^n \to T_\theta \widetilde{\gamma}_t$ be

$$w \mapsto \sum_{i=1}^{\hat{d}} \sum_{j=1}^{\hat{d}} G^{-1}_{ij}(\theta, t) \langle w, v_i(t) \rangle_{g(\theta)} v_j(t), \tag{5.40}$$

where $G^{-1} = (G^{-1}_{ij})$ is the inverse of the matrix $(G(\theta, t)_{ij}) = \left(\langle v_i(t), v_j(t) \rangle_{g(\theta)}\right)$.

Now we would like to proceed as in the previous section and define the Stratonovich, Itô-jet and Itô-vector projections of the Wonham SDE in θ-coordinates (given by (5.26)) onto $\widetilde{\gamma}_t$. There are a couple of observations to make before doing so, however: the first one is that $\widetilde{\gamma}_t$ is time dependent, so we need to adapt our projections accordingly; the second is that, due to our choice of submanifold, the Stratonovich, Itô-jet and Itô-vector projections are all equivalent.

Let us start with this second remark, and assume for a moment that $\widetilde{\gamma}_t = \widetilde{\gamma}$, independent of time. Then we can define the three projected SDEs on γ as in (5.31), (5.32) and (5.33). Since the vectors $\mathbb{h}_k(t)$ are invariant under the orthogonal projection (5.40), by letting $\overline{\mathbb{h}}_k = \mathbb{h}_k$ in (5.31), (5.32) and (5.33) we see that the three SDEs are the same. In regard to the time-dependency of $\widetilde{\gamma}_t$, in [7, Example 4] it is shown how, to maintain tangency along the submanifold $\widetilde{\gamma}_t$ as it varies in time, one needs to add to the drifts of the projected SDEs (all three of them, Stratonovich, Itô-jet and Itô-vector) the time-derivative of the submersion $\mathfrak{q}_t$.

Therefore in the end we define the γ-*projection filter* as the solution to the SDE

$$\mathrm{d}\tilde{\theta}_t = \overline{a}(\tilde{\theta}_t, t)\,\mathrm{d}t - \frac{1}{2}\sum_{k=1}^{d}\overline{\mathrm{H}}_k(\tilde{\theta}_t, t)\,\mathrm{d}t + \sum_{k=1}^{d}\mathbb{h}_k(t)\circ\mathrm{d}Y_t^k + \dot{\mathfrak{q}}_t(\tilde{\theta}_t)\,\mathrm{d}t, \quad \tilde{\theta}_0 \in \widetilde{\gamma}_0, \quad (5.41)$$

which becomes, in Itô's form,

$$\mathrm{d}\tilde{\theta}_t = \overline{a}\,\mathrm{d}t - \frac{1}{2}\sum_{k=1}^{d}\overline{\mathrm{H}}_k\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=1}^{n}\frac{\partial^2\mathfrak{q}_t}{\partial\theta^i\partial\theta^j}\mathbb{h}_k^i\mathbb{h}_k^j\,\mathrm{d}t + \sum_{k=1}^{d}\mathbb{h}_k\,\mathrm{d}Y_t^k + \dot{\mathfrak{q}}_t(\tilde{\theta}_t)\,\mathrm{d}t, \quad \tilde{\theta}_0 \in \widetilde{\gamma}_0. \quad (5.42)$$

Remark 5.9. If the orthogonal projection (5.40) does not depend on $\theta \in \widetilde{\gamma}_t$ (so for example if g is taken to be the Euclidean metric on $\mathbb{R}^n$), then the Stratonovich (5.41) and Itô's form (5.42) of the SDE for the γ-projection filter are the same.

We can write the γ-projection filter as a $\hat{d}$-dimensional SDE for the parameter $\xi \in \mathbb{R}^{\hat{d}}$ of the statistical family γ_t. In the case where the tangent vectors $v_k(t) = v_k$ are independent of time, letting the solution to (5.41) be given by the γ-projection filter $\tilde{\theta}_t = \theta(\xi_t, \widehat{\theta}) \in \widetilde{\gamma}_t$, we apply the chain rule to see

$$\begin{aligned}\mathrm{d}\xi_t = G_{ij}^{-1}(\tilde{\theta}_t)\Big(\mathfrak{Q}_{\mathrm{vec}}(a(\tilde{\theta}_t, t), \tilde{\theta}_t)\,\mathrm{d}t + \frac{1}{2}\sum_{k=1}^{d}\mathfrak{Q}_{\mathrm{vec}}(\mathbb{H}_k(t), \tilde{\theta}_t)\,\mathrm{d}t + \sum_{k=1}^{d}\mathbb{h}_k(t)\circ\mathrm{d}Y_t^k\Big)\\ \xi_0 \in \mathbb{R}^{\hat{d}},\end{aligned} \quad (5.43)$$

where, for $\theta \in \widetilde{\gamma}_t$ and $w \in T_\theta\widetilde{\gamma}_t$, we denote by $\mathfrak{Q}_{\mathrm{vec}}(w, p)$ the $\hat{d}$-dimensional vector with each entry the component of w with respect to each basis vector of $T_\theta\widetilde{\gamma}$, i.e.

$$\mathfrak{Q}_{\mathrm{vec}}(w, \theta) = \begin{bmatrix}\langle w, v_1\rangle_{g(\theta)}\\ \vdots\\ \langle w, v_{\hat{d}}\rangle_{g(\theta)}\end{bmatrix}.$$

When $\{v_k(t)\}$ are time dependent, it should still be possible to obtain an SDE similar to (5.43) for the parameters ξ_t, but the presence of the drift term $\dot{\mathfrak{q}}_t$ makes the calculations more complicated (one needs to compute the components of $\dot{\mathfrak{q}}_t$ with respect to the basis $\{v_k(t)\}$).

We conclude this section with a theorem analogous to Theorem 4.1.2 for the error bounds of the γ-projection filter.

Theorem 5.5.2 (Pathwise Hilbert error bounds for the γ-projection filter)**.** *Let π_t be the solution to* (5.1) *and let $\tilde{\theta}_t$ be the solution to the SDE* (5.42)*. Assume $q_{ij}(t) > 0$ for all $i \neq j$ and all $t \geq 0$, and $\pi_0 = \mu \in \mathring{\mathcal{S}}^n$. Let $\tilde{\pi}_t = p(\tilde{\theta}_t)$ be the image of $\tilde{\theta}_t$ in $\mathring{\mathcal{S}}^n$*

under the inverse of the θ-chart given by (5.24)*, and let $\tilde{\pi}_0 = \nu \in \gamma_0$. Assume $\tilde{\pi}_t$ is observable. Let $u_t \in (0,1)$ be the unique solution to the ODE with random coefficients given by*

$$\begin{aligned} \frac{\mathrm{d}u_t}{\mathrm{d}t} &= -\tilde{\lambda}^\star(t,u_t)u_t + \frac{1}{4}\max_{i,j}\Big\{\mathcal{E}_t^{1,i} - \mathcal{E}_t^{1,j} - \frac{1}{2}\sum_{k=1}^{d}\big(\mathcal{E}_t^{2,k,i} - \mathcal{E}_t^{2,k,j}\big)\Big\}(1-u_t^2), \\ u_0 &= \tanh\left(\frac{\mathcal{H}(\mu,\nu)}{4}\right), \end{aligned} \tag{5.44}$$

