DESIGN AND ANALYSIS OF PARTICLE-BASED ALGORITHMS FOR NONLINEAR FILTERING AND SAMPLING

BY

AMIRHOSSEIN TAGHVAEI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mechanical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Associate Professor Prashant G. Mehta, Chair and Director of Research
Professor Bruce Hajek
Professor Richard S. Laugesen
Associate Professor Maxim Raginsky

Abstract

This thesis is concerned with the design and analysis of particle-based algorithms for two problems: (i) the nonlinear filtering problem; (ii) and the problem of sampling from a target distribution. The contributions for these two problems appear in Part I and Part II of the thesis.

For the nonlinear filtering problem, the focus is on the feedback particle filter (FPF) algorithm. In the FPF algorithm, the empirical distribution of the particles is used to approximate the posterior distribution of the nonlinear filter. The particle update is implemented using a feedback control law that is designed such that the distribution of the particles, in the mean-field limit ($N = \infty$), is exactly equal to the posterior distribution. In Part I of this thesis, three separate problems related to the FPF methodology and algorithm are addressed.

The first problem, addressed in Chapter 2 of the thesis, is concerned with gain function approximation in the FPF algorithm. The exact gain function is the solution of a Poisson equation involving a probability-weighted Laplacian. The numerical problem is to approximate this solution using *only* particles sampled from the probability distribution. A diffusion map-based algorithm is presented for this problem. The algorithm is based on a reformulation of the Poisson equation as a fixed-point equation that involves the diffusion semigroup corresponding to the weighted Laplacian. The fixed-point problem is approximated with a finite-dimensional problem in two steps: In the first step, the semigroup is approximated with a Markov operator referred to as diffusion map. In the second step, the diffusion map is approximated empirically, using particles, as a Markov matrix. A procedure for carrying out error analysis of the approximation is introduced and certain asymptotic estimates for bias and variance error are derived. Some comparative numerical experiments are performed for a problem with non-Gaussian distribution. The algorithm is applied and illustrated for a numerical filtering example.

As part of the error analysis, some new results about the diffusion map approximation are obtained. These include (i) new error estimates between the diffusion map and the exact semigroup, based on the Feynman-Kac representation of the semigroup; (ii) a spectral gap for the diffusion map, based on the Foster-Lyapunov function method from the theory of stability of Markov processes; (ii) and error estimates for the empirical approximation of the diffusion map.

The second problem, addressed in Chapter 3 of the thesis, is motivated by the so-called uniqueness issue of FPF control law. The control law in FPF is designed such that the distribution of the particles, in the mean-field limit, is exactly equal to the posterior distribution. However, it has been noted in the literature that the FPF control law is not unique. The objective of this research is to describe a systematic approach

to obtain a unique control law for FPF. In Chapter 3, the optimality criteria from optimal transportation theory is used to identify a unique control law for FPF, in the linear Gaussian setting. Two approaches are outlined. The first approach is based on a time-stepping optimization procedure. We consider a time discretization of the problem, construct a discrete-time stochastic process by composition of sequence of optimal transport maps between the posterior distributions of two consecutive time instants, and then take the limit as the time step-size goes to zero. We obtain explicit formula for the resulting control law in the linear Gaussian setting. The control law is deterministic and requires the covariance matrix of the resulting stochastic process to be invertible. We present an alternative approach, which allows for singular covariance matrices. The resulting control law has additional stochastic terms, which vanish when the covariance matrix is non-singular. The second construction is important for finite-N implementation of the algorithm, where the empirical covariance matrix might be singular.

The third problem, addressed in Chapter 4, is concerned with the convergence and the error analysis for FPF algorithm. It is known that in the mean-field limit, the distribution of the particles is equal to the posterior distribution. However little is known about the convergence of the finite-N algorithm to the mean-field limit. We consider the linear Gaussian setting, and study two types of FPF algorithm: The deterministic linear FPF and the stochastic linear FPF. The important question in the linear Gaussian setting is about convergence of the empirical mean and covariance of the particles to the exact mean and covariance given by the Kalman filter. We derive equations for the evolution of empirical mean and covariance for the finite-N system for both algorithms. Remarkably, for the deterministic linear FPF, the equations for the mean and variance are identical to the Kalman filter. This allows strong conclusions on convergence and error properties under the assumption that the linear system is controllable and observable. It is shown that the error converges to zero even with finite number of particles. For the stochastic linear FPF, the equations for the empirical mean and covariance involve additional stochastic terms. Error estimates are obtained for the empirical mean and covariance under the stronger assumption that the linear system is stable and fully observable. We also presents propagation of chaos error analysis for both algorithms.

The Part II of the thesis is concerned with the sampling problem, where the objective is to sample from a unnormalized target probability distribution. The problem is formulated as an optimization problem on the space of probability distributions, where the objective is to minimize the relative entropy with respect to the target distribution. The gradient flow with respect to the Riemannian metric induced by the Wasserstein distance, is known to lead to Markov Chain Monte-Carlo (MCMC) algorithms based on the Langevin equation. The main contribution is to present a methodology and numerical algorithms for constructing accelerated gradient flows on the space of probability distributions. In particular, the recent variational formulation of accelerated methods in [Wibisono et al., 2016] is extended from vector valued variables to probability distributions. The variational problem is modeled as a mean-field optimal control problem. The maximum principle of optimal control theory is used to derive Hamilton's equations for the optimal gradient flow. A quantitative estimate on the asymptotic convergence rate is provided based on a Lyapunov function construction, when the objective functional is displacement convex. Two numerical approximations are presented to implement the Hamilton's equations as a system of N interacting particles. The algorithm is numerically illustrated and compared with the MCMC and Hamiltonian MCMC algorithms.

Acknowledgements

My objective, in this section, is to express my gratitude toward the people who helped me throughout my PhD experience and who were most influential in the completion of this PhD thesis.

Indeed, a huge portion of this gratitude belongs to my PhD advisor Prashant Mehta. This gratitude is not only because of his contribution to this thesis, but also for his significant role in me becoming "a researcher with my own taste". I am grateful to Prashant for the following four main reasons: (i) He provided me with a continuous and reliable source of funding, which is the basic concern of any international student, and supported me to attend all sorts of workshops and conferences (18 total); (ii) He was always available for questions, consult and advice about research, academia and life. Basically, his office door was literally and metaphorically open; (iii) He gave me the freedom to pursue my own research interests, and guided me with his particular vision and intuition; (iv) And he conveyed his passion to learn mathematics to me by still reading textbooks, attending classes, and making statements which I quote from him here: "Mathematics is nice, even if it does not work. Also true with love", "Work on a nice problem and let the math guide you, don't be afraid" and a quote he shared with me, which works well after receiving bad reviews "... the only thing we can totally control, and therefore the only thing we should ever worry about, is our own judgment about what is good."

My second portion of gratitude goes to my PhD examination committee: Prof. Sean Meyn, Prof. Bruce Hajek, Prof. Richard Laugesen, and Prof. Maxim Raginsky. Specially, I am thankful to Sean Meyn, whom I had the honor to work with, because of his contribution to main results in Chapter 2 of this thesis. I am also grateful to Sean for inviting me and giving me the opportunity to present my work at the 5th and 6th workshop on cognition and control at University of Florida. His words were always encouraging and boosted my confidence. I am also thankful to Prof. Bruce Hajek, because of his brilliant teaching of Random processes, that made me understand I am at the right place in my first semester as graduate srtudnet. I am also honored to be his teacher assistant for statistical learning. His passion and care for CSL community was always inspiring. I am also thankful to Prof. Richard Laugesen for his amazing lecture notes on linear pdes and his class on spectral analysis of pdes. I am also grateful to Prof. Maxim Raginsky for his insightful comments and feedback during the PhD exam, and about the writing of the thesis. I also appreciate Prof. Matus Telgarsky for being in my preliminary exam committee, for his insightful class on Machine learning theory, and our friendly conversations.

I also like to thank my other collaborators, Prof. Sebastian Reich for his insights and his offer to visit him at University of Potsdam, which unfortunately did not happen; my labmate Chi Zhang, for our fruitful

collaborations, friendship, and letting me stay at his place whenever I visit San Diego; my labmate Jin Kim, for his extremely helpful feedback, comments, and finding errors in my writings; and Jana de Wiljes for our collaboration. I am also grateful to Amin Jalali for being my supervisor during my internship at Technicolor AI lab, that turned out to be influential in my future research plans. I like to thank Prof. Krishnaprasad and Simone Surace for their insightful feedback and comments.

I was lucky to spend most of this past six years in an amazing building called coordinated science laboratory (CSL). I enjoyed vast variety of talks happening weekly at CSL, the annual CSL student conference and the CSL social hour, which I proudly took part in organizing them, and all the great faculty members, post-docs, and PhD students that I got to know by being in CSL. I am thankful to our CSL staff, specially our office manager Angie (Angelia Ellis), for her help with travel arrangements and dedication to work. I also like to thank administration and staff in Department of Mechanical Science and Engineering, and specially Kathy Smith, for being patient in answering all my questions. I also acknowledge the Computational Science and Engineering fellowship that supported my research in 2016-2017.

I am also thankful to all my colleagues whom I shared office with at CSL 348. I thank Sahand Hariri for helping me settle in as I began graduate school and saving me from harsh winter of 2014. I thank Adam Tilton, for giving me the excitement of working in a start-up, and his support in my job search. I thank Rohan Arora, for all his help, care, and sharing his lunchtime with me. I thank Mayank Baranwal for being an amazing classmate, helpful college, and his gift after he graduated. I thank Ashok Makkuva, for our literature discussions and our random walks around CSL. I was happy to spend my last year with my new labmates, Heng-Sheng Chang and Tixian Wang, and wish them the best for their future. I am also extremely thankful to all my great friends outside CSL that I shared very enjoyable moments with them during past six years, whom I can not name all here.

I also have two uncommon but important thank you statements. I am grateful to my first qualification exam committee that failed my exam, and made me question the problem that I was working on, and made me realize that I should and I can work on a problem that I am really passionate about. And I thank Cedric Villani for his amazing book topics of optimal transportation, which was given to me by my advisor at this transition period, and shaped my research career.

Last but not the least, I would like to thank my family: my lovely parents, who extended their unconditional love and did their best for me to get where I am, my two brilliant brothers, Mani and Meysam, who were always supportive and encouraging, and my aunt, who mailed me packages full of love during past six years so that I do not miss home.

Table of Contents

Part I

Nonlinear filtering

Chapter 1

Introduction

The Part I of this thesis concerns a class of particle-based algorithms to approximate the solution of the nonlinear filtering problem. Although, in this thesis, we consider the filtering problem in continuous-time setting, we briefly discuss the filtering problem in discrete-time setting below. The purpose is to describe the relevant particle-based algorithms, and to motivate the questions we seek to answer in this thesis by making analogy to their discrete-time counterparts.

For the filtering problem in discrete-time setting, the main task that a particle-based algorithm performs is to convert a sample of N particles $\{X_k^i\}_{i=1}^N$ from the filter distribution π_k at time t_k to a sample of N particles $\{X_{k+1}^i\}_{i=1}^N$ from the filter distribution π_{k+1} at time t_{k+1}, without having access to the explicit form of the distributions. The filtering distributions satisfy the recursion $\pi_{k+1}(x) = \gamma \pi_k(x) l_k(x)$ where $l_k(x)$ is the known likelihood function, and γ is the normalization constant. The normalization constant is assumed to be unknown, and it is difficult to compute it in practice. In general, the recursion formula also involves the effect due to system dynamics, which is ignored here, as it is not challenging to numerically implement it, and it is not necessary for the purpose of this discussion. The particle-based algorithms are designed to carry out this task, in an online fashion, whenever they receive a new measurement.

The task of converting samples from the filter distribution π_k to π_{k+1} can be viewed as the problem of finding a coupling $\gamma(\cdot,\cdot)$ between the distributions π_k and π_{k+1} [Moral, 2004, Cheng and Reich, 2013]. A coupling can be expressed according to $\gamma(x,x') = T_k(x|x')\pi_k(x')$ where $T_k(\cdot|\cdot)$ is the transition kernel. The transition kernel satisfies the identities $\pi_{k+1}(x) = \int T_k(x|x')\pi_k(x')\mathrm{d}x$ and $\int T_k(x|x')\mathrm{d}x = 1$. Once the coupling is at hand, new samples are simply generated using the transition kernel, i.e. $X_{k+1}^i \sim T_k(\cdot|X_k^i)$. Given this viewpoint, the particle-based algorithms simulate the following stochastic update law for the system of particles:

$$X_{k+1}^i \sim T_k(\cdot|X_k^i) \tag{1.1}$$

There are infinitely many ways to couple two distributions. The simplest one is an independent coupling where $T_k(x|x') = \pi_{k+1}(x)$. However, the explicit form π_{k+1} is unknown. Sequential importance resampling (SIR) particle filters [Gordon et al., 1993, Doucet, 2009] numerically implement the independent coupling as follows: First, a weighted distribution of the particles is formed according to $\sum_{i=1}^N w_i \delta_{X_k^i}$ where the weights $w_i = \frac{l_k(X_k^i)}{\sum_{j=1}^N l_k(X_k^j)}$ and δ_x is the Dirac distribution located at x. The weighted distribution forms an approximation of the true filter distribution π_{k+1}. This step is called importance sampling. Then, N particles are independently sampled from the weighted distribution, i.e $X_{k+1}^i \sim \sum_{i=1}^N w_i \delta_{X_k^i}$, by sampling from a multinomial distribution with parameter vector $(N, \{w_i\}_{i=1}^N)$. This step is called resampling.

The transition kernel in an independent coupling is completely stochastic. This has motivated application of other forms of couplings that are more deterministic and are optimal with respect certain optimality criteria [Del Moral, 2004, Bain and Crisan, 2009]. An important example is to use optimal transportation theory to find an optimal coupling [Reich, 2011, Cheng and Reich, 2013, El Moselhy and Marzouk, 2012, Kim et al., 2013]. In the optimal transportation-based approach, one searches for deterministic transition kernels of the form $T_k(x|x') = \delta_{x=\nabla\Phi(x')}$ where the function Φ should be chosen such that $T_k(\cdot|\cdot)$ satisfies the consistency condition $\pi_{k+1}(x) = \int T_k(x|x')\pi_k(x')\mathrm{d}x'$. The resulting equation that Φ should satisfy is the Monge-Ampère pde [Evans, 1997, Villani, 2003]. A numerical approximation of this approach, based on the empirical distribution of particles, led to the development of ensemble transform particle filters [Cheng and Reich, 2013]. Another related approach is based on the coupling obtained from the Schrödinger bridge problem between π_k and π_{k+1} [Reich, 2018].

Based on the coupling of distributions viewpoint, there are three relevant and important questions that one would like to answer for any particle-based algorithm. The first question is about numerical approximation of the coupling. An exact coupling is impossible to compute, because the filter distributions π_k and π_{k+1} are not known. Only N independent samples from π_k and the likelihood function $l_k(\cdot)$ is available. So one can only hope for approximation of an exact coupling, with an approximation error that converges to zero as the number of samples increase. The second question is about the optimality measures that one should consider to obtain a coupling, among all the couplings that satisfy the consistency condition. And the third problem is about error analysis of the algorithm. The objective of the error analysis is to obtain bounds on the error, that is introduced because of the approximate coupling, and to study how the error propagates with time.

In this thesis, we consider the filtering problem in the continuous-time setting. In contrast to discrete-time setting, the filter distribution π_t continuously evolves with time, as a continuous stream of measurements arrives. The objective of particle-based algorithms, in the continuous-time setting, is to continuously update N particles $\{X_t^i\}_{i=1}^N$, such that they are distributed according to π_t. The feedback particle filter (FPF) algorithm [Yang et al., 2013, 2016] carries out this task, by updating the particles according to the (Stratonovich) stochastic differential equation:

$$\mathrm{d}X_t^i = \mathsf{K}_t(X_t^i) \circ (\mathrm{d}Z_t - \frac{h(X_t^i) + \hat{h}_t}{2}\mathrm{d}t) \tag{1.2}$$

where the vector-field $\mathsf{K}_t(\cdot)$ is referred to as gain function, Z_t is the observation process, $h(\cdot)$ is the observation function, and $\hat{h}_t$ is the expectation of h with respect to the filter distribution (detailed description of the terms appear in Sec. 1.2). The gain function is assumed to be of gradient form $\mathsf{K}_t(x) = \nabla\phi(x)$, where the function ϕ is the solution to a probability-weighted Poisson equation $\frac{1}{\rho_t}\nabla\cdot(\rho_t\nabla\phi) = \text{r.h.s}$, where ρ_t is the probability density function of particles in the mean-field limit $N = \infty$. In numerical implementation, the density ρ_t is not known, and the gain function should be approximated in terms of N particles.

A fascinating fact about FPF algorithm is that under certain approximation of the gain function, known as constant gain approximation, the FPF algorithm is similar to a classical algorithm called ensemble Kalman filter (EnKF) [Evensen, 1994, Reich, 2011, Yang et al., 2016]. EnKF is a popular choice in

high-dimensional application in geophysical sciences, because of its scalability with the problem dimension [Evensen, 1994, Reich, 2011]. However, EnKF does not provide a consistent approximation of the filter distribution for nonlinear systems and non-Gaussian distributions. FPF provides the generalization of EnKF for nonlinear problems, essentially by construction a better approximation of the gain function.

In this thesis, we seek to answer three important questions regarding the FPF algorithm, that are analogous to the three questions discussed above in discrete-time setting. The first question is about numerical approximations of the solution of the Poisson equation, when the probability density is not known, and only finite number of samples are given. This question is analogous to the question of finding a coupling based on finite number of samples in (1.1). Specially, the form of the Poisson equation is suggestive of the fact that it is analogous to the Monge-Ampére pde. The second question is about understanding the optimality properties of the FPF control law, and its relation to the optimal transport couplings discussed above. And the third question is about the error analysis of the FPF algorithm with finite number of particles. In particular, how does the error, introduced from numerical approximation of the Poisson equation, propagates with time.

A summary of contributions of Part I of this thesis is presented in Sec. 1.3, after introducing the filtering problem in Sec. 1.1, and the FPF algorithm in Sec. 1.2. The part II of the thesis, is presented in a self-contained manner in Chapter 5.

1.1 Stochastic filtering problem

Nonlinear filtering is a principled approach to extract useful information about the state of a dynamical system from noisy sensor measurements. It has wide range of engineering applications. For example, the problem of finding the state of the robot—its position, velocity, and orientation—based on the available data from sensors—camera, accelerator, and gyroscope—can be solved with a filtering approach [Montemerlo et al., 2002]. Other applications of the filtering include target tracking [Bar-Shalom et al., 2004], weather prediction [Chen and Majda, 2018], meteorological sciences [Reich and Cotter, 2015], GPS positioning [Gustafsson et al., 2002], and computer vision [Isard and Blake, 1998].

A mathematical formulation of the filtering problem consists of two stochastic processes: (i) A hidden Markov process that is used to model the state of a dynamical system; (ii) The observation process that is used to model the sensor information available from the dynamical system. The filtering problem is to compute the probability distribution of the hidden state, given the history of observations.

In this thesis, we consider the filtering problem in the setting of continuous-time dynamics and continuous-time observation. In this setting, the state and the observation process are denoted as $\{X_t;\ t \geq 0\}$ and $\{Z_t;\ t \geq 0\}$ respectively. These processes are modelled by the following Itô stochastic differential equations (sde):

$$\text{State process:} \quad \mathrm{d}X_t = a(X_t)\mathrm{d}t + \sigma_B(X_t)\mathrm{d}B_t, \quad X_0 \sim \pi_{\text{init}} \tag{1.3a}$$

$$\text{Observation process:} \quad \mathrm{d}Z_t = h(X_t)\mathrm{d}t + \mathrm{d}W_t, \tag{1.3b}$$

where $X_t \in \mathbb{R}^d$ is the (hidden) state at time t, $Z_t \in \mathbb{R}^p$ is the observation process at time t, and B_t, W_t are mutually independent standard Wiener processes taking values in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively. The mappings $a(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, $h(\cdot) : \mathbb{R}^d \to \mathbb{R}^p$, and $\sigma_B(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times q}$ are known continuously differentiable functions, and π_{init} is the prior probability distribution.

The objective in the filtering problem is to compute the conditional distribution of the state X_t, given the history of observations (filtration) $\mathscr{Z}_t := \sigma(Z_s : 0 \le s \le t)$. Let

$$\pi_t(\cdot) := \mathsf{P}(X_t \in \cdot | \mathscr{Z}_t)$$

denote the conditional probability distribution, also referred is posterior distribution. π_t is an optimal filter, in the sense that, for any integrable function f, $\pi_t(f) := \int f \mathrm{d}\pi_t$ provides the the best mean-squared error estimate of $f(X_t)$, among all $\mathscr{Z}_t$-measurable random variables.

The evolution of the posterior distribution π_t is given by the Kushner-Stratonovich equation [Xiong, 2008, Ch. 5]. In the special and important linear Gaussian setting, the equation admits a finite-dimensional closed-form solution given by the Kalman-Bucy filter. This solution is described next.

1.1.1 Linear Gaussian setting and Kalman-Bucy filter

The filtering problem (1.3a)-(1.3b) is linear Gaussian if $a(\cdot)$, and $h(\cdot)$ are linear functions, $\sigma_B(\cdot)$ is a constant function, and π_{init} is a Gaussian distribution. Under these assumptions, sdes (1.3a)-(1.3b) take the following form

$$\mathrm{d}X_t = AX_t + \sigma_B \mathrm{d}B_t, \quad X_0 \sim \mathscr{N}(m_{\text{init}}, \Sigma_{\text{init}}) \tag{1.4a}$$

$$\mathrm{d}Z_t = HX_t \mathrm{d}t + \mathrm{d}W_t \tag{1.4b}$$

where A, H, σ_B are matrices of appropriate dimensions.

In the linear-Gaussian setting, (X_t, Z_t) is a Gaussian process. Therefore, the conditional distribution of X_t is Gaussian, denoted as $\mathscr{N}(m_t, \Sigma_t)$, where m_t and Σ_t are conditional mean and covariance, respectively. Their evolution is described by the Kalman-Bucy filter [Kalman and Bucy, 1961]:

$$\mathrm{d}m_t = Am_t + \mathsf{K}_t(\mathrm{d}Z_t - Hm_t \mathrm{d}t), \quad m_0 = m_{\text{init}} \tag{1.5a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = \mathrm{Ricc}(\Sigma_t), \quad \Sigma_0 = \Sigma_{\text{init}} \tag{1.5b}$$

where $\mathsf{K}_t := \Sigma_t H^\top$ is referred to as Kalman gain, and the Riccati function

$$\mathrm{Ricc}(\Sigma) := A\Sigma + \Sigma A^\top + \Sigma_B - \Sigma H^\top H \Sigma \tag{1.6}$$

with $\Sigma_B := \sigma_B \sigma_B^\top$.

The Kalman filter is one of the most widely used algorithm in engineering. Although the filter describes the posterior only in linear Gaussian settings, it is often used as an approximate algorithm even in more

general settings, e.g., by defining the matrices A and H according to the Jacobians of the mappings a and h. The resulting algorithm is referred to as the extended Kalman filter (EKF).

Application of the Kalman filter and its extensions is limited because of the following two reasons:

1. Nonlinearities in the dynamics and in the observation models lead to a non-Gaussian multi-modal posterior probability distributions. For such cases, Kalman and extended Kalman filters are known to perform poorly [Gordon et al., 2004, Budhiraja et al., 2007].

2. Kalman filter becomes computationally expensive when the state-space dimension d is large. The reason is that the computational cost scales as $O(d^2)$, as it needs to store and propagate the $d \times d$ covariance matrix.

This has motivated development of the Monte-Carlo methods to approximate the filter, in nonlinear and high-dimensional setting. One such approach is the particle filter briefly reviewed next.

1.1.2 Importance sampling-based particle filter

Importance sampling-based particle filter is an example of a sequential Monte-Carlo algorithm to approximate the filter [Gordon et al., 1993, Doucet, 2009]. It is comprised of N particles denoted as $\{X_t^1, \ldots, X_t^N\}$, where $X_t^i \in \mathbb{R}^d$ denotes the state of the i-th particle at time t, and N importance weights $\{w_t^i\}_{i=1}^N$, where $w_t^i \geq 0$ is the weight corresponding to i-th particle. For the nonlinear filtering problem (1.3a)-(1.3b), the evolution of the particles and the weights is given by:

$$\begin{aligned} \mathrm{d}X_t^i &= a(X_t^i)\mathrm{d}t + \mathrm{d}B_t^i, \quad X_0^i \sim \pi_{\text{init}} \\ \mathrm{d}M_t^i &= M_t^i h(X_t^i)\mathrm{d}Z_t, \quad M_0^i = 1 \\ w_t^i &= \frac{M_t^i}{\sum_{j=1}^N M_t^j} \end{aligned}$$

where $M_t^i \in \mathbb{R}$ are referred to as the unnormalized importance weights. Particle filters approximate the posterior distribution via the weighted empirical distribution of particles

$$\pi_t(f) \approx \sum_{i=1}^N w_t^i f(X_t^i) \tag{1.7}$$

Thus, the particle filters implement the effect of conditioning (due the observations) by assigning importance weights. The issue with this approach is that, after a few time steps, there are only a few particles with significant weights, while most particle have nearly zero weights. This issue is referred to as the particle degeneracy [Ristic et al., 2004, Doucet, 2009]. To mitigate this issue, a resampling procedure is performed, where N independent particles are resampled from the weighted empirical distribution $\tilde{X}_t^i \sim \sum_{j=1}^N w_i \delta_{X_t^i}$. Theoretically, it is shown that the empirical approximation (1.7) becomes exact in the limit as $N \to \infty$ with error rate $O(N^{-1/2})$ [Del Moral and Guionnet, 2001, Cappé et al., 2009]. However, both empirically and theoretically, it was discovered that particle filters perform poorly in high dimensional problems. To maintain

the same amount of error, a particle filter is known to require a number of particles that scales exponentially with the dimension. This issue is referred to as the *curse of dimensionality* [Bengtsson et al., 2008, Beskos et al., 2014, Rebeschini and Van Handel, 2015].

In the next section, we describe a class of controlled interacting particle-based algorithms that do not involve importance sampling.

1.2 Feedback particle filter algorithm

Feedback particle filter (FPF) is a class of control interacting particle-based algorithms that are designed to approximate the solution to the nonlinear filtering prorblem [Yang et al., 2013, 2016]. Its construction is based on the following two steps:

Step 1: Construct a stochastic process, denoted by $\bar{X}_t \in \mathbb{R}^d$, whose distribution is equal to π_t, the posterior distribution of X_t;

Step 2: Simulate N stochastic processes, denoted by $\{X_t^i\}_{i=1}^N$, to empirically approximate the distribution of $\bar{X}_t$.

$$\underbrace{\pi_t(f) \overset{\text{Step 1}}{=} \mathsf{E}[f(\bar{X}_t)|\mathscr{Z}_t]}_{\text{exactness condition}} \overset{\text{Step 2}}{\approx} \frac{1}{N}\sum_{i=1}^N f(X_t^i)$$

The process $\bar{X}_t$ is referred to as mean-field process and the N processes $\{X_t^i\}_{i=1}^N$ are referred to as particles. We first present the mean-field process of FPF, and then discuss the particle-based simulation.

The mean-field process, for the filtering problem (1.3a)-(1.3b), evolves according to the sde given by

$$\text{FPF:} \quad \mathrm{d}\bar{X}_t = \underbrace{a(\bar{X}_t)\mathrm{d}t + \mathrm{d}\bar{B}_t}_{\text{propagation}} + \underbrace{\mathsf{K}_t(\bar{X}_t) \circ (\mathrm{d}Z_t - \frac{h(\bar{X}_t)+\hat{h}_t}{2}\mathrm{d}t)}_{\text{feedback control law}}, \quad \bar{X}_0 \sim \pi_{\text{init}} \tag{1.8}$$

where $\bar{B}_t$ is a standard Wiener processes independent of $\bar{X}_0$ and $\hat{h}_t := \mathsf{E}[h(\bar{X}_t)|\mathscr{Z}_t]$. The $\circ$ indicates that the sde is expressed in its Stratonovich form. The gain function $\mathsf{K}_t(x) := \nabla\phi_t(x)$ where ϕ_t is the solution of the Poisson equation:

$$\text{Poisson equation:} \quad -\frac{1}{\rho_t(x)}\nabla\cdot(\rho_t(x)\nabla\phi_t(x)) = -(h(x)-\hat{h}_t) \quad \forall\, x \in \mathbb{R}^d \tag{1.9}$$

where ∇ and $\nabla\cdot$ denote the gradient and the divergence operators, respectively, and ρ_t denotes the density of the distribution of $\bar{X}_t$.

It is known that the mean-field process is exact, i.e the conditional probability distribution of $\bar{X}_t$ is equal to the posterior distribution of the nonlinear filtering problem [Yang et al., 2016].

The particles $\{X_t^i\}_{i=1}^N$ evolve according to:

$$\mathrm{d}X_t^i = a(X_t^i)\mathrm{d}t + \mathrm{d}B_t^i + \mathsf{K}_t^{(N)}(X_t^i) \circ (\mathrm{d}Z_t - \frac{h(X_t^i)+\hat{h}_t^{(N)}}{2}\mathrm{d}t), \quad X_0^i \overset{\text{i.i.d}}{\sim} \pi_{\text{init}} \tag{1.10}$$

for $i = 1, \ldots N$, where $\{B_t^i\}_{i=1}^N$ are mutually independent Wiener processes, $\hat{h}_t^{(N)} := \frac{1}{N}\sum_{i=1}^N h(X_t^i)$, and $\mathsf{K}_t^{(N)}$ is the output of an algorithm that is used to approximates the solution to the Poisson equation (1.9):

$$\text{Gain function approximation:} \quad \mathsf{K}_t^{(N)} := \text{Algorithm}(\{X_t^i\}_{i=1}^N; h) \tag{1.11}$$

The notation is suggestive of the fact that algorithm is adapted to the ensemble $\{X_t^i\}_{i=1}^N$ and the function h; the density $\rho_t(x)$ is not known in an explicit manner. Development and analysis of numerical algorithms for gain function approximation is one of the main topics of this thesis.

The salient feature of the FPF, compared to the conventional particle filters, is that it replaces the importance sampling and resampling step with a feedback control law. Because of this difference, FPF does not suffer from the particle degeneracy issue. Also in various numerical evaluations and comparisons, it has been observed that FPF exhibits smaller simulation variance [Berntorp, 2015, Tilton et al., 2013, Yang et al., 2013, Stano et al., 2014] and better scaling properties with the problem dimension compared to particle filters [Surace et al., 2017, Yang et al., 2016].

In the special linear Gaussian setting, the FPF algorithm simplifies to a particular form of the Ensemble Kalman filter (EnKF) algorithm. This is described next.

1.2.1 Ensemble Kalman filter as special case of FPF

Consider the linear Gaussian setting, where $h(x) = Hx$ and $\bar{X}_t$ has a Gaussian distribution with mean $\bar{m}_t$ and variance $\bar{\Sigma}_t$. Then the solution of the Poisson equation (1.9) is known in an explicit form [Yang et al., 2016, Sec. D]. The resulting gain function is constant and equal to the Kalman gain:

$$\mathsf{K}_t(x) \equiv \bar{\Sigma}_t H^\top \quad \forall\, x \in \mathbb{R}^d \tag{1.12}$$

Therefore, the mean-field process (1.8) for the linear Gaussian problem is given by:

$$\mathrm{d}\bar{X}_t = A\bar{X}_t \mathrm{d}t + \mathrm{d}\bar{B}_t + \bar{\Sigma}_t H^\top (\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t), \quad \bar{X}_0 \sim \pi_{\text{init}} \tag{1.13}$$

Given the explicit form of the gain function (1.12), the empirical approximation of the gain is simply $\mathsf{K}_t^{(N)} = \Sigma_t^{(N)} H^\top$ where $\Sigma_t^{(N)}$ is the empirical covariance of the particles. Therefore, the evolution of the particles becomes:

$$\mathrm{d}X_t^i = AX_t^i \mathrm{d}t + \mathrm{d}B_t^i + \mathsf{K}_t^{(N)} (\mathrm{d}Z_t - \frac{HX_t^i + Hm_t^{(N)}}{2}\mathrm{d}t), \quad X_0^i \overset{\text{i.i.d}}{\sim} \pi_{\text{init}} \tag{1.14}$$

for $i = 1 \ldots, N$, where $m_t^{(N)}$ is the empirical mean of the particles. The empirical quantities are computed as follows:

$$m_t^{(N)} := \frac{1}{N}\sum_{j=1}^N X_t^i, \quad \Sigma_t^{(N)} := \frac{1}{N-1}\sum_{j=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^\top$$

The linear Gaussian FPF (1.14) is identical to the square-root form of the ensemble Kalman filter [Bergemann and Reich, 2012, Eq. 3.3].

Ensemble Kalman filter (EnKF) was first introduced in [Evensen, 1994, Whitaker and Hamill, 2002], in discrete time setting, and later developed in Reich [2011], in continuous-time setting. It is extensively applied in geophysical sciences as an alternative to extended Kalman filter. In these applications, the state dimension is typically very high. The main advantage of the EnKF, compared to the EKF, is that the computational cost of the EnKF scales as $O(Nd)$ whereas the computational cost of the EKF scales as $O(d^2)$.

1.3 Contributions of this thesis and outline

Part I of the thesis, is concerned with the analysis of the FPF algorithm. The contribution are divided into the following three topics:

1.3.1 Gain function approximation

The mathematical problem is to numerically approximate the solution of the Poisson's equation (1.9) introduced in Sec. 1.2 and also repeated below:

$$-\Delta_\rho \phi = h - \hat{h} \tag{1.15}$$

where the weighted Laplacian $\Delta_\rho \phi(x) := \frac{1}{\rho(x)} \nabla \cdot (\rho(x) \nabla \phi(x))$; $\rho(x)$ is assumed to be an everywhere positive probability density on $\mathbb{R}^d$; $h(x)$ is a real-valued function defined on $\mathbb{R}^d$ and $\hat{h} := \int h(x)\rho(x)\mathrm{d}x$. The function ϕ is referred to as the solution. Its gradient is referred to as the gain function and denoted as $\mathsf{K}(x) := \nabla \phi(x)$.

The numerical approximation problem is as follows: Given N samples $\{X^i\}_{i=1}^N$, drawn i.i.d. from ρ, approximate the gains $\mathsf{K}(X^i) = \nabla \phi(X^i)$. The density ρ is not known in an explicit form.

We make the following assumptions:

(i) The probability density ρ is of the form $\rho(x) = e^{-V(x)}$ where the function $V(x) = \frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m) + w(x)$ for some $m \in \mathbb{R}^d$, $\Sigma \succ 0$, and $w \in C_b^\infty(\mathbb{R}^d)$;

(ii) The function $h(\cdot)$ is differentiable such that $\int (|h(x)|^4 + \|\nabla h(x)\|_2^4)\rho(x)\mathrm{d}x < \infty$, where $\|\cdot\|_2$ denotes the Euclidean norm.

In Chapter 2, we developed a novel diffusion map-based algorithm is designed for the numerical gain function approximation problem. Derivation of the algorithm involves the following steps:

Step 1: The Poisson equation (1.9) is formulated as a fixed-point equation

$$\phi = P_t \phi + \int_0^t P_s(h - \hat{h})\mathrm{d}s \tag{1.16}$$

for some $t > 0$ where $P_t = e^{t\Delta_\rho}$ is the diffusion semigroup associated with the weighted Laplacian Δ_ρ. We show in Prop. 2.2 that the solution of the fixed-point equation (1.16) is also a solution of the Poisson equation (1.15), and vice versa.

Step 2: The diffusion map approximation, introduced in [Coifman and Lafon, 2006], is used to approximate the semigroup P_t. The diffusion map is a Markov operator defined as follows:

$$T_t f(x) := \frac{\int g_t(x,y) \frac{f(y)\rho(y)}{\sqrt{(g_t * \rho)(y)}} \mathrm{d}y}{\int g_t(x,y) \frac{\rho(y)}{\sqrt{(g_t * \rho)(y)}} \mathrm{d}y} \tag{1.17}$$

where $g_t(x,y) = \exp(-\frac{|x-y|^2}{4t})$ is the Gaussian kernel, and $g_t * \rho$ is the convolution of Gaussian kernel with the probability distribution ρ. The fixed point problem (1.16) is approximated in terms of T_t according to

$$\phi_\varepsilon = T_\varepsilon^n \phi_\varepsilon + \sum_{k=1}^{n} \varepsilon T_\varepsilon^k (h - \hat{h}_\varepsilon) \tag{1.18}$$

where $\varepsilon = \frac{t}{n}$, n is a suitably chosen arge number, and $\hat{h}_\varepsilon = \int h \rho_\varepsilon \mathrm{d}x$, where ρ_ε is the invariant density for the Markov operator T_ε. The invariant density is a probability density function such that $\int T_\varepsilon f(x) \rho_\varepsilon(x) \mathrm{d}x = \int f(x) \rho_\varepsilon(x) \mathrm{d}x$ for all bounded functions f.

Step 3: The Markov operator T_t is approximated empirically, in terms of particles, according to

$$T_t^{(N)} f(x) = \frac{\sum_{i=1}^N g_t(x, X^i) \frac{f(X^i)}{\sqrt{\frac{1}{N}\sum_{j=1}^N g_t(X^i, X^j)}}}{\sum_{i=1}^N g_t(x, X^i) \frac{1}{\sqrt{\frac{1}{N}\sum_{j=1}^N g_t(X^i, X^j)}}} \tag{1.19}$$

Using $T_t^{(N)}$, the fixed point problem (1.18) is approximated according to

$$\phi_\varepsilon^{(N)} = (T_\varepsilon^{(N)})^n \phi_\varepsilon^{(N)} + \sum_{i=1}^{n} \varepsilon (T_\varepsilon^{(N)})^k (h - \hat{h}_\varepsilon^{(N)}) \tag{1.20}$$

where $\hat{h}_\varepsilon^{(N)} = \int h \rho_\varepsilon^{(N)} \mathrm{d}x$, and $\rho_\varepsilon^{(N)}$ is the invariant density for the Markov operator $T_\varepsilon^{(N)}$. We show in Prop. 2.5 that the fixed-point equation (1.20) is finite dimensional, and present a numerical procedure, in Table 2.1, to solve it.

We carry out an analysis of the proposed algorithm. We describe the results below:

(i) We study the approximation of the diffusion semigroup P_t with the Markov operator T_t. Existing results depend on Taylor series expansions of the definition (1.17) around $t = 0$, that hold when the underlying space is bounded, and the second derivative of the function f is bounded. We present a new approximation results that holds in unbounded setting, under weaker conditions on

f. We show in Prop. 2.3 that for all functions f such that $f, \nabla f \in L^4(\rho)$:

$$\|T^n_{\frac{t}{n}} f - P_t f\|_{L^2(\rho)} \leq c\frac{\sqrt{t}}{n}(\|f\|_{L^4(\rho)} + \|\nabla f\|_{L^4(\rho)})$$

where c is a constant independent of f, t, and n. Here $\|f\|_{L^p(\rho)} = (\int f^p \rho \mathrm{d}x)^{\frac{1}{p}}$ denotes the L^p-norm with respect to density ρ. The analysis is based on a Feynman-Kac representation of the diffusion semigroup which, to the best of our knowledge, has not been exploited before for analysis of diffusion map approximation.

(ii) We study the contraction property of the Markov operator $T^n_{\frac{t}{n}}$. The contraction property is important to ensure existence of the solution to the approximated fixed-point problem (1.18), and obtain non-asymptotic error bounds for $\phi_\varepsilon \to \phi$. We use the stochastic Lyapunov function method from stability theory of Markov chains to show in Prop. 2.6 that the Markov operator $T^n_{\frac{t}{n}}$ admits a spectral gap of the following form

$$\|T^n_{\frac{t}{n}}\|_2 \leq 1 - \lambda, \quad \forall n \geq n_0$$

for some positive constant $\lambda \in (0,1)$, and a number $n_0 \in \mathbb{N}$. Here $\|\cdot\|_2$ denotes the operator norm with respect to $L^2(\rho)$ norm. Consequently, in Thm. 2.1, we obtain an error bound between the solution to the exact fixed point problem (1.16) and the approximated problem (1.18):

$$\|\phi_\varepsilon - \phi\|_{L^2(\rho)} \leq O(\varepsilon)$$

(iii) Finally, we study the empirical approximation of the integral operator T_t. We show in Prop. 2.4 that for any function f and $\delta \in (0,1)$,

$$\|T^{(N)}_t f - T_t f\|^2_{L^2(\rho)} = O(\frac{\log(\frac{N}{\delta})}{Nt^d})$$

with probability larger than $1-\delta$. We show, in Thm. 2.2, the asymptotic convergence of the solution $\phi^{(N)}_\varepsilon \to \phi_\varepsilon$ almost surely, over any compact subset of $\mathbb{R}^d$.

1.3.2 Optimal transport linear FPF

The objective is to develop a systematic procedure to construct a mean-field process (1.8) such that its distribution is equal to the posterior distribution, i.e

$$\bar{X}_t \sim \pi_t, \quad \forall t \geq 0, \tag{1.21}$$

This is motivated by the so-called uniqueness issue in particle-based algorithms: In particular, for the linear Gaussian filtering problem, the mean-field process described for the square-root of the EnKF algorithm (1.13) is not the only mean-field process that satisfies the exactness condition (1.21). In fact, one can

construct a family of sdes, where the marginal $\bar{X}_t \sim \pi_t$. The construction is described in Chapter 3.

The uniqueness issue can be explained using concepts from the optimal transportation theory. The exactness condition (1.21) is a constraint on the marginal distribution of $\bar{X}_t$ at each time $t \geq 0$. The constraint does not specify the joint distribution at two time instants. There are infinitely many ways to couple two marginal distribution. As a result, there are infinitely many stochastic process that satisfy the exactness condition. Optimal transportation theory provides a way to uniquely couple two distributions [Villani, 2003]. We use this idea to identify a unique mean-field process $\bar{X}_t$, for the linear Gaussian filtering problem. The resulting filter is referred to as the optimal transport linear FPF.

We present two approaches to construct the optimal transport linear FPF. The first approach is based on a time-stepping optimization procedure, which appears in Sec. 3.3. The resulting mean-field process is:

$$\mathrm{d}\bar{X}_t = A\bar{X}_t \mathrm{d}t + \frac{1}{2}\Sigma_B \bar{\Sigma}_t^{-1}\ (\bar{X}_t - \bar{m}_t)\mathrm{d}t - \frac{1}{2}\bar{\mathsf{K}}_t(\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t) + (\text{extra term}) \tag{1.22}$$

where the extra terms is deterministic and does not effect the marginal distribution. It serves to make the dynamics symmetric, in the sense that is made clear in Sec. 3.3, and optimal in the optimal transportation sense. Note that the stochastic term $\sigma_B \mathrm{d}\bar{B}_t$ in (1.13) is replaced with the deterministic term $\frac{1}{2}\Sigma_B\bar{\Sigma}_t^{-1}(\bar{X}_t - \bar{m}_t)\mathrm{d}t$. This makes the evolution of $\bar{X}_t$ completely deterministic.

The optimal transport linear FPF, obtained from the time-stepping procedure, requires the covariance matrix $\bar{\Sigma}_t$ to be invertible. We present an alternative approach, which allows for singular covariance matrices. The second approach is presented in Sec. 3.4. The resulting mean-field process is

$$\mathrm{d}\bar{X}_t = A\bar{X}_t \mathrm{d}t + \frac{1}{2}(\sigma_B + e_t)u_t^\top(\bar{X}_t - \bar{m}_t)\mathrm{d}t + e_t \mathrm{d}\bar{B}_t - \frac{1}{2}\bar{\mathsf{K}}_t(\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t) + (\text{extra terms}) \tag{1.23}$$

where $u_t = \mathrm{Proj}(\sigma_B; \mathrm{Range}(\bar{\Sigma}_t))$ is the projection of σ_B into the range of the matrix $\bar{\Sigma}_t$, $e_t = \sigma_B - \bar{\Sigma}_t u_t$ is the error in projection, and the extra terms serve the same purpose as in (1.22). In the case where $\bar{\Sigma}_t$ is invertible, the optimal transport sde (1.23) simplifies to (1.22). The development of the sde for the singular covariance matrix case is important for the finite-N implementation, when $N < d$ and the empirical covariance matrix is necessarily singular.

1.3.3 Error analysis of the linear FPF

The objective of this research is to analyze the finite-N system FPF algorithm (1.14), in the linear Gaussian setting. We consider two class of linear FPF algorithms, deterministic linear FPF and stochastic linear FPF.

The deterministic linear FPF algorithm is the finite-N particle approximation of (1.22). We assume that the linear system (1.4a)-(1.4b) is controllable and observable, and the number of particles $N > d$. We obtain the following results:

1. We show in Sec. 4.3 that the evolution of the empirical mean $m_t^{(N)}$ and the empirical covariance $\Sigma_t^{(N)}$ is exactly equal to the Kalman-Bucy filter.

2. We show in Prop. 4.1 that the empirical mean and empirical covariance converges almost surely to

the exact mean and covariance: for any finite $N > d$, $\exists \lambda_0 > 0$ such that:

$$\lim_{t\to\infty} e^{\lambda t}\|m_t^{(N)} - m_t\|_2 = 0 \quad \text{a.s}, \quad \lim_{t\to\infty} e^{2\lambda t}\|\Sigma_t^{(N)} - \Sigma_t\|_F = 0 \quad \text{a.s}$$

for all $\lambda \in (0, \lambda_0)$, where $\|\cdot\|_2$ denotes the Euclidean norm, and $\|\cdot\|_F$ denotes the Frobenius norm.

3. We also obtain non-asymptotic mean-squared error estimates in Prop. 4.1 such that for any $t > 0$:

$$\mathsf{E}[\|m_t^{(N)} - m_t\|_2^2] \le (\text{const.})e^{-2\lambda t}\frac{\mathrm{Tr}(\Sigma_0) + \|\Sigma_0\|_F^2}{N}$$
$$\mathsf{E}[\|\Sigma_t^{(N)} - \Sigma_t\|_F^2] \le (\text{const.})e^{-4\lambda t}\frac{\|\Sigma_0\|_F^2}{N}$$

for all $\lambda \in (0, \lambda_0)$ where the constant depends on spectral properties of the system and does not scale with the dimension d. $\mathrm{tr}(\cdot)$ denotes the trace operation.

4. Finally, we carry out a propagation of chaos error analysis to conclude the convergence of the empirical distribution to the mean-field distribution in Prop. 4.3: In particular, we show that for any Lipschitz function f:

$$\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(X_t^i) - \mathsf{E}[f(\bar{X}_t)|\mathscr{Z}_t]\right|^2\right] \le \frac{(\text{const})}{N} \tag{1.24}$$

The stochastic linear FPF algorithm is the finite-N particle approximation of (1.14). We make strong assumptions that the system is stable and fully observable. We show the following results:

1. We show in Sec. 4.3 that the evolution of the empirical mean and the empirical covariance is similar to the Kalman-Bucy filter, with additional stochastic terms that scale as $O(N^{-\frac{1}{2}})$.

2. We obtain non-asymptotic mean-squared error estimates in Prop. 4.2 such that for any $t > 0$ and $N \in \mathbb{N}$:

$$\mathsf{E}[\|\Sigma_t^{(N)} - \Sigma_t\|_F^2] \le \frac{(\text{const})}{N},$$
$$\mathsf{E}[\|m_t^{(N)} - m_t\|_2] \le \frac{(\text{const})}{\sqrt{N}}$$

where the constants are uniformly bounded in time.

3. Also, we carry out propagation of chaos error analysis and obtain a result in Prop. 4.4 similar to (1.24).

1.3.4 Outline of the Part I of the thesis

The contributions for the gain function approximation problem is presented in Chapter 2. The optimal transport formulation of FPF is presented in Chapter 3. The error analysis of the linear FPF algorithms is presented in Chapter 4.

Chapter 2

Gain function approximation*

2.1 Introduction

This chapter is concerned with the numerical solutions of the Poisson equation (1.9) that arises in the FPF algorithm. Given an everywhere positive probability density function ρ and a real-valued function h, the (weighted) Poisson equation is given by

$$-\Delta_\rho \phi = h - \hat{h} \tag{2.1}$$

where the weighted Laplacian $\Delta_\rho \phi(x) := \frac{1}{\rho(x)} \nabla \cdot (\rho(x) \nabla \phi(x))$; and $\hat{h} := \int h(x) \rho(x) \mathrm{d}x$. The function ϕ, if one exists, is referred to as the solution. In the context of the filtering problem, the probability density ρ represents the density of the posterior distribution of the filter and the function h represents the observation model (see Sec. 1.2). The gradient of the solution to the Poisson equation is the gain function used in the FPF algorithm and denoted as $\mathsf{K}(x) := \nabla \phi(x)$.

The numerical approximation problem is as follows: Given N samples $\{X^1, \ldots, X^i, \ldots, X^N\}$, drawn i.i.d. from ρ, approximate the gains $\{\mathsf{K}^1, \ldots, \mathsf{K}^i, \ldots, \mathsf{K}^N\}$, where $\mathsf{K}^i := \mathsf{K}(X^i) = \nabla \phi(X^i)$. The density ρ is not known in an explicit form.

Development and error analysis of gain function approximation algorithms is the subject of this chapter. Before describing the general case, it is useful to review the linear Gaussian case where the solution to the Poisson equation is explicitly known.

Linear Gaussian setting: Suppose $h(x) = Hx$ and ρ is a Gaussian density with mean m and variance Σ. Then the solution of the Poisson equation is known in an explicit form [Yang et al., 2016, Sec. D]. The solution is $\phi(x) = \mathsf{K}^\top (x - m)$, where

$$\mathsf{K} \equiv \Sigma H^\top \quad \forall\, x \in \mathbb{R}^d \tag{2.2}$$

equal to the Kalman gain. Therefore, the gain function is constant, equal to the Kalman gain. Given the explicit form of the gain function (2.2), the empirical approximation of the gain is simply $\mathsf{K}_t^{(N)} = \Sigma^{(N)} H^\top$ where $\Sigma^{(N)}$ is the empirical covariance of the particles.

One extension of the Kalman gain is the so called *constant gain approximation* formula whereby the gain K_t is approximated by its expected value (which represents the best least-squared approximation of the gain by a constant). Remarkably, the expected value admits a closed-form expression which is then readily

*The preliminary results concerning the contributions of this chapter appears in [Taghvaei and Mehta, 2016b, Taghvaei et al., 2017].

approximated empirically using the particles (see Remark 2.2):

$$\text{Const. gain approx:} \quad \mathsf{E}[\mathsf{K}(X)] = \int_{\mathbb{R}^d} (h(x) - \hat{h})\, x\, \rho(x) \mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} (h(X_t^i) - \hat{h}^{(N)})\, X_t^i \tag{2.3}$$

The constant gain approximation formula has been used in nonlinear extensions of the EnKF algorithm [de Wiljes et al., 2016]. The connection to the Poisson equation provides a justification for this formula. The formula is attractive because it provides a consistent (as the number of particles $N \to \infty$) approximation of the Kalman gain in the linear Gaussian setting.

Design and analysis of the gain function approximation algorithm in the general case is a challenging problem because of two reasons: (i) Apart from the Gaussian case, there are no known closed-form solutions of (2.1); (ii) The density $\rho(x)$ is not explicitly known. One only has samples $\{X^i\}_{i=1}^N$ i.i.d drawn from ρ. The assumption is justified because in the limit of large N, the particles are approximately i.i.d (by the propagation of chaos); cf., [Sznitman, 1991].

2.1.1 Related work

Apart from its direct relevance to numerical approximation of the FPF, there are three topics of current research interest that are relevant to the subject of this chapter: (i) ensemble Kalman filter; (ii) particle flow algorithms for nonlinear filtering; and (iii) optimal transport. Specifically, the algorithms for gain function approximation described in this paper are also directly applicable to these other topics. These relationships are briefly discussed next:

Ensemble Kalman filter: The EnKF algorithm was first developed in the discrete-time setting [Evensen, 1994]. In the continuous-time setting, two formulations of the EnKF have been developed: EnKF with perturbed observation, and the more recent square root EnKF [Bergemann and Reich, 2012, Reich and Cotter, 2015]. As has already been noted in Sec. 1.2, the square root EnKF is in fact identical to the FPF algorithm in the linear Gaussian setting [Bergemann and Reich, 2012, Taghvaei et al., 2018].

The EnKF algorithm provides a consistent approximation in the linear Gaussian setting. Compared to the Kalman filter, the main utility of EnKF is that it does not require propagation of the covariance matrix. This reduces the computational complexity from $O(d^2)$ for the Kalman filter to $O(Nd)$. This is clearly advantageous in high dimensional problems when $N << d$. This property has made EnKF popular in applications such as weather prediction in high dimensional settings [Kalnay, 2002, Oliver et al., 2008]. The disadvantage of the EnKF algorithm, of course, is that it does not provide a consistent approximation for nonlinear problems.

FPF represents a the generalization of the EnKF to the nonlinear non-Gaussian setting [Taghvaei et al., 2018]: With the constant gain approximation, the algorithms are identical. Given this parallel, the problem of improving the EnKF algorithm in more general nonlinear non-Gaussian settings is directly related to the problem of better approximating the gain function in the FPF. In an application software based on EnKF, it is a relatively simple matter to replace the constant gain formula for the gain by more sophisticated approximations described in this paper. Certain empirical evaluations on the performance of FPF in high-

dimensional settings are reported in [Surace et al., 2017, Stano et al., 2014, Stano, 2013, Berntorp, 2015].

Error analysis and stability of EnKF is an active area of research; see [Le Gland et al., 2009, Kwiatkowski and Mandel, 2015, Del Moral and Tugaut, 2016] for linear models and [de Wiljes et al., 2016, Del Moral et al., 2017, Kelly et al., 2014] for nonlinear models. The error analysis for the gain function approximation in this chapter is a step towards error analysis of the FPF along these lines.

Particle flow algorithms: The following first-order (and hence an under determined) form of the Poisson equation appears in most types of particle flow algorithms:

$$\nabla \cdot (p_t(x)\mathsf{K}(x)) = (\text{rhs})$$

where the righthand-side (rhs) is given and $\mathsf{K}(x)$ defines a vector field that must be obtained to implement the particle flow. The pde appears in the first interacting particle representation of the continuous-time filtering in [Crisan and Xiong, 2007, 2010] and the discrete-time filtering in [Daum et al., 2010]. Stochastic extensions of these have also recently appeared in [Daum et al., 2017] where approximate solutions are also described based on Gaussian assumption on the density. The algorithm described here represent an approximation of a particular gradient form solution of the first-order pde.

Optimal transport: The mean-field sde (1.8) represents a transport that maps the prior distribution at time 0 to the posterior distribution at an (arbitrary) future time $t > 0$. Synthesis of optimal transport maps for implementing the Bayes formula appears in [Reich, 2011, Cheng and Reich, 2013, El Moselhy and Marzouk, 2012, Taghvaei and Mehta, 2016a, Heng et al., 2015, Chen et al., 2016, Kim et al., 2013, Ma and Coleman, 2011]. The relationship with the Poisson equation is through the ensemble transform filter which relies on a linear programming construction to approximate the optimal transport map [Cheng and Reich, 2013]. As discussed in [Taghvaei et al., 2018, Sec. 5.5], the solution of the Poisson equation yields an infinitesimal optimal transport map from the "prior" $p_t(x)$ to an un-normalized "posterior" $p_t(x)\exp(-th(x))$. Another closely related approach is optimal transportation is through the Gibbs flow [Heng et al., 2015].

Directly related to the FPF, the Galerkin method for the numerical solution of the Poisson equation appeared in original papers [Yang et al., 2013, 2016]. The Galerkin algorithm represents the 'direct" pde approach to construct a numerical approximation. The constant gain approximation is a particular example of a Galerkin solution. In general, the main problem with the Galerkin approximation is that it requires a selection of basis functions. This becomes intractable in high dimensions. To mitigate this issue, a proper orthogonal decomposition (POD)-based procedure to select basis functions is introduced in [Berntorp and Grover, 2016] and a continuation scheme for approximation appears in [Matsuura et al., 2016]. Certain probabilistic approaches based on dynamic programming appear in [Radhakrishnan et al., 2014].

The diffusion map-based algorithm proposed and analyzed here is inspired by the spectral clustering literature [Belkin, 2003, Von Luxburg, 2007]. The particular form of the kernel proposed here was introduced in [Coifman and Lafon, 2006]. Convergence analysis for this operator appears in [Hein et al., 2005, Singer, 2006, Coifman and Lafon, 2006, Giné et al., 2006, Hein et al., 2007, Von Luxburg et al., 2008, Belkin and Niyogi, 2007].

2.2 Mathematical preliminaries

Notation: For vectors $x, y \in \mathbb{R}^d$, the dot product is denoted as $x \cdot y$ and the Euclidean norm is denoted as $\|x\|_2 := \sqrt{x \cdot x}$. $L^p(\rho)$ is used to denote the space of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int |f(x)|^p \rho(x) \mathrm{d}x < \infty$. The $L^p(\rho)$ norm of $f \in L^p(\rho)$ is denoted by $\|f\|_{L^p(\rho)} := (\int |f(x)|^p \rho(x) \mathrm{d}x)^{\frac{1}{p}}$. The inner product on $L^2(\rho)$ is defined by $\langle f, g \rangle_{L^2} := \int f(x) g(x) \rho(x) \mathrm{d}x$. The space $H^1(\rho)$ is the space functions $f \in L^2(\rho)$ whose derivative (defined in the weak sense) is in $L^2(\rho)$. For a differentiable function f, $\|\nabla f\|_{L^p(\rho)} := (\int \|\nabla f(x)\|^p \rho(x) \mathrm{d}x)^{\frac{1}{p}}$. For an integrable function f, $\hat{f} := \int f(x) \rho(x) \mathrm{d}x$ denotes the mean. $L^2_0(\rho) := \{f \in L^2(\rho) \mid \hat{f} = 0\}$ and $H^1_0(\rho) := \{f \in H^1(\rho) \mid \hat{f} = 0\}$ denote the co-dimension 1 subspace of functions whose mean is zero. L^∞ denotes the space of bounded functions on $\mathbb{R}^d$ with associated norm denoted as $\|\cdot\|_{L^\infty}$. The space of smooth and bounded functions on $\mathbb{R}^d$ is denoted as $C^\infty_b(\mathbb{R}^d)$. The Borel σ-algebra on $\mathbb{R}^d$ is denoted by $\mathscr{B}(\mathbb{R}^d)$. The variance of the random variable X is denoted as $\mathrm{Var}(X)$. The indicator function, for a measurable set $A \in \mathbb{B}(\mathbb{R}^d)$, is denoted as $\mathbf{1}_{[A]}(\cdot)$ such that $\mathbf{1}_{[A]}(x) = 1$ if $x \in A$ and $\mathbf{1}_{[A]}(x) = 0$ if $x \notin A$

Assumptions: The following assumptions are made throughout the paper:

(i) **Assumption A1:** The probability density ρ is of the form $\rho(x) = e^{-V(x)}$ where the function $V(x) = \frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m) + w(x)$ for some $m \in \mathbb{R}^d$, $\Sigma \succ 0$, and $w \in C^\infty_b(\mathbb{R}^d)$;

(ii) **Assumption A2:** The function $h(\cdot)$ is differentiable and $h, \nabla h \in L^4(\rho)$.

Remark 2.1. *Assumption A1 implies that the distribution ρ is Gaussian with a bounded perturbation. It is commonly used in the theory of functional inequalities to obtain log-Sobolev and Poincaré inequalities with constants that does not depend on dimension Villani [2003]. In this paper, it is used to prove approximation result (see Prop. 2.3) and spectral gap (see Prop. 2.6) for the diffusion map. The assumption is restrictive, as it is not satisfied for a mixture of Gaussians. Proving the results in this paper with a weaker assumption such as $\rho = \rho_g * w$, the convolution of a Gaussian density ρ_g and a density w with a compact support, is the subject of continuing work.*

2.2.1 Spectral gap and weak formulation

Under Assumption (A1), the weighted Laplacian Δ_ρ has a discrete spectrum with an ordered sequence of eigenvalues $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots$ and associated eigenfunctions $\{e_n\}$ that form a complete orthonormal basis of $L^2(\rho)$ [Bakry et al., 2013, Cor. 4.10.9]. The trivial eigenfunction $e_0(x) = 1$, and for $f \in L^2_0(\rho)$, the spectral representation yields:

$$-\Delta_\rho f = \sum_{m=1}^{\infty} \lambda_m \langle e_m, f \rangle e_m \tag{2.4}$$

The positivity of the smallest non-trivial eigenvalue ($\lambda_1 > 0$) is referred to as the Poincaré inequality (or the spectral gap condition) [Bakry et al., 2008]. The inequality is equivalently expressed as

$$\int_{\mathbb{R}^d} (f - \hat{f})^2 \rho \mathrm{d}x \leq \frac{1}{\lambda_1} \int_{\mathbb{R}^d} \|\nabla f\|_2^2 \, \rho \mathrm{d}x \qquad \forall \, f \in H^1(\rho)$$

where $\hat{f} = \int f \rho \mathrm{d}x$.

The Poincaré inequality is important to show that the Poisson equation is well-posed and a unique solution exists. The solution to the Poisson equation is defined using the weak formulation.

A function $\phi \in H_0^1(\rho)$ is said to be a weak solution of (2.1) if

$$\int \nabla\phi(x) \cdot \nabla\psi(x)\rho(x)\mathrm{d}x = \int (h(x) - \hat{h})\psi(x)\rho(x)\mathrm{d}x \qquad \forall\ \psi \in H^1(\rho) \tag{2.5}$$

Equation (2.5) is referred to as the weak-form of the Poisson's equation. The weak-form is expressed succinctly as $\langle \nabla\phi, \nabla\psi \rangle = \langle h - \hat{h}, \psi \rangle$ where $\langle \cdot, \cdot \rangle$ is the inner-product in $L^2(\rho)$. The existence and uniqueness of the solution to the weak-form of the Poisson equation is stated in the following Proposition.

Proposition 2.1. *[Laugesen et al., 2015, Thm. 2.2.] Suppose ρ satisfies Assumption (A1) and h satisfies Assumption (A2). Then there exists a unique function $\phi \in H_0^1(\rho)$ that satisfies the weak-form of the Poisson equation (2.5). The solution satisfies the bound:*

$$\int \|\nabla\phi(x)\|_2^2\, \rho(x)\mathrm{d}x \leq \frac{1}{\lambda_1} \int (h(x) - \hat{h})^2 \rho(x)\mathrm{d}x$$

Remark 2.2 (Constant gain approximation)**.** *The weak formulation has led to the Galerkin algorithm presented in the original FPF papers [Yang et al., 2016]. A special case of the Galerkin solution is the constant gain approximation formula (2.3). The formula is obtained from the weak formulation (2.5). Choose the test functions to be coordinate functions: $\psi_m(x) = x_m$ for $m = 1, 2, \ldots, d$. Then,*

$$\int \frac{\partial \phi}{\partial x_m}(x)\rho(x)\mathrm{d}x = \int (h(x) - \hat{h}) x_m \rho(x)\mathrm{d}x, \quad \textit{for} \quad m = 1, \ldots, d$$

concluding the formula (2.3).

2.2.2 Semigroup formulation

Let $\{P_t\}_{t \geq 0}$ be the semigroup associated with the weighted Laplacian Δ_ρ. The semigroup allows for a probabilistic interpretation which is described next. Consider the following reversible Markov process $\{S_t\}_{t \geq 0}$ evolving in $\mathbb{R}^d$:

$$\mathrm{d}S_t = -\nabla V(S_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

where $V(x) := -\log(\rho(x))$ and $\{B_t\}_{t \geq 0}$ is a standard Weiner process in $\mathbb{R}^d$. Then

$$P_t f(x) = \mathsf{E}[f(S_t) | S_0 = x]$$

It is straightforward to verify that $P_t : L^2(\rho) \to L^2(\rho)$ is symmetric, i.e., $\langle P_t f, g \rangle = \langle f, P_t g \rangle$ for all $f, g \in L^2(\rho)$ and $\rho(x) = e^{-V(x)}$ is its invariant density. The semigroup also admits a kernel representation:

$$P_t f(x) = \sum_{m=1}^{\infty} e^{-t\lambda_m} \langle e_m, f \rangle e_m(x) = \int_{\mathbb{R}^d} \bar{k}_t(x, y) f(y) \rho(y) \mathrm{d}y$$

where $\bar{k}_t(x,y) := \sum_{m=0}^{\infty} e^{-t\lambda_m} e_m(x) e_m(y)$.

The spectral gap implies that $\|P_t\|_{L_0^2(\rho)} = e^{-t\lambda_1} < 1$. Hence, P_t is a strict contraction on $L_0^2(\rho)$. For the special case of Gaussian density $\rho = \mathcal{N}(m,\Sigma)$, the eigenfunctions are given by the Hermite polynomials. This leads to an explicit formula for the kernel $\bar{k}_t(x,y)$ in the Gaussian case, as described in Appendix. 2.6.1.

Consider the heat equation

$$\frac{\partial u}{\partial t} = \Delta_\rho u + (h - \hat{h}), \quad u(0,x) = f(x)$$

Its solution is given in terms of the semigroup as follows:

$$u(t,x) = P_t f(x) + \int_0^t P_{t-s}(h - \hat{h})(x) \mathrm{d}s$$

Letting $f(x) = \phi(x)$ where ϕ solves the Poisson equation (2.1) yields the following fixed-point equation for $t = \varepsilon$:

$$\text{(exact fixed-point equation)} \quad \phi = P_\varepsilon \phi + \int_0^\varepsilon P_s(h - \hat{h}) \mathrm{d}s \tag{2.6}$$

Equation (2.6) is referred to as the semigroup form of the Poisson equation (2.1).

The following Proposition shows that the weak form (2.5) and the semigroup form (2.6) are equivalent. The proof appears in the Appendix. 2.6.2.

Proposition 2.2. *Suppose ρ satisfies Assumption (A1) and h satisfies Assumption (A2). Then the unique solution $\phi \in H_0^1(\rho)$ to the weak form (2.5) is also the unique solution to the fixed-point equation (2.6).*

The semigroup formulation has led to the diffusion map-based algorithm which is the main focus of this chapter.

2.3 Diffusion map-based Algorithm

The diffusion map-based algorithm is based on a numerical approximation of the fixed-point equation (2.6). The main technique is to approximate the semigroup P_ε in the following three steps:

1. **Diffusion map approximation:** A family of Markov operators $\{T_\varepsilon\}_{\varepsilon>0}$ are defined as follows:

$$T_\varepsilon f(x) := \frac{1}{n_\varepsilon(x)} \int_{\mathbb{R}^d} k_\varepsilon(x,y) f(y) \rho(y) \mathrm{d}y \tag{2.7}$$

   where $n_\varepsilon(x) := \int k_\varepsilon(x,y)\rho(y)\mathrm{d}y$ is the normalization factor,

$$k_\varepsilon(x,y) := \frac{g_\varepsilon(x,y)}{\sqrt{\int g_\varepsilon(x,z)\rho(z)\mathrm{d}z}\sqrt{\int g_\varepsilon(y,z)\rho(z)\mathrm{d}z}}$$

and $g_\varepsilon(x,y) := \exp(-\frac{|x-y|^2}{4\varepsilon})$ is the Gaussian kernel in $\mathbb{R}$. For small positive values of ε, the Markov operator T_ε is referred to as the *diffusion map* approximation of the exact semigroup P_ε [Coifman and Lafon, 2006, Hein et al., 2005]. The precise statement of this approximation is contained in Prop. 2.3. For the special case of Gaussian density, an explicit formula for the diffusion map appears in the Appendix. 2.6.1.

2. **Empirical approximation:** The operator T_ε is approximated empirically by $\{T_\varepsilon^{(N)}\}_{\varepsilon>0,N\in\mathbb{N}}$ defined as follows:

$$T_\varepsilon^{(N)} f(x) := \frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^{N} k_\varepsilon^{(N)}(x,X^j) f(X^j) \tag{2.8}$$

where $n_\varepsilon^{(N)}(x) := \sum_{i=1}^N k_\varepsilon(x,X^i)$ is the normalization factor and

$$k_\varepsilon^{(N)}(x,y) := \frac{g_\varepsilon(x,y)}{\sqrt{\sum_{j=1}^N g_\varepsilon(x,X^j)}\sqrt{\sum_{j=1}^N g_\varepsilon(y,X^j)}}$$

Recall that $X^i \overset{\text{i.i.d}}{\sim} \rho$ for $i = 1,\dots,N$. So, by law of large numbers (LLN), $T_\varepsilon^{(N)} f$ represents an empirical approximation of the diffusion map T_ε. The precise statement of the empirical approximation is contained in Prop. 2.4.

3. **Approximation as Markov matrix:** An $N \times N$ Markov matrix T is defined with (i,j)-th element given by

$$\mathsf{T}_{ij} = \frac{1}{n_\varepsilon^{(N)}(X^i)} K_\varepsilon^{(N)}(X^i,X^j) \tag{2.9}$$

Finite-dimensional fixed-point equation: Using the three steps above, the original infinite-dimensional fixed-point equation (2.6) is approximated as a finite dimensional fixed-point equation

$$\Phi = \mathsf{T}\Phi + \varepsilon(\mathsf{h} - \pi(h)) \tag{2.10}$$

where $\mathsf{h} := (h(X^1),\dots,h(X^N))$ is a $N \times 1$ column vector, and $\pi(h) = \sum_{i=1}^N \pi_i h(X^i)$ where the probability vector $\pi_i = \frac{n_\varepsilon^{(N)}(X^i)}{\sum_{j=1}^N n_\varepsilon^{(N)}(X^j)}$ is the unique stationary distribution of the Markov matrix T. The solution Φ is used to define an approximation to the solution of the Poisson equation as follows:

$$\phi_\varepsilon^{(N)}(x) := \frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^{N} k_\varepsilon^{(N)}(x,X^j)\Phi_j + \varepsilon(h(x) - \pi(h)) \tag{2.11}$$

The approximation for the gain function is as follows:

$$\mathsf{K}_\varepsilon^{(N)}(x) = \nabla \left[\frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^{N} k_\varepsilon^{(N)}(x,X^j)(\Phi_j + \varepsilon \mathsf{h}_j) \right] \tag{2.12}$$

Upon evaluating the gradient in closed-form, the following linear formula results for the gain function

evaluated at particle locations:

$$\mathsf{K}^i := \mathsf{K}_\varepsilon^{(N)}(X^i) = \sum_{j=1}^{N} s_{ij} X^j \tag{2.13}$$

where

$$s_{ij} := \frac{1}{2\varepsilon}\mathsf{T}_{ij}(r_j - \sum_{k=1}^{N}\mathsf{T}_{ik}r_k), \quad r_j := \Phi_j + \varepsilon \mathsf{h}_j \tag{2.14}$$

The details of the calculation leading to the linear formula appear in the Appendix. 2.6.3.

Remark 2.3 (Numerical procedure)**.** *The fixed-point problem* (2.10) *is proposed to be solved in an iterative manner. The vector* Φ *is initialized at* $\Phi_0 = (0,\ldots,0) \in \mathbb{R}^N$ *and it is updated according to*

$$\Phi_{n+1} = \mathsf{T}\Phi_n + \varepsilon(\mathsf{h} - \pi(h)).$$

for $n = 0,1,\ldots,L$*. The procedure is guaranteed to converge as* $L \to \infty$*, with a geometric convergence rate, because* T *is a strict contraction on* $L_0^2(\pi)$ *(see Prop. 2.5-(ii)). In a filtering application, the procedure is initialized with the vector value* Φ *that is obtained from the previous filter iteration. The proposed iterative procedure, to solve the fixed-point equation* (2.10)*, is proffered to other numerical procedures because (i) it is numerically more efficient than solving a system of N linear equations; (ii) and it allows to use the solution obtained from the last filter iteration, as the initial value for the current filter iteration. The overall algorithm is tabulated in Algorithm 2.1.*

Algorithm 2.1 diffusion map based algorithm for gain function approximation

Input: $\{X^i\}_{i=1}^N$, $\{h(X^i)\}_{i=1}^N$, Φ_{prev}, ε, L
Output: $\{\mathsf{K}^i\}_{i=1}^N$

Calculate $g_{ij} := \exp(-|X^i - X^j|^2/4\varepsilon)$ for $i,j = 1$ to N
Calculate $k_{ij} := \frac{g_{ij}}{\sqrt{\sum_l g_{il}}\sqrt{\sum_l g_{jl}}}$ for $i,j = 1$ to N
Calculate $d_i = \sum_j k_{ij}$ for $i = 1$ to N
Calculate $\mathsf{T}_{ij} := \frac{k_{ij}}{d_i}$ for $i,j = 1$ to N
Calculate $\pi_i = \frac{d_i}{\sum_j d_j}$ for $i = 1$ to N
Calculate $\hat{\mathsf{h}} = \sum_{i=1}^N \pi_j h(X^i)$
Initialize $\Phi = \Phi_{\text{prev}}$
for $t = 1$ to L **do**
    $\Phi_i = \sum_{j=1}^N \mathsf{T}_{ij}\Phi_j + \varepsilon(\mathsf{h} - \hat{\mathsf{h}})$ for $i = 1$ to N
end for
Calculate $r_i = \Phi_i + \varepsilon \mathsf{h}_i$ for $i = 1$ to N
Calculate $s_{ij} = \frac{1}{2\varepsilon}\mathsf{T}_{ij}(r_j - \sum_{k=1}^N \mathsf{T}_{ik}r_k)$ for $i,j = 1$ to N
Calculate $\mathsf{K}_i = \sum_j s_{ij} X^j$ for $i = 1$ to N

Remark 2.4. The computational complexity of the diffusion map based algorithm is $O(N^2)$ because of the need to assemble the $N \times N$ matrix T. The computational complexity may be reduced using the sparsity

structure of the matrix T and sub-sampling techniques. Compared to the Galerkin algorithm with computational complexity of $O(Nd^3)$, the diffusion-map algorithm is advantageous in high-dimensional problems where $d >> N$.

2.3.1 Approximation results

The notation $G_\varepsilon(f)(x) := \int g_\varepsilon(x,y)f(y)\mathrm{d}y$ is used to denote the heat semigroup with a Gaussian kernel $g_\varepsilon(x,y)$, and

$$U_\varepsilon := \frac{1}{2}\log(\frac{G_\varepsilon(\rho)}{\rho^2}), \quad U := -\frac{1}{2}\log(\rho) \tag{2.15a}$$

$$W_\varepsilon := \frac{1}{\varepsilon}\log(e^{U_\varepsilon}G_\varepsilon(e^{-U_\varepsilon})), \quad W := |\nabla U|^2 - \Delta U \tag{2.15b}$$

The proof of the following proposition appears in Appendix. 2.6.5.

Proposition 2.3. *Consider the family of Markov operators $\{T_\varepsilon\}_{\varepsilon>0}$ defined according to (2.7). Let $n \in \mathbb{N}$, $t \in (0,t_0)$ with $t_0 < \infty$, and $\varepsilon = \frac{t}{n}$. Then,*

(i) *The semigroup P_t and the operator T_ε^n admit the following representations:*

$$P_t f(x) = e^{U(x)}\mathsf{E}[e^{-\int_0^t W(B^x_{2s})\mathrm{d}s}e^{-U(B^x_{2t})}f(B^x_{2t})] \tag{2.16}$$

$$T_\varepsilon^n f(x) = e^{U_\varepsilon(x)}\mathsf{E}[e^{-\varepsilon\sum_{k=0}^{n-1}W_\varepsilon(B^x_{2k\varepsilon})}e^{-U_\varepsilon(B^x_{2n\varepsilon})}f(B^x_{2n\varepsilon})] \tag{2.17}$$

for all $x \in \mathbb{R}^d$ where B_t^x is the Brownian motion with initial condition $B_0^x = x$.

(ii) *In the asymptotic limit as $\varepsilon \to 0$:*

$$U_\varepsilon(x) = U(x) + 2\varepsilon W(x) + \varepsilon\Delta V(x) + \varepsilon^2 r_\varepsilon^{(1)}(x) \tag{2.18a}$$

$$W_\varepsilon(x) = W(x) + \varepsilon r_\varepsilon^{(2)}(x) \tag{2.18b}$$

where $|r_\varepsilon^{(1)}(x)|, |r_\varepsilon^{(2)}(x)| = O(|x|^2)$ and $|\nabla r_\varepsilon^{(1)}(x)| = O(|x|)$ as $|x| \to \infty$.

(iii) *For all functions f such that $f, \nabla f \in L^4(\rho)$:*

$$\|(T^n_{\frac{t}{n}} - P_t)f\|_{L^2(\rho)} \leq (const.)\frac{\sqrt{t}}{n}(\|f\|_{L^4(\rho)} + \|\nabla f\|_{L^4(\rho)}) \tag{2.19}$$

where the constant only depends on t_0 and ρ.

The proof of the following proposition appears in Appendix. 2.6.8.