where for $t \geq 0$ and $\alpha \in \mathbf{N}$ the error terms are

$$\begin{aligned} \mathcal{E}_t^{1,\alpha} &= \left(\sum_{r=0}^{n} q_{r\alpha}(t)\frac{\tilde{\pi}_t^r}{\tilde{\pi}_t^\alpha}\right) - \widehat{\delta}_{\alpha 0}\left(\overline{a}^\alpha(\tilde{\theta}_t) + \frac{1}{2}\sum_{k=1}^{d}\sum_{i,j=1}^{n}\frac{\partial^2 \mathfrak{q}_t^\alpha(\tilde{\theta}_t)}{\partial\theta^i\partial\theta^j}\mathbb{h}_k^i(t)\mathbb{h}_k^j(t) + \dot{\mathfrak{q}}_t^\alpha(\tilde{\theta}_t)\right), \\ \mathcal{E}_t^{2,k,\alpha} &= \widehat{\delta}_{\alpha 0}\Big[\mathbb{H}_k^\alpha(t) - \overline{\mathbb{H}}_k^\alpha\big(\tilde{\theta}_t, t\big)\Big], \quad \textit{for } k = 1,\dots,d, \end{aligned} \tag{5.45}$$

and we have defined $\widehat{\delta}_{ij} := 1 - \delta_{ij}$ (with δ_{ij} the Kronecker delta). The decay coefficient is

$$\tilde{\lambda}^\star(t,u_t) = \min_{i\neq k}\left\{\left(q_{ik}(t)\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^k} + \sum_{\substack{j\neq i,k,\\ j\notin\tilde{\mathcal{J}}_k^i(t,u_t)}} q_{jk}(t)\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^k}\right)\frac{1+u_t}{1-u_t} + \left(q_{ki}(t)\frac{\tilde{\pi}_t^k}{\tilde{\pi}_t^i} + \sum_{\substack{j\neq i,k,\\ j\in\tilde{\mathcal{J}}_k^i(t,u_t)}} q_{ji}(t)\frac{\tilde{\pi}_t^j}{\tilde{\pi}_t^i}\right)\frac{1-u_t}{1+u_t}\right\},$$

and $\tilde{\mathcal{J}}_k^i(t,u_t) := \Big\{j \in \mathbf{N} : \frac{q_{jk}(t)}{\tilde{\pi}_t^k} \geq \frac{q_{ji}(t)}{\tilde{\pi}_t^i}\Big(\frac{1-u_t}{1+u_t}\Big)^2\Big\}$. Then for all $t < \infty$,

$$\tanh\left(\frac{\mathcal{H}(\pi_t,\tilde{\pi}_t)}{4}\right) \leq u_t.$$

Proof. For $i,j \in \mathbf{N}$, write the dynamics of the difference processes $\Delta_{ij}(t) = \log\frac{\pi_t^i}{\pi_t^j} - \log\frac{\tilde{\pi}_t^i}{\tilde{\pi}_t^j}$ as in (5.37), replacing the appropriate coefficients with the drift and diffusion coefficients of (5.42). Letting $\overline{\mathbb{h}}_k = \mathbb{h}_k$ for all $k = 1,\dots,d$, we see that the stochastic term of (5.37) vanishes. Then exactly the same arguments as in the proof of Theorem 4.1.2 yield the result. □

Before proceeding to implement the γ-projection filter and test it numerically, we make one final remark. The way we constructed the primary submanifold γ_t is similar to how the exponential families for the exponential projection filter are selected in Brigo, Hanzon and Le Gland [19, Sec. 6]. In particular, in discrete-time stochastic filtering we can view the filter as a sequence of prediction and correction steps. Bending the terminology a little bit to fit our continuous-time setting, what both us and the authors of [19] suggest is to choose the statistical family for the projection filter so that there is no error in the correction step, as the stochastic part of the filter equations are matched perfectly.

5.6 Numerics and future directions

In this final section we present some simple numerical examples for the γ-projection filter, and discuss a few ideas (to be explored in more detail in the future) to systematically augment the dimension of the submanifold γ_t until sufficient precision for the projection filter is achieved.

We consider the same examples as in Section 4.4. Recall the homogenous 3-state and 6-state Markov chains with transition matrices given by (4.25). As in Section 4.4, for the 3-state chain, we take the initial law of the signal X to be given by its ergodic distribution, i.e. $\mu = \text{law}(X_0) = \big(0.3, 0.3, 0.4\big)$. For the 6-state chain, we start X close to the boundary of the simplex, and take its initial law to be $\mu = \big(0.25, 0.1, 0.06, 0.07, 0.22, 0.3\big)$. The Wonham filter π_t is initialized at $\pi_0 = \mu$ in both cases. We take γ to be the (in this case) time-independent manifold given by (5.38). We choose as reference point $\widehat{\theta}$ for γ the stationary distribution of X (expressed in θ-coordinates), in both the 3-state and 6-state case. Since Y is 1-dimensional, γ must also be 1-dimensional, and in particular the representation of γ in the space of natural parameters is given by the line

$$\widetilde{\gamma} = \{\theta \in \mathbb{R}^n \,:\, \widehat{\theta} + \mathbb{h}\xi, \text{ for } \xi \in \mathbb{R}\}, \tag{5.46}$$

for $n = 2, 5$. We now need to choose the projection operator (5.40). For simplicity, we begin by fixing the metric to be the standard Euclidean metric in $\mathbb{R}^n$.

At this point, it only remains to determine the initial conditions for the γ-projection filter $\tilde{\pi}_t$. In the 3-state case, since $\mu \in \gamma$, we can start π_t from μ as well, so $\tilde{\pi}_0 = \mu$. In the 6-state case, $\pi_0 = \mu \notin \gamma$, so we take $\tilde{\pi}_0$ to be the projection of μ onto γ, i.e. $\tilde{\pi}_0 \approx \big(0.79, 0.03, 0.07, 0.06, 0.01, 0.02\big)$. We simulate 100 paths for the Wonham filter and the projection filter, and compute their Hilbert errors. In Figure 5.1 we plot the results. As we have already discussed extensively in Section 3.3.4 and Section 4.4, we know that our error bounds are far from tight, mostly because our contraction rate is far too low for how fast the stability error of the filter contracts in reality. However, we see that the γ-projection filter performs well, in both the examples we consider.

We now consider examples in slightly higher dimensions. We take a 10-state and a 20-state Markov chain, with randomly generated transition matrices Q and sensor functions h. We take their initial law μ also to be randomly sampled, from $\mathcal{S}^9$ and $\mathcal{S}^{19}$ respectively. We implement the γ-projection filter in each case, projecting μ on γ to obtain the initial conditions for $\tilde{\pi}_t$. In Figure 5.2 we plot a realization of the

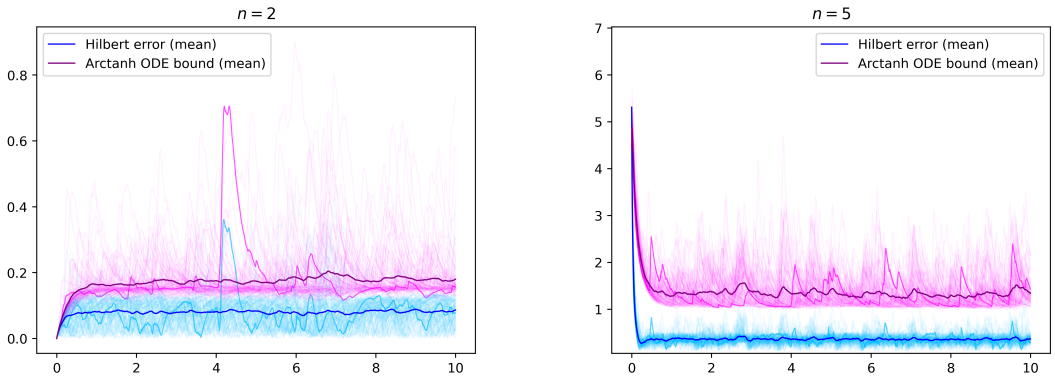

Figure 5.1: For dimensions $n = 2, 5$, we test our error bounds from Theorem 5.5.2 against the Hilbert error between the Wonham filter and the γ-projection filter. In each case, we plot 100 realizations of the Hilbert projective error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$ (faded, light blue) and of the ODE bound given by $4 \operatorname{arctanh}(u_t)$, where u_t solves (5.44) (faded, fuchsia). We highlight one sample path of the Hilbert error at random, together with its corresponding pathwise bound. In blue and purple we plot the sample means of the errors and of the bounds.

Hilbert error between the Wonham filter and the γ-projection filter for each of this systems. The performance of the γ-projection filter, as expected, is worse then in the previous examples, although the Hilbert error does not change much in magnitude from the 9- to the 19-dimensional case. Our error bounds, sadly, become virtually useless.

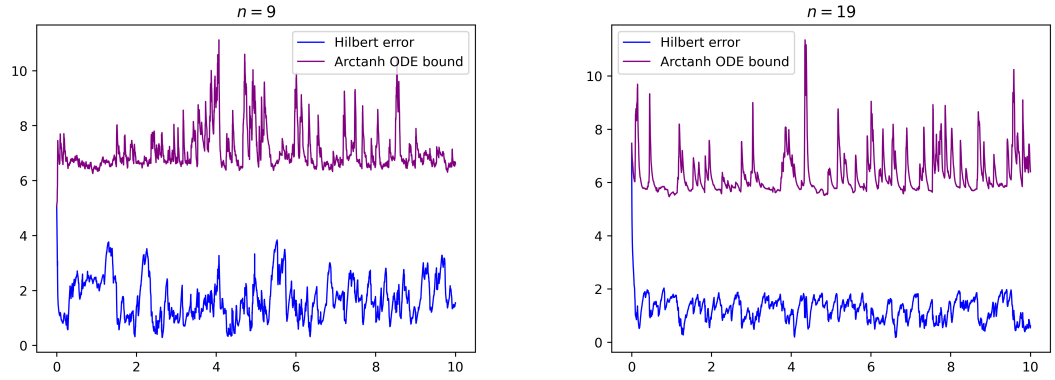

Figure 5.2: For dimensions $n = 9, 19$ we plot a realization of the Hilbert error (blue) between the Wonham filter and the γ-projection filter, and of the the ODE bound given by $4 \operatorname{arctanh}(u_t)$, where u_t solves (5.44) (purple).

Given these numerical results, there are a few approaches that we might take to try and make the γ-projection filter a viable alternative to the Wonham filter in

high dimension. The main idea is to systematically enlarge the submanifold γ_t so that the error bounds stay within a certain acceptable range. One way we might do this is to sample a few points at random from γ_t and to compute, at each of these points, the error of the projection operator (5.29). Running a PCA algorithm over the resulting error vectors, we can then select the directions along which the projection error is largest, and add them to γ_t. Note that this procedure can be done offline, before starting to run the projection filter, or adaptively: we can run the projection filter on the primary submanifold γ_t for a short amount of time, keeping track of the projection errors, and then add directions to γ_t based on a PCA on the projection errors at points of γ_t that the projection filters has actually visited.

We plan to implement this method, and variations of it, in the near future, and test it for high-dimensional problems. For now, we conclude the thesis with a geometric remark, as a final nod to the Hilbert projective metric, which has been such an instrumental part of the thesis.

Our work in this final chapter has focused on the use of Riemannian projections to construct projection filters. However, this is somewhat at odds with our analysis of the error using the Hilbert projective norm. A first reason for this is that, as the ball in the Hilbert geometry is not strictly convex, one cannot define a unique minimal-distance projection in Hilbert distance. We illustrate this issue in Figure 5.3, where we see (in red) the projection of a point P on a variety of subspaces l – in the latter two examples, there is an interval of points on l equidistant from P in the Hilbert geometry. A second reason is to better connect with existing work – the analysis of the three projection methods we have considered is only in the context of Riemannian manifolds. A final reason is computational, as there are simple, explicit methods of computing projections in inner product spaces, which can then be applied in a Riemannian context.

Nevertheless, the fact that we focus on the Hilbert geometry may suggest that alternatives to the Euclidean inner product are appropriate. Considering the apparent geometry of the Hilbert ball, we see that an ellipsoid with primary axis given by the vector $\mathbf{1} = (1, 1, 1)$, and all other axes symmetric may form a good approximation to the Hilbert ball. In fact, for 2-dimensional θ, the ellipse can be chosen to pass through all six vertices of the Hilbert ball. In higher dimension it can be chosen as the minimal ellipse of this type containing (or equivalently contained within) the Hilbert ball. We illustrate this ellipse in green in Figure 5.3.

Using this ellipse, with the corresponding inner product, we can define an inner product on θ vectors, and the metric this induces should approximate the Hilbert

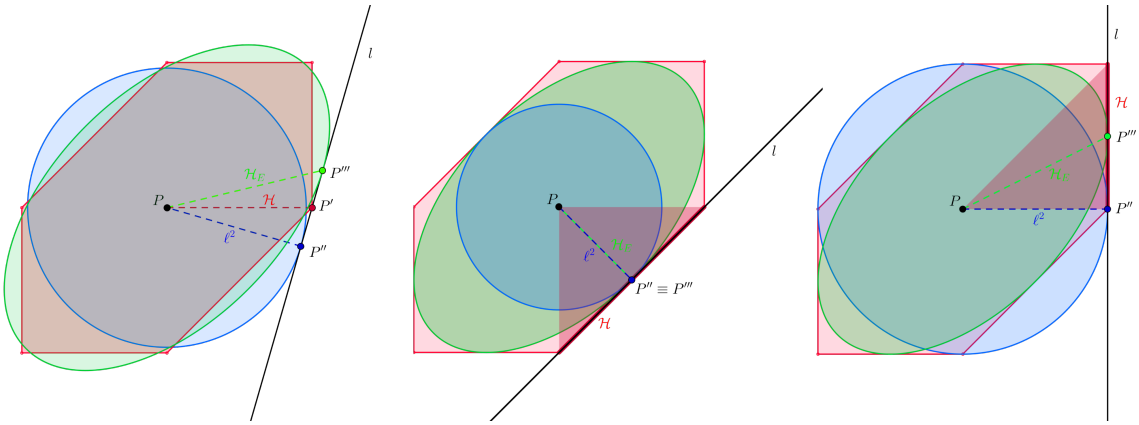

Figure 5.3: We compare the projections in Hilbert metric (red), Euclidean metric (blue), and elliptic-$\mathcal{H}$ (green) of a point $P \in \mathbb{R}^2$ onto a line l with different slopes. When l is parallel to the y-axis or the diagonal, the Hilbert projection is not unique. The Euclidean and elliptic-$\mathcal{H}$ projections are not generally the same unless the slope of l is 1 or -1. If the slope is -1, all three projections agree (not shown).

metric well, particularly when compared with the ℓ^2 geometry (shown in blue in Figure 5.3). We call this the elliptic-$\mathcal{H}$ inner product, and will consider this as an alternative projection geometry in our future analysis.