Proposition 2.4. *Consider the diffusion map kernel $\{T_\varepsilon\}_{\varepsilon>0}$, and its empirical approximation $\{T_\varepsilon^{(N)}\}_{\varepsilon>0,N\in\mathbb{N}}$. Then for any bounded continuous function $f \in C_b(\mathbb{R}^d)$:*

(i) *(Almost sure convergence) For all* $x \in \mathbb{R}^d$

$$\lim_{N\to\infty} T_\varepsilon^{(N)} f(x) = T_\varepsilon f(x) \quad a.s$$

(ii) *(Convergence rate) For any* $\delta \in (0,1)$*, in the asymptotic limit as* $N \to \infty$*,*

$$\int |T_\varepsilon^{(N)} f(x) - T_\varepsilon f(x)|^2 \rho(x) \mathrm{d}x = O(\frac{\log(\frac{N}{\delta})}{N\varepsilon^d})$$

with probability higher than $1-\delta$*.*

Remark 2.5 (Related work)**.** *The key idea in the proof of the Prop. 2.3 is the Feynman-Kac representation of the semigroup (2.16). To the best of our knowledge, this representation has not been used before in the analysis of the diffusion map approximation. Most of the existing results concerning the convergence of the diffusion map are based on a Taylor series expansion that would lead to a convergence of the form* $\lim_{\varepsilon\to 0} \frac{f(x)-T_\varepsilon f(x)}{\varepsilon} = \Delta_\rho f(x)$ *for each* $x \in \mathbb{R}^d$ *[Hein et al., 2005, Coifman and Lafon, 2006, Giné et al., 2006]. Convergence results of the form* $\lim_{n\to\infty} \|T_{\frac{t}{n}}^n f - P_t f\|_2 = 0$ *appear in [Coifman and Lafon, 2006, Ting et al., 2011], based on functional analytic arguments. The Taylor series type arguments typically require the distribution to be supported on a compact manifold which is not assumed here.*

2.4 Convergence and error analysis

The analysis of the diffusion map algorithm involves the consideration of the following four fixed point problems:

$$\text{(exact)} \qquad \phi = P_\varepsilon \phi + \int_0^\varepsilon P_s(h-\hat{h})\mathrm{d}s \tag{2.20}$$

$$\text{(diffusion map approx.)} \qquad \phi_\varepsilon = T_\varepsilon \phi_\varepsilon + \varepsilon(h-\hat{h}_\varepsilon) \tag{2.21}$$

$$\text{(empirical approx.)} \qquad \phi_\varepsilon^{(N)} = T_\varepsilon^{(N)} \phi_\varepsilon^{(N)} + \varepsilon(h-\pi(h)) \tag{2.22}$$

$$\text{(finite-dim.)} \qquad \Phi = \mathsf{T}\Phi + \varepsilon(\mathsf{h}-\pi(\mathsf{h})) \tag{2.23}$$

where $\hat{h}_\varepsilon := \int h(x)\rho_\varepsilon(x)\mathrm{d}x$ and $\rho_\varepsilon(x) := \frac{n_\varepsilon(x)\rho(x)}{\int n_\varepsilon(x)\rho(x)\mathrm{d}x}$ is the density of the invariant probability distribution associated with the Markov operator T_ε.

In practice, the finite-dimensional problem (2.23) is solved. The existence and uniqueness of the solution for this problem is the subject of the following proposition whose proof appears in Appendix. 2.6.4.

Proposition 2.5. *Consider the finite-dimensional fixed point equation (2.23).*
Then almost surely

(i) T *is a reversible Markov matrix with a unique stationary distribution*

$$\pi_i := \frac{n_\varepsilon^{(N)}(X^i)}{\sum_{j=1}^N n_\varepsilon^{(N)}(X^j)} \tag{2.24}$$

for $i = 1, \ldots, N$.

(ii) T *is a strict contraction on* $L_0^2(\pi) = \{v \in \mathbb{R}^N; \sum \pi_i v_i = 0\}$. *Hence the fixed point equation* (2.23) *has a unique solution* $\Phi \in L_0^2(\pi)$.

(iii) *The (empirical approx.) fixed point equation* (2.22) *has a unique solution given by (see* (2.11)*)*

$$\phi_\varepsilon^{(N)}(x) = \frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^{N} k_\varepsilon^{(N)}(x, X^j)\Phi_j + \varepsilon(h(x) - \pi(h))$$

Based on the results in Prop. 2.2 and Prop. 2.5, the exact solution ϕ and the numerical solution $\phi_\varepsilon^{(N)}$ are both well-defined. The remaining task is to show the convergence of $\phi_\varepsilon^{(N)} \to \phi$ as $N \to \infty$ and $\varepsilon \to 0$. We break the convergence analysis into two parts, bias and variance:

$$\phi_\varepsilon^{(N)} \underset{\text{(variance)}}{\overset{N\uparrow\infty}{\longrightarrow}} \phi_\varepsilon \underset{\text{(bias)}}{\overset{\varepsilon\downarrow 0}{\longrightarrow}} \phi$$

Before describing the general result, it is useful to first introduce an example that helps illustrate the bias-variance trade-off in this problem.

2.4.1 Example - the scalar case

In the scalar case (where $d = 1$), the Poisson equation is:

$$-\frac{1}{\rho(x)}\frac{\mathrm{d}}{\mathrm{d}x}(\rho(x)\frac{\mathrm{d}\phi}{\mathrm{d}x}(x)) = h(x) - \hat{h}$$

Integrating twice yields the solution explicitly

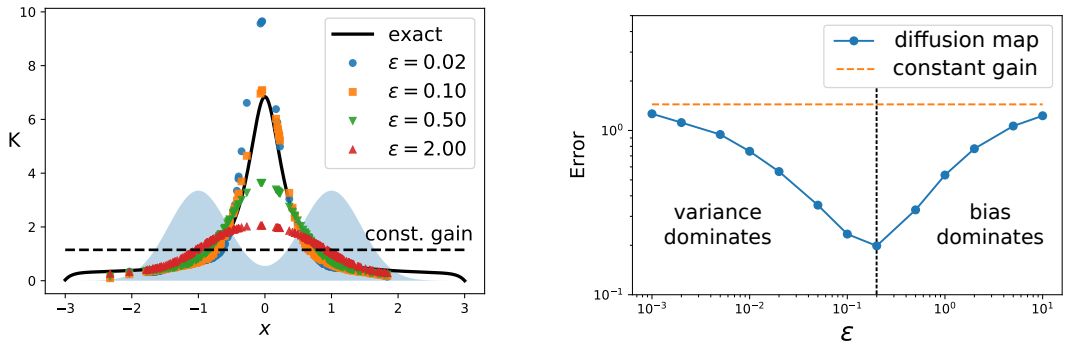
$$\mathsf{K}_{\text{exact}}(x) = \frac{\mathrm{d}\phi}{\mathrm{d}x}(x) = -\frac{1}{\rho(x)}\int_{-\infty}^{x} \rho(z)(h(z) - \hat{h})\mathrm{d}z \tag{2.25}$$

For the choice of ρ as the sum of two Gaussians $\mathcal{N}(-1, \sigma^2)$ and $\mathcal{N}(+1, \sigma^2)$ with $\sigma^2 = 0.2$ and $h(x) = x$, the solution obtained using (2.25) is depicted in Fig. 2.1 (a). Also depicted is the approximate solution obtained using the diffusion map algorithm with $N = 200$. As $\varepsilon \to \infty$ the approximate gain converges to the constant gain approximation. As ε becomes smaller, the approximation becomes more accurate. However, for very small values of ε the approximation is poor due to the variance error.

The bias-variance trade-off while varying the the parameter ε is depicted in Fig. 2.1 (b). The L^2 error is computed as a Monte-Carlo average:

$$\text{error} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}|\mathsf{K}^{(m)}(X^i) - \mathsf{K}_{\text{exact}}(X^i)|^2 \tag{2.26}$$

Fig. 2.1 (b) depicts the error obtained from averaging over $M = 1000$ simulations as a function of the parameter ε. It is observed that for a fixed number of particles N, there is an optimal value of ε that



(a) (b)

Figure 2.1: Simulation results for the diffusion-map algorithm for the scalar bimodal example: (a) Approximate gain function for different choices of ε compared to the exact gain function (solid line). The shaded area in the background is the bimodal probability density function ρ. The dashed line is the constant gain approximation solution; (b) Gain function approximation error of the diffusion-map algorithm as a function of the parameter ε. All the results are with $N = 200$ particles.

minimizes the error.

The vector counterpart of this example appears in Sec. 2.5.1.

2.4.2 Bias

The analysis of bias has two parts:

1. To show that the (diffusion map) fixed-point equation (2.21) admits a unique solution ϕ_ε for *all* positive choices of ε;

2. To show that $\phi_\varepsilon \to \phi$ as $\varepsilon \downarrow 0$.

For $n \in \mathbb{N}$, iterate the fixed-point equation (2.21) n times to obtain:

$$\phi_\varepsilon = T_\varepsilon^n \phi_\varepsilon + \sum_{k=0}^{n-1} \varepsilon T_\varepsilon^k (h - \hat{h}_\varepsilon) \tag{2.27}$$

We let $\varepsilon = \frac{t}{n}$ for some $t > 0$ and study the solution of this fixed-point equation as $n \to \infty$. Note that the solution to the iterated fixed-point equation (2.27) is identical to the solution to the fixed-point equation (2.21).

The fixed-point equation (2.27) is the (discrete) Poisson equation that appears in the theory of Markov chain simulation [Glynn and Meyn, 1996, Meyn and Tweedie, 2009] and stochastic control [Meyn, 2008, Ch. 9]. Theory presented in these references illustrates how bounds on the solution are obtained under a Foster-Lyapunov drift condition. A similar strategy is adopted here.

In the following proposition, an existence-uniqueness result is described for the fixed-point equation (2.27). The technical step in the proof involves a Foster-Lyapunov condition known as DV(3) [Kontoyiannis et al., 2005]. The proof appears in Appendix. 2.6.6.

Proposition 2.6. *Consider the family of Markov operators* $\{T_\varepsilon\}_{\varepsilon>0}$ *defined in* (2.7). *Let* $n \in \mathbb{N}$, $t \in (0, t_0)$, *and* $\varepsilon = \frac{t}{n}$, *with* $t_0 < \infty$. *Then there exists positive constants a, b, R,* δ, *a probability measure* ν, *and a number* $n_0 \in \mathbb{N}$ *such that for all* $n > n_0$:

$$\log(e^{-U_\varepsilon} T_\varepsilon^n e^{U_\varepsilon}) \leq -atU_\varepsilon + bt \tag{2.28a}$$

$$T_\varepsilon^n \mathbf{1}_{[A]}(x) \geq \delta \nu(A) \quad \forall |x| \leq R, \quad \forall A \in \mathscr{B}(\mathbb{R}^d) \tag{2.28b}$$

Consequently,

(i) *The chain with transition kernel* T_ε^n *is geometrically ergodic with invariant density*

$$\rho_\varepsilon(x) := \frac{n_\varepsilon(x)\rho(x)}{\int n_\varepsilon(x)\rho(x)\mathrm{d}x} \tag{2.29}$$

(ii) T_ε^n *is reversible with respect to the density* ρ_ε *It admits a spectral gap as a linear operator* $T_\varepsilon^n : L_0^2(\rho_\varepsilon) \to L_0^2(\rho_\varepsilon)$ *that is uniform with respect to* ε. *The spectral gap is denoted as* λ.

(iii) *There exists a solution to* (2.27) *with the bound*

$$\|\phi_\varepsilon\|_{L^2(\rho_\varepsilon)} \leq \frac{t\|h\|_{L^2(\rho_\varepsilon)}}{\lambda}$$

The proof of the following main result appears in Appendix. 2.6.7.

Theorem 2.1. *Suppose the assumptions (A1)-(A2) hold for the density* ρ *and the function h, and* ϕ *denotes the exact solution of* (2.20). *Consider the approximation of this problem defined by the (diffusion map) fixed-point equation* (2.21). *For the approximate problem:*

1. ***Existence-Uniqueness:*** *For each fixed* $\varepsilon > 0$, *there exists a unique solution* ϕ_ε.
2. ***Convergence:*** *In the asymptotic limit as* $\varepsilon \to 0$

$$\|\phi_\varepsilon - \phi\|_{L^2(\rho_\varepsilon)} = O(\varepsilon) \tag{2.30}$$

2.4.3 Variance

The analysis of the variance concerns the (empirical) fixed-point equation (2.22) whose solution is denoted as $\phi_\varepsilon^{(N)}$. The parameter ε is assumed to be positive and fixed and N is assumed to be finite but large.

The existence-uniqueness of $\phi_\varepsilon^{(N)}$ has already been shown as part of Prop. 2.5. The convergence has only been shown only for the case where the density has a compact support.

Assumption A3: The distribution ρ has compact support given by $\Omega \subset \mathbb{R}^d$.

Theorem 2.2. *Suppose the assumptions (A2)-(A3) hold for the density* ρ *and the function h, and* ϕ_ε *denotes the solution of the (kernel) fixed-point equation* (2.21) *for a fixed positive parameter* ε. *Consider the*

approximation of this problem defined by the (empirical) fixed-point equation (2.22). For the approximate problem:

1. ***Existence-Uniqueness:*** *For each finite N, there exists (almost surely) a unique solution* $\phi_\varepsilon^{(N)}$.
2. ***Convergence:*** *The approximate solution* $\phi_\varepsilon^{(N)}$ *converges to the kernel solution* ϕ_ε

$$\lim_{N\to\infty} \|\phi_\varepsilon^{(N)} - \phi_\varepsilon\|_\infty = 0, \quad a.s \tag{2.31}$$

Remark 2.6. *(related work) The proof of the convergence* $\phi_\varepsilon^{(N)} \to \phi_\varepsilon$ *is based on similar results in the numerical analysis of integral equations on a grid [Anselone, 1971, Atkinson, 1976, Baker, 1977]. A related approach is used in [Von Luxburg et al., 2008] to show the consistency of spectral clustering.*

2.4.4 Relationship to the constant gain approximation

Although the convergence and error analysis pertains to the $\varepsilon \downarrow 0$ limit, an important property of the diffusion map approximation is that the numerical procedure yields a unique solution for arbitrary values of ε (see Prop. 2.5). In fact, more can be said: one recovers the constant gain approximation formula in the $\varepsilon \to \infty$ limit.

Before stating the result, it is useful to recall the three formulae for the gain:

(i) **Exact formula:** $\mathsf{K} = \nabla\phi$ is defined using the exact solution ϕ;

(ii) **Kernel formula:** K_ε is defined using the solution ϕ_ε to the (diffusion-map) approximation fixed-point equation:

$$\mathsf{K}_\varepsilon(x) := \nabla_x \left[\frac{1}{n_\varepsilon(x)} \int k_\varepsilon(x,y)(\phi_\varepsilon(y) + \varepsilon h(y))\rho_\varepsilon(y)\mathrm{d}y \right] \tag{2.32}$$

(iii) **Empirical formula:** $K_\varepsilon^{(N)}$ is the empirical version of the kernel formula. It was defined in (2.12) using the solution Φ of the finite-dimensional fixed-point problem.

The proof of the following Proposition appears in the Appendix. 2.6.10.

Proposition 2.7. *Consider the fixed-point problems (2.21) and (2.22) in the limit as* $\varepsilon \to \infty$.

(i) *The kernel formula of the gain is given by*

$$\lim_{\varepsilon\to\infty} \mathsf{K}_\varepsilon = \int (h(x) - \hat{h})\rho(x)\mathrm{d}x$$

(ii) *For any finite N, the empirical formula of the gain is given by*

$$\lim_{\varepsilon\to\infty} \mathsf{K}_\varepsilon^{(N)} = \frac{1}{N}\sum_{i=1}^N (h(X^i) - \hat{h}^{(N)})X^i \quad a.s.$$

This result serves to highlight the connection between the FPF and the EnKF: With the diffusion map approximation of the gain, the FPF approaches EnKF in the limit of large ε. The parameter ε can then be regarded as the tuning parameter to "improve" the gain. Of course, for any finite value of N, this can only be done up to a point – where variance becomes dominant (see Fig. 2.1).

2.5 Numerics

2.5.1 Example - the vector case

A vector generalization of the scalar example in Sec. 2.4.1 is obtained by considering the following form of the probability density function in d-dimensions:

$$\rho(x) = \rho_B(x_1) \prod_{n=2}^{d} \rho_G(x_n), \quad \text{for} \quad x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$$

where ρ_B is the bimodal distribution $\frac{1}{2}\mathcal{N}(-1,\sigma^2) + \frac{1}{2}\mathcal{N}(+1,\sigma^2)$ introduced in Sec. 2.4.1 , and ρ_G is the Gaussian distribution $\mathcal{N}(0,\sigma^2)$. Also suppose the function $h(x) = x_1$. The simple example is illustrative of realistic application scenarios where the density has non-Gaussian features along certain (not necessarily apriori known) low-dimensional subspace. The directions orthogonal to this subspace are modelled here as Gaussian noise.

For this problem, the exact gain function is easily obtained as

$$\mathsf{K}_{\text{exact}}(x) = (\mathsf{K}_{\text{exact}}(x_1), 0, \ldots, 0)$$

where the function $\mathsf{K}_{\text{exact}}(x_1)$ is given by the formula (2.25) in Sec. 2.4.1. The exact solution is used to compute error properties as dimension increases.

The diffusion map algorithm (Table 2.1) is simulated to approximate the gain function for this problem. For each particle $X^i = (X_1^i, \ldots, X_d^i)$, the first coordinate $X_1^i \overset{\text{i.i.d}}{\sim} \frac{1}{2}\mathcal{N}(-1,\sigma^2) + \frac{1}{2}\mathcal{N}(+1,\sigma^2)$ and other the coordinates $X_n^i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\sigma^2)$ for $n = 2, \ldots, d$.

Fig. 2.2 depicts the Monte-Carlo error (2.26) computed from running $M = 100$ simulations. A summary of these results is as follows:

1. Fig. 2.2-(a) depicts the M.C. error as a function of the parameters ε and d for a fixed number of particles $N = 1000$. Also depicted is the error with the constant gain approximation. The constant gain error serves here as baseline.

   For large values of ε, the error asymptotes to the error for the constant-gain approximation. This is because (see Prop. 2.7) the kernel gain approaches the constant gain as $\varepsilon \to \infty$.

   The other aspect to note is the bias-variance trade-off first illustrated in Fig. 2.1 for the scalar case. As the dimension increases, the error due to the variance becomes dominant at relatively larger values of ε.

2. Fig. 2.2-(b) depicts the bias-variance trade-off as a function of number of particles N for the fixed $d = 1$. It is not a surprise that the error gets better, for all choices of ε, as the number of particles increase. However, the optimal value of ε – at which the error is the smallest – is relatively insensitive to changes in N.

3. Fig. 2.2-(c) depicts the error as function of N for different values of ε. The dimension $d = 1$ is fixed. The error goes down as $O(\frac{1}{N})$ and asymptotes to the $O(\varepsilon)$ bias. The $O(\frac{1}{N})$ is a LLN type estimate and $O(\varepsilon)$ bias error is consistent with the conclusion of the Thm. 2.1.

4. Fig. 2.2-(d) depicts the run time comparison between the diffusion-map algorithm and the constant gain algorithm. The scaling for the diffusion-map algorithm is $O(N^2)$ which is significantly more expensive than the $O(N)$ scaling of the constant gain approximation.

Remark 2.7 (Selection of ε)**.** *The numerical results, in 2.2, suggest that there is an optimal value of ε that minimizes the error. However, in practice, computing the optimal value of ε for different problems is difficult. Instead, in the literature involving kernel methods, it is proposed to set $\varepsilon = \frac{4med^2}{\log(N)}$ where med is the median value of all pairwise distances $\{\|X^i - X^j\|\}_{i\neq j}$ [Chaudhuri et al., 2017]. The justification is that, with such a choice, the $N \times N$ matrix whose i,j-th entry is $g_\varepsilon(X^i, X^j)$ is relatively away from the identity matrix.*

2.5.2 Filtering example

Consider the following filtering problem:

$$\begin{aligned} \mathrm{d}X_t &= 0, \quad X_0 \sim p_0 \\ \mathrm{d}Z_t &= h(X_t)\mathrm{d}t + \sigma_w \mathrm{d}W_t \end{aligned}$$

where $X_t \in \mathbb{R}$, $Z_t \in \mathbb{R}$, $\sigma_W > 0$, and $\{W_t\}$ is standard Brownian motion, independent of X_t. The prior distribution p_0 is Gaussian $\mathcal{N}(0,1)$ and the observation function $h(x) = |x|$. For the static filtering problem, the posterior distribution is explicitly given by:

$$p_t^\star(x) = (\text{const.}) p_0(x) \exp\left(\frac{1}{\sigma_w^2}(h(x)Z_t - \frac{1}{2}h^2(x)t)\right)$$

For comparative purposes, the FPF algorithm with the diffusion-map gain approximation and the constant gain approximation are implemented. With the latter approximation, the FPF is an EnKF algorithm. The simulation parameters are as follows: The measurement noise $\sigma_w = 0.1$. The simulation is carried out for $T = 500$ time-steps with step-size $\Delta t = 0.001$. Both the algorithms use $N = 200$ particles with identical initialization. For the diffusion-map approximation, the kernel bandwidth $\varepsilon = 0.1$.

The numerical results are depicted in Fig. 2.3. The distribution of the particles along with the exact posterior distribution are depicted in Fig. 2.3-(a). It is observed that the FPF algorithm with the diffusion

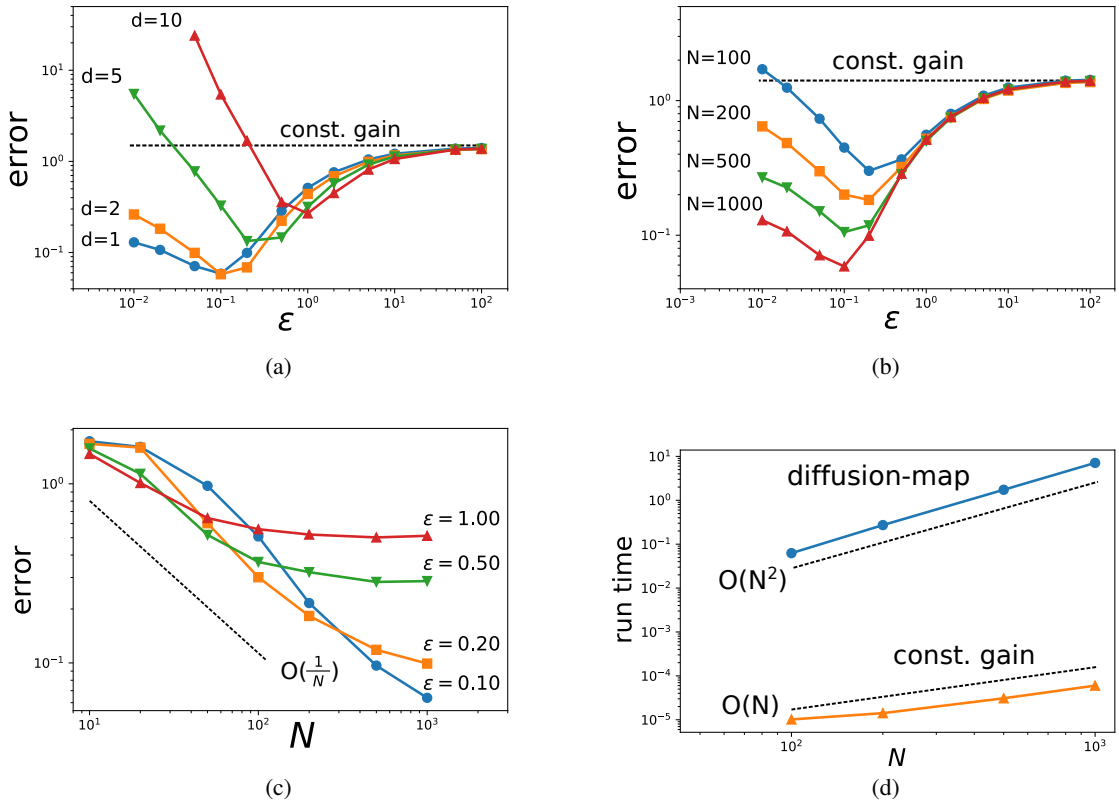


Figure 2.2: Simulation results for the diffusion-map algorithm for the vector bimodal example: (a) Gain function approximation error as a function of ε for $d \in \{1,2,5,10\}$. (b) Error as a function of ε for $N \in \{100,200,500,1000\}$. (c) Error as a function of N for $\varepsilon \in \{0.1,0.2,0.5,1.0\}$; (d) Comparison of the run-time

map approximation provides a more accurate approximation of the posterior distribution. In contrast, the constant-gain approximation fails to reproduce the bimodal nature of the posterior distribution.

A quantitative estimate of the performance is provided in terms of a mean squared error (m.s.e.). in estimating the conditional expectation of the function $\psi(x) = x\mathbf{1}_{[x\leq 0]}$. A Monte Carlo estimate of the m.s.e. is depicted in Fig. 2.3-(b) with $M = 100$ runs. At time t, it is calculated according to

$$\text{m.s.e.}_t = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N}\psi(X_t^{m,i}) - \int \psi(x)p_t^\star(x)\mathrm{d}x\right)^2$$

At time $t = 0$, the empirical distribution of the particles is an accurate approximation of the prior distribution, because the particles are sampled i.i.d. from the prior distribution. Therefore, the m.s.e at $t = 0$ is small. As time progress, the difference between the empirical distribution and the exact posterior becomes larger because the filter update is not exact. As the time-step Δt is small, the main source of the m.s.e. error is due to the error in the gain function approximation. Therefore, the diffusion map FPF with its more accurate approximation of the gain yields better m.s.e., compared to the EnKF using the constant gain approximation.

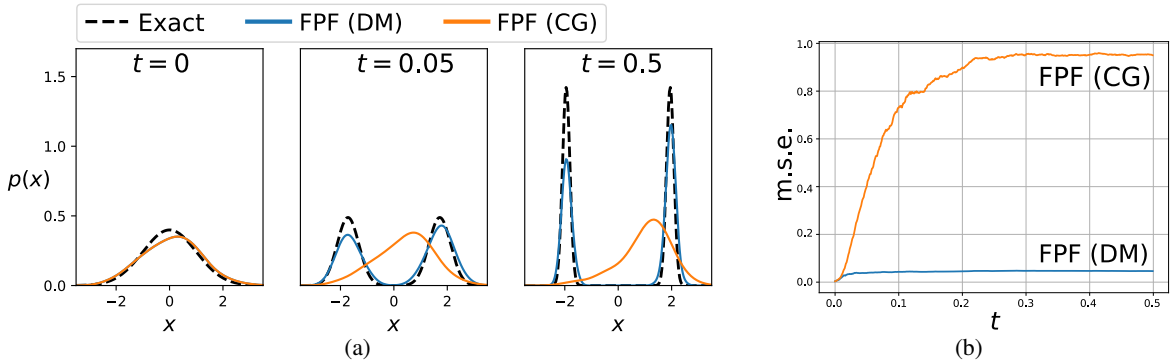


Figure 2.3: Simulation results for the FPF algorithm for the filtering example: (a) The distribution of the particles obtained using the diffusion-map approximation and the constant gain approximation as compared to the exact distribution (dashed line); (b) Plot of the mean squared error in estimating the conditional expectation of the function $\psi(x) = x\mathbf{1}_{[x<0]}$.

2.6 Proof of the main results

2.6.1 Exact semigroup and and its diffusion map approximation for the Gaussian case

In this section, we provide explicit formulae for the exact semigroup P_t and its diffusion map approximation T_ε, for the special case when the density ρ is a Gaussian $\mathcal{N}(m,\Sigma)$. For the Gaussian case, the semigroup is the Ornstein-Uhlenbeck semigroup [Bakry et al., 2013, Sec. 2.7.1] an its spectral representation is obtained in terms of the Hermite polynomials. For notational ease, after an appropriate change of coordinates, we assume $m = 0$ and $\Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)$ where $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_d^2 > 0$ are ordered eigenvalues of Σ.

Definition 2.1. *The* Hermite polynomials *are recursively defined as*

$$\hbar_{n+1}(x) = x\hbar_n(x) - \hbar_n'(x), \quad \hbar_0(x) = 1,$$

where the prime $'$ denotes the derivative.

Proposition 2.8. *Suppose the density ρ is Gaussian $\mathcal{N}(0,\Sigma)$ with the variance $\Sigma = diag(\sigma_1^2,\ldots,\sigma_d^2)$ and $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_d^2 > 0$. Then*

(i) *The exact semigroup P_t and the diffusion map T_ε admit the following integral representations:*

$$P_t f(x) = \int_{\mathbb{R}^d} \prod_{j=1}^d \frac{1}{(2\pi\sigma_j^2(1-e^{-2t\sigma_j^{-2}}))^{1/2}} \exp\Big(-\frac{|y_j - e^{-t\sigma_j^{-2}}x_j|^2}{2\sigma_j^2(1-e^{-2t\sigma_j^{-2}})}\Big) f(y)\mathrm{d}y \tag{2.33}$$

$$T_\varepsilon f(x) = \int_{\mathbb{R}^d} \prod_{j=1}^d \frac{1}{(4\pi\varepsilon(1-\delta_j))^{1/2}} \exp\Big(-\frac{|y_j - (1-\delta_j)x_j|^2}{4\varepsilon(1-\delta_j)}\Big) f(y)\mathrm{d}y \tag{2.34}$$

where $\delta_j := \varepsilon\frac{\sigma_j^2+4\varepsilon}{\sigma_j^4+3\sigma_j^2\varepsilon+4\varepsilon^2}$ for $j = 1,\ldots,d$.

(ii) *The operators P_t and T_ε each have a unique invariant Gaussian density given by $\mathcal{N}(0,\Sigma)$ and $\mathcal{N}(0,\Sigma_\varepsilon)$, respectively, where $\Sigma_\varepsilon = diag(\sigma_{\varepsilon,1}^2,\ldots,\sigma_{\varepsilon,d}^2)$ with $\sigma_{\varepsilon,j}^2 = \frac{2\varepsilon(1-\delta_j)}{\delta_j(2-\delta_j)}$ for $j = 1,\ldots,d$.*

(iii) *The eigenvalues and the associated eigenfunctions are as follows:*

$$\textit{Spectrum of the semigroup } P_t: \quad \lambda_n = \prod_{j=1}^d e^{-t\frac{n_j}{\sigma_j^2}}, \quad e_n(x) = \prod_{j=1}^d \hbar_{n_j}\Big(\frac{x_j}{\sigma_j}\Big)$$

$$\textit{Spectrum of the diffusion map } T_\varepsilon: \quad \lambda_n = \prod_{j=1}^d (1-\delta_j)^{n_j}, \quad e_n(x) = \prod_{j=1}^d \hbar_{n_j}\Big(\frac{x_j}{\sigma_{\varepsilon,j}}\Big)$$

for $n = (n_1,\ldots,n_d) \in \mathbb{Z}_+^d$.

(iv) *The operator norm $\|P_t\|_{L^2(\rho)} = e^{-\frac{t}{\sigma_1^2}}$ and $\|T_\varepsilon\|_{L^2(\rho_\varepsilon)} = 1-\delta_1$.*

Proof. (i) The explicit formula (2.33) for the exact semigroup is given in [Bakry et al., 2013, Sec. 2.7]. The explicit formula (2.34) is obtained by evaluating the definition (2.7) for the Gaussian

case. Consider first the simpler scalar case. Let $\rho_G(x-m;\sigma^2)$ denote the probability density function for the Gaussian distribution $\mathcal{N}(m,\sigma^2)$. Then by the convolution property

$$\int_{\mathbb{R}^d} g_\varepsilon(x,y)\rho_G(y;0,\sigma^2)\mathrm{d}y = \rho_G(x;0,\sigma^2+2\varepsilon)$$

Hence

$$\begin{aligned}
n_\varepsilon(x) &= \frac{1}{\sqrt{\rho_G(x;0,\sigma^2+2\varepsilon)}}\int g_\varepsilon(x,y)\frac{\rho_G(y;0,\sigma^2)}{\sqrt{\rho_G(y;0,\sigma^2+2\varepsilon)}}\mathrm{d}y \\
&= \frac{1}{\sqrt{4\pi\varepsilon}}\frac{\sqrt{2\pi(\sigma^2+2\varepsilon)}}{\sqrt{2\pi\sigma^2}}\frac{1}{e^{-\frac{x^2}{4(\sigma^2+2\varepsilon)}}}\int e^{-\frac{|y-x|^2}{4\varepsilon}}\frac{e^{-\frac{y^2}{2\sigma^2}}}{e^{-\frac{y^2}{4(\sigma^2+2\varepsilon)}}}\mathrm{d}y \\
&= \frac{1}{\sqrt{4\pi\varepsilon}}\frac{\sqrt{(\sigma^2+2\varepsilon)}}{\sqrt{\sigma^2}}e^{\frac{x^2}{4(\sigma^2+2\varepsilon)}}\int e^{-\frac{|y-\frac{a}{2\varepsilon}x|^2}{2a}}e^{\frac{ax^2}{8\varepsilon^2}-\frac{x^2}{4\varepsilon}}\mathrm{d}y \\
&= \frac{1}{\sqrt{4\pi\varepsilon}}\frac{\sqrt{(\sigma^2+2\varepsilon)}}{\sqrt{\sigma^2}}e^{\frac{x^2}{4(\sigma^2+2\varepsilon)}+\frac{ax^2}{8\varepsilon^2}-\frac{x^2}{4\varepsilon}}\sqrt{2\pi a}
\end{aligned}$$

where $\frac{1}{a} = \frac{1}{2\varepsilon}+\frac{1}{\sigma^2}-\frac{1}{2(\sigma^2+2\varepsilon)}$.

Now, using the definition (2.7),

$$\begin{aligned}
T_\varepsilon f(x) &= \frac{1}{n_\varepsilon(x)}\int g_\varepsilon(x,y)\frac{\rho_G(y;0,\sigma^2)}{\sqrt{\rho_G(x;0,\sigma^2+2\varepsilon)}\sqrt{\rho_G(y;0,\sigma^2+2\varepsilon)}}f(y)\mathrm{d}y \\
&= \frac{1}{n_\varepsilon(x)}\frac{1}{\sqrt{4\pi\varepsilon}}\frac{\sqrt{(\sigma^2+2\varepsilon)}}{\sqrt{\sigma^2}}e^{\frac{x^2}{4(\sigma^2+2\varepsilon)}}\int e^{-\frac{|y-\frac{a}{2\varepsilon}x|^2}{2a}}e^{\frac{ax^2}{8\varepsilon^2}-\frac{x^2}{4\varepsilon}}f(y)\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi a}}\int e^{-\frac{|y-\frac{a}{2\varepsilon}x|^2}{2a}}f(y)\mathrm{d}y
\end{aligned}$$

Writing $\frac{a}{2\varepsilon} = 1-\delta$ where $\delta = \varepsilon\frac{\sigma^2+4\varepsilon}{\sigma^4+3\varepsilon\sigma^2+4\varepsilon^2}$ gives the following succinct formula for the scalar case:

$$T_\varepsilon f(x) = \frac{1}{\sqrt{4\pi\varepsilon(1-\delta)}}\int e^{-\frac{|y-(1-\delta)x|^2}{4\varepsilon(1-\delta)}}f(y)\mathrm{d}y$$

The extension to the vector case is straightforward. By Assumption, the covariance matrix is diagonal. Hence, one can write the Gaussian density as a product of Gaussian marginals. Also, the Gaussian kernel $g_\varepsilon(x,y)$ can be expressed as a product of Gaussian kernels for the coordinates. Therefore, the kernel for the vector case is obtained as a product of kernels for the coordinates.

(ii) The invariant probability density and the spectrum are obtained by comparison of the formula (2.34) for the approximate kernel to the formula (2.33) for the exact kernel. Indeed, define

the parameters τ_j and $\sigma_{\varepsilon,j}$ as follows:

$$1-\delta_j = e^{-\tau_j \sigma_{j,\varepsilon}^{-2}}$$
$$2\varepsilon(1-\delta_j) = \sigma_{\varepsilon,j}^2(1-e^{-2\tau_j\sigma_{\varepsilon,j}^{-2}})$$

for $j=1,\ldots,d$. Then one can express

$$T_\varepsilon = \prod_{j=1}^{d} e^{\tau_j \Delta_{q_j}}$$

where Δ_{q_j} is the one-dimensional weighted Laplacian for the $j-$th coordinate x_j, and q_j is a Gaussian $\mathcal{N}(0,\sigma_{\varepsilon,j}^2)$. Hence, the formulae for the invariant probability density and the spectrum follow from explicit results known for $e^{\tau_j\Delta_{q_j}}$ from [Bakry et al., 2013]:

$$\lambda_n = \prod_{j=1}^{d} e^{-t_j \frac{n_j}{\sigma_{\varepsilon,j}^2}} = \prod_{j=1}^{d}(1-\delta_j)^{n_j}, \quad e_n = \prod_{j=1}^{d} \hbar_{n_j}(\frac{x_j}{\sigma_{\varepsilon,j}})$$

(iii) The operator norm $\|T_\varepsilon\|_{L_0^2(\rho_\varepsilon)}$ is given by the maximum non-trivial eigenvalue of T_ε which is equal to $1-\min_j(\delta_j)$.

□

2.6.2 Proof of Prop. 2.2

Based on the use of the spectral representation (2.4), the weak solution of the Poisson equation is readily seen to be

$$\phi = \sum_{m=1}^{\infty} \frac{1}{\lambda_m}\langle e_m, h\rangle e_m \tag{2.35}$$

This solution (2.35) also satisfies the fixed-point equation (2.6) because

$$\begin{aligned} P_t\phi + \int_0^t P_s(h-\hat{h})\mathrm{d}s &= \sum_{m=1}^{\infty} e^{-t\lambda_m}\langle e_m,\phi\rangle e_m + \int_0^t \sum_{m=1}^{\infty} e^{-s\lambda_m}\langle e_m,h\rangle e_m \mathrm{d}s \\ &= \sum_{m=1}^{\infty} \frac{e^{-t\lambda_m}}{\lambda_m}\langle e_m,h\rangle e_m + \sum_{m=1}^{\infty}\frac{1-e^{-t\lambda_m}}{\lambda_m}\langle e_m,h\rangle e_m = \phi \end{aligned}$$

The uniqueness of the solution to the fixed-point equation (2.6) follows from the contraction mapping principle because $\|P_t\|_{L_0^2(\rho)} = e^{-t\lambda_1} < 1$.

2.6.3 Derivation of the linear form of the gain (2.13)

By a direct calculation,

$$\nabla_x \frac{k_\varepsilon^{(N)}(x,X^j)}{\sum_{l=1}^N k_\varepsilon^{(N)}(x,X^l)} = \frac{\frac{X^j-x}{2\varepsilon} k_\varepsilon^{(N)}(x,X^j)}{\sum_l k_\varepsilon^{(N)}(x,X^l)} - \frac{\sum_{l=1}^N \frac{X^l-x}{2\varepsilon} k_\varepsilon^{(N)}(x,X^l)}{\sum_l k_\varepsilon^{(N)}(x,X^l)} \frac{k_\varepsilon^{(N)}(x,X^j)}{\sum_l k_\varepsilon^{(N)}(x,X^l)}$$

which evaluated at $x = X^i$ yields

$$\nabla_x \left(\frac{k_\varepsilon^{(N)}(x,X^i)}{\sum_{j=1}^N k_\varepsilon(x,X^j)} \right) \Bigg|_{x=X^i} = \frac{1}{2\varepsilon} \left(X^j \mathsf{T}_{ij} - \sum_{l=1}^N X^l \mathsf{T}_{il} \mathsf{T}_{ij} \right)$$

Using the definitions (2.12) for $\mathsf{K}_\varepsilon^{(N)}$, and (2.14) for r and s,

$$\begin{aligned}
\mathsf{K}_\varepsilon^{(N)}(X^i) &= \nabla_x \left(\frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^N k_\varepsilon^{(N)}(x,X^j)(\Phi_j + \varepsilon \mathsf{h}_j) \right) \Bigg|_{x=X^i} \\
&= \nabla_x \left(\frac{\sum_{j=1}^N k_\varepsilon^{(N)}(x,X^j) r_j}{\sum_{j=1}^N k_\varepsilon^{(N)}(x,X^j)} \right) \Bigg|_{x=X^i} \\
&= \frac{1}{2\varepsilon} \left(\sum_{j=1}^N X^j \mathsf{T}_{ij} (r_j - \sum_{l=1}^N \mathsf{T}_{il} r_l) \right) = \sum_{j=1}^N s_{ij} X^j
\end{aligned}$$

2.6.4 Proof of Prop. 2.5

1. T is a Markov matrix because $\mathsf{T}_{ij} = \frac{1}{n_\varepsilon^{(N)}(X^i)} k_\varepsilon^{(N)}(X^i,X^j) > 0$ a.s. and

$$\sum_{j=1}^N \mathsf{T}_{ij} = \frac{1}{n_\varepsilon^{(N)}(X^i)} \sum_{i=1}^N k_\varepsilon^{(N)}(X^i,X^j) = \frac{n_\varepsilon^{(N)}(X^i)}{n_\varepsilon^{(N)}(X^i)} = 1$$

   The stationary distribution is π because

$$\begin{aligned}
\sum_{i=1}^N \pi_i \mathsf{T}_{ij} &= \sum_{i=1}^N \frac{n_\varepsilon^{(N)}(X^i)}{\sum_{k=1}^N n_\varepsilon^{(N)}(X^k)} \frac{k_\varepsilon^{(N)}(X^i,X^j)}{n_\varepsilon^{(N)}(X^i)} \\
&= \frac{\sum_{i=1}^N k_\varepsilon^{(N)}(X^i,X^j)}{\sum_{k=1}^N n_\varepsilon^{(N)}(X^k)} = \frac{n_\varepsilon^{(N)}(X^j)}{\sum_{k=1}^N n_\varepsilon^{(N)}(X^k)} = \pi_j
\end{aligned}$$

All entries of the Markov matrix are positive. Hence the Markov chain is irreducible and aperiodic. Therefore, the stationary distribution is unique. It is reversible because

$$\begin{aligned}\pi_i \mathsf{T}_{ij} &= \frac{n_\varepsilon^{(N)}(X^i)}{\sum_{k=1}^N n_\varepsilon^{(N)}(X^k)} \frac{k_\varepsilon^{(N)}(X^i,X^j)}{n_\varepsilon^{(N)}(X^i)} = \frac{k_\varepsilon^{(N)}(X^j,X^i)}{\sum_{k=1}^N n_\varepsilon^{(N)}(X^k)} \\ &= \frac{n_\varepsilon^{(N)}(X^j)}{\sum_{k=1}^N n_\varepsilon^{(N)}(X^k)} \frac{k_\varepsilon^{(N)}(X^j,X^i)}{n_\varepsilon^{(N)}(X^j)} = \pi_j \mathsf{T}_{ji}\end{aligned}$$

2. Denote $\delta := \min_{ij} \mathsf{T}_{ij}$. Then $\delta > 0$ a.s. Therefore, $\|\mathsf{T}\|_{L_0^2(\pi)} \le 1 - \frac{N\delta}{2} < 1$, and is thus a contraction on $L_0^2(\pi)$ [Stroock, 2013, Ch. 5]). It follows, from the contraction mapping principle, that the fixed point equation (2.10) has a unique solution.

3. Evaluating the definition (2.11) at $x = X^i$ concludes $\phi_\varepsilon^{(N)}(X^i) = \Phi_i$ because,

$$\begin{aligned}\phi_\varepsilon^{(N)}(X^i) &= \frac{1}{n_\varepsilon^{(N)}(X^i)} \sum_{j=1}^n k_\varepsilon^{(N)}(X^i,X^j)\Phi_j + \varepsilon(h(X^i) - \pi(h)) \\ &= \sum_{j=1}^N \mathsf{T}_{ij}\Phi_j + \varepsilon(\mathsf{h}_i - \pi(h)) = \Phi_i\end{aligned}$$

Therefore $\phi_\varepsilon^{(N)}$ solves the fixed-point equation (2.22), because

$$\begin{aligned}T_\varepsilon^{(N)}\phi_\varepsilon^{(N)}(x) &= \frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^n k_\varepsilon^{(N)}(x,X^j)\phi_\varepsilon^{(N)}(X^j) \\ &= \frac{1}{n_\varepsilon^{(N)}(x)} \sum_{j=1}^n k_\varepsilon^{(N)}(x,X^j)\Phi_j \\ &\overset{(2.11)}{=} \phi_\varepsilon^{(N)}(x) - \varepsilon(h(x) - \pi(h))\end{aligned}$$

2.6.5 Proof of the Prop. 2.3

Proof. (i) Let $U = -\frac{1}{2}\log(\rho)$ and $W = |\nabla U|^2 - \Delta U$ as defined in (2.15a) (2.15b). To obtain the representation (2.16) for the semigroup P_t, consider the unitary transformation [Bakry et al., 2013, Sec. 1.15.7]:

$$e^{-U}\Delta_\rho\, e^{U} = \Delta - W \tag{2.36}$$

Therefore, for any function $f \in C_b(\mathbb{R}^d)$,

$$e^{-U}\, P_t\, e^{U}(f) = e^{t(\Delta - W)}(f) = \mathsf{E}[e^{-\int_0^t W(B_{2s}^x)\mathrm{d}s} f(B_{2t}^x)]$$

where the stochastic representation (second equality) follows from the Feynman-Kac formula; B_t^x is a Brownian motion initialized at x. Setting $f(x) = e^{-U(x)}g(x)$,

$$P_t g(x) = e^{U(x)} e^{t(\Delta - W)}(e^{-U} g)(x) = e^{U(x)} \mathsf{E}[e^{-\int_0^t W(B_{2s}^x)\mathrm{d}s} e^{-U(B_{2t}^x)} g(B_{2t}^x)]$$

which is the representation (2.16).

Next, the representation (2.17) is obtained. Using the definitions, (2.7) of T_ε and (2.15a) (2.15b) of U_ε and W_ε,

$$\begin{aligned} T_\varepsilon f(x) = \frac{G_\varepsilon(f e^{-U_\varepsilon})(x)}{G_\varepsilon(e^{-U_\varepsilon})(x)} &= e^{U_\varepsilon(x) - \varepsilon W_\varepsilon(x)} G_\varepsilon(e^{-U_\varepsilon} f)(x) \\ &= e^{U_\varepsilon(x) - \varepsilon W_\varepsilon(x)} \mathsf{E}[e^{-U_\varepsilon(B_{2\varepsilon}^x)} f(B_{2\varepsilon}^x)] \end{aligned}$$

where the final equality follows from using the stochastic representation of the heat semigroup G_ε. The representation (2.17) is obtained by iterating this formula n times.

(ii) Without loss of generality, upon a change of coordinates, assume $m = 0$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ in Assumption A1. Using the definitions

$$U_\varepsilon(x) = U(x) - \frac{1}{2}\log(\rho(x)) + \frac{1}{2}\log(G_\varepsilon(\rho)(x)) \tag{2.37}$$

Now, $\log(\rho(x)) = \log(\rho_g(x;\Sigma)) + w(x)$. So, the main calculation is to approximate $\log(G_\varepsilon(\rho))$. Using the definition

$$\begin{aligned} G_\varepsilon(\rho)(x) &= \int_{\mathbb{R}^d} g_\varepsilon(x,y)\rho_g(y;\Sigma)e^{-w(y)}\mathrm{d}y \\ &= \int_{\mathbb{R}^d} \frac{e^{-\sum_{n=1}^d \frac{|x_n - y_n|^2}{4\varepsilon}}}{(4\pi\varepsilon)^{d/2}} \frac{e^{-\sum_{n=1}^d \frac{|y_n|^2}{2\sigma_n^2} - w(y)}}{\prod_{n=1}^d (2\pi\sigma_n^2)^{1/2}}\mathrm{d}y \\ &= \frac{e^{-\frac{1}{2}\sum_{n=1}^d \frac{|x_n|^2}{2(\sigma_n^2 + 2\varepsilon)}}}{\prod_{n=1}^d (2\pi(\sigma_n^2 + 2\varepsilon))^{1/2}} \int_{\mathbb{R}^d} \frac{e^{-\sum_{n=1}^d \frac{|y_n - (1-\delta_n)x_n|^2}{4\varepsilon(1-\delta_n)}}}{\prod_{n=1}^d (4\pi\varepsilon(1-\delta_n))^{1/2}} e^{-w(y)}\mathrm{d}y \\ &= \rho_g(x;\Sigma + 2\varepsilon I) G_\varepsilon^{(\delta)}(e^{-w})((I-\delta)x) \end{aligned}$$

where $\delta_n = \frac{2\varepsilon}{\sigma_n^2 + 2\varepsilon}$, $\delta = \mathrm{diag}(\delta_1, \ldots, \delta_d)$ and $G_\varepsilon^{(\delta)}$ is the semigroup associated with the pde $\frac{\partial}{\partial t} G_t^{(\delta)} f = G_t^{(\delta)}(\mathrm{tr}((I-\delta)\nabla^2 f))$.

The Taylor expansion of $G_\varepsilon^{(\delta)}(e^{-w})$, about $\varepsilon = 0$, is expressed as

$$G_\varepsilon^{(\delta)}(e^{-w})(x) = e^{-w(x)} + \varepsilon \mathrm{Tr}((I-\delta)\nabla^2 e^{-w})(x) + \underbrace{\int_0^\varepsilon \int_0^\tau (\sum_{m,n=1}^d (1-\delta_m)^2(1-\delta_n)^2 G_s^{(\delta)} \partial_n^2 \partial_m^2 e^{-w})(x) \mathrm{d}s\mathrm{d}\tau}_{\varepsilon^2 r_\varepsilon(x)}$$

where $\partial_m^2 := \frac{\partial^2}{\partial x_n^2}$.

Using the property that $G_s^{(\delta)} \partial_n f = \partial_n G^{(\delta)} f$, $\|G_t^{(\delta)}(f)\|_{L^\infty} \leq \|f\|_{L^\infty}$ and the assumption (A1) that $w \in C_b^\infty(\mathbb{R}^d)$, we conclude that $r_\varepsilon \in C_b^\infty(\mathbb{R}^d)$. Therefore,

$$\log(G_\varepsilon \rho(x)) = \log(\rho_g(x; \Sigma + 2\varepsilon I)) - \underbrace{(w - \log(1 + \varepsilon \mathrm{tr}((I-\delta)e^w \nabla^2 e^{-w}) + \varepsilon^2 e^w r_\varepsilon))|_{(I-\delta)x}}_{w_\varepsilon^{(1)}(x)}$$

The asymptotic expansion of $w_\varepsilon^{(1)}$, as $\varepsilon \to 0$, is obtained as

$$w_\varepsilon^{(1)}(x) = w(x) - 2\varepsilon x^\top \Sigma^{-1} \nabla w(x) - \varepsilon e^w \Delta e^{-w}(x) + O(\varepsilon^2)$$

where the remainder term has at most linear growth as $|x| \to \infty$.

Substituting the asymptotic expression for $\log(G_\varepsilon \rho(x))$ in (2.37),

$$\begin{aligned} U_\varepsilon(x) &= U(x) - \frac{1}{2}\log(\rho_g(x;\Sigma)) + \frac{1}{2}w(x) + \frac{1}{2}\log(\rho_g(x;\Sigma+2\varepsilon I)) - \frac{1}{2}w_\varepsilon^{(1)}(x) \\ &= U(x) + \frac{\varepsilon}{2} x^\top \Sigma^{-1}(\Sigma + 2\varepsilon I)^{-1} x - \frac{\varepsilon}{2}\mathrm{Tr}(\Sigma^{-1}) \\ &\quad + \varepsilon x^\top \Sigma^{-1} \nabla w(x) + \frac{\varepsilon}{2}(\|\nabla w(x)\| - 2^2 - \Delta w(x)) + O(\varepsilon^2) \\ &= U(x) + \underbrace{\frac{\varepsilon}{2}\|\Sigma^{-1}x + \nabla w(x)\|_2^2 - \frac{\varepsilon}{2}(\mathrm{Tr}(\Sigma^1) + \Delta w(x))}_{2\varepsilon W(x) + \frac{\varepsilon}{2}\Delta V(x)} + O(\varepsilon^2) \end{aligned}$$

where the remainder $O(\varepsilon^2)$ error term has at most quadratic growth as $\|x\|_2 \to \infty$. This concludes the proof of approximation (2.18a).

Based on this above calculation, the following estimate for an upper bound of the function U is obtained (it is used in the proof of Prop. 2.6):

$$\begin{aligned} U_\varepsilon(x) &\leq \frac{1}{4}x^\top \Sigma^{-1} x + \varepsilon(\|\Sigma^{-1}x\|_2^2 + \|\nabla w\|_{L^\infty}^2 + \|\Delta V\|_{L^\infty}) + \varepsilon^2(C_1\|x\|_2^2 + C_2) \\ &\leq \frac{1}{8\sigma_1^2}\|x\|_2^2 + \frac{\sigma_1^2}{8}(\|\nabla w\|_{L^\infty}^2 + \|\Delta V\|_{L^\infty} + \frac{C_2\sigma_1^2}{8}) \end{aligned} \tag{2.38}$$

where recall $\sigma_1^2 = \lambda_{\min}(\Sigma)$.

Next, the approximation (2.18b) is derived. Using the definition

$$\varepsilon W_\varepsilon(x) = U_\varepsilon(x) + \log(G_\varepsilon e^{-U_\varepsilon}(x))$$

By repeating the steps, just used to approximate $\log(G_\varepsilon(\rho))$, it is shown

$$\log(G_\varepsilon(e^{-U_\varepsilon})) = \log(\rho_g(x; 2\Sigma(I+\delta)^{-1} + 2\varepsilon I)) - w_\varepsilon^{(2)}(x)$$

where

$$w_\varepsilon^{(2)}(x) = w(x) - \frac{1}{2}w_\varepsilon^{(1)}(x) - \frac{\varepsilon}{2}x^\top \Sigma^{-1}\nabla w(x) - \varepsilon(\frac{1}{4}\|\nabla w(x)\|_2^2 - \frac{1}{2}\Delta w(x)) + O(\varepsilon^2)$$

Therefore,

$$\begin{aligned}
\varepsilon W_\varepsilon(x) &= -\log(\rho_g(x; 2\Sigma(I+\delta)^{-1})) + \log(\rho_g(x; 2\Sigma(I+\delta)^{-1} + 2\varepsilon I)) \\
&\quad + w(x) - \frac{1}{2}w_\varepsilon^{(1)}(x) - w_\varepsilon^{(2)}(x) + O(\varepsilon^2) \\
&= \frac{2\varepsilon}{4}x^\top (I+\delta)\Sigma^{-1}(2\Sigma(I+\delta)^{-1} + 2\varepsilon I)^{-1}x - \frac{\varepsilon}{2}\mathrm{Tr}(\Sigma^{-1}) \\
&\quad + \varepsilon x^\top \Sigma^{-1}\nabla w(x) + \frac{\varepsilon}{2}(\frac{1}{4}\|\nabla w(x)\|_2^2 - \frac{1}{2}\Delta w(x)) + O(\varepsilon^2) \\
&= \underbrace{\varepsilon\,(\frac{1}{4}\|\Sigma^{-1}x + \nabla w(x)\|_2^2 - \frac{1}{2}(\mathrm{Tr}(\Sigma^{-1}) + \Delta w(x)))}_{W(x)} + O(\varepsilon^2)
\end{aligned}$$

where the error term has at most quadratic growth as $|x| \to \infty$. This concludes the proof of the approximation (2.18b).

Based on this above calculation, the following estimate for a lower bound of the function W_ε is obtained (it is used in the proof of Prop. 2.6):

$$\begin{aligned}
W_\varepsilon(x) &= \frac{1}{4}\|\Sigma^{-1}x + \nabla w(x)\|_2^2 - \frac{1}{2}(\mathrm{Tr}(\Sigma^{-1}) + \Delta w(x)) + \varepsilon r_\varepsilon^{(2)}(x) \\
&\geq \frac{1}{8}\|\Sigma^{-1}x\|_2^2 - \frac{1}{2}(\|\nabla w\|_{L^\infty}^2 + \mathrm{Tr}(\Sigma^{-1}) + \|\Delta w\|_{L^\infty}) - \varepsilon(C_1\|x\|_2^2 + C_2) \\
&\geq \alpha\|x\|_2^2 - \beta
\end{aligned} \tag{2.39}$$

where $\alpha = \frac{1}{16\sigma_d^4}$, $\beta = \frac{1}{2}(\|\nabla w\|_{L^\infty}^2 + \mathrm{Tr}(\Sigma^{-1}) + \|\Delta w\|_{L^\infty} + \frac{C_2}{8\sigma_1^2})$ and $\varepsilon \leq \frac{1}{16C_1\sigma_d^4}$ (where recall $\sigma_d^2 = \lambda_{\max}(\Sigma)$).

(iii) Let $\tilde{P}_t$ denote the semigroup for the weighted Laplacian Δ_q with the density $q(x) = e^{-2U_\varepsilon(x)}$. We break the error into two parts:

$$\|T_\varepsilon^n f - P_t f\|_{L^2(\rho)} \leq \|T_\varepsilon^n f - \tilde{P}_t f\|_{L^2(\rho)} + \|\tilde{P}_t f - P_t f\|_{L^2(\rho)}$$

The bounds for the two terms on the right-hand side are derived in the following two steps:

Step 1. Using the stochastic representation (2.16)-(2.17),

$$(T_\varepsilon^n - \tilde{P}_t) f(x) = e^{U_\varepsilon(x)} \mathsf{E}\left[e^{-U_\varepsilon(B_{2t}^x)} f(B_{2t}^x)\, \zeta_t \right]$$

where $\zeta_t := e^{-\varepsilon \sum_{k=0}^{n-1} W_\varepsilon(B_{2k\varepsilon_n}^x)} - e^{-\int_0^t W(B_{2s}^x)\mathrm{d}s}$. By the Cauchy-Schwartz inequality

$$|(T_\varepsilon^n - \tilde{P}_t) f(x)| \le e^{U_\varepsilon(x)}\, \mathsf{E}[|f(B_{2t}^x)|^2 e^{-2U_\varepsilon(B_{2t}^x)}]^{\frac{1}{2}}\, \mathsf{E}\left[|\zeta_t|^2\right]^{\frac{1}{2}}$$

Next we obtain a bound for ζ_t. Upon using the inequality $|e^{-x} - e^{-y}| \le e^{-\min(x,y)}|x-y|$,

$$|\zeta_t| \le e^{-C} \left| \sum_{k=0}^{n-1} \varepsilon (W_\varepsilon(B_{2k\varepsilon}^x) - W(B_{2k\varepsilon}^x)) \right| + e^{-C} \left| \int_0^t W(B_{2s}^x)\mathrm{d}s - \sum_{k=0}^{n-1} \varepsilon W(B_{2k\varepsilon}^x) \right| \tag{2.40}$$

where $C = t \min(\min_{x\in\mathbb{R}^d} W(x), \min_{x\in\mathbb{R}^d} W_\varepsilon(x))$. Now, C is finite because, as $|x| \to \infty$, $W(x) \to \infty$ (Assumption A1) and $W_\varepsilon(x) \to \infty$ (by (2.18b)).

The expectation of the first term on the right-hand side of (2.40) is bounded as follows:

$$\begin{aligned}
\mathsf{E}\left[\left| \sum_{k=0}^{n-1} \varepsilon (W_\varepsilon(B_{2k\varepsilon}^x) - W(B_{2k\varepsilon}^x)) \right|^2 \right]^{\frac{1}{2}} &\le \sum_{k=0}^{n-1} \varepsilon \mathsf{E}[|W_\varepsilon(B_{2k\varepsilon}^x) - W(B_{2k\varepsilon}^x)|^2]^{\frac{1}{2}} \\
&\le \sum_{k=0}^{n-1} \varepsilon^2 \mathsf{E}[(C_1 \|x + B_{2k\varepsilon}^x\|_2^2 + C_2)^2]^{\frac{1}{2}} \\
&\le \sum_{k=0}^{n-1} \varepsilon^2 (2C_1 \|x\|_2^2 + 2C_1 \mathsf{E}[|B_{2k\varepsilon}^x|^4]^{\frac{1}{2}} + C_2) \\
&\le \varepsilon t \left[2C_1 \|x\|_2^2 + 6C_1 t + C_2 \right]
\end{aligned}$$

where the second inequality follows from the bound $|W_\varepsilon(x) - W(x)| = \varepsilon |r_\varepsilon^{(2)}(x)| \le \varepsilon(C_1 \|x\|_2^2 + C_2)$ for some constants C_1, C_2 (see (2.18b)).

The expectation of the second term in (2.40) is bounded as follows:

$$\begin{aligned}
\mathsf{E}\left[\left| \int_0^t W(B_{2s}^x)\mathrm{d}s - \sum_{k=0}^{n-1} \varepsilon W(B_{2k\varepsilon}^x) \right|^2 \right]^{\frac{1}{2}} &\le \varepsilon (\mathsf{E}\left[\int_0^t \|\nabla W(B_{2s}^x)\|_2^2 \mathrm{d}s \right]^{\frac{1}{2}} + t \|\Delta W\|_{L^\infty}) \\
&\le \varepsilon (\mathsf{E}\left[\int_0^t |C_3 \|x + B_s\|_2 + C_4|^2 \mathrm{d}s \right]^{\frac{1}{2}} + t C_5) \\
&\le \varepsilon t^{\frac{1}{2}} (C_3 \|x\|_2 + C_3 t + C_4) + \varepsilon C_5 t
\end{aligned}$$

where the Taylor expansion of $W(x)$ is used to obtain the first inequality, and for the second inequality, Assumption (A1) is used to bound $|\nabla W(x)| \le \|\Sigma^{-1}\| |x| + \|\nabla w\|_{L^\infty} = C_3 |x| + C_4$ and

$\|\Delta W\|_{L^\infty} \le \|\Sigma^{-1}\| + \|\Delta w\|_{L^\infty} = C_5$.

Putting together the two expectation bounds ,

$$|(T^n_{\varepsilon_n} - \tilde{P}_t)f(x)| \le e^{U_\varepsilon(x)} \mathsf{E}[f^2(B^x_{2t}) e^{-2U_\varepsilon(B^x_{2t})}]^{\frac{1}{2}} C\varepsilon t^{\frac{1}{2}} (\|x\|_2^2 + 1)$$

where C is a constant that only depends on t_0. Upon taking the $L^2(\rho)$ norm

$$\begin{aligned}
&\|T^n_\varepsilon - \tilde{P}_t f\|^2_{L^2(\rho)} \\
&\quad \le C\varepsilon^2 t \int\int f^2(y)\rho_g(x-y;2t)(\|x\|_2^2+1)^2 e^{2U_\varepsilon(x)-2U_\varepsilon(y)}\rho(x)\mathrm{d}y\mathrm{d}x \\
&\quad \le C\varepsilon^2 t \int\int f^2(y)\rho_g(x-y;2t)(\|x\|_2^4+1) e^{2\varepsilon(2W(x)+\frac{1}{2}\Delta V(x)+\varepsilon r^{(1)}_\varepsilon(x))}\rho(y)\mathrm{d}y\mathrm{d}x \\
&\quad \le C\varepsilon^2 t \int f^2(y)(\|y\|_2^4 + 12t^2 + 1) e^{8\varepsilon W(y)+O(\varepsilon^2)}\rho(y)\mathrm{d}y \\
&\quad \le C\varepsilon^2 t \left[\int (\|x\|_2^4 + 12t^2+1)^2 e^{8\varepsilon W(x)+O(\varepsilon^2)}\rho(x)\mathrm{d}x\right]^{1/2} \left[\int f^4(x)\rho(x)\mathrm{d}x\right]^{1/2} \\
&\quad \le C\varepsilon^2 t \|f\|^2_{L^4(\rho)}
\end{aligned}$$

Step 2. Because P_t and $\tilde{P}_t$ are semigroups with generators Δ_ρ and Δ_q, respectively, we have the identity: $P_t f - \tilde{P}_t f = \int_0^t P_{t-s}(\Delta_\rho - \Delta_q)\tilde{P}_s f \mathrm{d}s$. Upon taking the $L^2(\rho)$ norm of both sides, using the triangle inequality, because P_t is contraction on $L^2(\rho)$,

$$\|P_t f - \tilde{P}_t f\|_{L^2(\rho)} \le \int_0^t \|(\Delta_\rho - \Delta_q)\tilde{P}_s f\|_{L^2(\rho)}\mathrm{d}s$$

Now,

$$\begin{aligned}
&\|(\Delta_\rho - \Delta_q)\tilde{P}_s f\|^2_{L^2(\rho)} = 4\int |(\nabla U(x) - \nabla U_\varepsilon(x))\cdot\nabla(\tilde{P}_s f)(x)|^2 \rho(x)\mathrm{d}x \\
&\quad \le 4\left[\int \|\nabla U(x) - \nabla U_\varepsilon(x)\|_2^4 \frac{\rho(x)}{q(x)}\rho(x)\mathrm{d}x\right]^{1/2} \left[\int |\nabla \tilde{P}_s f(x)|^4 q(x)\mathrm{d}x\right]^{1/2} \\
&\quad \le 4\varepsilon^2 \left[\int |C_1\|x\|_2 + C_2|^4 e^{-2U(x)+2U_\varepsilon(x)}\rho(x)\mathrm{d}x\right]^{1/2} \|\nabla f\|^2_{L^4(q)} \\
&\quad \le C\varepsilon^2 \|\nabla f\|^2_{L^4(\rho)}
\end{aligned}$$

where the identity $\Delta_\rho f - \Delta_q f = 2\nabla U\cdot\nabla f - 2\nabla U_\varepsilon\cdot\nabla f$ is used in the first step, the Cauchy-Schwartz inequality in the second step, and the bounds $\|\nabla U_\varepsilon(x) - \nabla U(x)\|_2 \le \varepsilon(C_1\|x\|_2 + C_2)$ and $\|\nabla \tilde{P}_s f\|_{L^4(q)} \le \|\nabla f\|_{L^4(q)}$ in the third step.

Combining the two sets of bounds in steps 1 and 2, one obtains (2.19).

□

2.6.6 Proof of the Prop. 2.6

Proof. (i) The Lyapunov condition (2.28a), known as DV(3) of [Kontoyiannis et al., 2005], is the necessary and sufficient condition for geometric ergodicity (and in fact the stronger U_ε-uniform ergodicity) [Meyn and Tweedie, 2009, Thm. 15.0.1]. The distribution ρ_ε is invariant because $\forall f \in C_b(\mathbb{R}^d)$,

$$\begin{aligned}\int T_\varepsilon f(x)\rho_\varepsilon(x)\mathrm{d}x &= \int\int \frac{1}{n_\varepsilon(x)} k_\varepsilon(x,y) f(y)\rho(y)\mathrm{d}y \frac{n_\varepsilon(x)\rho(x)}{\int n_\varepsilon(z)\rho(z)\mathrm{d}z}\mathrm{d}x \\ &= \frac{1}{\int n_\varepsilon(z)\rho(z)\mathrm{d}z}\int\int k_\varepsilon(x,y)\rho(x)\mathrm{d}x f(y)\rho(y)\mathrm{d}y \\ &= \frac{1}{\int n_\varepsilon(z)\rho(z)\mathrm{d}z}\int f(y) n_\varepsilon(y)\rho(y)\mathrm{d}y = \int f(x)\rho_\varepsilon(x)\mathrm{d}x\end{aligned}$$

(ii) The invariant density ρ_ε is reversible because $\forall f, g \in C_b(\mathbb{R}^d)$

$$\begin{aligned}\int g(x) T_\varepsilon f(x)\rho_\varepsilon(x)\mathrm{d}x &= \int\int g(x)\frac{k_\varepsilon(x,y)}{n_\varepsilon(x)} f(y)\rho(y)\frac{n_\varepsilon(x)\rho(x)}{\int n_\varepsilon(z)\rho(z)\mathrm{d}z}\mathrm{d}y\mathrm{d}x \\ &= \frac{1}{\int n_\varepsilon(z)\rho(z)\mathrm{d}z}\int T_\varepsilon g(y) n_\varepsilon(y) f(y)\rho(y)\mathrm{d}y \\ &= \int f(y) T_\varepsilon g(y)\rho_\varepsilon(y)\mathrm{d}y\end{aligned}$$

The spectral gap follows from Lyapunov condition (2.28a) and the fact that the chain is reversible [Roberts and Rosenthal, 1997, Thm 2.1]. The spectral gap is denoted as λ.

(iii) The solution ϕ_ε satisfies the bound:

$$\|\phi\|_{L^2(\rho_\varepsilon)} \le \frac{\|\sum_{k=0}^{n-1}\varepsilon T_\varepsilon^k(h-\hat{h}_\varepsilon)\|_{L^2(\rho_\varepsilon)}}{1-\|T_\varepsilon^n\|_{L_0^2(\rho_\varepsilon)}} \le \frac{\varepsilon n\|h\|_{L^2(\rho_\varepsilon)}}{1-\|T_\varepsilon^n\|_{L_0^2(\rho_\varepsilon)}} \le \frac{t\|h\|_{L^2(\rho_\varepsilon)}}{\lambda}$$

It remains to verify the Lyapunov condition (2.28a): Using (2.17)

$$e^{-U_\varepsilon} T_\varepsilon^n e^{U_\varepsilon}(x) = \mathsf{E}[e^{-\varepsilon\sum_{k=0}^{n-1} W_\varepsilon(B^x_{2k\varepsilon})}] \le \mathsf{E}[e^{-\varepsilon\sum_{k=0}^{n-1}(\alpha\|B^x_{2k\varepsilon_n}\|_2^2-\beta)}]$$

where the second inequality follows from using the lower bound $W_\varepsilon(x) \ge \alpha|x|^2 - \beta$ derived in (2.39).

We now claim that

$$\mathsf{E}[e^{-\varepsilon\sum_{k=0}^{m-1}(\alpha\|B^x_{2k\varepsilon}\|_2^2-\beta)}] = e^{-\alpha_m\|x\|_2^2+\beta_m} \tag{2.41}$$

for $m = 1,\dots,n$ where $\{\alpha_m\}_{m=1}^n$ and $\{\beta_m\}_{m=1}^n$ are defined using the recursions:

$$\begin{aligned}\alpha_{m+1} &= \alpha\varepsilon + \frac{\alpha_m}{1+4\varepsilon\alpha_m}, \quad \alpha_1 = \alpha\varepsilon \\ \beta_{m+1} &= \beta_m + \beta\varepsilon - \frac{1}{2}\log(1+4\varepsilon\alpha_m), \quad \beta_1 = \beta\varepsilon\end{aligned}$$

Assuming for now that the claim is true

$$\log(e^{-U}T_{\varepsilon_n}^n e^U(x)) \le \log(\mathsf{E}[e^{-\varepsilon\sum_{k=0}^{n-1}(\alpha\|B^x_{2k\varepsilon_n}\|_2^2-\beta)}]) = -\alpha_n\|x\|_2^2+\beta_n$$

An upper-bound for β_n and a lower-bound for α_n are obtained as follows:

1. For the sequence $\{\beta_m\}_{m=1}^n$,

$$\beta_{m+1} \le \beta_m + \beta\varepsilon, \quad \Rightarrow \quad \beta_n \le \beta_1 + (n-1)\beta\varepsilon = \beta t$$

2. For the sequence $\{\alpha_m\}_{m=1}^n$,

$$\alpha_{m+1} \le \alpha_m + \alpha\varepsilon, \quad \Rightarrow \quad \alpha_m \le \alpha_1 + (n-1)\alpha\varepsilon = \alpha t$$

Therefore,

$$\alpha_{m+1} \ge \frac{\alpha_m}{1+4\varepsilon\alpha t} + \alpha\varepsilon, \quad \alpha_1 = \alpha\varepsilon$$

It then follows

$$\alpha_n \ge \alpha t e^{-4\alpha t^2}$$

Upon using the two bounds

$$\log(e^{-U_\varepsilon}T_{\varepsilon_n}^n e^{U_\varepsilon}(x)) \le -\alpha t e^{-4\alpha t^2}\|x\|_2^2 + \beta t \le -atU_\varepsilon(x) + bt$$

where the second inequality follows from using the upper bound $U_\varepsilon(x) \le \frac{1}{8\sigma_1^2}\|x\|_2^2 + \frac{\sigma_1^2}{8}C$ derived in (2.38). The following estimates are obtained for constants

$$a = 8\sigma_1^2\alpha e^{-4\alpha t_0^2}, \qquad b = \beta + C\sigma_1^4\alpha e^{-4\alpha t_0}$$

It remains to prove the claim (2.41). The constants α_1 and β_1 for $m=1$ are easily verified by direct evaluation and for $m>1$,

$$\begin{aligned}\mathsf{E}[e^{-\varepsilon\sum_{k=0}^{m}(\alpha\|B^x_{2k\varepsilon}\|_2^2-\beta)}] &= \mathsf{E}[e^{-\varepsilon\alpha\|x\|_2^2+\beta\varepsilon}e^{-\alpha_m\|B_{2\varepsilon}\|_2^2+\beta_m}] \\ &= e^{-\varepsilon\alpha\|x\|_2^2 - \frac{\alpha_m}{1+4\varepsilon\alpha_m}\|x\|_2^2 + \varepsilon\beta+\beta_m - \frac{1}{2}\log(1+4\varepsilon\alpha_m)}\end{aligned}$$

The minorization inequality (2.28b) is obtained next. For $|x| \le R$:

$$\begin{aligned}T_{\frac{t}{n}}^n \mathbf{1}_{[A]}(x) &= e^{U_\varepsilon(x)}\mathsf{E}[e^{-\varepsilon\sum_{k=0}^{n-1}W_\varepsilon(B^x_{2k\varepsilon})}e^{-U_\varepsilon(B^x_{2t})}\mathbf{1}_{[B^x_{2t}\in A]}] \\ &\ge \frac{e^{\min_{\|x\|_2\le R}U_\varepsilon(x)}}{e^{\max_{\|x\|_2\le R+10}(U_\varepsilon(x)+tW_\varepsilon(x))}}\mathsf{P}([B^x_{2t}\in A]\cap[\sup_{s\in[0,2t]}\|B_s\|_2 \le 10]) \ge \delta\nu(A)\end{aligned}$$

where

$$\nu(A) = \mathsf{P}(\{B_{2t}^x \in A\} | \{ \sup_{s\in[0,2t]} \|B_s\|_2 \le 10\})$$
$$\delta = \frac{e^{\min_{|x|\le R,\varepsilon\in(0,1)} U_\varepsilon(x)}}{e^{\max_{|x|\le R+10,\varepsilon\in(0,1)}(U_\varepsilon(x)+tW_\varepsilon(x))}}(1-2e^{-\frac{50}{t_0}})$$

because $\mathsf{P}(\sup_{s\in[0,2t]} \|B_s\|_2 \ge 10) \le e^{-\frac{100}{2t}} \le e^{-\frac{50}{t_0}}$. □

2.6.7 Proof of the Thm. 2.1

Proof. (i) The existence of the solution is proved in Prop. 2.6.

(ii) We break the error into two parts:

$$\|\phi_\varepsilon - \phi\|_{L^2(\rho_\varepsilon)} \le \|\phi_\varepsilon - \tilde\phi\|_{L^2(\rho_\varepsilon)} + \|\tilde\phi - \phi\|_{L^2(\rho_\varepsilon)}$$

where $\tilde\phi$ is the solution to the fixed point equation $\tilde\phi = P_\varepsilon\tilde\phi + \varepsilon(h-\hat h)$ with the exact semigroup P_ε. The bounds for the two terms on the right-hand side are derived in the following two steps:

Step 1. Iterating the formula $\tilde\phi = P_\varepsilon\tilde\phi + \varepsilon(h-\hat h)$ for $n = \lfloor\frac{1}{\varepsilon}\rfloor$ times yields,

$$\tilde\phi = P_\varepsilon^n\tilde\phi + \sum_{k=0}^{n-1}\varepsilon P_\varepsilon^k(h-\hat h)$$

and subtracting this from (2.27) gives

$$\phi_\varepsilon - \tilde\phi = T_\varepsilon^n(\phi_\varepsilon - \tilde\phi) + (T_\varepsilon^n - P_\varepsilon^n)\tilde\phi + \sum_{k=0}^{n-1}\varepsilon(T_\varepsilon^k - P_\varepsilon^k)h + t(\hat h - \hat h_\varepsilon)$$

This forms a (discrete) Poisson equation whose solution exists and is bounded according to Prop. 2.6:

$$\begin{aligned}\|\phi_\varepsilon - \tilde\phi\|_{L^2(\rho_\varepsilon)} &\le \frac{n\varepsilon}{\lambda}\left(\|(T_\varepsilon^n - P_\varepsilon^n)\tilde\phi\|_{L^2(\rho_\varepsilon)} + \|\sum_{k=0}^{n-1}\varepsilon(T_\varepsilon^k - P_\varepsilon^k)h\|_{L^2(\rho_\varepsilon)} + n\varepsilon|\hat h - \hat h_\varepsilon|\right)\\ &\le \frac{Cn\varepsilon}{\lambda}\left(\|(T_\varepsilon^n - P_\varepsilon^n)\tilde\phi\|_{L^2(\rho)} + \|\sum_{k=0}^{n-1}\varepsilon(T_\varepsilon^k - P_\varepsilon^k)h\|_{L^2(\rho)} + n\varepsilon|\hat h - \hat h_\varepsilon|\right)\end{aligned} \tag{2.42}$$

where we used $\|\cdot\|_{L^2(\rho_\varepsilon)} \le C\|\cdot\|_{L^2(\rho)}$ in the second step. This is true because $\rho_\varepsilon(x) = e^{-U_\varepsilon(x)}G_\varepsilon(e^{-U_\varepsilon})(x) = \rho(x)e^{-3\varepsilon W(x)-\varepsilon\Delta V(x)+O(\varepsilon^2)} \le C\rho(x)$ using the formula (2.18a).