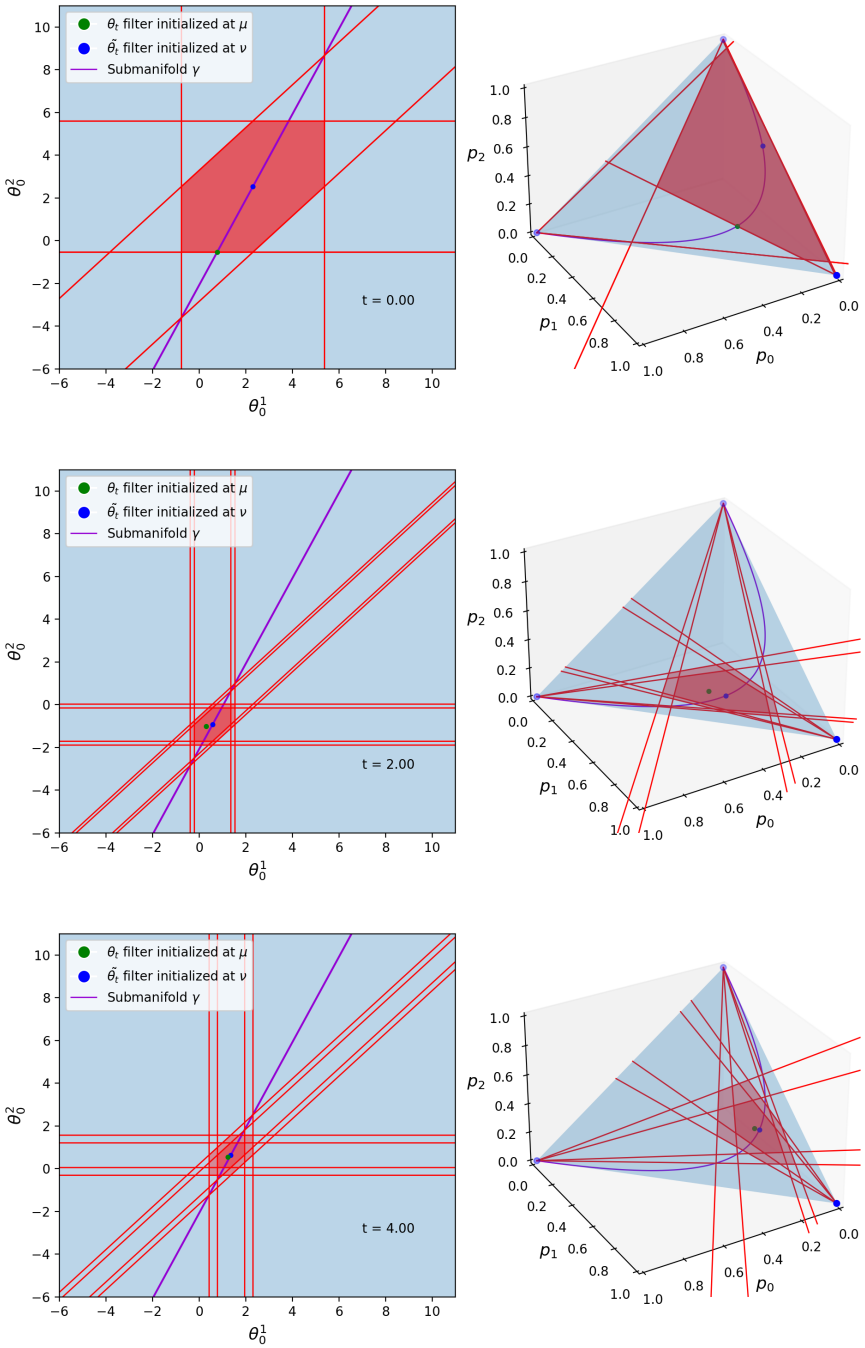

Figure 5.4: Evolution in time of the Wonham filter π_t and of the γ-projection filter $\tilde{\pi}_t$, shown side-by-side in $\mathcal{S}^2$ and in $\mathbb{R}^2$, together with the bounding Hilbert ball for the error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$, centered at $\tilde{\pi}_t$, and with radius given by $4\,\mathrm{arctanh}(u_t)$, where u_t solves (5.44).

Appendix A

Auxiliary results

A.1 The maximum process of a family of semimartingales

In this appendix we use an appropriate smooth approximation to study the dynamics of the maximum of a family of continuous stochastic processes driven by a common Brownian motion.

Recall the following smooth approximations of the maximum and the argmax. Let $\alpha \in (0, \infty)$ and let $\mathbf{x} = \{x_i\}_{i=0}^n$ be a sequence of real numbers. We define the *LogSumExp* function $LSE_\alpha(\mathbf{x})$ as

$$LSE_\alpha(\mathbf{x}) = \frac{1}{\alpha} \log \sum_k e^{\alpha x_k},$$

and the *SmoothMax* function $S_\alpha(\mathbf{x})$ as

$$S_\alpha(\mathbf{x}) = \frac{\sum_j x_j e^{\alpha x_j}}{\sum_k e^{\alpha x_k}}.$$

Given a family $\mathbf{c} = \{c_i\}_{i=0}^n$ of real-valued coefficients, we also define the *Soft-ArgMax* (or *SoftMax*) function $S_\alpha^{arg}(\mathbf{x}, \mathbf{c})$ as

$$S_\alpha^{arg}(\mathbf{x}, \mathbf{c}) = \frac{\sum_j c_j e^{\alpha x_j}}{\sum_k e^{\alpha x_k}}.$$

We start by proving a few simple lemmata.

Notation. Let $\mathcal{I}$ be the argmax of $\mathbf{x}$, i.e. $\mathcal{I} := \{j \in \mathbf{N} \,:\, x_j = \max_{i \in \mathbf{N}} x_i\} \subset \mathbf{N}$.

Lemma A.1.1 (Convergence to maximum)**.**

$$\lim_{\alpha \to \infty} LSE_\alpha(\mathbf{x}) = \lim_{\alpha \to \infty} S_\alpha(\mathbf{x}) = \max_{i \in \{0,\dots,n\}} x_i$$

Proof. Let $M = \max_{i\in\{0,\dots,n\}} x_i$. We have that

$$M = \frac{1}{\alpha}\log e^{\alpha M} \leq LSE_\alpha(\mathbf{x}) \leq \frac{1}{\alpha}\log\left((n+1)e^{\alpha M}\right) = \frac{\log(n+1)}{\alpha} + M,$$

and taking the limit as $\alpha \to \infty$ yields the result. For the *SmoothMax* function, consider $\mathcal{I}$, the argmax of $\mathbf{x}$, and let $|\mathcal{I}| = d \geq 1$ be its size. Then

$$\begin{aligned} S_\alpha(\mathbf{x}) &= \sum_{j\in\mathcal{I}} \frac{x_j}{d + \sum_{k\notin\mathcal{I}} e^{\alpha(x_k - x_j)}} + \sum_{j\notin\mathcal{I}} \frac{x_j}{1 + \sum_{k\neq j} e^{\alpha(x_k - x_j)}} \\ &= \frac{dM}{d + \sum_{k\notin\mathcal{I}} e^{-\alpha(M - x_k)}} + \sum_{j\notin\mathcal{I}} \frac{x_j}{1 + \sum_{\substack{k\neq j \\ k\in\mathcal{I}}} e^{\alpha(M - x_k)} + \sum_{\substack{k\neq j \\ k\notin\mathcal{I}}} e^{\alpha(x_k - x_j)}} \xrightarrow{\alpha\to\infty} M. \end{aligned}$$

□

Lemma A.1.2. *Let $\mathcal{I}$ be the argmax of $\mathbf{x}$ and $|\mathcal{I}| = d \geq 1$ be its size. Then*

$$\lim_{\alpha\to\infty} S_\alpha^{arg}(\mathbf{x}, \mathbf{c}) = \frac{1}{d}\sum_{j\in\mathcal{I}} c_j.$$

Proof. Similar to Lemma A.1.1. □

Lemma A.1.3 (Derivatives of $LSE_\alpha(\mathbf{x})$)**.**

$$\begin{aligned} \frac{\partial}{\partial x_i} LSE_\alpha(\mathbf{x}) &= \frac{e^{\alpha x_i}}{\sum_k e^{\alpha x_k}}, \\ \frac{\partial^2}{\partial x_i^2} LSE_\alpha(\mathbf{x}) &= \alpha \sum_{j\neq i} \frac{e^{\alpha(x_i + x_j)}}{\left(\sum_k e^{\alpha x_k}\right)^2}, \\ \frac{\partial^2}{\partial x_i \partial x_j} LSE_\alpha(\mathbf{x}) &= -\alpha \frac{e^{\alpha(x_i + x_j)}}{\left(\sum_k e^{\alpha x_k}\right)^2}. \end{aligned}$$

Proof. Easy calculations. □

Now consider the function

$$f_\alpha(\mathbf{x}) = \sum_i \frac{\partial^2}{\partial x_i^2} LSE_\alpha(\mathbf{x}) = -\sum_i \sum_{j\neq i} \frac{\partial^2}{\partial x_i \partial x_j} LSE_\alpha(\mathbf{x}) = \alpha \sum_i \sum_{j\neq i} \frac{e^{\alpha(x_i + x_j)}}{\left(\sum_k e^{\alpha x_k}\right)^2}.$$

Lemma A.1.4. *If $\max_i x_i$ is unique, i.e. if $\exists!\, j^\star$ such that $\max_i x_i = x_{j^\star}$, then*

$$\lim_{\alpha\to\infty} f_\alpha(\mathbf{x}) = 0.$$

Proof. Let $x_{j^\star} := \max_i x_i$. Since $x_{j^\star}$ is the unique maximizer, there exists $\varepsilon_j > 0$ such that $x_j = x_{j^\star} - \varepsilon_j$ for all $j \neq j^\star$. Then we have

$$\begin{aligned} f_\alpha(\mathbf{x}) &= \alpha\left[\sum_{j\neq j^\star} \frac{e^{\alpha(x_{j^\star}+x_j)}}{\left(\sum_k e^{\alpha x_k}\right)^2} + \sum_{i\neq j^\star} \frac{e^{\alpha(x_i+x_j^\star)}}{\left(\sum_k e^{\alpha x_k}\right)^2} + \sum_{i\neq j^\star}\sum_{\substack{j\neq i\\ j\neq j^\star}} \frac{e^{\alpha(x_i+x_j)}}{\left(\sum_k e^{\alpha x_k}\right)^2}\right] \\ &= \frac{\alpha}{\left(e^{\alpha x_{j^\star}} + \sum_{k\neq j^\star} e^{\alpha(x_{j^\star}-\varepsilon_k)}\right)^2}\left[2\sum_{j\neq j^\star} e^{\alpha(2x_{j^\star}-\varepsilon_j)} + \sum_{i\neq j^\star}\sum_{j\neq i} e^{\alpha(2x_{j^\star}-\varepsilon_i-\varepsilon_j)}\right] \\ &= \frac{\alpha}{\left(1 + \sum_{k\neq j^\star} e^{-\alpha\varepsilon_k}\right)^2}\left[2\sum_{j\neq j^\star} e^{-\alpha\varepsilon_j} + \sum_{i\neq j^\star}\sum_{j\neq i} e^{-\alpha(\varepsilon_i+\varepsilon_j)}\right], \end{aligned}$$

and since ε_j is strictly positive for all $j \neq j^\star$, in the limit as $\alpha \to \infty$ the negative exponentials $e^{-\alpha\varepsilon_j}$ dominate α, and $f_\alpha \to 0$. □

Lemma A.1.5. *Consider the function*

$$g_\alpha(x) = \alpha\frac{e^{\alpha x}}{(1+e^{\alpha x})^2}.$$

We have that $g_\alpha(x)\,\mathrm{d}x \to \delta_0$ as $\alpha \to \infty$ in the sense of weak convergence of measures, where δ_0 denotes the Dirac mass at 0.

Proof. First, note that for all $\alpha > 0$

$$\int_{\mathbb{R}} g_\alpha(x)\,\mathrm{d}x = 1.$$

Consider any continuous bounded function $\varphi(x) \in C_b(\mathbb{R})$. For all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\begin{aligned} \Big|\int_{\mathbb{R}} \varphi(x)g_\alpha(x)\,\mathrm{d}x - \varphi(0)\Big| &\leq \int_{\mathbb{R}} g_\alpha(x)\big|\varphi(x)-\varphi(0)\big|\,\mathrm{d}x \\ &\leq \varepsilon\int_{-\delta}^{\delta} g_\alpha(x)\,\mathrm{d}x + \int_{-\infty}^{-\delta} \alpha e^{-\alpha\delta}\big|\varphi(x)-\varphi(0)\big|\,\mathrm{d}x \\ &\quad + \int_{\delta}^{\infty} \frac{\alpha}{1+e^{\alpha\delta}}\big|\varphi(x)-\varphi(0)\big|\,\mathrm{d}x \\ &\leq \varepsilon + \int_{-\infty}^{-\delta} \alpha e^{-\alpha\delta}\big|\varphi(x)-\varphi(0)\big|\,\mathrm{d}x + \int_{\delta}^{\infty} \frac{\alpha}{1+e^{\alpha\delta}}\big|\varphi(x)-\varphi(0)\big|\,\mathrm{d}x. \end{aligned}$$

Taking the limit as $\alpha \to \infty$, the last two integrals go to 0. Hence the limit of the left-hand side is less then ε for any $\varepsilon > 0$, so we are done. □

We now move on to studying the dynamics of the maximum of a family of continuous semimartingales driven by a common Brownian motion. Note that we specifically deal with semimartingales which have absolutely continuous finite variation part, which implies that their local times have a bicontinuous modification in $t \in \mathbb{R}^+$ and $a \in \mathbb{R}$ (see Definition 4.2.4). This is the case for all stochastic processes which can be written as the solution of an Itô SDE with integrable drift and stochastic term driven by a semimartingale with absolutely continuous finite variation.

Consider a family of $\mathbb{R}$-valued continuous semimartingales $\mathbf{X}_t = \{X_t^i\}_{i=0}^n$ with dynamics

$$\mathrm{d}X_t^i = b_t^i \,\mathrm{d}t + \sigma_t^i \,\mathrm{d}B_t, \tag{A.1}$$

where b_t^i and σ_t^i are (real, predictable, stochastically integrable) drift and diffusion coefficients for all $i = 0, \ldots, n$, and B_t is a standard Brownian motion.