It remains to bound the three terms inside the bracket in (2.42):

$$\|T_\varepsilon^n\tilde{\phi} - P_{n\varepsilon}\tilde{\phi}\|_{L^2(\rho)} \le C\varepsilon\sqrt{n\varepsilon}(\|\tilde{\phi}\|_{L^4(\rho)} + \|\nabla\tilde{\phi}\|_{L^4(\rho)})$$
$$\|\sum_{k=0}^{n-1}\varepsilon(T_\varepsilon^k - P_\varepsilon^k)h\|_{L^2(\rho)} \le C\varepsilon(n\varepsilon)\sqrt{n\varepsilon}(\|h\|_{L^4(\rho)} + \|\nabla h\|_{L^4(\rho)})$$
$$|\hat{h}_\varepsilon - \hat{h}| \le \int |h(x)|\rho(x)|e^{-3\varepsilon W(x) - \varepsilon\Delta V(x) + O(\varepsilon^2)} - 1|\mathrm{d}x \le \varepsilon C\|h\|_{L^2(\rho)}$$

by using the error estimates Prop. 2.3-(iii). Therefore,

$$\|\phi_\varepsilon - \tilde{\phi}\|_{L^2(\rho_\varepsilon)} \le \varepsilon C(\|h\|_{L^4(\rho)} + \|\nabla h\|_{L^4(\rho)} + \|\tilde{\phi}\|_{L^4(\rho)} + \|\nabla\tilde{\phi}\|_{L^4(\rho)})$$

Step 2. Both ϕ and $\tilde{\phi}$ are solutions with the exact semigroup P_ε. Using the spectral representation (2.4),

$$\phi = \sum_{m=1}^{\infty}\frac{1}{\lambda_m}\langle h, e_m\rangle e_m, \quad \tilde{\phi} = \sum_{m=1}^{\infty}\frac{\varepsilon}{1-e^{-\varepsilon\lambda_m}}\langle h, e_m\rangle e_m$$

Therefore,

$$\|\tilde{\phi} - \phi\|_{L^2(\rho)}^2 = \varepsilon^2\sum_{m=1}^{\infty}(\frac{1}{1-e^{-\varepsilon\lambda_m}} - \frac{1}{\varepsilon\lambda_m})^2|\langle h, e_m\rangle|^2 \le \varepsilon^2\|h\|_{L^2(\rho)}^2$$

and thus $\|\tilde{\phi} - \phi\|_{L^2(\rho_\varepsilon)} \le C\|\tilde{\phi} - \phi\|_{L^2(\rho)} \le \varepsilon^2 C\|h\|_{L^2(\rho)}^2$.

Combining the estimates from steps 1 and 2,

$$\|\phi_\varepsilon - \phi\|_{L^2(\rho_\varepsilon)} \le \varepsilon C(\|h\|_{L^4(\rho)} + \|\nabla h\|_{L^4(\rho)} + \|\tilde{\phi}\|_{L^4(\rho)} + \|\nabla\tilde{\phi}\|_{L^4(\rho)})$$

□

2.6.8 Proof of the Prop. 2.4

Proof. Denote $\eta_j = (\sqrt{\frac{(g_\varepsilon * \rho)(X^j)}{\frac{1}{N}\sum_{l=1}^N g_\varepsilon(X^j, X^l)}} - 1)$ and express:

$$T_\varepsilon^{(N)} f(x) = \frac{\int k_\varepsilon(x,y)f(y)\rho(y)\mathrm{d}y + \xi_1^{(N)} + \zeta_1^{(N)}}{n_\varepsilon(x) + \xi_2^{(N)} + \zeta_2^{(N)}}$$

where

$$\xi_1^{(N)} = \frac{1}{N}\sum_{j=1}^{N} k_\varepsilon(x,X^j)f(X^j) - \mathsf{E}[k_\varepsilon(x,X^j)f(X^j)],\;\; \zeta_1^{(N)} = \frac{1}{N}\sum_{j=1}^{N} k_\varepsilon(x,X^j)f(X^j)\eta_j$$

$$\xi_2^{(N)} = \frac{1}{N}\sum_{j=1}^{N} k_\varepsilon(x,X^j) - \mathsf{E}[k_\varepsilon(x,X^j)],\quad \zeta_2^{(N)} = \frac{1}{N}\sum_{j=1}^{N} k_\varepsilon(x,X^j)\eta_j$$

(i) To prove the part-(i) of the Prop. 2.4, the strategy is to show that as $N\to\infty$ the stochastic terms $\xi_1^{(N)},\xi_2^{(N)},\zeta_1^{(N)},\zeta_2^{(N)}$ converge to zero almost surely. We do this in two steps below, $\xi_1^{(N)},\xi_2^{(N)}$ in step 1, and $\zeta_1^{(N)},\zeta_2^{(N)}$ in step 2.

Step 1: Convergence of $\xi_1^{(N)}$ and $\xi_1^{(N)}$ follows from direct application of the strong law of large numbers (SLLN). The SLLN applies because the summand for $\xi_1^{(N)}$ and $\xi_2^{(N)}$ are independent and identically distributed (i.i.d) and moreover have finite variance:

$$\mathrm{Var}\left(k_\varepsilon(x,X)f(X)\right) \le \frac{C}{\varepsilon^{d/2}}\frac{\|f\|_{L^\infty}^2\rho(x)}{(g_\varepsilon*\rho)^2(x)} \tag{2.43}$$

$$\mathrm{Var}\left(k_\varepsilon(x,X)\right) \le \frac{C}{\varepsilon^{d/2}}\frac{\rho(x)}{(g_\varepsilon*\rho)^2(x)} \tag{2.44}$$

where we used $g_\varepsilon^2(x,y)\le C\varepsilon^{-d/2}g_{\varepsilon/2}(x,y)$.

Step 2: In order to show the almost sure convergence of $\zeta_1^{(N)}$ and $\zeta_1^{(N)}$ to zero, we first show that in the limit as $N\to\infty$,

$$|\eta_i| \le C\sqrt{\frac{\log(\frac{N}{\delta})}{N\varepsilon^{d/2}q_\varepsilon(X^i)}},\quad \forall i=1,\dots,N \tag{2.45}$$

with probability larger than $1-\delta$ for any arbitrary choice of $\delta\in(0,1)$. Assuming for now that the claim is true, it then follows

$$\zeta_1^{(N)} \le \sqrt{\frac{C\log(\frac{N}{\delta})}{N\varepsilon^{d/2}}}\left(\frac{1}{N}\sum_{j=1}^{N} k_\varepsilon(x,X^j)\frac{|f(X^j)|}{\sqrt{g_\varepsilon*\rho(X^j)}}\right) \tag{2.46}$$

with probability larger than $1-\delta$. The term inside the bracket converges almost surely to its limit $\mathsf{E}[k_\varepsilon(x,X)\frac{|f(X)|}{\sqrt{g_\varepsilon*\rho(X)}}]$, by SLLN, because

$$\mathsf{E}\left(k_\varepsilon(x,X)\frac{|f(X)|}{\sqrt{g_\varepsilon*\rho(X)}}\right) \le \frac{C\|f\|_{L^\infty}\rho(x)}{(g_\varepsilon*\rho)^{3/2}(x)}$$

The proof that $\zeta_1^{(N)}\xrightarrow{\text{a.s.}}0$ is completed by an application of the Borel-Cantelli lemma. Indeed, choose a sequence $\{\delta_N\}_{N=1}^\infty$ given by $\delta_N=\frac{1}{N^2}$. Then $\sum_{N=1}^\infty \mathsf{P}(\zeta_1^{(N)}>\varepsilon_N)\le\sum_{N=1}^\infty\delta_N<\infty$ where $\varepsilon_N=\sqrt{\frac{C\log(N^3)}{N\varepsilon^{d/2}}}$. Because $\varepsilon_N\to 0$, then $\zeta_1^{(N)}\xrightarrow{\text{a.s}}0$. The proof of $\zeta_2^{(N)}\xrightarrow{\text{a.s}}0$ is identical.

It remains to prove the claim (2.45). It follows from the Bernstein inequality. We have for any $a > 0$:

$$\begin{aligned}
\mathsf{P}(\eta_i \geq a) &= \mathsf{P}\left(\sqrt{\frac{(g_\varepsilon * \rho)(X^j)}{\frac{1}{N}\sum_{l=1}^N g_\varepsilon(X^j, X^l)}} \geq 1 + a\right) \\
&\leq \mathsf{P}\left(\frac{(g_\varepsilon * \rho)(X^j)}{\frac{1}{N}\sum_{l=1}^N g_\varepsilon(X^j, X^l)} \geq 1 + a\right) \\
&= \mathsf{P}\left(\frac{(g_\varepsilon * \rho)(X^j) - \frac{1}{N}\sum_{l=1}^N g_\varepsilon(X^j, X^l)}{(g_\varepsilon * \rho)(X^j)} \geq \frac{a}{1+a}\right)
\end{aligned}$$

The random variables $g_\varepsilon(X^i, X^j)$ are i.i.d, bounded by $(4\pi\varepsilon)^{-\frac{d}{2}}$, and the variance

$$\mathsf{E}\left[|g_\varepsilon(X^i, X^j)|^2 | X^j\right] \leq \frac{1}{(8\pi\varepsilon)^{d/2}}(g_{\varepsilon/2} * \rho)(X^j)$$

Therefore by Bernstein inequality,

$$|\eta_i| \leq C\sqrt{\frac{(g_{\varepsilon/2} * \rho)(X^j)\log(\frac{2}{\delta})}{N(8\pi\varepsilon)^{d/2}(g_\varepsilon * \rho)(X^j)^2}}$$

with probability higher than $1 - \delta$. The result is obtained by union bound for $i = 1, \ldots, N$ and $\|\frac{g_{\varepsilon/2} * \rho}{g_\varepsilon * \rho}\|_{L^\infty} < \infty$.

(ii) Collecting the estimates (2.43)-(2.44)-(2.46) and application of the Bernstein inequality yields:

$$\begin{aligned}
|\xi_1^{(N)}| &\leq \sqrt{\frac{C\|f\|_\infty^2 \log(\frac{1}{\delta})\rho(x)}{N\varepsilon^{d/2}(g_\varepsilon * \rho)^2(x)}}, \quad |\xi_2^{(N)}| \leq \sqrt{\frac{C\log(\frac{1}{\delta})\rho(x)}{N\varepsilon^{d/2}(g_\varepsilon * \rho)^2(x)}} \\
|\zeta_1^{(N)}| &\leq \sqrt{\frac{C\|f\|_\infty^2 \log(\frac{N}{\delta})\rho^2(x)}{N\varepsilon^{d/2}(g_\varepsilon * \rho)^3(x)}}, \quad |\zeta_2^{(N)}| \leq \sqrt{\frac{C\log(\frac{N}{\delta})\rho^2(x)}{N\varepsilon^{d/2}(g_\varepsilon * \rho)^3(x)}}
\end{aligned}$$

with probability larger than $1 - 4\delta$. Therefore one obtains the bound:

$$|T_\varepsilon^{(N)} f(x) - T_\varepsilon f(x)| \leq \sqrt{\frac{C\log(\frac{N}{\delta})\rho(x)}{N\varepsilon^{d/2}(g_\varepsilon * \rho)^2(x)n_\varepsilon^2(x)}}$$

with probability larger than $1 - 4\delta$. Upon squaring and integrating both sides with respect to $\rho(x)$ proves the rate:

$$\begin{aligned}
\|T_\varepsilon^{(N)} f - T_\varepsilon f\|_2 &\leq \sqrt{\frac{C\log(\frac{N}{\delta})}{N\varepsilon^{d/2}}}\left(\int \frac{\rho(x)}{(g_\varepsilon * \rho)^2(x)n_\varepsilon^2(x)}\rho(x)\mathrm{d}x\right)^{1/2} \\
&\leq \sqrt{\frac{C\log(\frac{N}{\delta})}{N\varepsilon^{d/2}}}\left(\int e^{-2\varepsilon|\nabla V(x)|^2 + \frac{3}{2}\varepsilon|\nabla V(x)|^2}\mathrm{d}x\right)^{1/2} \leq \sqrt{\frac{C\log(\frac{N}{\delta})}{N\varepsilon^d}}
\end{aligned}$$

□

2.6.9 Proof of the Thm. 2.2

In the proof of Thm. 2.2, the function space of interest is $C(\Omega)$, the Banach space of continuous functions on (a compact set) $\Omega \subset \mathbb{R}^d$ equipped with the $\|\cdot\|_{L^\infty}$ norm. The space $C_0(\Omega) := \{f \in C(\Omega) \mid \int f\rho_\varepsilon = 0\}$. Consider T_ε and $T_\varepsilon^{(N)}$ as linear operators from $C(\Omega)$ to $C(\Omega)$.

Part-(i) has already been proved as part of the Prop. 2.5. The proof of part (ii) relies on the verification of the following three conditions:

(i) The family of operators $\{T_\varepsilon^{(N)}\}_{N=1}^{\infty}$ is collectively compact, as linear operators on $C(\Omega)$.

(ii) For any function $f \in C(\Omega)$,

$$\lim_{N\to\infty} \|T_\varepsilon^{(N)} f - T_\varepsilon f\|_{L^\infty} = 0, \quad \text{a.s.} \tag{2.47}$$

(iii) The operator $(I - T_\varepsilon)^{-1}$ is a bounded operator on $C_0(\Omega)$.

Once these three conditions have been verified, the convergence result (2.31) follows from a standard result in the approximation theory of the numerical solutions of integral equations [Hutson et al., 2005, Thm. 7.6.6].

The proof of the three conditions is as follows:

(i) The collective compactness holds if the set $S = \{T_\varepsilon^{(N)} f;\ \forall f \in C(\Omega), \|f\|_{L^\infty} \le 1, N \in \mathbb{N}\}$ is relatively compact. Relative compactness follows from an application of the Arzela-Ascoli theorem. In order to apply Arzela-Ascoli theorem, we need to show that S is uniformly bounded and equicontinuous. The two conditions hold because

$$\begin{aligned} &\text{(unif. boundedness)} \quad |T_\varepsilon^{(N)} f(x)| \le \|f\|_{L^\infty} \frac{\sum_{i=1}^N k_\varepsilon^{(N)}(x, X^i)}{\sum_{i=1}^N k_\varepsilon^{(N)}(x, X^i)} \le 1 \\ &\text{(equicontinuous)} \quad |T_\varepsilon^{(N)} f(x) - T_\varepsilon^{(N)} f(x')| \le \frac{L}{\varepsilon} |x - x'| e^{\frac{L}{2\varepsilon}|x-x'|} \end{aligned} \tag{2.48}$$

for all $x, x' \in \Omega$ and f such that $\|f\|_{L^\infty} \le 1$. The detailed calculation to obtain the second inequality appears at the end of the proof.

(ii) Fix a function $f \in C(\Omega)$. From Prop. 2.4-(i), we know that $T_\varepsilon^{(N)} f(x)$ converges to $T_\varepsilon f(x)$ almost surely pointwise for all $x \in \Omega$. Because Ω is compact and $\{T_\varepsilon^{(N)} f\}$ is equicontinuous, pointwise convergence implies uniform convergence (2.47).

(iii) From parts (i) and (ii) above, it can be concluded that T_ε is a compact operator. Therefore, using the Fredholm alternative theorem, in order to show $(I - T_\varepsilon)^{-1}$ is bounded, it is enough to show that $I - T_\varepsilon$ is injective. The injectivity property is shown by contradiction. Suppose there exists a function $f \in C_0(\Omega)$ such that $f - T_\varepsilon f = 0$. Let $x_0 \in \Omega$ be a point that achieves the maximum of the

function f. Such a point exists because f is continuous and Ω is compact. Evaluating $f - T_\varepsilon f = 0$ at $x = x_0$ yields

$$0 = f(x_0) - T_\varepsilon f(x_0) = \frac{1}{n_\varepsilon(x)} \int k_\varepsilon(x_0, y)(f(x_0) - f(y))\mathrm{d}y$$

Because $k_\varepsilon(x_0, y) > 0$ and $f(y) \leq f(x_0)$, this implies $f(y) = f(x_0)$ for all $y \in \Omega$. Therefore, the function f is a constant. But the only constant function in $C_0(\Omega)$ is zero. Hence $I - T_\varepsilon$ is injective and its inverse $(I - T_\varepsilon)^{-1}$ is bounded.

It remains to prove the equicontinuity inequality (2.48) which is done next:

$$\begin{aligned}
|T_\varepsilon^{(N)} f(x) - T_\varepsilon^{(N)} f(x')| &\leq |\frac{\sum_{i=1}^N k_\varepsilon^{(N)}(x, X^i) f(X^i)}{\sum_{i=1}^N k_\varepsilon^{(N)}(x, X^i)} - \frac{\sum_{i=1}^N k_\varepsilon^{(N)}(x', X^i) f(X^i)}{\sum_{i=1}^N k_\varepsilon^{(N)}(x', X^i)}| \\
&\leq 2\|f\|_{L^\infty} \frac{\sum_{i=1}^N k_\varepsilon^{(N)}(x, X^i) |1 - \frac{k_\varepsilon(x', X^i)}{k_\varepsilon(x, X^i)}|}{\sum_{i=1}^N k_\varepsilon(x, X^i)} \\
&\leq 2 \max_{i=1,\ldots,N} |1 - \frac{k_\varepsilon(x', X^i)}{k_\varepsilon(x, X^i)}| \leq \frac{L}{\varepsilon} \|x - x'\|_2 e^{\frac{L}{2\varepsilon}\|x - x'\|_2}
\end{aligned}$$

where the last inequality is obtained as follows

$$|1 - \frac{k_\varepsilon(x', X^i)}{k_\varepsilon(x, X^i)}| = |1 - \frac{g_\varepsilon(x', X^i)}{g_\varepsilon(x, X^i)}| = |1 - e^{-\frac{(x' - x)\cdot(x' + x - 2X^i)}{4\varepsilon}}| \leq \frac{L}{2\varepsilon} \|x - x'\|_2 e^{\frac{L}{2\varepsilon}\|x - x'\|_2}$$

where $L = \max_{x,y \in \Omega} \|x - y\|_2$ is the diameter of Ω.

2.6.10 Proof of Prop. 2.7

1. Consider first the finite-N case. In the asymptotic limit as $\varepsilon \to \infty$, we have $(2\pi\varepsilon)^{d/2} g_\varepsilon(x, y) = 1 + O(\frac{1}{\varepsilon})$. Therefore,

$$\begin{aligned}
k_\varepsilon^{(N)}(x, y) &= \frac{g_\varepsilon(x, y)}{\sqrt{\frac{1}{N}\sum_{j=1}^N g_\varepsilon(x, X^j)} \sqrt{\frac{1}{N}\sum_{j=1}^N g_\varepsilon(y, X^j)}} = 1 + O(\frac{1}{\varepsilon}) \\
n_\varepsilon^{(N)}(x) &= \frac{1}{N} \sum_{i=1}^N k_\varepsilon^{(N)}(x, X^i) = 1 + O(\frac{1}{\varepsilon})
\end{aligned}$$

and

$$T_\varepsilon^{(N)} f(x) = \frac{\frac{1}{N}\sum_{j=1}^N k_\varepsilon(x, X^j) f(X^j)}{n_\varepsilon^{(N)}(x)} = \frac{1}{N} \sum_{j=1}^N f(X^j) + O(\frac{1}{\varepsilon})$$

It is also easy to see, e.g., by using a Neumann series solution, that in the asymptotic limit as $\varepsilon \to \infty$,

the solution of the fixed-point equation (2.23) is given by

$$\Phi = \varepsilon(\mathsf{h} - \frac{1}{N}\sum_{l=1}^{N}\mathsf{h}_l) + O(1)$$

Therefore,

$$r = \Phi + \varepsilon\mathsf{h} = 2\varepsilon\mathsf{h} - \varepsilon(\frac{1}{N}\sum_{l=1}^{N}\mathsf{h}_l) + O(1)$$
$$s_{ij} = \frac{1}{2\varepsilon}\mathsf{T}_{ij}(r_j - \sum_{k=1}^{N}\mathsf{T}_{ik}r_k) = \frac{1}{N}(\mathsf{h}_j - \frac{1}{N}\sum_{l=1}^{N}\mathsf{h}_l) + O(\frac{1}{\varepsilon})$$

and using the gain approximation formula (2.13),

$$\mathsf{K}_i = \sum_{j=1}^{N} s_{ij}X^j = \frac{1}{N}\sum_{j=1}^{N}(\mathsf{h}_j - \frac{1}{N}\sum_{l=1}^{N}\mathsf{h}_l)X^j + O(\frac{1}{\varepsilon})$$

2. The calculations for the kernel formula are entirely analogous. In the asymptotic limit as $\varepsilon \to \infty$,

$$T_\varepsilon f(x) = \int f(x)\rho(x)\mathrm{d}x + O(\frac{1}{\varepsilon})$$
$$\phi_\varepsilon(x) = \varepsilon(h(x) - \hat{h}) + O(1)$$

and, using $\theta(x) = x$ to denote the coordinate function and $\cdot$ to denote function multiplication, the gain approximation formula (2.32) evaluates to

$$\begin{aligned}
\mathsf{K}_\varepsilon(x) &= \frac{1}{2\varepsilon}\left[T_\varepsilon(\theta\cdot\phi_\varepsilon + \varepsilon(h-\hat{h})) - T_\varepsilon(\theta)\,T_\varepsilon(\phi_\varepsilon + \varepsilon(h-\hat{h}))\right] \\
&= \frac{1}{2}T_\varepsilon(\theta\cdot\frac{\phi_\varepsilon}{\varepsilon} + h - \hat{h}) - \frac{1}{2}T_\varepsilon(\theta)\,T_\varepsilon(\frac{\phi_\varepsilon}{\varepsilon} + h - \hat{h}) + O(\frac{1}{\varepsilon}) \\
&= T_\varepsilon(\theta\cdot h - \hat{h}) - T_\varepsilon(\theta)\,T_\varepsilon(h-\hat{h}) + O(\frac{1}{\varepsilon}) \\
&= \int x(h(x)-\hat{h})\rho(x)\mathrm{d}x + O(\frac{1}{\varepsilon})
\end{aligned}$$

Chapter 3

Optimal Transport FPF*

3.1 Introduction

Consider the filtering problem (1.3a)-(1.3b) introduced in Sec. 1.1, and the FPF algorithm introduced in Sec. 1.2. This chapter is concerned with the design of the mean-field process (1.8) in the FPF algorithm.

Consider a general representation for the mean-field process $\{\bar{X}_t\}_{t\geq 0}$ in terms of a sde:

$$\mathrm{d}\bar{X}_t = u_t(\bar{X}_t)\mathrm{d}t + \mathsf{K}_t(\bar{X}_t)\mathrm{d}Z_t + v_t(\bar{X}_t)\mathrm{d}\bar{B}_t, \quad \bar{X}_0 \sim \pi_{\text{init}} \tag{3.1}$$

where $\{Z_t\}$ is the observation process, $\{\bar{B}_t\}$ is standard Brownian motion, and $u_t(\cdot)$, $\mathsf{K}_t(\cdot)$, and $v_t(\cdot)$ are the control terms. The control terms should be measurable with respect to the filtration $\mathscr{Z}_t := \sigma(\{Z_s;\, s\in[0,t]\})$. The control problem is to choose $u_t(\cdot)$, $\mathsf{K}_t(\cdot)$, and $v_t(\cdot)$ such that $\bar{X}_t$ is distributed according to posterior distribution π_t, i.e.,

$$\bar{X}_t \sim \pi_t \quad \forall t \geq 0 \tag{3.2}$$

where π_t denotes the conditional probability distribution of X_t. When the condition (3.2) is true, the filter is said to be exact.

There are infinitely many choices of control law that all lead to exact filters. This is not surprising: The condition (3.2) specifies only the marginal distribution of the stochastic process $\{\bar{X}\}_{t\geq 0}$ at each times $t \geq 0$. This is not enough to uniquely identify a stochastic process, e.g the joint distributions at two time instants is not known. The non-uniqueness issue can be also understood through the lens of optimal transportation theory: interpret the sde (3.1) as transporting the initial distribution π_0 at time $t = 0$ (prior) to the conditional distribution π_t at time t (posterior). Clearly, there are infinitely many maps that transport one distribution into another.

The mean-field process in the FPF algorithm (1.8), is a specific choice so that the filter is exact for the general nonlinear non-Gaussian problem. In the special linear Gaussian case, there are two established forms of mean-field process that are exact. The first one is the mean-field limit of the EnKF with perturbed observation [Reich, 2011, Del Moral and Tugaut, 2016]. The second one is the square-root form of EnKF introduced in Sec. 1.2.1.

The goal of this chapter thus is to highlight and address the issue of uniqueness in design of mean-field process. Although the issue is relevant more generally, the focus of this chapter is on the linear Gaussian

*The preliminary results concerning the contributions of this chapter appears in [Taghvaei and Mehta, 2016a].

problem.

Notation: The space of positive symmetric definite matrices of size $d \times d$ is denoted by S^d_{++}. $\mathcal{N}(m,\Sigma)$ is a Gaussian probability distribution with mean m and covariance $\Sigma \in S^d_{++}$. For a vector m, $\|m\|_2$ denotes the Euclidean norm. For a square matrix Σ, $\|\Sigma\|_F$ denotes the Frobenius norm, $\|\Sigma\|_2$ is the spectral norm, $\Sigma^\top$ is the matrix-transpose, $\mathrm{Tr}(\Sigma)$ is the matrix-trace, and $\mathrm{Ker}(\Sigma)$ denotes the null-space.

3.1.1 Related work

The technical approach based on optimal transportation has its roots in the optimal transportation theory [Evans, 1997, Villani, 2003]. These methods have been widely applied for uncertainty propagation. This includes synthesis of optimal transport maps for implementing the Bayes rules as a special case [Reich, 2011, Cheng and Reich, 2013, El Moselhy and Marzouk, 2012, Heng et al., 2015]. Also, related to the optimal transportation, is the Schrödinger bridge problem which is proposed for implementing the Bayes rule [Reich, 2018].

3.2 The Non-uniqueness Issue

Consider the linear Gaussian filtering problem:

$$\mathrm{d}X_t = AX_t\mathrm{d}t + \sigma_B \mathrm{d}B_t \quad X_0 \sim \mathcal{N}(m_{\text{init}}, \Sigma_{\text{init}}) \tag{3.3a}$$

$$\mathrm{d}Z_t = HX_t\mathrm{d}t + \mathrm{d}W_t \tag{3.3b}$$

where $X_t \in \mathbb{R}^d$ is the state at time t, $Z_t \in \mathbb{R}^m$ is the observation process, B_t, W_t are mutually independent Wiener processes taking values in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively, and A, H, σ_B are matrices of appropriate dimension. Without loss of generality, it is assumed that the covariance matrices of B_t and W_t are identity matrices. The initial condition X_0 is assumed to have a Gaussian distribution $\mathcal{N}(m_{\text{init}}, \Sigma_{\text{init}})$ with $\Sigma_{\text{init}} \succ 0$. The filtering problem is to compute the posterior distribution $\pi_t(\cdot) := \mathsf{P}(X_t \in \cdot | \mathscr{Z}_t)$, where $\mathscr{Z}_t = \sigma(Z_s;\, 0 \le s \le t)$.

The following is assumed throughout the remainder of this chapter:

Assumption (A1): The system (A,H) is detectable and (A,σ_B) is stabilizable.

In this linear Gaussian case, the posterior distribution π_t is Gaussian $\mathcal{N}(m_t,\Sigma_t)$, whose mean m_t and variance Σ_t evolve according to the Kalman-Bucy filter [Kalman and Bucy, 1961]:

$$\mathrm{d}m_t = Am_t\mathrm{d}t + \mathsf{K}_t(\mathrm{d}Z_t - Hm_t\mathrm{d}t), \quad m_0 = m_{\text{init}} \tag{3.4a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = \mathrm{Ricc}(\Sigma_t) := A\Sigma_t + \Sigma_t A^\top + \Sigma_B - \Sigma H^\top H \Sigma_t, \quad \Sigma_0 = \Sigma_{\text{init}} \tag{3.4b}$$

where $\mathsf{K}_t := \Sigma_t H^\top$ is the Kalman gain and $\Sigma_B := \sigma_B \sigma_B^\top$.

The linear FPF [Yang et al., 2016] (and also the square-root form of the EnKF [Reich, 2011]) is described

by the Mckean-Vlasov sde:

$$\mathrm{d}\bar{X}_t = A\bar{X}_t\mathrm{d}t + \sigma_B\mathrm{d}\bar{B}_t + \bar{\mathsf{K}}_t\big(\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t\big), \tag{3.5}$$

where $\bar{\mathsf{K}}_t := \bar{\Sigma}_t H^\top$ is the Kalman gain, $\bar{B}_t$ is a standard Wiener process, $\bar{m}_t := \mathsf{E}[\bar{X}_t|\mathscr{Z}_t]$, $\bar{\Sigma}_t := \mathsf{E}[(\bar{X}_t - \bar{m}_t)(\bar{X}_t - \bar{m}_t)^\top|\mathscr{Z}_t]$ are the mean-field terms, and $\bar{X}_0 \sim \mathscr{N}(m_{\text{init}}, \Sigma_{\text{init}})$. According to the following Theorem, the mean-field process $\bar{X}_t$ is exact, i.e the distribution of $\bar{X}_t$ is equal to the posterior distribution. The proof appears in the Appendix 3.5.1.

Theorem 3.1. (Exactness of linear FPF) *Consider the linear Gaussian filtering problem* (3.3a)-(3.3b) *and the linear FPF* (3.5)*. If* $\bar{X}_0 \sim \pi_{init}$*, then*

$$\bar{X}_t \sim \pi_t, \quad \forall t \geq 0 \tag{3.6}$$

The proof of exactness involves showing that the conditional mean and covariance of $\bar{X}_t$ evolve according to the Kalman filter equations for mean and covariance. Formally, upon taking the mean of the sde (3.5), the evolution of the conditional mean $\bar{m}_t$ is easily seen to be the same as the Kalman filter equation (3.4a). For the covariance, define the error process $\xi_t = \bar{X}_t - \bar{m}_t$. Then, the equation for ξ_t is obtained by subtracting (3.5) from (3.4a). This gives,

$$\mathrm{d}\xi_t = (A - \frac{1}{2}\bar{\Sigma}_t H^T H)\xi_t + \mathrm{d}\bar{B}_t$$

The equation for the variance of ξ_t is now given by the Lyapunov equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\Sigma}_t = (A - \frac{1}{2}\bar{\Sigma}_t H^T H)\bar{\Sigma}_t + \bar{\Sigma}_t(A - \frac{1}{2}\bar{\Sigma}_t H^T H)^\top + \Sigma_B = \mathrm{Ricc}(\bar{\Sigma}_t)$$

which is identical to (3.4b). The arguments of the exactness proof suggests a general procedure to construct an exact $\bar{X}_t$ process. In particular, express $\bar{X}_t$ as a sum of two terms:

$$\bar{X}_t = \bar{m}_t + \xi_t$$

Let $\bar{m}_t$ evolve according to the Kalman filter equation (3.4a). The evolution of ξ_t is defined by the sde

$$\mathrm{d}\xi_t = G_t\xi_t\mathrm{d}t + \sigma_t\mathrm{d}\bar{B}_t$$

where G_t and σ_t are solutions to the matrix equation

$$G_t\bar{\Sigma}_t + \bar{\Sigma}_t G_t^T + \sigma_t\sigma_t^\top = \mathrm{Ricc}(\bar{\Sigma}_t) \tag{3.7}$$

By construction, the equation for the variance is given by the Riccati equation (3.4b). In general, there are infinitely many solutions for (3.7). Below, we describe three solutions that lead to three established form of EnKF and linear FPF:

(i) EnKF with perturbed observation:

$$G_t = A - \bar{\Sigma}_t H^\top H, \quad \sigma_t = \begin{bmatrix} \bar{\Sigma}_t H^\top & \sigma_B \end{bmatrix}$$

(ii) Stochastic linear FPF:

$$G_t = A - \frac{1}{2}\bar{\Sigma}_t H^\top H, \quad \sigma_t = \sigma_B$$

(iii) Deterministic linear FPF:

$$G_t = A - \frac{1}{2}\bar{\Sigma}_t H^\top H + \frac{1}{2}\bar{\Sigma}_t^{-1}\Sigma_B, \quad \sigma_t = 0$$

Moreover, from a solution G_t, one can construct a family of solutions $G_t + \bar{\Sigma}_t^{-1}\Omega_t$, where Ω_t is any skew-symmetric matrix. In Sec. 3.3, we describe how to uniquely identify a solution, using optimal transportation theory.

3.2.1 Finite-N implementation

In a numerical implementation of the linear FPF algorithm (3.5), one simulates N stochastic processes (particles) $\{X_t^i : 1 \le i \le N\}$, where X_t^i is the state of the i^{th}-particle at time t. The evolution of X_t^i is obtained upon empirically approximating the mean-field terms. The finite-N filter for the linear FPF (3.5) is an interacting particle system:

$$\mathrm{d}X_t^i = AX_t^i \mathrm{d}t + \sigma_B \mathrm{d}B_t^i + \mathsf{K}_t^{(N)}(\mathrm{d}Z_t - \frac{HX_t^i + Hm_t^{(N)}}{2}\mathrm{d}t) \tag{3.8}$$

where $\mathsf{K}_t^{(N)} := \Sigma_t^{(N)} H^\top$; $\{B_t^i\}_{i=1}^N$ are independent copies of B_t; $X_0^i \overset{\text{i.i.d}}{\sim} \mathcal{N}(m_0, \Sigma_0)$ for $i = 1, 2, \ldots, N$; and the empirical approximations of the two mean-field terms are as follows:

$$\begin{aligned} m_t^{(N)} &:= \frac{1}{N}\sum_{j=1}^N X_t^i, \\ \Sigma_t^{(N)} &:= \frac{1}{N-1}\sum_{j=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^\top \end{aligned} \tag{3.9}$$

3.3 Optimal Transport FPF

3.3.1 Background on Optimal transportation

Let μ_X and μ_Y be two given probability measures on $\mathbb{R}^d$ with finite second moments. The Monge optimal transportation problem is to minimize

$$\min_T \ \mathsf{E}[(T(X) - X)^2] \tag{3.10}$$

over all measurable maps $T : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$X \sim \mu_X,\ T(X) \sim \mu_Y \tag{3.11}$$

If it exists, the minimizer T is called the *optimal transport map* between μ_X and μ_Y. The optimal cost is referred to as L^2-Wasserstein distance between μ_X and μ_Y.

The explicit form of the optimal transport map is known when the marginal distributions are Gaussians. The explicit form of the optimal transport map is given in the following Theorem. This explicit form of optimal transport map is used in our proposed time-stepping procedure in Sec. 3.3.2.

Theorem 3.2. *(Optimal map between Gaussians [Givens et al., 1984, Prop. 7]) Consider the optimization problem (3.10), with constraint (3.11). Suppose μ_X and μ_Y are Gaussian distributions, $\mathcal{N}(m_X, \Sigma_X)$ and $\mathcal{N}(m_Y, \Sigma_Y)$, with $\Sigma_X, \Sigma_Y \succ 0$. Then the optimal transport map between μ_X and μ_Y is given by*

$$T(x) = m_Y + F(x - m_X) \tag{3.12}$$

where $F = \Sigma_Y^{\frac{1}{2}}(\Sigma_Y^{\frac{1}{2}}\Sigma_X\Sigma_Y^{\frac{1}{2}})^{-\frac{1}{2}}\Sigma_Y^{\frac{1}{2}}$.

3.3.2 The Time stepping optimization procedure

The following time stepping optimization procedure is proposed to obtain the optimal transport FPF:

1. Divide the time interval $[0, T]$ into $n \in \mathbb{N}$ equal time steps with the time instants $t_0 = 0 < t_1 < \ldots < t_n = T$.

2. Initialize a discrete time random process $\{\bar{X}_{t_k}\}_{k=1}^n$ according to the initial distribution (prior) of X_0,

$$\bar{X}_{t_0} \sim \pi_0$$

3. For each time step $[t_k, t_{k+1}]$, evolve the process $\bar{X}_{t_k}$ according to

$$\bar{X}_{t_{k+1}} = T_k(\bar{X}_{t_k}), \quad \text{for} \quad k = 0, \ldots, n-1 \tag{3.13}$$

   where the map T_k is the optimal transport map between two probability measures π_{t_k} and $\pi_{t_{k+1}}$.

4. Take the limit as $n \to \infty$ to obtain the continuous-time process $\bar{X}_t$ and the sde:

$$\mathrm{d}\bar{X}_t = u_t(\bar{X}_t)\mathrm{d}t + \bar{\mathsf{K}}_t(\bar{X}_t)\mathrm{d}Z_t \tag{3.14}$$

The procedure leads to the control laws u_t and $\bar{\mathsf{K}}_t$ that depend upon π_t. Since π_t is unknown, one simply replaces it with $\tilde{\pi}_t(\cdot) := \mathsf{P}(\bar{X}_t \in \cdot | \mathscr{Z}_t)$ – as the two are identical by construction. The resulted sde (3.14) is referred to as the *optimal transport FPF*. Explicit formula for the optimal transport FPF in the linear Gaussian case is the subject of the following. The proof appears in Appendix 3.5.2.

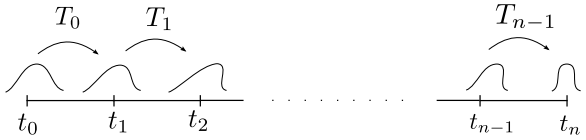


Figure 3.1: The time stepping optimization procedure.

Proposition 3.1. *Consider the linear Gaussian filtering problem* (3.3a)-(3.3b)*. Assume* $\Sigma_0 \succ 0$ *and Assumption A1 holds. Then the optimal transport FPF is given by*

$$\mathrm{d}\bar{X}_t = A\bar{m}_t \mathrm{d}t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - H\bar{m}\mathrm{d}t) + G_t(\bar{X}_t - \bar{m}_t)\mathrm{d}t \tag{3.15}$$

where $\bar{\mathsf{K}}_t := \bar{\Sigma}_t H^T$, $\bar{m}_t = \mathsf{E}[\bar{X}_t|\mathscr{Z}_t]$, $\bar{\Sigma}_t = \mathsf{E}[(\bar{X}_t - \bar{m}_t)(\bar{X}_t - \bar{m}_t)^\top|\mathscr{Z}_t]$, $\bar{X}_0 \sim \mathscr{N}(m_0, \Sigma_0)$*, and* G_t *is the (unique) symmetric matrix that solves the matrix equation*

$$G_t\bar{\Sigma}_t + \bar{\Sigma}_t G_t = Ricc(\bar{\Sigma}_t) \tag{3.16}$$

The filter is exact. That is, the conditional distribution of $\bar{X}_t$ *is Gaussian* $\mathscr{N}(\bar{m}_t, \bar{\Sigma}_t)$ *with* $\bar{m}_t = m_t$ *and* $\bar{\Sigma}_t = \Sigma_t$*.*

Remark 3.1. The unique symmetric solution to the matrix equation (3.16) is given by:

$$G_t = \int_0^\infty e^{-s\bar{\Sigma}_t}\mathrm{Ricc}(\bar{\Sigma}_t)e^{-s\bar{\Sigma}_t}\mathrm{d}s$$

For the purpose of comparison to the original form of the FPF algorithm, the solution can be expressed as:

$$G_t = A - \frac{1}{2}\bar{\Sigma}_t H^\top H + \frac{1}{2}\Sigma_B\bar{\Sigma}_t^{-1} + \Omega_t\bar{\Sigma}_t^{-1}$$

where Ω_t is the (unique) skew-symmetric matrix that solves the matrix equation

$$\Omega_t\bar{\Sigma}_t^{-1} + \bar{\Sigma}_t^{-1}\Omega_t = (A^\top - A) + \tfrac{1}{2}(\bar{\Sigma}_t H^\top H - H^\top H\bar{\Sigma}_t) + \tfrac{1}{2}(\Sigma_B\bar{\Sigma}_t - \bar{\Sigma}_t\Sigma_B) \tag{3.17}$$

Using this form of the solution, the optimal transport sde (3.15) is expressed as

$$\mathrm{d}\bar{X}_t = A\bar{X}_t\mathrm{d}t + \frac{1}{2}\Sigma_B\bar{\Sigma}_t^{-1}\,(\bar{X}_t - \bar{m}_t)\mathrm{d}t - \bar{\mathsf{K}}_t(\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t) + \Omega_t\bar{\Sigma}_t^{-1}(\bar{X}_t - \bar{m}_t)\mathrm{d}t \tag{3.18}$$

Compared to the original (linear Gaussian) FPF (3.5), the optimal transport FPF (3.18) has two differences:

(i) The stochastic term $\mathrm{d}\bar{B}_t$ is replaced with the deterministic term $\frac{1}{2}\Sigma_B\bar{\Sigma}_t^{-1}(\bar{X}_t - \bar{m}_t)\mathrm{d}t$. Given a Gaussian prior, the two terms yield the same posterior. However, in a finite-N implementation,

the difference becomes significant. The stochastic term serves to introduce an additional variance error of order $O(\frac{1}{\sqrt{N}})$.

(ii) The sde (3.18) has an extra term involving the skew-symmetric matrix Ω_t. The extra term does not effect the posterior distribution. This term is viewed as a correction term that serves to make the dynamics symmetric and hence optimal in the optimal transportation sense. It is noted that for the scalar ($d = 1$) case, the skew-symmetric term is zero. Therefore, in the scalar case, the update formula in the original FPF (3.5) is optimal. In the vector case, it is optimal iff $\Omega_t \equiv 0$.

Remark 3.2 (Finite-N implementation)**.** *A finite-N implementation of the optimal transport linear FPF* (3.15) *requires empirical approximation of the mean-field terms $\bar{m}_t$ and $\bar{\Sigma}_t$. However, with the empirical approximation, the assumption $\Sigma_0^{(N)} \succ 0$ may not be satisfied. In particular, when $N < d$, the empirical covariance matrix is of rank $N < d$, hence singular. If the covariance matrix is singular, the optimal transport FPF can not be implemented, because a solution to the Lyapunov equation* (3.16) *may not exist. In contrast, the stochastic linear FPF* (3.5) *does not have any terms involving $\bar{\Sigma}^{-1}$ and can be approximated for any choice of N. In Sec. 3.4, we propose an alternative approach that allows for singular covariance matrix.*

3.4 The singular covariance case

The derivation of the optimal transport linear FPF (3.15) crucially relies on the assumption that $\bar{\Sigma}_0 \succ 0$ which in turn implies that, in the time-stepping procedure, $\bar{\Sigma}_{t_k}^{(N)} \succ 0$ for $k = 0, 1, \ldots, n-1$. In the proof of Prop. 3.1, the assumption is used to derive the optimal transport map T_k (see (3.12)). In general, when the covariance of Gaussian random variables $\bar{X}_{t_k}$ or $\bar{X}_{t_{k+1}}$ is singular, the optimal transport map T_k may not exist. In the singular case, the relaxed form of the optimal transportation problem, first introduced by Kantorovich, is used to search for optimal (stochastic) couplings instead of (deterministic) transport maps [Villani, 2003].

$$\min_{\pi} \ \mathsf{E}_{(X,Y)\sim\pi}[|X-Y|^2] \tag{3.19}$$

where π is a joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$, with marginals equal to μ_X and μ_Y.

Example 3.1. *Consider Gaussian random variable X and Y with distributions, $\mathcal{N}(m_X, \Sigma_X)$ and $\mathcal{N}(m_Y, \Sigma_Y)$, respectively. Suppose*

$$m_X = m_Y = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma_X = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}, \quad \Sigma_Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

where $\varepsilon \geq 0$ is small. If $\varepsilon > 0$, the optimal transportation map exists, and is given by

$$Y = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{\varepsilon}} \end{bmatrix} X$$

If $\varepsilon = 0$, then there is no transport map that satisfies the constraints of the optimal transportation problem.

However, a (stochastic) coupling that minimizes the Kantorovich problem (3.19) exists, given by

$$Y = X + \begin{bmatrix} 0 \\ 1 \end{bmatrix} B$$

where $B \sim \mathcal{N}(0,1)$ is independent of X.

In order to do consider the singular covariance case, the time-stepping procedure should be modified to consider a sequence of optimal stochastic couplings, instead of deterministic optimal transport maps. However, carrying out the time-stepping construction with stochastic maps become complicated due to lack of a unique and explicit form of the stochastic couplings. Instead, we begin with a suitable general form of the sde for the mean-field process

$$\mathrm{d}\bar{X}_t = G_t(\bar{X}_t - \bar{m}_t)\mathrm{d}t + \mathrm{d}v_t + \sigma_t \mathrm{d}\bar{B}_t \tag{3.20}$$

and then choose G_t, v_t and σ_t such that the stochastic map $\bar{X}_t \to \bar{X}_{t+\Delta t} = \bar{X}_t + \int_t^{t+\Delta t} G_s(\bar{X}_s - \bar{m}_s)\mathrm{d}s + (v_{t+\Delta t} - v_t) + \int_t^{t+\Delta t} \sigma_s \mathrm{d}B_s$ is optimal, in the optimal transportation sense, in the asymptotic limit as $\Delta t \to 0$. In Appendix 3.5.3, we show that $\mathrm{d}v_t = A\bar{m}_t\mathrm{d}t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - H\bar{m}_t)\mathrm{d}t$, and σ_t and G_t are the solutions to the following optimization problem:

Optimization problem: Define the sets

$$\mathscr{D}_\Sigma := \{(\sigma, G) \in \mathbb{R}^{d\times d_B} \times \mathbb{R}^{d\times d};\ G\Sigma + \Sigma G^\top + \sigma\sigma^\top = \mathrm{Ricc}(\Sigma)\}$$
$$\mathscr{D}_\Sigma|_{\sigma^*} := \{G \in \mathbb{R}^{d\times d};\ G\Sigma + \Sigma G^\top + \sigma^*\sigma^{*\top} = \mathrm{Ricc}(\Sigma)\}$$

The pair $(\sigma^*, G^*) \in \mathscr{D}_\Sigma$ is optimal if

$$\mathrm{Tr}(\sigma^*(\sigma^*)^\top) = \min_{(\sigma,G)\in\mathscr{D}_\Sigma} \mathrm{Tr}(\sigma\sigma^\top), \quad \mathrm{Tr}(G^*\bar{\Sigma}_t G^*) = \min_{G\in\mathscr{D}_\Sigma|_{\sigma^\star}} \mathrm{Tr}(G\bar{\Sigma}_t G^\top) \tag{3.21}$$

The justification for the optimization problem appears in Appendix 3.5.3 where the following proposition concerning its solution is also proved.

Proposition 3.2. *Consider the optimization problem (3.21). Let*

$$u_t = \arg\min_{u\in\mathbb{R}^{d\times d_B}} \|\bar{\Sigma}_t u - \sigma_B\| = Proj(\sigma_B; Range(\bar{\Sigma}_t))$$

Then, $\sigma_t^* = \sigma_B - \bar{\Sigma}_t u_t$ and the unique symmetric matrix $G_t^\star$ that solves the Lyapunov equation

$$G_t^*\bar{\Sigma}_t + \bar{\Sigma}_t G_t^* = Ricc(\bar{\Sigma}_t) - \sigma_t^*(\sigma_t^*)^\top \tag{3.22}$$

minimize the optimization problem (3.21)*. The resulting optimal transport linear FPF is*

$$\mathrm{d}\bar{X}_t = A\bar{m}_t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - H\bar{m}_t\mathrm{d}t) + G_t^*(\bar{X}_t - \bar{m}_t)\mathrm{d}t + \sigma_t^*\mathrm{d}\bar{B}_t \tag{3.23}$$

Remark 3.3. *The solution to the Lyapunov equation* (3.22) *can be expressed as*

$$G_t = A - \frac{1}{2}\bar{\Sigma}_t HH^\top + \frac{1}{2}(\sigma_B + \sigma_t^*)u_t^\top + \Omega_t^{(0)}P_K + \Omega_t^{(1)}(P_R\bar{\Sigma}_t P_R)^{-1} \tag{3.24}$$

where $\Omega^{(0)}$ is an arbitrary $d\times d$ matrix, $\Omega^{(1)}$ is a skew-symmetric $d\times d$ matrix, P_R is the projection into range of $\bar{\Sigma}_t$, and P_K is projection into the kernel of $\bar{\Sigma}_t$. The matrices $\Omega^{(0)}$ and $\Omega^{(1)}$ are chosen such that the matrix G_t symmetric. Using this form of solution, the optimal transport linear FPF (3.23) *is as follows*

$$\begin{aligned}\mathrm{d}\bar{X}_t =& A\bar{X}_t\mathrm{d}t + \frac{1}{2}(\sigma_B+\sigma_t^*)u_t^\top(\bar{X}_t-\bar{m}_t)\mathrm{d}t + \sigma_t^*\mathrm{d}\bar{B}_t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t)\\ &+ (\Omega_t^{(0)}P_K + \Omega_t^{(1)}(P_R\bar{\Sigma}_tP_R)^{-1})(\bar{X}_t-\bar{m}_t)\mathrm{d}t\end{aligned} \tag{3.25}$$

It is noted that in the formula (3.25) *reduces to the formula* (3.18) *when the covariance matrix is non-singular, because $u_t = Prroj(\sigma_B; Range(\bar{\Sigma}_t)) = \sigma_B$, $\sigma_t^* = 0$, $P_K = 0$ and $P_R = I$. In particular, the stochastic term $\sigma_t\mathrm{d}B_t$ is zero when $\sigma_B \in Range(\bar{\Sigma}_t)$. The role of the stochastic term $\sigma_t^*\mathrm{d}\bar{B}_t$ is to account for the the effect of the process noise $\sigma_B\mathrm{d}B_t$ that can not be captured with the linear term $G_t(\bar{X}_t-\bar{m}_t)\mathrm{d}t$.*

3.4.1 Finite-N implementation

In a numerical implementation of the optimal transport linear FPF algorithm (3.23), one simulates N particles by empirically approximating the mean-field terms. The evolution of the particles is given by the sde

$$\mathrm{d}X_t^i = Am_t^{(N)}\mathrm{d}t + \mathsf{K}_t^{(N)}(\mathrm{d}Z_t - Hm_t^{(N)}\mathrm{d}t) + G_t^{(N)}(X_t^i - m_t^{(N)})\mathrm{d}t + \sigma_t\mathrm{d}B_t^i, \quad X_0^i \sim \mathcal{N}(m_0,\Sigma_0) \tag{3.26}$$

where $\mathsf{K}_t^{(N)} := \Sigma_t^{(N)}H^\top$; $\{B_t^i\}_{i=1}^N$ are independent copies of B_t, $\sigma_t = \sigma_B - u_t$, $u_t = \arg\min_u \|\Sigma_t^{(N)}u - \sigma_B\|$, $G_t^{(N)}$ is the unique symmetric matrix solution to the Lyapunov equation

$$G_t^{(N)}\Sigma_t^{(N)} + \Sigma_t^{(N)}G_t^{(N)} = \mathrm{Ricc}(\Sigma_t^{(N)}) - \sigma_t\sigma_t^\top$$

and $m_t^{(N)}$ and $\Sigma_t^{(N)}$ are empirical mean and covariance defined in (3.9). Note that the stochastic term is zero when $\sigma_B \in \mathrm{Range}(\Sigma_t^{(N)})$, which is true when $\sigma_B \in \mathrm{span}\{X_t^1,\ldots,X_t^N\}$.

3.5 Proof of the main results

3.5.1 Proof of Theorem 3.1

Proof. First, we show that the conditional mean and covariance of $\bar{X}_t$ evolve according to Kalman filtering equations. Express the sde (3.5) in integral form,

$$\bar{X}_t = \bar{X}_0 + \int_0^t \sigma_b \mathrm{d}\bar{B}_s + \int_0^t \bar{\mathsf{K}}_s(\mathrm{d}Z_s - \frac{H\bar{X}_s + H\bar{m}_s}{2}\mathrm{d}s)$$

Upon taking the conditional expectation of both sides

$$\begin{aligned}
\bar{m}_t &= \mathsf{E}[\bar{X}_0|\mathscr{Z}_t] + \mathsf{E}[\int_0^t \sigma_b \mathrm{d}\bar{B}_s|\mathscr{Z}_t] \\
&\quad + \mathsf{E}[\int_0^t \bar{\mathsf{K}}_s(\mathrm{d}Z_s - \frac{H\bar{X}_s + H\bar{m}_s}{2}\mathrm{d}s)|\mathscr{Z}_t] \\
&= \mathsf{E}[\bar{X}_0|\mathscr{Z}_0] + \int_0^t \mathsf{E}[\bar{\mathsf{K}}_s|\mathscr{Z}_s]\mathrm{d}Z_s - \int_0^t \mathsf{E}[\bar{\mathsf{K}}_s\frac{H\bar{X}_s + H\bar{m}_s}{2}|\mathscr{Z}_s]\mathrm{d}s \\
&= \bar{m}_0 + \int_0^t \bar{\mathsf{K}}_s(\mathrm{d}Z_s - H\bar{m}_s\mathrm{d}s)
\end{aligned}$$

where we used the fact that $\bar{X}_t$ is adapted to the filteration $\mathscr{Z}_t$ to obtain the second line (see [Xiong, 2008, Lemma 5.4]). As a result, the sde for the conditional mean is

$$\mathrm{d}\bar{m}_t = A\bar{m}_t\mathrm{d}t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - H\bar{m}_t\mathrm{d}t) \tag{3.27}$$

Define the error ξ_t according to $\xi_t := \bar{X}_t - \bar{m}_t$. The equation for ξ_t is obtained by simply subtracting (3.27) from (3.5). This gives,

$$\mathrm{d}\xi_t = (A - \frac{\bar{\Sigma}_t H^T H}{2})\xi_t + \sigma_B \mathrm{d}\bar{B}_t$$

By application of the Itô rule

$$\begin{aligned}
\mathrm{d}(\bar{\xi}_t\bar{\xi}_t^\top) =& (A - \frac{1}{2}\bar{\Sigma}_t H^\top H)\bar{\xi}_t\bar{\xi}_t^\top \mathrm{d}t + \sigma_B \mathrm{d}\bar{B}_t\bar{\xi}_t^\top \\
&+ \bar{\xi}_t\bar{\xi}_t^\top(A^\top - \frac{1}{2}H^\top H\bar{\Sigma}_t)\mathrm{d}t + \bar{\xi}_t(\sigma_B\mathrm{d}\bar{B}_t)^\top + \Sigma_B\mathrm{d}t
\end{aligned}$$

which concludes the sde for the conditional covariance $\bar{\Sigma}_t = \mathsf{E}[\bar{\xi}_t\bar{\xi}_t^\top|\mathscr{Z}_t]$ following the same procedure as for the conditional mean,

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\Sigma}_t = A\bar{\Sigma}_t + \bar{\Sigma}_t A^T + \Sigma_B - \bar{\Sigma}_t H^T H\bar{\Sigma}_t$$

which is identical to the Ricatti equation (3.4b). Hence $\bar{\Sigma}_t = \Sigma_t$ for all $t \geq 0$ because $\bar{\Sigma}_0 = \Sigma_0$. This also implies $\bar{\mathsf{K}}_t = \mathsf{K}_t$, which further implies that the sde for conditional mean (3.27) is identical to the Kalman filter equation (3.4a). Therefore, $\bar{m}_t = m_t$ for all $t \geq 0$ because $\bar{m}_0 = m_0$.

Given $\bar{\Sigma}_t = \Sigma_t$ and $\bar{m}_t = m_t$, the mean-field terms in the McKean-Vlasov sde (3.5) can be treated as exogenous processes. Therefore, the McKean-Vlasov sde (3.5) simplifies to a Ornstein-Uhlenbeck sde.

Because the distribution of the initial condition $\bar{X}_0$ is Gaussian, the distribution of $\bar{X}_t$ is also Gaussian given by $\mathcal{N}(m_t, \Sigma_t)$ which is equal to the posterior distribution given by Kalman filter and concludes the proof. □

3.5.2 Proof of Proposition 3.1

The key step in the proof is the following Lemma,

Lemma 3.1. *Consider the ode (3.4b). Let Σ_t be the solution for $t \in [0,T]$. Then the following relationship holds ,*

$$\Sigma_{t+\Delta t}^{\frac{1}{2}}(\Sigma_{t+\Delta t}^{\frac{1}{2}}\Sigma_t\Sigma_{t+\Delta t}^{\frac{1}{2}})^{-\frac{1}{2}}\Sigma_{t+\Delta t}^{\frac{1}{2}} = I + G_t\Delta t + O(\Delta t^2), \tag{3.28}$$

where G_t is the solution to the matrix equation,

$$G_t\Sigma_t + \Sigma_t G_t = A\Sigma_t + \Sigma_t A^T + I - \Sigma_t H^T H \Sigma_t, \tag{3.29}$$

and the second order term is uniformly bounded for all $t \in [0,T]$.

Proof. The solution Σ_t is positive and bounded since the system is observable [Ocone and Pardoux, 1996]. Fix $t \in [0,T]$, and define

$$F(s) := \Sigma_{t+s}^{\frac{1}{2}}(\Sigma_{t+s}^{\frac{1}{2}}\Sigma_t\Sigma_{t+s}^{\frac{1}{2}})^{-\frac{1}{2}}\Sigma_{t+s}^{\frac{1}{2}}.$$

The relationship (3.28) is obtained by considering the Taylor series of $F(s)$ at $s = 0$,

$$F(\Delta t) = I + \dot{F}(0)\Delta t + \frac{1}{2}\ddot{F}(\tau)\Delta t^2,$$

for some $\tau \in [0,\Delta t]$, and showing that $\dot{F}(0) = G_t$. This is verified by considering,

$$F(s)\Sigma_t F(s) = \Sigma_{t+s}.$$

On evaluating the derivative with respect to s at $s = 0$,

$$\dot{F}(0)\Sigma_t + \Sigma_t\dot{F}(0) = A\Sigma_t + \Sigma_t A^T + I - \Sigma_t H^\top H \Sigma_t.$$

Since the solution to the Lyapunov equation (3.29) is unique, $\dot{F}(0) = G_t$. Also the second order derivative is uniformly bounded for all $t \in [0,T]$, by the observability assumption. □

Proof. (Prop. 3.1) The proof of exactness is similar to the proof of Theorem 3.1 and is omitted. In order to obtain the optimal transport sde, the time stepping procedure is used. The key step in the procedure is to obtain the optimal transport map T_k. The optimal map is between two Gaussians, $\mathcal{N}(m_{t_k}, \Sigma_{t_k})$ and $\mathcal{N}(m_{t_{k+1}}, \Sigma_{t_{k+1}})$. By Theorem 3.2-(ii), the optimal map is,

$$\bar{X}_{t_{k+1}} = m_{t_{k+1}} + F_k(\bar{X}_{t_k} - m_{t_k}),$$

where $F_k = \Sigma_{t_{k+1}}^{\frac{1}{2}}(\Sigma_{t_{k+1}}^{\frac{1}{2}}\Sigma_{t_k}\Sigma_{t_{k+1}}^{\frac{1}{2}})^{-\frac{1}{2}}\Sigma_{t_{k+1}}^{\frac{1}{2}}$. Using Lemma 3.1,

$$\bar{X}_{t_{k+1}} = \bar{m}_{t_{k+1}} + (\bar{X}_{t_k} - m_{t_k}) + G_k(\bar{X}_{t_k} - m_{t_k})\Delta t + O(\Delta t^2).$$

To obtain the sde, take a sum over $k = 0, 1, \ldots, n-1$,

$$\bar{X}_{t_n} = \bar{X}_{t_0} + \bar{X}_{t_n} - m_{t_0} + \sum_{k=0}^{n-1}\left[G_k(\bar{X}_{t_k} - m_{t_k})\Delta t + O(\Delta t^2)\right].$$

In the limit as $\Delta t \to 0$,

$$\bar{X}_{t_n} = \bar{X}_{t_0} + m_{t_n} - \hat{X}_{t_0} + \int_0^t G_s(\bar{X}_s - m_s)\mathrm{d}s.$$

where the uniform boundedness of the second order term is used. The associated sde is,

$$\mathrm{d}\bar{X}_t = \mathrm{d}m_t + G_t(\bar{X}_t - m_t)\mathrm{d}t,$$

where $\mathrm{d}m_t$ is given by (3.4a). Finally one obtains (3.15) by replacing m_t and Σ_t with $\bar{m}_t$ and $\bar{\Sigma}_t$ respectively, which are identical by exactness. □

3.5.3 Justification for the optimization problem (3.21) and proof of Prop. 3.2

The stochastic map $\bar{X}_t \to \bar{X}_{t+\Delta t}$ is equal to

$$\begin{aligned}\bar{X}_{t+\Delta t} &= \bar{X}_t + \int_t^{t+\Delta t} G_s(\bar{X}_s - \bar{m}_s)\mathrm{d}s + (v_{t+\Delta t} - v_t) + \int_t^{t+\Delta t}\sigma_s \mathrm{d}B_s \\ &= \bar{X}_t + \Delta t G_t(\bar{X}_t - \bar{m}_t) + v_{t+\Delta t} - v_t + \sqrt{\Delta t}\sigma_t\zeta + o(\Delta t)\end{aligned}$$

where $\zeta \sim \mathcal{N}(0,1)$. In order for the stochastic map to be optimal, it should satisfy the marginal constant $\bar{X}_{t+\Delta t} \sim \mathcal{N}(m_{t+\Delta t}, \Sigma_{t+\Delta t})$ if $\bar{X}_t \sim \mathcal{N}(m_t, \Sigma_t)$, and minimize the cost $\mathsf{E}[|\bar{X}_{t+\Delta t} - \bar{X}_t|^2]$. The marginal constraint is satisfied if

$$\begin{aligned} m_t + v_{t+\Delta t} - v_t &= m_{t+\Delta t} \\ (I + \Delta t G_t)\Sigma_t(I + \Delta t G_t) + \Delta t\sigma_t\sigma_t^\top + o(\Delta t) &= \Sigma_{t+\Delta t}\end{aligned}$$

The first constraint implies $\mathrm{d}v_t = \mathrm{d}m_t = Am_t\mathrm{d}t + \mathsf{K}_t(\mathrm{d}Z_t - Hm_t\mathrm{d}t)$. Dividing the second constraint by Δt and taking the limit as $\Delta t \to 0$ concludes

$$G_t\Sigma_t + \Sigma_t G_t^\top + \sigma_t\sigma_t^\top = \mathrm{Ricc}(\Sigma_t) \tag{3.30}$$

The optimal transportation cost is

$$\mathsf{E}[|\bar{X}_{t+\Delta t} - \bar{X}_t|^2] = |m_{t+\Delta t} - m_t|^2 + \Delta t\underbrace{\mathrm{Tr}(\sigma_t\sigma_t^\top)}_{f_1(\sigma_t)} + (\Delta t)^2\underbrace{\mathrm{Tr}(G_t\Sigma_t G_t^\top)}_{f_2(G_t)} + o(\Delta t^2)$$

Taking the limit as $\Delta t \to 0$ implies that one should minimize $f_1(\sigma_t)$ first, and then $f_2(G_t)$, under the constraint (3.30). This justifies the optimization problem (3.21).

In order to prove Proposition 3.2, decompose $\sigma_B = \sigma_t^* + \Sigma_t u_t$, where $u_t = \arg\min_u \|\Sigma_t u - \sigma_B\|^2$. As a result, $\sigma_t^* \in \mathrm{Ker}(\Sigma_t)$. Express $G_t = A - \frac{1}{2}\Sigma_t H^\top H + (\sigma_t^* + \frac{1}{2}\Sigma_t u_t)u_t^\top + \tilde{G}_t$, where $\tilde{G}_t$ is the new optimization variable. With the new variable, the constraint of the optimization problem becomes

$$\sigma\sigma^\top + \tilde{G}\Sigma_t + \Sigma_t \tilde{G}^\top = \sigma_t^*(\sigma_t^*)^\top$$

Multiply both sides from left and right by the projection operator P_K, into the kernel of Σ_t, to obtain

$$P_K \sigma\sigma^\top P_K = \sigma_t^*(\sigma_t^*)^\top$$

where we used $P_K\Sigma_t = 0$ and $P_K\sigma_t^* = \sigma_t^*$. Then, it is clear that the minimizer of $\mathrm{Tr}(\sigma\sigma^\top)$ under the constraint $P_K\sigma\sigma^\top P_K = \sigma_t^*(\sigma_t^*)^\top$ is equal to $\sigma = \sigma_t^*$. The new constraint, with $\sigma = \sigma_t^*$, is

$$G\Sigma_t + \Sigma_t G^\top = \mathrm{Ricc}(\Sigma_t) - \sigma_t^*(\sigma_t^*)$$

and the optimization problem is to minimize $\mathrm{Tr}(G\Sigma_t G^\top)$. Using the spectral representation $\Sigma_t = \sum_{m=1}^d \lambda_m u_m u_m^\top$ where u_m are orthogonal eigenvectors, and $\lambda_m \geq 0$ are eigenvalues, the constraint and optimization problem is

$$\begin{aligned} \text{minimize} \quad & \sum_{m,n=1}^{d} \lambda_m G_{mn} G_{nm} \\ \text{subject to} \quad & \lambda_n G_{mn} + \lambda_m G_{nm} = R_{mn}, \quad \text{for} \quad m,n = 1,\ldots,d \end{aligned}$$

where $G_{nm} = u_n^\top G u_m$ and $R = \mathrm{Ricc}(\Sigma_t) - \sigma_t^*(\sigma_t^*)$. The solution to the optimization problem is $G_{nm}^* = \frac{R_{mn}}{\lambda_m + \lambda_n}$. Therefore, the matrix G_t^* is symmetric, and unique symmetric solution to the Lyapunov equation (3.22).

Chapter 4

Finite-N system error analysis*

4.1 Introduction

This chapter is concerned with error analysis and long-term stability analysis of the FPF algorithm. In the mean-field ($N = \infty$) limit, the FPF is known to be exact, i.e, the conditional probability distribution of the particles is equal to the posterior distribution. However, little is known about the convergence of the finite-N system to the mean-field limit. The objective of this chapter is to address some of these questions in the linear Gaussian setting.

In the linear Gaussian setting, the FPF algorithm is similar to the ensemble Kalman filter algorithm (see Sec. 1.2.1). Ensemble Kalman filter (EnKF) was first introduced in [Evensen, 1994], in discrete time setting, as an alternative to the extended Kalman filter (EKF) for applications in geophysical sciences. In these applications, the state dimension is typically very high. The main advantage of the EnKF, compared to the EKF, is that the computational cost of the EnKF scales linearly with the state dimension whereas the computational cost of the EKF scales as the dimension squared.

Since its introduction, the EnKF has evolved into different formulations. The most two well-known formulations are (i) EnKF based on perturbed observation [Evensen, 2003] and (ii) the square root EnKF [Whitaker and Hamill, 2002]. For a review of the different discrete time formulations of the EnKF see [Reich and Cotter, 2015, Ch. 6-7] [Law et al., 2015, Ch. 4]. The two aforementioned discrete time formulations of the EnKF algorithm have been extended to the continuous time setting [Bergemann and Reich, 2012]. The continuous time formulation of the EnKF is usually referred to as the ensemble Kalman-Bucy filter (EnKBF). For a recent review of the EnKBF algorithm and its connection to the FPF algorithm see [Taghvaei et al., 2018]. The EnKBF algorithm and the linear FPF have the following three established formulations:

(i) EnKBF with perturbed observation [Bergemann and Reich, 2012] [Del Moral and Tugaut, 2016];

(ii) Stochastic linear FPF [Yang et al., 2016, Eq. (26)] which is same as the square root EnKBF [Bergemann and Reich, 2012];

(iii) Deterministic linear FPF [Taghvaei and Mehta, 2016a, Eq. (15)] [de Wiljes et al., 2016];

In this chapter, the stochastic linear FPF and the deterministic linear FPF are studied. Both the formulations are exact in the following sense: In the mean-field limit the distribution of the particles equals the posterior

*The preliminary results concerning the contributions of this chapter appears in [Taghvaei and Mehta, 2018a,b].

distribution of the filter. The main difference between the two formulations is that the process noise term in the stochastic FPF is replaced with a deterministic term in the deterministic FPF.

The goal of this chapter is to characterize the error properties of the FPF in the limit when the number of particles N is large but finite. The error metrics of interest include the mean-squared error between the finite-N estimates (empirical mean and the empirical covariance) and their mean-field limits (conditional mean and covariance). Additionally, it is of interest to investigate the convergence of the empirical distribution of the interacting particle system to the conditional distribution obtained in the mean-field limit.

4.1.1 Literature review on error analysis of EnKF

Theoretical error and convergence analysis of the EnKF algorithm is an active area of research. In the discrete time setting, it is shown that the ensemble distribution converges to the mean-field limit with the convergence rate $O(\frac{1}{\sqrt{N}})$ for any finite time [Le Gland et al., 2009] [Mandel et al., 2011]. The asymptotic (in time) stability analysis is more difficult. It is shown that if the system dynamics is stable and admits a Lyapunov function, and the observation model satisfies the "observable energy criterion" (which holds under full state observation), then the system is ergodic and it is stable with respect to initial conditions [Tong et al., 2016]. The well-posedness of the EnKF and its accuracy using the variance inflation technique is studied in [Kelly et al., 2014]. Related finite-time results on the convergence of the discrete-time square root EnKF appear in [Kwiatkowski and Mandel, 2015]. The analysis in [Kwiatkowski and Mandel, 2015] is simpler as the model is deterministic and the update formula exactly equals the Kalman filter update formula.

The analysis for EnKBF and linear FPF is more recent. For EnKBF with perturbed observation, under certain assumptions (stable and fully observable), it has been shown that the empirical distribution of the ensemble converges to the mean-field distribution uniformly for all time with the rate $O(\frac{1}{\sqrt{N}})$ [Del Moral and Tugaut, 2016]. This result has been extended to the nonlinear setting for the case with Langevin type dynamics with a strongly convex potential and full linear observation [Del Moral et al., 2017].

4.1.2 Notation

For a vector m, $\|m\|_2$ denotes the Euclidean norm. For a square matrix Σ, $\|\Sigma\|_F$ denotes the Frobenius norm, $\|\Sigma\|_2$ is the spectral norm, $\Sigma^\top$ is the matrix-transpose, $\mathrm{tr}(\Sigma)$ is the matrix-trace, and $\mathrm{cond}(\Sigma) = \|\Sigma\|_2\|\Sigma^{-1}\|_2$ is the condition number. The space of symmetric positive definite matrices is denoted by S^d_{++}. $\mathcal{N}(m,\Sigma)$ denotes a Gaussian probability distribution with mean m and covariance $\Sigma \in S^d_{++}$. There are three types of stochastic process considered in this chapter: (i) X_t denotes the state of the (hidden) signal at time t; (ii) X^i_t denotes the state of the i^{th} particle in a population of N particles; and (iii) $\bar{X}_t$ denotes the state of the McKean-Vlasov model obtained in the mean-field limit ($N = \infty$). The mean and the covariance for these are denoted as follows: (i) (m_t, Σ_t) is the conditional mean and the conditional covariance pair for X_t; (ii) $(m^{(N)}_t, \Sigma^{(N)}_t)$ is the empirical mean and the empirical covariance for the ensemble $\{X^i_t\}_{i=1}^N$; and (iii) $(\bar{m}_t, \bar{\Sigma}_t)$ is the conditional mean and the conditional covariance for the mean-field process $\bar{X}_t$.

4.2 Problem formulation

We consider the linear Gaussian filtering problem (1.4a)-(1.4b), where the solution is given by the Kalman-Bucy filter (1.5a)-(1.5b). We consider two types of linear FPF algorithms: The deterministic linear FPF, and the stochastic linear FPF. The two algorithms are described next.

4.2.1 Deterministic linear FPF

The mean-field process $\bar{X}_t$ evolves according to:

$$\mathrm{d}\bar{X}_t = A\bar{m}_t\mathrm{d}t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - H\bar{m}_t\mathrm{d}t) + \sqrt{\mathrm{Ricc}}(\bar{\Sigma}_t)(\bar{X}_t - \bar{m}_t)\mathrm{d}t, \quad \bar{X}_0 \sim \mathcal{N}(m_{\text{init}}, \Sigma_{\text{init}}) \tag{4.1}$$

where $\bar{\mathsf{K}}_t = \bar{\Sigma}_t H^\top$ is the Kalman gain, $\bar{m}_t = \mathsf{E}[\bar{X}_t|\mathscr{Z}_t]$ is the mean, and $\bar{\Sigma}_t = \mathsf{E}[(\bar{X}_t - \bar{m}_t)(\bar{X}_t - \bar{m}_t)^\top|\mathscr{Z}_t]$ is the covariance, and

$$\sqrt{\mathrm{Ricc}}(Q) := A - \frac{1}{2}QH^\top H + \frac{1}{2}\Sigma_B Q^{-1} + \Omega Q^{-1} \tag{4.2}$$

for any $Q \in S^d_{++}$ where Ω is *any* skew symmetric $d \times d$ matrix. The optimal transport FPF formula (3.15) is obtained by using a particular choice of the skew-symmetric matrix Ω_t as specified in Prop. 3.1. The more general case is considered here because the error analysis results are more generally applicable to the model (4.3). This filter is referred to as the deterministic linear FPF.

The evolution of the particles X^i_t is given by the sde:

$$\mathrm{d}X^i_t = Am^{(N)}_t\mathrm{d}t + \mathsf{K}^{(N)}_t(\mathrm{d}Z_t - Hm^{(N)}_t\mathrm{d}t) + \sqrt{\mathrm{Ricc}}(\Sigma^{(N)}_t)(X^i_t - m^{(N)}_t)\mathrm{d}t, \quad X^i_0 \sim \mathcal{N}(m_{\text{init}}, \Sigma_{\text{init}}) \tag{4.3}$$

where $\mathsf{K}^{(N)}_t := \Sigma^{(N)}_t H^\top$; and empirical approximations of mean and variance are

$$\begin{aligned} m^{(N)}_t &:= \frac{1}{N}\sum_{j=1}^N X^i_t, \\ \Sigma^{(N)}_t &:= \frac{1}{N-1}\sum_{j=1}^N (X^i_t - m^{(N)}_t)(X^i_t - m^{(N)}_t)^\top \end{aligned} \tag{4.4}$$

4.2.2 Stochastic linear FPF

The mean-field process is given by the sde

$$\mathrm{d}\bar{X}_t = A\bar{X}_t\mathrm{d}t + \sigma_B\mathrm{d}\bar{B}_t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - \frac{H\bar{X}_t + H\bar{m}_t}{2}\mathrm{d}t), \quad \bar{X}_0 \sim \mathcal{N}(m_{\text{init}}, \Sigma_{\text{init}}) \tag{4.5}$$

where $\bar{\mathsf{K}}_t = \bar{\Sigma}_t H^\top$ is the Kalman gain, $\bar{m}_t = \mathsf{E}[\bar{X}_t|\mathscr{Z}_t]$ is the mean, and $\bar{\Sigma}_t = \mathsf{E}[(\bar{X}_t - \bar{m}_t)(\bar{X}_t - \bar{m}_t)^\top|\mathscr{Z}_t]$ is the covariance.

The evolution of the particles is given by

$$\mathrm{d}X_t^i = AX_t^i\mathrm{d}t + \sigma_B\mathrm{d}B_t^i + \mathsf{K}_t^{(N)}(\mathrm{d}Z_t - \frac{HX_t^i + Hm_t^{(N)}}{2}\mathrm{d}t) \tag{4.6}$$

where $\mathsf{K}_t^{(N)} := \Sigma_t^{(N)}H^\top$; $\{B_t^i\}_{i=1}^N$ are independent copies of B_t; and the empirical approximations of the two mean-field terms are given by (4.4).

Remark 4.1 (Comparison of the deterministic and the stochastic FPF)**.** *In the deterministic FPF, there is no explicit Wiener process for the process noise. For example, with the choice of $\Omega_t = 0$, the deterministic linear FPF* (4.1) *has the same terms as the stochastic linear FPF* (4.5)*, except that the process noise term $\sigma_B\mathrm{d}\bar{B}_t$ in* (4.5) *is replaced by $\frac{1}{2}\Sigma_B\bar{\Sigma}_t^{-1}(\bar{X}_t - \bar{m}_t)\mathrm{d}t$ in* (4.1)*. With any Gaussian prior, the term serves to simulate the effect of the process noise.*

The sde (4.1) and (4.5) represents the mean-field limit of the interacting particle system (4.3) and (4.6) respectively. These models are referred to as McKean-Vlasov SDEs [McKean, 1966] and their analysis is referred to as propagation of chaos [Sznitman, 1991].

The convergence and error analysis relies closely on the classical results on stability of the Kalman filter, which appears in Appendix 4.6.2.