We apply Itô's Lemma to derive the dynamics of $LSE_\alpha(\boldsymbol{X}_\cdot)(t)$ as

$$\begin{aligned}
&\mathrm{d}LSE_\alpha(\boldsymbol{X}_\cdot)(t) \\
&= \sum_{i=0}^n \frac{e^{\alpha X_t^i}}{\sum_k e^{\alpha X_t^k}} \,\mathrm{d}X_t^i + \frac{1}{2}\sum_{i=0}^n \sum_{j=0}^n \alpha e^{\alpha X_t^i} \left(\frac{\delta_{ij}}{\sum_k e^{\alpha X_t^k}} - \frac{e^{\alpha X_t^j}}{(\sum_k e^{\alpha X_t^k})^2} \right) \mathrm{d}\langle X_\cdot^i, X_\cdot^j \rangle_t \\
&= \sum_{i=0}^n \frac{e^{\alpha X_t^i}}{\sum_k e^{\alpha X_t^k}} b_t^i \,\mathrm{d}t + \sum_{i=0}^n \frac{e^{\alpha X_t^i}}{\sum_k e^{\alpha X_t^k}} \sigma_t^i \,\mathrm{d}B_t + \frac{1}{2} f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(t) \,\mathrm{d}t,
\end{aligned} \tag{A.2}$$

where we have written δ_{ij} for the Kronecker delta and defined the function

$$f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(t) := \alpha \sum_i \sum_{j \neq i} \frac{e^{\alpha(X_t^i + X_t^j)}}{\left(\sum_k e^{\alpha X_t^k}\right)^2} \left((\sigma_t^i)^2 - \sigma_t^i \sigma_t^j \right). \tag{A.3}$$

We rewrite (A.2) in integral form as follows, for all $s \leq t$,

$$\begin{aligned}
LSE_\alpha(\boldsymbol{X}_\cdot)(t) = LSE_\alpha(\boldsymbol{X}_\cdot)(s) &+ \int_s^t S_\alpha^{arg}(\mathbf{X}_\cdot, \mathbf{b}_\cdot)(r) \,\mathrm{d}r + \int_s^t S_\alpha^{arg}(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(r) \,\mathrm{d}B_r \\
&+ \frac{1}{2}\int_s^t f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(r) \,\mathrm{d}r.
\end{aligned} \tag{A.4}$$

We are interested in the limit of the above when we send α to infinity. For each time t, define the argmax of $\mathbf{X}_t$ by $\mathcal{I}_t = \{j \in \mathbf{N} : X_t^j \geq X_t^i \ \forall i \in \mathbf{N}\}$. Since $S_\alpha^{arg}(\mathbf{X}_\cdot, \mathbf{b}_\cdot)(r) \leq \max_i b_t^i$, and b^i is integrable for all i by assumption, we can apply dominated convergence to yield

$$\lim_{\alpha \to \infty} \int_0^t S_\alpha^{arg}(\mathbf{X}_\cdot, \mathbf{b}_\cdot)(s) \,\mathrm{d}s = \int_0^t \frac{1}{|\mathcal{I}_s|} \sum_{j \in \mathcal{I}_s} b_s^j \,\mathrm{d}s.$$

Similarly, $\max_i \sigma_t^i$ is integrable against B_t, so we can apply dominated convergence for stochastic integrals and get

$$\lim_{\alpha\to\infty} \int_s^t S_\alpha^{arg}(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(r)\, \mathrm{d}B_r = \int_s^t \frac{1}{|\mathcal{I}_r|} \sum_{j\in\mathcal{I}_r} \sigma_r^j \, \mathrm{d}B_r.$$

The last integral on the right hand side of (A.4) is trickier to deal with.

Proposition A.1.6. *Consider a family of continuous semimartingales $\mathbf{X}_t = \{X_t^i\}_{i=0}^n$ with dynamics given by* (A.1)*. Let f_α be defined as in* (A.3)*. Then for all $s \le t$*

$$\lim_{\alpha\to\infty} \int_s^t f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(r)\, \mathrm{d}r \le \sum_i \sum_{j>i} \Big(L_t^0(X_\cdot^i - X_\cdot^j) - L_s^0(X_\cdot^i - X_\cdot^j) \Big) \quad a.s.$$

Proof. Exploiting symmetry, we start by rewriting $f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(t)$ as

$$f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(t) = \frac{1}{2}\alpha \sum_i \sum_{j\neq i} \frac{e^{\alpha(X_t^i + X_t^j)}}{\big(\sum_k e^{\alpha X_t^k}\big)^2} \big(\sigma_t^i - \sigma_t^j\big)^2,$$

and hence note that the last integral on the right-hand side of (A.4) is always positive. Moreover, with g_α as in Lemma A.1.5,

$$\begin{aligned}
\frac{\alpha e^{\alpha(X_t^i + X_t^j)}}{\big(\sum_k e^{\alpha X_t^k}\big)^2} &= \frac{\alpha}{2 + e^{\alpha(X_t^i - X_t^j)} + e^{\alpha(X_t^j - X_t^i)} + \sum_{k\neq i,j} \sum_{l\neq i,j} e^{\alpha(X_t^k + X_t^l - X_t^i - X_t^j)}} \\
&\le \frac{\alpha e^{\alpha(X_t^i - X_t^j)}}{(1 + e^{\alpha(X_t^i - X_t^j)})^2} = g_\alpha(X_\cdot^i - X_\cdot^j)(t).
\end{aligned}$$

The occupation time formula (see e.g. [74, Chapter 6, Corollary 1.6]) yields

$$\begin{aligned}
\lim_{\alpha\to\infty} \int_s^t f_\alpha(\mathbf{X}_\cdot, \boldsymbol{\sigma}_\cdot)(r)\, \mathrm{d}r &\le \lim_{\alpha\to\infty} \frac{1}{2} \int_s^t \sum_i \sum_{j\neq i} g_\alpha(X_\cdot^i - X_\cdot^j)(r) \big(\sigma_r^i - \sigma_r^j\big)^2 \, \mathrm{d}r \\
&= \sum_i \sum_{j\neq i} \lim_{\alpha\to\infty} \frac{1}{2} \int_s^t g_\alpha(X_\cdot^i - X_\cdot^j)(r)\, \mathrm{d}\langle X_\cdot^i - X_\cdot^j \rangle_r \\
&= \sum_i \sum_{j\neq i} \lim_{\alpha\to\infty} \frac{1}{2} \int_{\mathbb{R}} g_\alpha(z) \Big(L_t^z(X_\cdot^i - X_\cdot^j) - L_s^z(X_\cdot^i - X_\cdot^j) \Big) \, \mathrm{d}z \\
&= \frac{1}{2} \sum_i \sum_{j\neq i} \Big(L_t^0(X_\cdot^i - X_\cdot^j) - L_s^0(X_\cdot^i - X_\cdot^j) \Big)
\end{aligned}$$

almost surely, where the final equality relied on the weak convergence of $g_\alpha(z)\, \mathrm{d}z$ to a Dirac mass at 0 by Lemma A.1.5. Note that the $g_\alpha(z)$ are not compactly supported (compare with Definition 4.2.4), but this is fine since the local time $L_t^z(X_\cdot^i - X_\cdot^j)$ is bounded in z a.s. (see Barlow and Yor [11, Corollary 5.2.2]). □

A.2 Numerical experiments

In this appendix we provide some details about the simulations for the plots in Figure 3.2. For the sake of comparison between the different dimensions, we give the rate matrix Q a fixed structure, and keep the contraction coefficient constant across dimensions.

For $n = 2, 20, 50, 100$, we take the signal process X to be a Markov chain on $n+1$ states $\{0, \dots, n\}$ such that if X is at state i at some time t, it will be equally likely to jump to state $i+1$ or $i-1$, while it will only jump to state $j \neq i \pm 1$ with much lower probability. In other words, the chain switches quickly between a state and its two closest neighbours, but it only mixes slowly with the states further away. We let the jump rate from state i grow with the dimension of the chain: for $n \geq 3$, we set the off-tridiagonal entries of $Q = (q_{ij})$ to be 1, the upper and lower diagonals to be $n+1$, and therefore the diagonal to be $-3n$, i.e.

$$(q_{ij}) = \begin{cases} n+1, & \text{if } j \equiv i \pm 1 \pmod{n}, \\ -3n, & \text{if } j = i, \\ 1, & \text{otherwise.} \end{cases}$$

For $n = 2$, we simply take Q to be the symmetric matrix with -2 on the diagonal and 1 in the other entries. By fixing Q this way for all n, we have that the contraction rate from Theorem 3.2.1 is $\lambda = 2$, and does not change across all dimensions.

The chain X has uniform stationary distribution, denoted by $\mu = (\frac{1}{n}, \dots, \frac{1}{n})$. This is the point at the centre of the probability simplex $\mathcal{S}^n$. We take $\text{law}(X_0) = \mu$. Finally, we set the sensor function $h \in \mathbb{R}^{n+1}$ to be a randomly generated vector such that, for each $i \in \mathbf{N}$, $h^i = z_i + x_i$, where z_i is a random integer in $\{-10, \dots, 10\}$, and x_i is a realization of a uniform random variable in $[0, 1]$.

The initial condition for the optimal filter π_t is $\pi_0 = \mu$. The 'wrong' Wonham filter $\tilde{\pi}_t$ is initialized at $\nu \neq \mu$: to determine ν, we perturb μ by adding/subtracting $\frac{1}{2} \min_i \mu_i$ from all the components of μ according to $n+1$ independent Bernoulli random variables, and renormalizing.

Having fixed all these parameters, we generate 300 sample paths for the signal and the observation processes, and compute the optimal and 'wrong' Wonham filters by solving the Zakai equation (see e.g. [10, Remark 3.26]) with a simple Euler scheme and renormalizing after each step. We plot the realizations of the Hilbert error $\mathcal{H}(\pi_t, \tilde{\pi}_t)$, together with the bounds from Theorem 3.2.1 and Proposition 3.3.11. Note that the bounds from Proposition 3.3.11 are path-dependent (to compute them we need to observe $\tilde{\pi}_t$), so for each realization of the Hilbert error

we have corresponding realizations of the bounds from Proposition 3.3.11. They are also expressed as bounds for $\tanh(\mathcal{H}(\pi_t, \tilde{\pi}_t)/4)$ (as opposed to $\mathcal{H}(\pi_t, \tilde{\pi}_t)$). Taking arctanh on both sides of (3.25), and multiplying by 4, yields that $\mathcal{H}(\pi_t, \tilde{\pi}_t) \leq 4 \operatorname{arctanh}(u_t)$, where u_t solves (3.23); given the potential for the dynamics of u to have very large Lipschitz coefficients, we use a tamed Euler scheme (see e.g. Hutzenthaler, Jentzen and Kloeden [48]) to solve the ODE numerically. Concavity and monotonicity of tanh yield $\mathcal{H}(\pi_t, \tilde{\pi}_t) \leq \mathcal{H}(\mu, \nu) e^{-\int_0^t \tilde{\lambda}_s \, \mathrm{d}s}$ from (3.27); we compute $\tilde{\lambda}_t = 2 \min_{i \neq k} \left(q_{ik} q_{ki} + \sum_{j \neq i,k} \tilde{\pi}_t^j \min\{ q_{ji} q_{ik} / \tilde{\pi}_t^k, q_{jk} q_{ki} / \tilde{\pi}_t^i \} \right)^{1/2}$ at each timestep and perform numerical integration to plot the bound.

Bibliography

[1] S. Amari. Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.*, 10(2):357–385, 1982.

[2] S. Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.

[3] S. Amari. *Information geometry and its applications*, volume 194 of *Applied Mathematical Sciences*. Springer, Tokyo, 2016.

[4] J. Armstrong and D. Brigo. Nonlinear filtering via stochastic PDE projection on mixture manifolds in L^2 direct metric. *Mathematics of Control, Signals, and Systems*, 28(1):5, 2016.

[5] J. Armstrong and D. Brigo. Intrinsic stochastic differential equations as jets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2210):20170559, 2018.

[6] J. Armstrong, D. Brigo, and E. Rossi Ferrucci. Optimal approximation of SDEs on submanifolds: the Itô-vector and Itô-jet projections. *Proceedings of the London Mathematical Society*, 119(1):176–213, 2019.

[7] J. Armstrong, D. Brigo, and E. Rossi Ferrucci. Projections of sdes onto submanifolds. *Information Geometry*, pages 1–31, 2023.

[8] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.*, 33(6):697–725, 1997.

[9] R. Atar and O. Zeitouni. Lyapunov exponents for finite state nonlinear filtering. *SIAM J. Control Optim.*, 35(1):36–55, 1997.

[10] A. Bain and D. Crisan. *Fundamentals of stochastic filtering*, volume 60 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2009.

[11] M. T. Barlow and M. Yor. Semi-martingale inequalities via the Garsia-Rodemich-Rumsey lemma, and applications to local times. *Journal of Functional Analysis*, 49(2):198–229, 1982.

[12] P. Baxendale, P. Chigansky, and R. Liptser. Asymptotic stability of the Wonham filter: ergodic and nonergodic signals. *SIAM J. Control Optim.*, 43(2):643–669, 2004.

[13] Y. Benoist. Convexes hyperboliques et fonctions quasisymétriques. *Publ. Math. Inst. Hautes Études Sci.*, (97):181–237, 2003.

[14] G. Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85:219–227, 1957.

[15] G. Birkhoff. *Lattice theory*. American Mathematical Society Colloquium Publications, Vol. XXV. American Mathematical Society, Providence, R.I., third edition, 1967.

[16] V. I. Bogachev. *Measure theory. Vol. I, II*. Springer-Verlag, Berlin, 2007.

[17] P. Brémaud. *Discrete probability models and methods*. Springer, 2017.

[18] D. Brigo, B. Hanzon, and F. Le Gland. A differential geometric approach to nonlinear filtering: the projection filter. Research Report 2598, INRIA, june 1995. hal-02101519.

[19] D. Brigo, B. Hanzon, and F. Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5(3):495–534, 1999.

[20] A. Budhiraja. Asymptotic stability, ergodicity and other asymptotic properties of the nonlinear filter. *Ann. Inst. H. Poincaré Probab. Statist.*, 39(6):919–941, 2003.

[21] A. Budhiraja and H. J. Kushner. Robustness of nonlinear filters over the infinite time interval. *SIAM J. Control Optim.*, 36(5):1618–1637, 1998.

[22] P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Arch. Rational Mech. Anal.*, 52:330–338, 1973.

[23] Y. Chen, T. Georgiou, and M. Pavon. Entropic and displacement interpolation: a computational approach using the Hilbert metric. *SIAM J. Appl. Math.*, 76(6):2375–2396, 2016.

[24] P. Chigansky. Stability of nonlinear filters: A survey. *Lecture notes, Petropolis, Brazil*, 2006.

[25] P. Chigansky, R. Liptser, and R. Van Handel. Intrinsic methods in filter stability. In *The Oxford handbook of nonlinear filtering*, pages 319–351. Oxford University Press, Oxford, 2011.

[26] P. Chigansky and R. Van Handel. Model robustness of finite state nonlinear filtering over the infinite time horizon. *Ann. Appl. Probab.*, 17(2):688–715, 2007.

[27] Ş. Cobzaş, R. Miculescu, and A. Nicolae. *Lipschitz functions*, volume 2241 of *Lecture Notes in Mathematics*. Springer, Cham, 2019.

[28] S. N. Cohen and R. J. Elliott. *Stochastic calculus and applications*. Probability and its Applications. Springer, Cham, second edition, 2015.

[29] S. N. Cohen and E. Fausti. Exponential contractions and robustness for approximate Wonham filters. *arXiv:2305.02256*, 2023.

[30] S. N. Cohen and E. Fausti. Hyperbolic contractivity and the hilbert metric on probability measures. *arXiv:2309.02413*, 2023.