4.3 Evolution equations for mean and covariance

Consider the finite-N filters – Eq. (4.6) for the stochastic FPF and Eq. (4.3) for the deterministic FPF. The empirical mean and covariance are defined in Eq. (4.4). The error is defined as

$$\xi_t^i := X_t^i - m_t^{(N)} \quad \text{for } i = 1, 2, \ldots, N$$

Deterministic linear FPF: For the finite-N filter (4.3), the evolution equations for the mean, covariance, and error are as follows:

$$\mathrm{d}m_t^{(N)} = Am_t^{(N)}\mathrm{d}t + \mathsf{K}_t^{(N)}(\mathrm{d}Z_t - Hm_t^{(N)}\mathrm{d}t) \tag{4.7a}$$

$$\frac{\mathrm{d}\Sigma_t^{(N)}}{\mathrm{d}t} = A\Sigma_t^{(N)} + \Sigma_t^{(N)}A^\top + \sigma_B\sigma_B^\top - \Sigma_t^{(N)}H^\top H\Sigma_t^{(N)} \tag{4.7b}$$

$$\frac{\mathrm{d}\xi_t^i}{\mathrm{d}t} = \sqrt{\mathrm{Ricc}}(\Sigma_t^{(N)})\xi_t^i$$

The calculations leading to the derivation of these equations appear in the Appendix 4.6.1.

Stochastic linear FPF: For the finite-N filter (4.6), the evolution equations for the mean, covariance, and

error are as follows:

$$\mathrm{d}m_t^{(N)} = Am_t^{(N)}\mathrm{d}t + \mathsf{K}_t^{(N)}(\mathrm{d}Z_t - Hm_t^{(N)}\mathrm{d}t) + \frac{\sigma_B}{\sqrt{N}}\mathrm{d}\tilde{B}_t \tag{4.8a}$$

$$\mathrm{d}\Sigma_t^{(N)} = (A\Sigma_t^{(N)} + \Sigma_t^{(N)}A^\top + \Sigma_B - \Sigma_t^{(N)}H^\top H\Sigma_t^{(N)})\mathrm{d}t + \frac{\mathrm{d}M_t}{\sqrt{N}} \tag{4.8b}$$

$$\mathrm{d}\xi_t^i = (A - \frac{1}{2}\mathsf{K}_t^{(N)}H)\xi_t^i\mathrm{d}t + \sigma_B\mathrm{d}B_t^i - \frac{\sigma_B}{\sqrt{N}}\mathrm{d}\tilde{B}_t \tag{4.8c}$$

where $\tilde{B}_t := \frac{1}{\sqrt{N}}\sum_{i=1}^N B_t^i$ is a standard Wiener process and $\mathrm{d}M_t = \frac{\sqrt{N}}{N-1}\sum_{i=1}^N(\xi_t^i\mathrm{d}B_t^{i\top}\sigma_B^\top + \sigma_B\mathrm{d}B_t^i\xi_t^{i\top})$ is a matrix-valued martingale with $\mathsf{E}[\mathrm{d}M_t\mathrm{d}M_t^\top] = (\frac{N}{N-1})^2(\Sigma_t^{(N)}\Sigma_B + \Sigma_B\Sigma_t^{(N)} + 2\mathrm{Tr}(\Sigma_B)\Sigma_t^{(N)})\mathrm{d}t$. The details of derivation of these equations appear in Appendix 4.6.1.

Even though the fluctuations scale as $O(N^{-\frac{1}{2}})$, the analysis is challenging as has been noted in literature (see the remark after Theorem 3.1 in [Del Moral and Tugaut, 2016]). Error analysis of the ensemble Kalman filter with noise terms appears in [Del Moral and Tugaut, 2016, Bishop et al., 2017, Bishop and Del Moral, 2018] under the additional assumption that the matrix $H^\top H$ is positive definite which is equivalent to the observation matrix H be full-rank. Analysis of the deterministic FPF closely follows the stability theory for Kalman filter. Related analysis appears in the recent work [de Wiljes et al., 2016].

4.4 Error Analysis

4.4.1 Deterministic linear FPF

The following is assumed throughout the rest of this Section:
Assumption A1: The system (A,H) is detectable and (A,σ_B) is stabilizable.
Assumption A2: Assume $N > d$ and the initial empirical covariance matrix $\Sigma_0^{(N)} \in S_{++}^d$.

The main result for the finite-N deterministic linear FPF is as follows with the proof given in Appendix 4.6.3.

Proposition 4.1. *Consider the Kalman filter (3.4a)-(3.4b) initialized with the prior $\mathcal{N}(m_{init},\Sigma_{init})$ and the finite-N deterministic FPF (4.3) initialized with $X_0^i \overset{i.i.d}{\sim} \mathcal{N}(m_{init},\Sigma_{init})$ for $i = 1,2,\ldots,N$. Under Assumption (A1)-(A2), the following characterizes the convergence and error properties of the empirical mean and covariance $(m_t^{(N)},\Sigma_t^{(N)})$ obtained from the finite-N filter to the mean and covariance (m_t,Σ_t) obtained from the Kalman filter:*

(i) *Convergence: For any finite N, as $t \to \infty$:*

$$\lim_{t\to\infty} e^{\lambda t}\|m_t^{(N)} - m_t\|_2 = 0 \quad a.s$$
$$\lim_{t\to\infty} e^{2\lambda t}\|\Sigma_t^{(N)} - \Sigma_t\|_F = 0 \quad a.s$$

for all $\lambda \in (0,\lambda_0)$.

(ii) *Mean-squared error: For any $t > 0$, as $N \to \infty$:*

$$\mathsf{E}[\|m_t^{(N)} - m_t\|_2^2] \leq (const.)e^{-2\lambda t}\frac{Tr(\Sigma_0) + \|\Sigma_0\|_F^2}{N} \tag{4.9a}$$

$$\mathsf{E}[\|\Sigma_t^{(N)} - \Sigma_t\|_F^2] \leq (const.)e^{-4\lambda t}\frac{\|\Sigma_0\|_F^2}{N} \tag{4.9b}$$

for all $\lambda \in (0, \lambda_0)$ where λ_0 is defined in Theorem 4.1. The constant depends on λ, $\|\Sigma_0 - \Sigma_\infty\|_2$, and $\|H\|_2$.

Remark 4.2. *Asymptotically (as $t \to \infty$) the empirical mean and variance of the finite-N filter becomes exact. This is because of the stability of the Kalman filter whereby the filter forgets the initial condition. In fact, the i.i.d assumption on the initial condition X_0^i is not necessary to obtain this conclusion.*

Remark 4.3. *(Scaling with dimension) If the parameters of the linear Gaussian filtering problem (1.4a)-(1.4b) scale with the dimension in a way that the spectral norms $\|\Sigma_0\|_2$, $\|\Sigma_\infty\|_2$, $\|H\|_2$, and λ_0 do not change, then the constant in the error bounds (4.9a)-(4.9b) do not change. The only term that scales with the dimension is $\|\Sigma_0\|_F^2$ and $Tr(\Sigma_0)$. For example, if one assumes $\Sigma_0 = \sigma_0^2 I_{d\times d}$, then $|\Sigma_0\|_F^2 = d\sigma_0^4$ and $Tr(\Sigma_0) = d\sigma_0^2$. Therefore, the error estimates scale linearly with d.*

Remark 4.4. *The error estimates (4.9a)-(4.9b) holds for any skew-symmetric choice of Ω_t in (4.3). Therefore, the optimal choice of Ω_t does not effect the error estimates for mean and variance. In Sec. 4.5, we study the error for estimating expectation of an arbitrary function f, where the choice of Ω can be effective.*

4.4.2 Stochastic linear FPF

Assumption A3: The matrix A is stable, i.e $\mu(A) := \lambda_{\max}(\frac{A+A^\top}{2}) < 0$. And the matrix $H^\top H = I$ identity matrix.

The main result regarding the convergence of the empirical mean and empirical covariance is the following Proposition. The proof appears in the Appendix 4.6.5.

Proposition 4.2. *Consider the mean-field system (3.5), and the finite-N system (4.6) under Assumption (A3).*

(i) *For any $t > 0$, and as $N \to \infty$:*

$$\mathsf{E}[\|\Sigma_t^{(N)} - \Sigma_t\|_F^2] \leq \frac{3\|\Sigma_0\|_F^2}{N}e^{-(4\mu(A)-\frac{1}{N})t} + \frac{c_{var}}{N} \tag{4.10}$$

where $c_{var} = \frac{1}{2\mu(A)}(Tr(\Sigma_B^2) + \frac{1}{2}Tr(\Lambda_{max}^2) + \frac{Tr(\Sigma_B)^2}{2\mu(A)} + 2Tr(\Sigma_B)Tr(\Sigma_0))$, and Λ_{max} is an upper-bound on the solution to the exact Riccati equation (see Lemma 4.1).

(ii) *For any $t > 0$ and as $N \to \infty$:*

$$\mathsf{E}[\|m_t^{(N)} - m_t\|_2] \leq \frac{Tr(\Sigma_0)^{1/2}}{\sqrt{N}}e^{-\mu(A)t} + \frac{c_{mean}}{\sqrt{N}} \tag{4.11}$$

where the constant $c_{mean} = (\frac{3\|\Sigma_0\|_F^2 + c_{var}}{2\mu(A)})^{\frac{1}{2}} + (\frac{Tr(\Sigma_B)}{2\mu(A)})^{\frac{1}{2}}$.

4.5 Propagation of chaos

At the initial time $t = 0$, the particles $\{X_0^i\}_{i=1}^N$ are sampled i.i.d. from the prior distribution. In any finite-N implementation of the filter, the i.i.d. property is destroyed for $t > 0$ because of the interactions: For the linear FPFs (4.6) and (4.3), the interaction terms are a function of the empirical mean $m_t^{(N)}$ and the empirical covariance $\Sigma_t^{(N)}$. Since these terms depend upon all the particles, the i^{th} particle in the population is coupled to/interacts with (the randomness of) all other particles. Even though the particles are no longer i.i.d for any finite choice of N, one (formally) expects the particles to become approximately i.i.d (in a sense that needs to be made precise) for large N. Intuitively, this is because as $N \to \infty$, $m_t^{(N)} \to m_t$ and $\Sigma_t^{(N)} \to \Sigma_t$. And for the limiting mean-field model, the particles are i.i.d for $t > 0$ provided they are i.i.d. at the initial time $t = 0$. The phenomenon is referred to as the *propagation of chaos* whereby the chaos (i.i.d property of the population) propagates through time.

The mathematical definitions are as follows: Denote $E := \mathbb{R}^d \times [0,\infty)$. Let μ_N be the probability measure on E^N associated with the process $(X^1,\ldots,X^N)$. Let $\bar{\mu}$ be the probability measure on E associated with the mean-field solution $\bar{X}$. Then μ_N is said to be $\bar{\mu}$*-chaotic* if

$$\pi_k \mu_N \xrightarrow{\text{weak}} \bar{\mu}^{(k)} \quad \text{as} \quad N \to \infty$$

where $\pi_k \mu_N$ is the k-marginal distribution, $\bar{\mu}^{(k)}$ is the k-fold product, and the convergence is in the weak sense. A somewhat easier formulation of this condition appears in [Sznitman, 1991, Proposition 2.2] as

$$\lim_{N\to\infty} \mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(X^i) - \mathsf{E}[f(\bar{X})]\right|^2\right] = 0 \tag{4.12}$$

for all bounded functionals $f : E \to \mathbb{R}$.

Remark 4.5. *Some difficulties in carrying out the propagation of chaos analysis for the FPF are as follows: (i) The drift term in the evolution equation for the covariance is not Lipschitz; For the stochastic FPF* (4.6)*, the noise terms (the martingale* M_t*) depend upon the state. In our analysis, we circumvent some of these difficulties by limiting to the linear Gaussian setting, and using our analysis of the empirical mean and empirical covariance from Sec.* 4.2*. Even in this special case, we show the convergence for the marginal distribution* only *for fixed time* $t > 0$*. That is, we show*

$$\lim_{N\to\infty} \mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(X_t^i) - \mathsf{E}[f(\bar{X}_t)]\right|^2\right] = 0 \tag{4.13}$$

for all bounded functions $f : \mathbb{R}^d \to \mathbb{R}$.

4.5.1 Deterministic linear FPF

Derivation of error estimates involve construction of N independent copies of the mean-field equation (4.1) corresponding to the deterministic FPF (4.3). Consistent with our convention to denote mean-field variables with a bar, the stochastic processes are denoted as $\{\bar{X}_t^i : 1 \le i \le N\}$ where $\bar{X}_t^i$ denotes the state of the i^{th} particle at time t. The particle evolves according to the mean-field equation (4.1) as

$$\mathrm{d}\bar{X}_t^i = A\bar{m}_t \mathrm{d}t + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - H\bar{m}_t \mathrm{d}t) + \bar{G}_t(\bar{X}_t^i - \bar{m}_t)\mathrm{d}t \tag{4.14}$$

where the initial condition $\bar{X}_0^i = X_0^i$ – the initial condition of the i^{th} particle in the finite-N FPF (4.3). The mean-field process $\bar{X}_t^i$ is thus coupled to X_t^i through the initial condition. The following Proposition characterizes the error between X_t^i and $\bar{X}_t^i$ (the estimate is essential for the propagation of chaos analysis). The proof appears in the Appendix 4.6.4.

Proposition 4.3. *Consider the stochastic processes X_t^i and $\bar{X}_t^i$ whose evolution is defined according to the deterministic FPF (4.3) and its mean-field model (4.14), respectively. The initial condition $X_0^i \overset{i.i.d}{\sim} \mathcal{N}(m_0, \Sigma_0)$ for $i = 1, 2, \ldots, N$ Then under Assumptions (A1)-(A2):*

$$\mathsf{E}[\|X_t^i - \bar{X}_t^i\|_2^2]^{1/2} \le \frac{(const.)}{\sqrt{N}} \tag{4.15}$$

The estimate (4.15) is used to prove the following important result that the empirical distribution of the particles in the linear FPF converges weakly to the true posterior distribution. Its proof appears in the Appendix 4.6.4.

Corollary 4.1. *Consider the linear filtering problem (1.4a)-(1.4b) and the finite-N deterministic FPF (4.3). The initial condition $X_0^i \overset{i.i.d}{\sim} \mathcal{N}(m_0, \Sigma_0)$ for $i = 1, 2, \ldots, N$ and the dimension $d = 1$. Under Assumptions (A1) and (A2), for any Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, in the asymptotic limit as $N \to \infty$*

$$\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(X_t^i) - \mathsf{E}[f(X_t)|\mathscr{Z}_t]\right|^2\right]^{1/2} \le \frac{(const.)}{\sqrt{N}}$$

4.5.2 Stochastic linear FPF

In this subsection, we carry out the propagation of chaos error analysis for the stochastic linear FPF (4.6). Introduce N independent copies of the mean-field process (4.5) denoted by $\{\bar{X}_t^i;\ i = 1, \ldots, N\}$ such that

$$\mathrm{d}\bar{X}_t^i = A\bar{X}_t^i \mathrm{d}t + \mathrm{d}B_t^i + \bar{\mathsf{K}}_t(\mathrm{d}Z_t - \frac{H\bar{X}_t^i + H\bar{m}_t}{2}\mathrm{d}t), \quad \bar{X}_0^i = X_0^i \tag{4.16}$$

for $i = 1, \ldots, N$. Note that $\bar{X}_t^i$ and X_t^i are coupled through the same initial condition and the same process noise $\mathrm{d}B_t^i$. The result regarding the convergence of the empirical distribution is the following Proposition. The proof appears in Appendix 4.6.6.

Proposition 4.4. *Consider the mean-field system (4.5), the finite-N system (4.6), and the stochastic processes $\bar{X}_t^i$ defined in (4.16) under Assumption (A3).*

(i) *Particles: For any $t > 0$ and as $N \to \infty$:*

$$\mathsf{E}[\|X_t^i - \bar{X}_t^i\|_2] \leq \frac{(const.)}{\sqrt{N}} \tag{4.17}$$

for $i = 1, \ldots, N$

(ii) *For any Lipschitz function f:*

$$\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(X_t^i) - \mathsf{E}[f(\bar{X}_t)|\mathscr{Z}_t]\right|\right] \leq \frac{(const)}{\sqrt{N}} \tag{4.18}$$

4.6 Proof of the main results

4.6.1 Derivation of evolution equations in Sec. 4.3

(A) Finite-N stochastic FPF: Consider Eq. (4.6) for the i^{th} particle. Summing up over the index $i = 1, \ldots, N$ and dividing by N, Eq. (4.8a) for the mean is obtained. To obtain (4.8b), first define $\xi_t^i := X_t^i - m_t^{(N)}$. Therefore,

$$\mathrm{d}\xi_t^i = (A - \frac{1}{2}\mathsf{K}_t^{(N)} H)\xi_t^i \mathrm{d}t + \sigma_B \mathrm{d}B_t^i - \frac{1}{N}\sum_{j=1}^{N} \sigma_B \mathrm{d}B_t^j$$

and

$$\begin{aligned}
\mathrm{d}(\xi_t^i \xi_t^{i\top}) =& (A - \frac{1}{2}\mathsf{K}_t^{(N)} H)\xi_t^i \xi_t^{i\top} \mathrm{d}t + \xi_t^i \xi_t^{i\top} (A - \frac{1}{2}\mathsf{K}_t^{(N)} H)^\top \mathrm{d}t \\
&+ \frac{N-1}{N}\sigma_B \sigma_B^\top \mathrm{d}t + \xi_t^i (\mathrm{d}B_t^i - \frac{1}{N}\sum_{j=1}^{N} \mathrm{d}B_t^j)^\top \sigma_B^\top \\
&+ \sigma_B (\mathrm{d}B_t^i - \frac{1}{N}\sum_{j=1}^{N} \mathrm{d}B_t^j) \xi_t^{i\top}
\end{aligned}$$

Summing over $i = 1, \ldots, N$ and dividing by $(N-1)$ gives

$$\begin{aligned}
\mathrm{d}\Sigma_t^{(N)} =& (A - \frac{1}{2}\mathsf{K}_t^{(N)} C)\Sigma_t^{(N)} \mathrm{d}t + \Sigma_t^{(N)} (A - \frac{1}{2}\mathsf{K}_t^{(N)} H)^\top \mathrm{d}t \\
&+ \sigma_B \sigma_B^\top \mathrm{d}t + \frac{1}{N-1}\sum_{i=1}^{N} \xi_t^i \mathrm{d}B_t^{i\top} \sigma_B^\top + \frac{1}{N-1}\sum_{i=1}^{N} \sigma_B \mathrm{d}B_t^i \xi_t^{i\top}
\end{aligned}$$

which is Eq. (4.8b) for the covariance.

(B) Finite-N deterministic FPF: Eq. (4.3) is obtained as before by summing up Eq. (4.3) for the i^{th} particle from $i = 1, \ldots, N$. The equation for the empirical mean is simply obtained by summing up the equations (4.3)

for $i = 1, \ldots, N$. To obtain the equation for the empirical covariance, first define $\xi_t^i := X_t^i - m_t^{(N)}$. Therefore, $\mathrm{d}\xi_t^i = G_t^{(N)} \xi_t^i \mathrm{d}t$ and

$$\mathrm{d}(\xi^t \xi_t^{i\top}) = G_t^{(N)} \xi_t^i \xi_t^{i\top} \mathrm{d}t + \xi_t^i \xi_t^{i\top} G_t^{(N)\top} \mathrm{d}t$$

Summing over $i = 1, \ldots, N$ and dividing by $(N-1)$ gives

$$\frac{\mathrm{d}\Sigma_t^{(N)}}{\mathrm{d}t} = G_t^N \Sigma_t^{(N)} + \Sigma_t^{(N)} G_t^{(N)\top}$$

which is Eq. (4.7b) for the covariance.

It is noted that $G_t^{(N)}$ is well-defined because $\Sigma_0^{(N)}$ and thus $\Sigma_t^{(N)}$ is invertible because of Assumption (A2).

4.6.2 Background on the Stability of the Kalman filter

Theorem 4.1 (Lemma 2.2 and Theorem 2.3 in [Ocone and Pardoux, 1996]). *Consider the Kalman filter* (3.4a)-(3.4b) *with initial condition* (m_0, Σ_0). *Then, under Assumption (A1):*

(i) *There exists a solution* $\Sigma_\infty \succ 0$ *to the algebraic Riccati equation (ARE)*

$$A\Sigma_\infty + \Sigma_\infty A^\top + \sigma_B \sigma_B^\top - \Sigma_\infty H^\top H \Sigma_\infty = 0 \tag{4.19}$$

such that $A - \Sigma_\infty H^\top H$ *is Hurwitz. Let*

$$0 < \lambda_0 = min\{-Real\ \lambda : \lambda\ is\ an\ eigenvalue\ of\ A - \Sigma_\infty H^\top H\} \tag{4.20}$$

(ii) *The error covariance* $\Sigma_t \to \Sigma_\infty$ *exponentially fast for* any *initial condition* Σ_0 *(not necessarily the prior): for all* $\lambda \in (0, \lambda_0)$, *there exists a constant* c_λ *such that*

$$\lim_{t\to\infty} \|\Sigma_t - \Sigma_\infty\|_F \le c_\lambda e^{-2\lambda t} \to 0$$

(iii) *Starting from two initial conditions* (m_0, Σ_0) *and* $(\tilde{m}_0, \tilde{\Sigma}_0)$, *the means converge in the following senses:*

$$\begin{aligned} \lim_{t\to\infty} \mathsf{E}[\|m_t - \tilde{m}_t\|^2] &\le (const.)\, e^{-2\lambda t} \to 0 \\ \lim_{t\to\infty} \|m_t - \tilde{m}_t\| e^{\lambda t} &= 0 \quad a.s. \end{aligned}$$

for all $\lambda \in (0, \lambda_0)$.

Throughout this paper, the notation Σ_∞ is used to denote the positive definite solution of the ARE (4.19) and λ_0 is used to denote the spectral bound as defined in (4.20).

Lemma 4.1 (Theorem 2.2 in [Bishop and Del Moral, 2017]). *Consider the Riccati equation (3.4b) under Assumption (I). Assume the initial matrix $\Sigma_0 \in S^d_{++}$ is positive symmetric definite (p.s.d). Then, there exists matrices $\Lambda_{min}, \Lambda_{max} \in S^d_{++}$ such that the solution Σ_t satisfies*

$$\Lambda_{min} \preceq \Sigma_t \preceq \Lambda_{max}$$

4.6.3 Proof of the Prop. 4.1

Since the equations for the empirical mean (4.7a) and the empirical covariance (4.7b) are identical to the Kalman filter (3.4a)-(3.4b), the a.s. convergence of mean and variance follows from the filter stability theory (see Theorem 4.1). We present the proof for the mean-squared estimates in the following steps:

1. First, we form an estimate for the spectral norm of the transition matrix e^{tF_∞}. From Theorem 4.1 we know that all eigenvalues of F_∞ have negative real parts smaller than $-\lambda_0$. Consider the Jordan decomposition $J = T^{-1}F_\infty T$. In this case, we have the estimate

$$\|e^{tF_\infty}\|_2 \leq \|T\|_2\|T^{-1}\|_2 \left(\max_{0\leq k\leq n} \frac{t^k}{k!}\right) e^{-\lambda_0 t}, \quad \forall t > 0 \tag{4.21}$$

   where n the maximum multiplicity of eigenvalues of F_∞. As a result, for all $\lambda < \lambda_0$, there exists a constant $c'_\lambda := \|T\|_2\|T^{-1}\|_2 \sup_{t\geq 0} e^{-(\lambda_0-\lambda)t}\left(\max_{0\leq k\leq n} \frac{t^k}{k!}\right)$ such that

$$\|e^{tF_\infty}\|_2 \leq c'_\lambda e^{-\lambda t} \tag{4.22}$$

2. Next, we show contraction properties of the transition matrix $\Phi_{t,s}$ corresponding to the linear system $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{t,s} = F_t\Phi_{t,s}$, $\Phi_{s,s} = I$ where $F_t = A - \Sigma_t H^\top H$. For $x_t = \Phi_{t,s}x_s$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}x_t = F_t x_t = F_\infty x_t + (\Sigma_\infty - \Sigma_t)H^\top H x_t$$

   Therefore

$$x_t = e^{tF_\infty}x_s + \int_s^t e^{(t-\tau)F_\infty}(\Sigma_\infty - \Sigma_\tau)H^\top H x_\tau \mathrm{d}\tau$$

   Upon taking the norm and using the triangle inequality

$$\|x_t\|_2 \leq c_\lambda e^{-t\lambda}\|x_s\|_2 + \int_s^t c_\lambda e^{-(t-\tau)\lambda}\|\Sigma_\tau - \Sigma_\infty\|_2\|H^\top H\|_2\|x_\tau\|_2 \mathrm{d}\tau$$

   Using the Gronwall inequality

$$\|x_t\|_2 \leq c'_\lambda e^{-\lambda(t-s)}\|x_s\|_2 e^{c'_\lambda\|H^\top H\|_2\int_s^t\|\Sigma_\tau-\Sigma_\infty\|_2\mathrm{d}\tau}$$

   concluding

$$\|\Phi_{t,s}\|_2 \leq c'_\lambda e^{-\lambda(t-s)} e^{c'_\lambda\|H^\top H\|_2\int_s^t\|\Sigma_\tau-\Sigma_\infty\|_2\mathrm{d}\tau}$$

Then, using the exponential convergence $\|\Sigma_t - \Sigma_\infty\|_2 \le c_\lambda e^{-2\lambda t}$ from Lemma 4.1, we get

$$\|\Phi_{t,s}\|_2 \le c'_\lambda e^{-\lambda(t-s)} e^{c'_\lambda \|H^\top H\|_2 \frac{c_\lambda \|\Sigma_0 - \Sigma_\infty\|_2}{2\lambda}}$$

which we express as $\|\Phi_{t,s}\|_2 \le c_\lambda e^{-\lambda(t-s)}$ by redefining the constant c_λ. Similarly, the spectral bound $\|\Phi^{(N)}_{t,s}\|_2 \le c_\lambda e^{-\lambda(t-s)}$ holds when one replaces Σ_t with $\Sigma^{(N)}_t$, because $\Sigma^{(N)}_t$ also evolves according to the Riccati equation and converges exponentially to Σ_∞.

3. In this step, we prove the estimate for the error $\Sigma^{(N)}_t - \Sigma_t$. From the Ricatti equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Sigma^{(N)}_t - \Sigma_t) = (A - \Sigma_t HH^\top)(\Sigma^{(N)}_t - \Sigma_t) + (\Sigma^{(N)}_t - \Sigma_t)(A - \Sigma^{(N)}_t HH^\top)^\top$$

The solution satisfies

$$\Sigma^{(N)}_t - \Sigma_t = \Phi_{t,0}(\Sigma^{(N)}_0 - \Sigma_0)(\Phi^{(N)}_{t,0})^\top$$

where $\Phi_{t,0}$ and $\Phi^{(N)}_{t,0}$ are the state transition matrices that were defined in step 2. Then,

$$\begin{aligned}\|\Sigma^{(N)}_t - \Sigma_t\|_F &\le \|\Phi^{\Sigma_t}_t\|_2 \|\Phi^{\Sigma^{(N)}_t}_t\|_2 \|\Sigma^{(N)}_0 - \Sigma_0\|_F \\ &\le c^2_\lambda e^{-2\lambda t} \|\Sigma^{(N)}_0 - \Sigma_0\|_F\end{aligned}$$

where we applied the upper-bound on the spectral norm of the transition matrix from step 2. Upon squaring and taking the expectation of both sides

$$\begin{aligned}\mathsf{E}[\|\Sigma^{(N)}_t - \Sigma_t\|^2_F] &\le c^4_\lambda e^{-4\lambda t} \mathsf{E}[\|\Sigma^{(N)}_0 - \Sigma_0\|^2_F] \\ &= c^4_\lambda e^{-4\lambda t} \mathsf{E}[\mathrm{Tr}((\Sigma^{(N)}_0 - \Sigma_0)^2)] \\ &= c^4_\lambda e^{-4\lambda t} \mathsf{E}[\mathrm{Tr}((\frac{1}{N}\sum_{i=1}^N \xi^i_0 {\xi^i_0}^\top - \Sigma_0)^2)] \\ &= c^4_\lambda e^{-4\lambda t} \frac{1}{N} \mathsf{E}[\mathrm{Tr}((\xi^i_0 {\xi^i_0}^\top - \Sigma_0)^2)] \\ &\le c^4_\lambda e^{-4\lambda t} \frac{\mathsf{E}[\|\xi_0\|^4_2]}{N} \\ &= c^4_\lambda e^{-4\lambda t} \frac{3\|\Sigma_0\|^2_F}{N}\end{aligned}$$

4. Finally, we prove the estimate for the mean. Subtracting the equation (3.4a) for Kalman mean from the equation (4.7a) for empirical mean yields:

$$\mathrm{d}m^{(N)}_t - \mathrm{d}m_t = (A - \Sigma^{(N)}_t H^\top H)(m^{(N)}_t - m_t)\mathrm{d}t + (\Sigma^{(N)}_t - \Sigma_t)H^\top H \mathrm{d}I_t$$

where $\mathrm{d}I_t = \mathrm{d}Z_t - Hm_t\mathrm{d}t$ is the innovation process. The solution satisfies the equation:

$$m_t^{(N)} - m_t = \Phi_t^{(N)}(m_0^{(N)} - m_0) + \int_0^t \Phi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)H^\top H \mathrm{d}I_s$$

The mean-squared norm of the first term is bounded by:

$$\begin{aligned}\mathsf{E}[\|\Phi_t^{(N)}(m_0^{(N)} - m_0)\|_2^2] &\le c_\lambda^2 e^{-2\lambda t}\mathsf{E}[\|m_0^{(N)} - m_0\|_2^2] \\ &\le c_\lambda^2 e^{-2\lambda t}\frac{\mathrm{Tr}(\Sigma_0)}{N}\end{aligned}$$

where we used the fact that the innovation process $\mathrm{d}I_t$ is a Brownian motion [Xiong, 2008, Lemma 5.6]. The mean-squared norm of the first term is bounded by:

$$\begin{aligned}\mathsf{E}\left[\left\|\int_0^t \Phi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)H^\top \mathrm{d}I_s\right\|_2^2\right] &= \int_0^t \mathsf{E}\left[\mathrm{Tr}\left(\Phi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)H^\top H(\Sigma_s^{(N)} - \Sigma_s)\Phi_{t,s}^{(N)\top}\right)\right]\mathrm{d}s \\ &\le \int_0^t \mathsf{E}[\|\Phi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)\|_F^2]\|H\|_2^2\mathrm{d}s \\ &\le \|H\|_2^2\int_0^t c_\lambda^2 e^{-2\lambda(t-s)}c_\lambda^4 e^{-4\lambda s}\mathsf{E}[\|\Sigma_0^{(N)} - \Sigma_0\|_F^2]\mathrm{d}s \\ &= c_\lambda^6\|H\|_2^2\frac{3\|\Sigma_0\|_F^2}{N}\frac{e^{-2\lambda t}}{2\lambda}\end{aligned}$$

The bounds on the first term and second term add together to conclude

$$\mathsf{E}[\|m_t - m_t^{(N)}\|_2^2] \le e^{-2\lambda t}\frac{2c_\lambda^2\mathrm{Tr}(\Sigma_0)}{N} + e^{-2\lambda t}\frac{6c_\lambda^6\|H\|_2^2\|\Sigma_0\|_F^2}{2\lambda N}$$

4.6.4 Proofs of the Prop. 4.3 and Cor. 4.1

Proof. Use the decomposition

$$X_t^i = m_t^{(N)} + \xi_t^i, \quad \bar{X}_t^i = m_t + \bar{\xi}_t^i$$

to express the error as

$$\mathsf{E}[\|X_t^i - \bar{X}_t^i\|_2^2]^{1/2} \le \mathsf{E}[\|m_t^{(N)} - \bar{m}_t\|_2^2]^{1/2} + \mathsf{E}[\|\xi_t^i - \bar{\xi}_t^i\|_2^2]^{1/2}$$

The error $m_t^{(N)} - \bar{m}_t$ is already obtained in Proposition 4.1 as (4.9a). The key step is to analyze the error $\xi_t^i - \bar{\xi}_t^i$. By definition,

$$\mathrm{d}\xi_t^i = \sqrt{\mathrm{Ricc}}(\Sigma_t^{(N)})\xi_t^i\mathrm{d}t, \quad \mathrm{d}\bar{\xi}_t^i = \sqrt{\mathrm{Ricc}}(\Sigma_t)\bar{\xi}_t^i\mathrm{d}t$$

Let $\Psi_{t,s}^{(Q_t)}$ be the state transition matrix corresponding to the linear system

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi_{t,s}^{(Q_t)} = \sqrt{\mathrm{Ricc}}(Q_t)\Psi_{t,s}^{(Q_t)}, \quad \Psi_{s,s}^{(Q_t)} = I$$

Therefore, $\xi_t^i = \Psi_{t,0}^{(\Sigma_t^{(N)})}\xi_0^i$ and $\bar{\xi}_t^i = \Psi_{t,0}^{(\Sigma_t)}\bar{\xi}_0^i$. We bound the error $\xi_t^i - \bar{\xi}_t^i$ in the following steps:

1) First, we bound the spectral norm of the transition matrix $\Psi_{t,s}^{(\Sigma_\infty)} = e^{(t-s)\sqrt{\mathrm{Ricc}}(\Sigma_\infty)}$. Consider the linear system:

$$\frac{\mathrm{d}}{\mathrm{d}t}y_t = \sqrt{\mathrm{Ricc}}(\Sigma_\infty)^\top y_t$$

and the Lyapunov function $V(y) = y^\top \Sigma_\infty y$. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V(y_t) = 0, \quad &\Rightarrow \quad y_t^\top \Sigma_\infty y_t = y_s^\top \Sigma_\infty y_s \\ &\Rightarrow \quad |y_t|^2 \le \frac{\lambda_{\max}(\Sigma_\infty)}{\lambda_{\min}(\Sigma_\infty)}|y_s|^2\end{aligned}$$

as a result

$$\|e^{(t-s)\sqrt{\mathrm{Ricc}}(\Sigma_\infty)^\top}\|_2 \le c, \quad \forall t \ge s \ge 0$$

where $c = \sqrt{\frac{\lambda_{\max}(\Sigma_\infty)}{\lambda_{\min}(\Sigma_\infty)}}$. Therefore $\|\Psi_{t,s}^{(\Sigma_\infty)}\|_2 = \|e^{(t-s)\sqrt{\mathrm{Ricc}}(\Sigma_\infty)^\top}\|_2 \le c$.

2) Next, we bound the spectral norm of the transition matrix $\Psi_{t,s}^{(\Sigma_t)}$. Consider the linear system

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x_t &= \sqrt{\mathrm{Ricc}}(\Sigma_t)x_t \\ &= \sqrt{\mathrm{Ricc}}(\Sigma_\infty)x_t + (\sqrt{\mathrm{Ricc}}(\Sigma_t) - \sqrt{\mathrm{Ricc}}(\Sigma_\infty))x_t\end{aligned}$$

which satisfies

$$x_t = \Psi_{t,0}^{(\Sigma_\infty)}x_0 + \int_0^t \Psi_{t,s}^{(\Sigma_\infty)}(\sqrt{\mathrm{Ricc}}(\Sigma_s) - \sqrt{\mathrm{Ricc}}(\Sigma_\infty))x_s\mathrm{d}s$$

and the bound,

$$\|x_t\|_2 \le c\|x_0\|_2 + c\int_0^t \|\sqrt{\mathrm{Ricc}}(\Sigma_s) - \sqrt{\mathrm{Ricc}}(\Sigma_\infty)\|_2\|x_s\|_2\mathrm{d}s$$

where we used $\|\Psi_{t,s}^{(\Sigma_\infty)}\|_2 \le c$ from step 1. By application of Gronwall inequality

$$\|x_t\|_2 \le c\|x_0\|_2 e^{c\int_0^t \|\sqrt{\mathrm{Ricc}}(\Sigma_t) - \sqrt{\mathrm{Ricc}}(\Sigma_\infty)\|_2\mathrm{d}s}$$

Next, use

$$\begin{aligned}\|\sqrt{\mathrm{Ricc}}(\Sigma_t) - \sqrt{\mathrm{Ricc}}(\Sigma_\infty)\|_2 &\le \|\frac{1}{2}\Sigma_B(\Sigma_t^{-1} - \Sigma_\infty^{-1}) - \frac{1}{2}(\Sigma_t - \Sigma_\infty)H^\top H\|_2 \\ &\le \frac{1}{2}\|\Sigma_B\|_2\|\Sigma_\infty^{-1}\|_2\|\Sigma_t^{-1}\|_2\|\Sigma_t - \Sigma_\infty\|_2 + \frac{1}{2}\|H^\top H\|_2\|\Sigma_t - \Sigma_\infty\|_2 \\ &\le \frac{1}{2}(\|\Sigma_B\|_2\|\Sigma_\infty^{-1}\|_2\|\Lambda_{\min}^{-1}\|_2 + \|H^\top H\|_2)c_\lambda e^{-2\lambda t}\end{aligned}$$

where we used $\|\Sigma_t - \Sigma_\infty\|_2 \le c_\lambda e^{-2\lambda t}$ from Theorem 4.1 and $\|\Sigma_t^{-1}\| \le \|\Lambda_{\min}^{-1}\|_2$ from Lemma 4.1, to conclude

$$\|x_t\|_2 \le c\|x_0\|_2 e^{c\frac{c_\lambda}{4\lambda}(\|\Sigma_B\|_2\|\Sigma_\infty^{-1}\|_2\|\Lambda_{\min}^{-1}\|_2 + \|H^\top H\|_2)} =: c'\|x_0\|$$

concluding $\|\Psi_{t,s}^{(\Sigma_t)}\| \le c'$ for all $t \ge s \ge 0$.