[31] J. B. Conway. *A course in functional analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1990.

[32] D. Crisan. The stochastic filtering problem: a brief historical account. *Journal of Applied Probability*, 51(A):13–22, 2014.

[33] D. Crisan and B. Rozovskiĭ, editors. *The Oxford handbook of nonlinear filtering*. Oxford University Press, Oxford, 2011.

[34] P. de la Harpe. On Hilbert's metric for simplices. In *Geometric group theory, Vol. 1 (Sussex, 1991)*, volume 181 of *London Math. Soc. Lecture Note Ser.*, pages 97–119. Cambridge Univ. Press, Cambridge, 1993.

[35] B. Delyon and O. Zeitouni. Lyapunov exponents for filtering problems. In *Applied stochastic analysis (London, 1989)*, volume 5 of *Stochastics Monogr.*, pages 511–521. Gordon and Breach, New York, 1991.

[36] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2001.

[37] L. Dubois. Projective metrics and contraction principles for complex cones. *J. Lond. Math. Soc. (2)*, 79(3):719–737, 2009.

[38] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication.

[39] M. Emery. *Stochastic Calculus in Manifolds*. Universitext. Springer-Verlag Berlin Heidelberg, 1 edition, 1989.

[40] S.N. Ethier and T.G. Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[41] G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[42] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra Appl.*, 114/115:717–735, 1989.

[43] B. Hanzon. A differential-geometric approach to approximate nonlinear filtering. *Geometrization of statistical theory*, pages 219–223, 1987.

[44] M. Hazewinkel, S. I. Marcus, and H. J. Sussmann. Nonexistence of finite-dimensional filters for conditional statistics of the cubic sensor problem. *Systems & control letters*, 3(6):331–340, 1983.

[45] D. Hilbert. Ueber die gerade linie als kürzeste verbindung zweier punkte. *Mathematische Annalen*, (46):91–96, 1895.

[46] E. Hopf. An inequality for positive linear integral operators. *J. Math. Mech.*, 12:683–692, 1963.

[47] E. P. Hsu. *Stochastic analysis on manifolds*, volume 38 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[48] M. Hutzenthaler, A. Jentzen, and P. E. Kloeden. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4):1611–1641, 2012.

[49] J. Jacod and A.N. Shiryaev. *Limit theorems for stochastic processes*. Springer, 2nd edition, 2013.

[50] R. Jentzsch. Über Integralgleichungen mit positivem Kern. *J. Reine Angew. Math.*, 141:235–244, 1912.

[51] S. J. Julier and J. K. Uhlmann. New extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI*, volume 3068, pages 182–193. International Society for Optics and Photonics, 1997.

[52] O. Kallenberg. *Foundations of modern probability*. Springer, 1997.

[53] R. E. Kálmán. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[54] R. E. Kálmán and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of basic Engineering*, 83:95–108, 1961.

[55] J. W. Kim and P. G. Mehta. A dual characterization of the stability of the wonham filter. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1621–1628. IEEE, 2021.

[56] J. W. Kim, P. G Mehta, and S. Meyn. The conditional poincaré inequality for filter stability. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1629–1636. IEEE, 2021.

[57] E. Kohlberg and J. W. Pratt. The contraction mapping approach to the Perron-Frobenius theory: why Hilbert's metric? *Math. Oper. Res.*, 7(2):198–210, 1982.

[58] M. G. Kreĭn and M. A. Rutman. Linear operators leaving invariant a cone in a Banach space. *Uspehi Matem. Nauk (N.S.)*, 3(1(23)):3–95, 1948.

[59] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of markov processes. *Journal of Multivariate Analysis*, 1(4):365–393, 1971.

[60] H. J. Kushner. On the differential equations satisfied by conditional probablitity densities of Markov processes, with applications. *J. SIAM Control Ser. A*, 2:106–119, 1964.

[61] F. Le Gland and L. Mevel. Basic properties of the projective product with application to products of column-allowable nonnegative matrices. *Math. Control Signals Systems*, 13(1):41–62, 2000.

[62] F. Le Gland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems*, 13(1):63–93, 2000.

[63] F. Le Gland and N. Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Ann. Appl. Probab.*, 14(1):144–187, 2004.

[64] J. M. Lee. *Introduction to Riemannian manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, 2018. Second edition of [MR1468735].

[65] B. Lemmens and R. Nussbaum. *Nonlinear Perron-Frobenius theory*, volume 189 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2012.

[66] R. S. Liptser and A. N. Shiryayev. *Statistics of random processes. I.* Applications of Mathematics, Vol. 5. Springer-Verlag, New York-Heidelberg, 1977. General theory, Translated by A. B. Aries.

[67] P. S. Maybeck. *Stochastic models, estimation, and control*. Academic press, 1982.

[68] L. A. McGee, S. F. Schmidt, and G. L. Smith. Applications of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle. *NASA Technical Report R-135, Tech. Rep*, 1962.

[69] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Cambridge, 2nd edition, 2009.

[70] F. Nielsen and K. Sun. Non-linear embeddings in hilbert simplex geometry. *arXiv preprint arXiv:2203.11434*, 2022.

[71] D. Ocone and E. Pardoux. Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM Journal on Control and Optimization*, 34(1):226–243, 1996.

[72] O. Perron. Zur Theorie der Matrices. *Math. Ann.*, 64(2):248–263, 1907.

[73] B. B. Phadke. A triangular world with hexagonal circles. *Geometriae Dedicata*, 3:511–520, 1974/75.

[74] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[75] E. Rossi Ferrucci. *Rough path perspectives on the Itô-Stratonovich dilemma*. PhD thesis, Imperial College London, 2021.

[76] H. H. Rugh. Cones and gauges in complex spaces: spectral gaps and complex Perron-Frobenius theory. *Ann. of Math. (2)*, 171(3):1707–1752, 2010.

[77] H. Samelson. On the Perron-Frobenius theorem. *Michigan Math. J.*, 4:57–59, 1957.

[78] E. Seneta. *Non-negative matrices and Markov chains*. Springer Series in Statistics. Springer, New York, 2006. Revised reprint of the second (1981) edition [Springer-Verlag, New York; MR0719544].

[79] E. Seneta and S. Sheridan. Strong ergodicity of nonnegative matrix products. *Linear Algebra Appl.*, 37:277–292, 1981.

[80] R. L. Stratonovič. Conditional Markov processes. *Teor. Verojatnost. i Primenen.*, 5:172–195, 1960.

[81] G. Teschl. *Ordinary differential equations and dynamical systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[82] A. C. Thompson. On certain contraction mappings in a partially ordered vector space. *Proc. Amer. Math. Soc.*, 14:438–443, 1963.

[83] R. Van Handel. *Filtering, stability, and robustness*. PhD thesis, California Institute of Technology, 2007.

[84] R. Van Handel. Observability and nonlinear filtering. *Probab. Theory Related Fields*, 145(1-2):35–74, 2009.

[85] R. Van Handel. The stability of conditional Markov processes and Markov chains in random environments. *Ann. Probab.*, 37(5):1876–1925, 2009.

[86] R. Van Handel. Uniform observability of hidden Markov models and filter stability for unstable signals. *Ann. Appl. Probab.*, 19(3):1172–1199, 2009.

[87] W. M. Wonham. Some applications of stochastic differential equations to optimal nonlinear filtering. *J. SIAM Control Ser. A*, 2:347–369 (1965), 1965.

[88] F. Zhang. *Matrix theory*. Universitext. Springer, New York, second edition, 2011. Basic results and techniques.