3) Finally, we aim to bound the error $\xi_t^i - \bar{\xi}_t^i$. By definition,

$$\mathrm{d}\xi_t^i - \mathrm{d}\bar{\xi}_t^i = \sqrt{\mathrm{Ricc}}(\Sigma_t^{(N)})(\xi_t^i - \bar{\xi}_t^i)\mathrm{d}t + (\sqrt{\mathrm{Ricc}}(\Sigma_t) - \sqrt{\mathrm{Ricc}}(\Sigma_t^{(N)}))\bar{\xi}_t^i \mathrm{d}t$$

Therefore

$$\xi_t^i - \bar{\xi}_t^i = \Psi_{t,0}^{(\Sigma_t^{(N)})}(\xi_0^i - \bar{\xi}_0^i) + \int_0^t \Psi_{t,s}^{(\Sigma_s^{(N)})}(\sqrt{\mathrm{Ricc}}(\Sigma_s) - \sqrt{\mathrm{Ricc}}(\Sigma_s^{(N)}))\bar{\xi}_s^i \mathrm{d}s$$

Taking the norm ,

$$\|\xi_t^i - \bar{\xi}_t^i\|_2 \le c'\|\xi_0^i - \bar{\xi}_0^i\|_2 + c'\int_0^t \|(\sqrt{\mathrm{Ricc}}(\Sigma_s) - \sqrt{\mathrm{Ricc}}(\Sigma_s^{(N)}))\|_2 \|\bar{\xi}_s^i\| \mathrm{d}s$$

where we used the bound $\|\Psi_{t,s}^{(\Sigma_t^{(N)})}\| \le c'$ from step (3). Upon taking the mean-squared norm of both sides and using the triangle inequality and Cauchy-Schwartz inequality,

$$\mathsf{E}[\|\xi_t^i - \bar{\xi}_t^i\|_2^2]^{1/2} \le c'\mathsf{E}[\|\xi_0^i - \bar{\xi}_0^i\|_2^2]^{1/2} + c'\int_0^t \mathsf{E}[\|(\sqrt{\mathrm{Ricc}}(\Sigma_s) - \sqrt{\mathrm{Ricc}}(\Sigma_s^{(N)}))\|_2^2]^{1/2}\mathsf{E}[\|\bar{\xi}_s^i\|^2]^{1/2}\mathrm{d}s$$

Using the identity $\xi_0^i - \bar{\xi}_0^i = m_0 - m_0^{(N)}$, the bound

$$\begin{aligned}\|\sqrt{\mathrm{Ricc}}(\Sigma_t^{(N)}) - \sqrt{\mathrm{Ricc}}(\Sigma_t)\|_2 &\le \frac{1}{2}(\|\Sigma_B\|_2\|\Sigma_t^{-1}\|_2\|\Sigma_t^{(N)^{-1}}\|_2 + \|H^\top H\|_2)\|\Sigma_t^{(N)} - \Sigma_t\|_2 \\ &\le \frac{1}{2}(\|\Sigma_B\|_2\|\Lambda_{\min}^{-1}\|_2^2 + \|H^\top H\|_2)e^{-2\lambda t}\frac{(\text{const.})}{\sqrt{N}}\end{aligned}$$

and the identity $\mathsf{E}[\|\bar{\xi}_2^i\|_2^2] = \mathrm{Tr}(\Sigma_t)$ yields

$$\mathsf{E}[\|\xi_t^i - \bar{\xi}_t^i\|_2^2]^{1/2} \le \frac{(\text{const.})}{\sqrt{N}}\int_0^t e^{-2\lambda s}\mathrm{d}s \le \frac{(\text{const.})}{\sqrt{N}}$$

concluding the result. □

Proof of the Corollary 4.1. Using the triangle inequality,

$$\begin{aligned}\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(X_t^i) - \mathsf{E}[f(X_t)|\mathscr{Z}_t]\right|^2\right]^{1/2} \le &\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(X_t^i) - \frac{1}{N}\sum_{i=1}^N f(\bar{X}_t^i)\right|^2\right]^{1/2} \\ &+ \mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(\bar{X}_t^i) - \mathsf{E}[f(X_t)|\mathscr{Z}_t]\right|^2\right]^{1/2}\end{aligned}$$

The second term is given by

$$\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(\bar{X}_t^i) - \mathsf{E}[f(X_t)|\mathscr{Z}_t]\right|^2\right]^{1/2} = \frac{\mathrm{Var}(f(X_t)|\mathscr{Z}_t)}{\sqrt{N}}$$

because $\bar{X}_t^i$ are i.i.d with distribution equal to the conditional distribution. It only remains to bound the first term:

$$\begin{aligned}\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(X_t^i) - \frac{1}{N}\sum_{i=1}^{N} f(\bar{X}_t^i)\right|^2\right]^{1/2} &\leq \frac{1}{N}\sum_{i=1}^{N}\mathsf{E}\left[\left|f(X_t^i) - f(\bar{X}_t^i)\right|^2\right]^{1/2} \\ &\leq \frac{(\text{const.})}{N}\sum_{i=1}^{N}\mathsf{E}\left[\left|X_t^i - \bar{X}_t^i\right|^2\right]^{1/2} \leq \frac{(\text{const.})}{\sqrt{N}}\end{aligned}$$

where we used triangle inequality in the first step, the Lipschitz property of f in the second step, and the estimate (4.15) in the last step. □

4.6.5 Proof of the Proposition 4.2

Proof. (i) Recall the evolutions for Σ_t in (3.4b) and the evolution for $\Sigma_t^{(N)}$ in (4.8b) stated below under the Assumption that $H^\top H = I$:

$$\begin{aligned}\mathrm{d}\Sigma_t &= (A\Sigma_t + \Sigma_t A^\top + \Sigma_B - \Sigma_t^2)\mathrm{d}t \\ \mathrm{d}\Sigma_t^{(N)} &= (A\Sigma_t^{(N)} + \Sigma_t^{(N)}A^\top + \Sigma_B - (\Sigma_t^{(N)})^2)\mathrm{d}t + \frac{\mathrm{d}M_t}{\sqrt{N}}\end{aligned}$$

First, we obtain a bound for $\mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)})]$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)})] &= \mathsf{E}[\mathrm{Tr}((A + A^\top - \Sigma_t^{(N)})\Sigma_t^{(N)})] + \mathrm{Tr}(\Sigma_B) \\ &\leq -2\mu(A)\mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)})] + \mathrm{Tr}(\Sigma_B)\end{aligned}$$

Therefore

$$\mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)})] \leq e^{-2\mu(A)t}\mathsf{E}[\mathrm{Tr}(\Sigma_0^{(N)})] + \frac{\mathrm{Tr}(\Sigma_B)}{2\mu(A)}$$

Next, we bound the error $\Sigma_t^{(N)} - \Sigma_t$. Subtracting the two equations for $\Sigma_t^{(N)}$ and Σ_t yields,

$$\mathrm{d}(\Sigma_t^{(N)} - \Sigma_t) = (A - \frac{\Sigma_t + \Sigma_t^{(N)}}{2})(\Sigma_t^{(N)} - \Sigma_t)\mathrm{d}t + (\Sigma_t^{(N)} - \Sigma_t)(A - \frac{\Sigma_t + \Sigma_t^{(N)}}{2})^\top \mathrm{d}t + \frac{\mathrm{d}M_t}{\sqrt{N}}$$

Define $R_t = \|\Sigma_t^{(N)} - \Sigma_t\|_F^2 = \mathrm{Tr}((\Sigma_t^{(N)} - \Sigma_t)^2)$ the squared of the Forbenious norm of the error. By Itó rule

$$\begin{aligned}\mathrm{d}R_t =& 2\mathrm{Tr}((A + A^\top - \Sigma_t - \Sigma_t^{(N)})(\Sigma_t^{(N)} - \Sigma_t)^2) + \frac{1}{\sqrt{N}}\mathrm{Tr}((\mathrm{d}M_t + \mathrm{d}M_t^\top)(\Sigma_t^{(N)} - \Sigma_t)) \\ &+ \frac{2}{N}(\mathrm{Tr}(\Sigma_t^{(N)}\Sigma_B) + \mathrm{Tr}(\Sigma_t^{(N)})\mathrm{Tr}(\Sigma_B))\mathrm{d}t\end{aligned}$$

Then upon taking the expectation

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}R_t &\le -4\mu(A)R_t + \frac{2}{N}(\mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)}\Sigma_B)] + \mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)})]\mathrm{Tr}(\Sigma_B)) \\ &\le -4\mu(A)R_t + \frac{2}{N}(\frac{1}{2}R_t + \frac{1}{2}\mathrm{Tr}(\Sigma_B^2) + \mathrm{Tr}(\Sigma_t\Sigma_B) + \mathrm{Tr}(\Sigma_B)\mathsf{E}[\mathrm{Tr}(\Sigma_t^{(N)})])\end{aligned}$$

Concluding the estimate:

$$R_t \le e^{-(4\mu(A)-\frac{1}{N})t}R_0 + \frac{1}{2\mu(A)N}(\mathrm{Tr}(\Sigma_B^2) + \frac{1}{2}\mathrm{Tr}(\Lambda_{\max}^2) + \frac{\mathrm{Tr}(\Sigma_B)^2}{2\mu(A)} + 2e^{-(2\mu(A)-\frac{1}{N})t}\mathrm{Tr}(\Sigma_B)\mathsf{E}[\mathrm{Tr}(\Sigma_0^{(N)})])$$

(ii) Recall the evolutions for m_t in (3.4a) and the evolution for $m_t^{(N)}$ in (4.8a) stated below under the Assumption that $H^\top H = I$:

$$\begin{aligned}\mathrm{d}m_t &= Am_t\mathrm{d}t + \Sigma_t H^\top \mathrm{d}I_t \\ \mathrm{d}m_t^{(N)} &= Am_t^{(N)}\mathrm{d}t + \Sigma_t^{(N)}H^\top \mathrm{d}I_t - \Sigma_t^{(N)}(m_t^{(N)} - m_t)\mathrm{d}t + \frac{1}{\sqrt{N}}\sigma_B\mathrm{d}\tilde{B}_t\end{aligned}$$

where the innovation process $\mathrm{d}I_t = \mathrm{d}Z_t - Hm_t\mathrm{d}t$. Subtracting the two equations yields

$$\mathrm{d}(m_t^{(N)} - m_t) = (A - \Sigma_t^{(N)})(m_t^{(N)} - m_t)\mathrm{d}t + (\Sigma_t^{(N)} - \Sigma_t)H^\top \mathrm{d}I_t + \frac{1}{\sqrt{N}}\sigma_B\mathrm{d}\tilde{B}_t$$

The solution satisfies

$$m_t^{(N)} - m_t = \Phi_{t,s}^{(N)}(m_0^{(N)} - m_0) + \int_0^t \Phi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)H^\top \mathrm{d}I_s + \frac{1}{\sqrt{N}}\int_0^t \Phi_{t,s}^{(N)}\sigma_B\mathrm{d}\tilde{B}_s$$

where $\Phi_{t,s}^{(N)}$ is the state transition matrix for the linear system $\frac{\mathrm{d}}{\mathrm{d}t}x_t = (A - \Sigma_t^{(N)})x_t$. The state transition matrix satisfies the bound $\|\Phi_{t,s}^{(N)}\| \le e^{-\mu(A)(t-s)}$ almost surely. Hence,

$$\begin{aligned}\mathsf{E}\|m_t^{(N)} - m_t\| &\le e^{-\mu(A)t}\mathsf{E}\|m_0^{(N)} - m_0\| + \mathsf{E}[(\int_0^t \Phi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)H^\top \mathrm{d}I_s)^2]^{1/2} + \frac{1}{\sqrt{N}}\mathsf{E}[(\int_0^t \Phi_{t,s}^{(N)}\sigma_B\mathrm{d}\tilde{B}_s)^2]^{1/2} \\ &\le e^{-\mu(A)t}\mathsf{E}\|m_0^{(N)} - m_0\| + \left(\int_0^t e^{-2\mu(A)(t-s)}\mathsf{E}[\|\Sigma_t^{(N)} - \Sigma_t\|_F^2]\mathrm{d}s\right)^{1/2} + \frac{1}{\sqrt{N}}\left(\int_0^t e^{-2\mu(A)(t-s)}\mathrm{Tr}(\Sigma_B)\mathrm{d}s\right)^{1/2} \\ &\le e^{-\mu(A)t}\mathsf{E}\|m_0^{(N)} - m_0\| + \left(\frac{3\|\Sigma_0\|_F^2 + c_{\mathrm{var}}}{2\mu(A)N}\right)^{1/2} + \left(\frac{\mathrm{Tr}(\Sigma_B)}{2\mu(A)N}\right)^{1/2}\end{aligned}$$

where $c_{\text{var}} = \frac{1}{2\mu(A)N}(\text{Tr}(\Sigma_B^2) + \frac{1}{2}\text{Tr}(\Lambda_{\max}^2) + \frac{\text{Tr}(\Sigma_B)^2}{2\mu(A)} + 2\text{Tr}(\Sigma_B)\mathsf{E}[\text{Tr}(\Sigma_0^{(N)})])$.

□

4.6.6 Proof of the Proposition 4.4

(i) Define $\bar{\xi}_t^i = \bar{X}_t^i - \bar{m}_t$. The evolution for $\bar{\xi}_t^i$ is

$$\mathrm{d}\bar{\xi}_t^i = (A - \frac{1}{2}\Sigma_t)\bar{\xi}_t^i + \sigma_B \mathrm{d}B_t^i, \quad \bar{\xi}_0^i = X_0^i - m_0$$

Subtract this from (4.8c) yields:

$$\mathrm{d}(\xi_t^i - \bar{\xi}_t^i) = (A - \frac{1}{2}\Sigma_t^{(N)})(\xi_t^i - \bar{\xi}_t)\mathrm{d}t - \frac{1}{2}(\Sigma_t^{(N)} - \Sigma_t)\bar{\xi}_t^i - \frac{1}{\sqrt{N}}\sigma_B \mathrm{d}\tilde{B}_t, \quad \xi_0^i - \bar{\xi}_0^i = m_0 - m_0^{(N)}$$

The solution satisfies the implicit integral equation

$$\xi_t^i - \bar{\xi}_t^i = \Psi_{t,0}^{(N)}(\xi_0^i - \bar{\xi}_0^i) - \frac{1}{2}\int_0^t \Psi_{t,s}^{(N)}(\Sigma_s^{(N)} - \Sigma_s)\bar{\xi}_s^i \mathrm{d}s - \frac{1}{\sqrt{N}}\int_0^t \Psi_{t,s}^{(N)}\sigma_B \mathrm{d}\tilde{B}_s$$

where $\Psi_{t,s}^{(N)}$ is the state transition matrix corresponding to the linear system $\frac{\mathrm{d}}{\mathrm{d}t}x_t = (A - \frac{1}{2}\Sigma_t^{(N)})x_t$ which satisfies the bound $\|\Psi_{t,s}^{(N)}\| \leq e^{-\mu(A)(t-s)}$. Upon taking the norm and expectation:

$$\begin{aligned}\mathsf{E}[\|\xi_t^i - \bar{\xi}_t^i\|] &\leq e^{-\mu(A)t}\mathsf{E}[\|m_0^{(N)} - m_0\|] + \frac{1}{2}\int_0^t e^{-\mu(A)(t-s)}\mathsf{E}[\|\Sigma_s^{(N)} - \Sigma_s\|^2]^{1/2}\mathsf{E}[\|\bar{\xi}_s^i\|^2]^{1/2}\mathrm{d}s + \sqrt{\frac{\text{Tr}(\Sigma_B)}{2\mu(A)N}} \\ &\leq e^{-\mu(A)t}\frac{\text{Tr}(\Sigma_0)^{\frac{1}{2}}}{\sqrt{N}} + \frac{(3\|\Sigma_0\|_F^2 + c_{\text{var}})^{\frac{1}{2}}\text{Tr}(\Lambda_{\max})^{\frac{1}{2}}}{2\mu(A)\sqrt{N}} + \sqrt{\frac{\text{Tr}(\Sigma_B)}{2\mu(A)N}}\end{aligned}$$

Combining this result with the estimate (4.11) and the inequality

$$\mathsf{E}[\|X_t^i - \bar{X}_t^i\|] \leq \mathsf{E}[\|m_t^{(N)} - \bar{m}_t\|] + \mathsf{E}[\|\xi_t^i - \bar{\xi}_t^i\|]$$

concludes the estimate (4.17).

(ii) Note that

$$\frac{1}{N}\sum_{i=1}^N f(X_t^i) - \mathsf{E}[f(\bar{X}_t)|\mathscr{Z}_t] = \frac{1}{N}\sum_{i=1}^N (f(X_t^i) - f(\bar{X}_t^i)) + \frac{1}{N}\sum_{i=1}^N f(\bar{X}_t^i) - \mathsf{E}[f(\bar{X}_t)|\mathscr{Z}_t]$$

Taking the norm and using the triangle inequality yields

$$\mathsf{E}\left[\left|\frac{1}{N}\sum_{i=1}^N f(X_t^i) - \mathsf{E}[f(\bar{X}_t)|\mathscr{Z}_t]\right|\right] \leq \frac{1}{N}\sum_{i=1}^N \text{Lip}(f)\mathsf{E}[\|\bar{X}_t^i - X_t^i\|] + \frac{\text{var}(f)}{\sqrt{N}}$$

where we used f is Lipschitz, and $\bar{X}_t^i$ are i.i.d. Using the result of part (i) concludes the proof.

Part II

Sampling

Chapter 5

Accelerated Flow for Probability Distributions*

5.1 Introduction

Optimization on the space of probability distributions is important to a number of machine learning models including variational inference [Blei et al., 2017], generative models [Goodfellow et al., 2014, Arjovsky et al., 2017], and policy optimization in reinforcement learning [Sutton et al., 2000]. A number of recent studies have considered solution approaches to these problems based upon a construction of gradient flow on the space of probability distributions [Liu and Wang, 2016, Richemond and Maginnis, 2017, Zhang et al., 2018, Frogner and Poggio, 2018, Chizat and Bach, 2018, Chen et al., 2018, Liu et al., 2018]. Such constructions are useful for convergence analysis as well as development of numerical algorithms.

In this chapter, we propose a methodology and numerical algorithms that achieve accelerated gradient flows on the space of probability distributions. The proposed numerical algorithms are related to yet distinct from the accelerated stochastic gradient descent [Jain et al., 2017] and Hamiltonian Markov chain Monte-Carlo (MCMC) algorithms [Neal et al., 2011, Cheng et al., 2017]. The proposed methodology extends the variational formulation of [Wibisono et al., 2016] from vector valued variables to probability distributions. The original formulation of [Wibisono et al., 2016] was used to derive and analyze the convergence properties of a large class of accelerated optimization algorithms, most significant of which is the continuous-time limit of the Nesterov's algorithm [Su et al., 2014]. The limit is referred to as the Nesterov's ordinary differential equation.

The extension proposed in our work is based upon a generalization of the formula for the Lagrangian in [Wibisono et al., 2016]: (i) the kinetic energy term is replaced with the expected value of kinetic energy; (ii) the potential energy term is replaced with a suitably defined functional on the space of probability distributions. The variational problem is to obtain a trajectory in the space of probability distributions that minimizes the action integral of the Lagrangian.

The variational problem is modeled as a mean-field optimal problem [Bensoussan et al., 2013, Carmona and Delarue, 2017]. The maximum principle of the optimal control theory is used to derive the Hamilton's equations which represent the first order optimality conditions. The Hamilton's equations provide a generalization of the Nesterov's ODE to the space of probability distributions. A Lyapunov function is proposed for the convergence analysis of the solution of the Hamilton's equations. In this way, quantitative estimates on convergence rate are obtained for the case when the objective functional is displacement convex [Mc-

*The content of this chapter is based on [Taghvaei and Mehta, 2019].

Cann, 1997]. Table 5.1 provides a summary of the relationship between the original variational formulation in [Wibisono et al., 2016] and the extension proposed in this chapter.

We also consider the important special case when the objective functional is the relative entropy functional $\mathsf{D}(\rho|\rho_\infty)$ defined with respect to a target probability distribution ρ_∞. In this case, the accelerated gradient flow is shown to be related to the continuous limit of the Hamiltonian Monte-Carlo algorithm [Cheng et al., 2017] (Remark 5.2). The Hamilton's equations are finite-dimensional for the special case when the initial and the target probability distributions are both Gaussian. In this case, the mean evolves according to the Nesterov's ODE. For the general case, the Lyapunov function-based convergence analysis applies when the target distribution is log-concave.

As a final contribution, the proposed methodology is used to obtain a numerical algorithm. The algorithm is an interacting particle system that empirically approximates the distribution with a finite but large number of N particles. The difficult part of this construction is the approximation of the interaction term between particles. For this purpose, two types of approximations are described: (i) Gaussian approximation which is asymptotically (as $N \to \infty$) exact in Gaussian settings; and (ii) Diffusion map approximation which is computationally more demanding but asymptotically exact for a general class of distributions.

5.1.1 Related work

Construction of accelerated flows for probability distribution was proposed in [Liu et al., 2018] based on the generalization of the Nesterov's method to Riemannian manifolds [Liu et al., 2017]. The procedure involves approximating the exponential map and parallel transport map for probability distributions in the Wasserstein space. Our construction of accelerated flow is different from [Liu et al., 2018] in several respects: i) we describe a variational formulation and make connection to mean-field control theory; ii) our variational construction yields a continuous-time algorithm providing a straightforward comparison to HMCMC; iii) we carry out convergence analysis based upon a Lyapunov function method; iv) and analysis in Gaussian setting shows we recover the Nesterov ode.

Another class of related work are the interacting particle-based numerical algorithms designed to sample from a target distribution. An example is the Stein variational gradient descent (SVGD) algorithm [Liu and Wang, 2016, Liu, 2017] based on the Riemannian construction of the gradient flow. Another example is the particle optimization method [Chen et al., 2018], whose update is obtained from a solution to an optimization problem based on the variational formulation of the Langevin dynamics. Interacting particle systems have also been shown to be useful for numerically solving the nonlinear filtering problem [Del Moral et al., 1998, Reich, 2011, Yang et al., 2016, Zhang et al., 2019].

Notation: The gradient and divergence operators are denoted as ∇ and $\nabla\cdot$ respectively. With multiple variables, ∇_z denotes the gradient with respect to the variable z. Therefore, the divergence of the vector field U is $\nabla \cdot U(x) = \sum_{n=1}^d \nabla_{x_n} U_n(x)$. The space of absolutely continuous probability measures on $\mathbb{R}^d$ with finite second moments is denoted by $\mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$. For a measure $\mu \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ and a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$, the push-forward of μ by T is denoted by $T\#\mu$. The second-order Wasserstein distance between any two measures $\mu, \nu \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ is denoted as $W_2(\mu,\nu)$. The Wasserstein gradient and the

	Vector	Probability distribution
State-space	$\mathbb{R}^d$	$\mathscr{P}_2(\mathbb{R}^d)$
Objective function	$f(x)$	$\mathsf{F}(\rho) := \mathsf{D}(\rho\|\rho_\infty)$
Lagrangian	$e^{\alpha_t+\gamma_t}\left(\frac{1}{2}\|e^{-\alpha_t}u\|^2 - e^{\beta_t}f(x)\right)$	$e^{\alpha_t+\gamma_t}\mathsf{E}\left[\frac{1}{2}\|e^{-\alpha_t}U\|^2 - e^{\beta_t}\log(\frac{\rho(X)}{\rho_\infty(X)})\right]$
Lyapunov funct.	$\frac{1}{2}\|x+e^{-\gamma_t}y-\bar{x}\|^2$ $+e^{\beta_t}(f(x)-f(\bar{x}))$	$\frac{1}{2}\mathsf{E}[\|X_t+e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t)\|^2]$ $+e^{\beta_t}(\mathsf{F}(\rho_t)-\mathsf{F}(\rho_\infty))$
Convergence rate	$f(x_t)-f(\bar{x}) \le O(e^{-\beta_t})$	$F(\rho_t)-F(\rho_\infty) \le O(e^{-\beta_t})$

Table 5.1: Summary of the variational formulations for vectors and probability distributions.

Gâteaux derivative of a functional F is denoted as $\nabla_\rho \mathsf{F}(\rho)$ and $\frac{\partial \mathsf{F}}{\partial \rho}(\rho)$ respectively (see Appendix 5.5.2 for definition). The probability distribution of a random variable Z is denoted as $\mathrm{Law}(Z)$.

5.2 Review of the variational formulation of [Wibisono et al., 2016]

The basic problem is to minimize a C^1 smooth convex function f on $\mathbb{R}^d$. The standard form of the gradient descent algorithm for this problem is an ODE:

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = -\nabla f(X_t), \quad t \ge 0 \tag{5.1}$$

Accelerated forms of this algorithm are obtained based on a variational formulation due to [Wibisono et al., 2016]. The formulation is briefly reviewed here using an optimal control formalism. The Lagrangian $L : \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$L(t,x,u) := e^{\alpha_t+\gamma_t}\left(\underbrace{\frac{1}{2}|e^{-\alpha_t}u|^2}_{\text{kinetic energy}} - \underbrace{e^{\beta_t}f(x)}_{\text{potential energy}}\right) \tag{5.2}$$

where $t \ge 0$ is the time, $x \in \mathbb{R}^d$ is the state, $u \in \mathbb{R}^d$ is the velocity or control input, and the time-varying parameters $\alpha_t, \beta_t, \gamma_t$ satisfy the following scaling conditions: $\alpha_t = \log p - \log t$, $\beta_t = p\log t + \log C$, and $\gamma_t = p\log t$ where $p \ge 2$ and $C > 0$ are constants.

The variational problem is

$$\begin{aligned} \underset{u}{\text{Minimize}} \quad & J(u) = \int_0^\infty L(t,X_t,u_t)\mathrm{d}t \\ \text{Subject to} \quad & \frac{\mathrm{d}X_t}{\mathrm{d}t} = u_t, \quad X_0 = x_0 \end{aligned} \tag{5.3}$$

where u_t is the control input chosen to minimize the objective function $J(u)$. over all control laws $\{u_t\}_{t>0}$ in $\mathbb{R}^d$.

The Hamiltonian function is

$$H(t,x,y,u) = y \cdot u - L(t,x,u) \tag{5.4}$$

where $y \in \mathbb{R}^d$ is dual variable and $y \cdot u$ denotes the dot product between vectors y and u.

According to the Pontryagin's Maximum Principle, the optimal control $u_t^* = \arg\max_v H(t, X_t, Y_t, v) = e^{\alpha_t - \gamma_t} Y_t$. The resulting Hamilton's equations are

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = +\nabla_y H(t, X_t, Y_t, u_t^*) = e^{\alpha_t - \gamma_t} Y_t, \; X_0 = x_0 \tag{5.5a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -\nabla_x H(t, X_t, Y_t, u_t^*) = -e^{\alpha_t + \beta_t + \gamma_t} \nabla f(X_t), \tag{5.5b}$$

The system (5.5) is an example of accelerated gradient descent algorithm. Specifically, if the parameters $\alpha_t, \beta_t, \gamma_t$ are defined using $p = 2$, one obtains the continuous-time limit of the Nesterov's accelerated algorithm. It is referred to as the Nesterov's ODE in this chapter.

For this system, a Lyapunov function is as follows:

$$V(t, x, y) = \frac{1}{2} \left| x + e^{-\gamma_t} y - \bar{x} \right|^2 + e^{\beta_t} (f(x) - f(\bar{x})) \tag{5.6}$$

where $\bar{x} \in \arg\min_x f(x)$. It is shown in [Wibisono et al., 2016] that upon differentiating along the solution trajectory, $\frac{\mathrm{d}}{\mathrm{d}t} V(t, X_t, Y_t) \leq 0$. This yields the convergence rate:

$$f(X_t) - f(\bar{x}) \leq O(e^{-\beta_t}), \quad \forall t \geq 0 \tag{5.7}$$

5.3 Variational formulation for probability distributions

5.3.1 Motivation and background

Let $\mathsf{F} : \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d) \to \mathbb{R}$ be a functional on the space of probability distributions. Consider the problem of minimizing $\mathsf{F}(\rho)$. The (Wasserstein) gradient flow with respect to $\mathsf{F}(\rho)$ is a curve ρ_t such that

$$\frac{\partial \rho_t}{\partial t}(x) = \nabla \cdot (\rho_t(x) \nabla_\rho \mathsf{F}(\rho_t)(x)) \tag{5.8}$$

where (the vector field) $\nabla_\rho \mathsf{F}(\rho) : \mathbb{R}^d \to \mathbb{R}^d$ is the Wasserstein gradient of F.

An important example is the relative entropy functional where $\mathsf{F}(\rho) = \mathsf{D}(\rho|\rho_\infty) := \int_{\mathbb{R}^d} \log(\frac{\rho(x)}{\rho_\infty(x)}) \rho(x) \mathrm{d}x$ where $\rho_\infty \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ is referred to as the target distribution. The gradient of relative entropy is given by $\nabla_\rho \mathsf{F}(\rho)(x) = \nabla \log(\frac{\rho(x)}{\rho_\infty(x)})$ (Ch. 8.3 in [Villani, 2003]). The gradient flow

$$\frac{\partial \rho_t}{\partial t}(x) = -\nabla \cdot (\rho_t(x) \nabla \log(\rho_\infty(x))) + \Delta \rho_t(x) \tag{5.9}$$

is the Fokker-Planck equation [Jordan et al., 1998]. The gradient flow achieves the density transport from an initial probability distribution ρ_0 to the target (here, also equilibrium) probability distribution ρ_∞; and underlies the construction and the analysis of Markov chain Monte-Carlo (MCMC) algorithms. The simplest

MCMC algorithm is the Langevin stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad X_0 \sim \rho_0$$

where B_t is the standard Brownian motion in $\mathbb{R}^d$.

The main problem of this chapter is to construct an accelerated form of the gradient flow (5.8). The proposed solution is based upon a variational formulation. As tabulated in Table 5.1, the solution represents a generalization of [Wibisono et al., 2016] from its original deterministic finite-dimensional to now probabilistic infinite-dimensional settings.

The variational problem can be expressed in two equivalent forms: (i) The probabilistic form and (ii) The partial differential equation (PDE)

5.3.2 Probabilistic form of the variational problem

Consider the stochastic process $\{X_t\}_{t\geq 0}$ that takes values in $\mathbb{R}^d$ and evolves according to:

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t, \quad X_0 \sim \rho_0$$

where the control input $\{U_t\}_{t\geq 0}$ also takes values in $\mathbb{R}^d$, and $\rho_0 \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ is the probability distribution of the initial condition X_0. It is noted that the randomness here comes only from the random initial condition.

Suppose the objective functional is of the form $\mathsf{F}(\rho) = \int \tilde{F}(\rho, x)\rho(x)\mathrm{d}x$. The Lagrangian $\mathsf{L} : \mathbb{R}^+ \times \mathbb{R}^d \times \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathsf{L}(t,x,\rho,u) := e^{\alpha_t+\gamma_t}\left(\underbrace{\frac{1}{2}|e^{-\alpha_t}u|^2}_{\text{kinetic energy}} - \underbrace{e^{\beta_t}\tilde{F}(\rho,x)}_{\text{potential energy}}\right) \tag{5.10}$$

This formula is a natural generalization of the Lagrangian (5.2) and the parameters $\alpha_t, \beta_t, \gamma_t$ are defined exactly the same as in the finite-dimensional case. The stochastic optimal control problem is:

$$\begin{aligned} \text{Minimize} \quad & \mathsf{J}(u) = \mathsf{E}\left[\int_0^\infty \mathsf{L}(t, X_t, \rho_t, U_t)\mathrm{d}t\right] \\ \text{Subject to} \quad & \frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t, \quad X_0 \sim \rho_0 \end{aligned} \tag{5.11}$$

where $\rho_t = \mathrm{Law}(X_t) \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ is the probability density function of the random variable X_t.

The Hamiltonian function $\mathsf{H} : \mathbb{R}^+ \times \mathbb{R}^d \times \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for this problem is given by (see Sec. 6.2.3 in [Carmona and Delarue, 2017]):

$$\mathsf{H}(t,x,\rho,y,u) := u \cdot y - \mathsf{L}(t,x,\rho,u) \tag{5.12}$$

where $y \in \mathbb{R}^d$ is the dual variable.

Remark 5.1. *The variational problem* (5.11) *is an example of a mean-field (McKean-Vlasov) optimal con-*

trol problem. This is because the Lagrangian depends also upon the law of the stochastic process; cf., Ch. 6 in [Carmona and Delarue, 2017].

5.3.3 PDE formulation of the variational problem

An equivalent pde formulation is obtained by considering the stochastic optimal control problem (5.11) as a deterministic optimal control problem on the space of the probability distributions. Specifically, the process $\{\rho_t\}_{t\geq 0}$ is a deterministic process that takes values in $\mathscr{P}_{\text{ac},2}(\mathbb{R}^d)$ and evolves according to the continuity equation

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t u_t)$$

where $u_t : \mathbb{R}^d \to \mathbb{R}^d$ is now a time-varying vector field. The Lagrangian $\mathscr{L} : \mathbb{R}^+ \times \mathscr{P}_{\text{ac},2}(\mathbb{R}^d) \times L^2(\mathbb{R}^d;\mathbb{R}^d) \to \mathbb{R}$ is defined as:

$$\mathscr{L}(t,\rho,u) := e^{\alpha_t+\gamma_t}\left[\int_{\mathbb{R}^d} \frac{1}{2}|e^{-\alpha_t}u(x)|^2\rho(x)\mathrm{d}x - e^{\beta_t}\mathsf{F}(\rho)\right] \tag{5.13}$$

The optimal control problem is:

$$\begin{aligned} &\text{Minimize} && \int_0^\infty \mathscr{L}(t,\rho_t,u_t)\mathrm{d}t \\ &\text{Subject to} && \frac{\partial \rho_t}{\partial t} + \nabla\cdot(\rho_t u_t) = 0 \end{aligned} \tag{5.14}$$

The Hamiltonian function $\mathscr{H} : \mathbb{R}^+ \times \mathscr{P}_{\text{ac},2}(\mathbb{R}^d) \times \mathscr{C}(\mathbb{R}^d;\mathbb{R}) \times L^2(\mathbb{R}^d;\mathbb{R}^d) \to \mathbb{R}$ is

$$\mathscr{H}(t,\rho,\phi,u) := \langle \nabla\phi, u\rangle_{L^2(\rho)} - \mathscr{L}(t,\rho,u) \tag{5.15}$$

where $\phi \in \mathscr{C}(\mathbb{R}^d;\mathbb{R})$ is the dual variable and the inner-product $\langle \nabla\phi, u\rangle_{L^2(\rho)} := \int_{\mathbb{R}^d} \nabla\phi(x)\cdot u(x)\rho(x)\mathrm{d}x$

5.3.4 Main result

Theorem 5.1. *Consider the variational problem* (5.11)-(5.14).

(i) *For the probabilistic form* (5.11) *of the variational problem, the optimal control* $U_t^* = e^{\alpha_t-\gamma_t}Y_t$, *where the optimal trajectory* $\{(X_t,Y_t)\}_{t\geq 0}$ *evolves according to the Hamilton's odes:*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = U_t^* = e^{\alpha_t-\gamma_t}Y_t, \quad X_0 \sim \rho_0 \tag{5.16a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t+\beta_t+\gamma_t}\nabla_\rho \mathsf{F}(\rho_t)(X_t), \quad Y_0 = \nabla\phi_0(X_0) \tag{5.16b}$$

where ϕ_0 *is a convex function, and* $\rho_t = Law(X_t)$.

(ii) *For the pde form* (5.14) *of the variational problem, the optimal control is* $u_t^* = e^{\alpha_t-\gamma_t}\nabla\phi_t(x)$,

where the optimal trajectory $\{(\rho_t, \phi_t)\}_{t\geq 0}$ *evolves according to the Hamilton's pdes:*

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \underbrace{e^{\alpha_t - \gamma_t} \nabla \phi_t}_{u_t^*}), \tag{5.17a}$$

$$\frac{\partial \phi_t}{\partial t} = -e^{\alpha_t - \gamma_t} \frac{|\nabla \phi_t|^2}{2} - e^{\alpha_t + \gamma_t + \beta_t} \nabla_\rho \mathsf{F}(\rho) \tag{5.17b}$$

(iii) *The solutions of the two forms are equivalent in the following sense:*

$$Law(X_t) = \rho_t, \quad U_t = u_t(X_t), \quad Y_t = \nabla \phi_t(X_t)$$

(iv) *Suppose additionally that the functional* F *is displacement convex and* ρ_∞ *is its minimizer. Define*

$$\begin{aligned} V(t) = &\frac{1}{2}\mathsf{E}[|X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)|^2] \\ &+ e^{\beta_t}(\mathsf{F}(\rho) - \mathsf{F}(\rho_\infty)) \end{aligned} \tag{5.18}$$

where the map $T_{\rho_t}^{\rho_\infty} : \mathbb{R}^d \to \mathbb{R}^d$ *is the optimal transport map from* ρ_t *to* ρ_∞*. Assume the dimension* $d = 1$*. Consequently, the following rate of convergence is obtained along the optimal trajectory*

$$\mathsf{F}(\rho_t) - \mathsf{F}(\rho_\infty) \leq O(e^{-\beta_t}), \quad \forall t \geq 0$$

Proof sketch. The Hamilton's equations are derived using the standard mean-field optimal control theory [Carmona and Delarue, 2017]. The Lyapunov function argument is based upon the variational inequality characterization of a displacement convex function (see Eq. 10.1.7 in [Ambrosio et al., 2008]). The detailed proof appears in the Appendix 5.5.1. We expect that the assumption that $d = 1$ is not necessary. This is the subject of the continuing work. □

5.3.5 Relative entropy as the functional

In the remainder of this chapter, we assume that the functional $\mathsf{F}(\rho) = \mathsf{D}(\rho|\rho_\infty)$ is the relative entropy where $\rho_\infty \in \mathscr{P}_{\text{ac},2}(\mathbb{R}^d)$ is a given target probability distribution. In this case the Hamilton's equations are given by

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = e^{\alpha_t - \gamma_t} Y_t, \quad X_0 \sim \rho_0 \tag{5.19a}$$

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t}(\nabla f(X_t) + \nabla \log(\rho_t(X_t))), \tag{5.19b}$$

with $Y_0 = \nabla \phi_0(X_0)$, where $\rho_t = \mathrm{Law}(X_t)$ and $f = -\log(\rho_\infty)$. Moreover, if f is convex (or equivalently ρ_∞ is log-concave), then F is displacement convex with the unique minimizer at ρ_∞ and the convergence estimate is given by $\mathsf{D}(\rho_t|\rho_\infty) \leq O(e^{-\beta_t})$.

Remark 5.2. *The Hamilton's equations* (5.19) *with the relative entropy functional is related to the under-damped Langevin equation [Cheng et al., 2017]. A basic form of the under-damped (or second order)*

Langevin equation is

$$\begin{aligned} \mathrm{d}X_t &= v_t \mathrm{d}t \\ \mathrm{d}v_t &= -\gamma v_t \mathrm{d}t - \nabla f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \end{aligned} \tag{5.20}$$

where $\{B_t\}_{t\geq 0}$ *is the standard Brownian motion.*

Consider next, the the accelerated flow (5.19)*. Denote* $v_t := e^{\alpha_t - \gamma_t} Y_t$*. Then, with an appropriate choice of scaling parameters (e.g.* $\alpha_t = 0$*,* $\beta_t = 0$ *and* $\gamma_t = -\gamma t$ *):*

$$\begin{aligned} \mathrm{d}X_t &= v_t \mathrm{d}t \\ \mathrm{d}v_t &= -\gamma v_t \mathrm{d}t - \nabla f(X_t)\mathrm{d}t - \nabla_x \log(\rho_t(X_t)) \end{aligned} \tag{5.21}$$

The scaling parameters are chosen here for the sake of comparison and do not satisfy the ideal scaling conditions of [Wibisono et al., 2016].

The sdes (5.20) *and* (5.21) *are similar except that the stochastic term* $\sqrt{2}\mathrm{d}B_t$ *in* (5.20) *is replaced with a deterministic term* $-\nabla_x \log(\rho_t(X_t))$ *in* (5.21)*. Because of this difference, the resulting distributions are different.*

5.3.6 Quadratic Gaussian case

Suppose the initial distribution ρ_0 and the target distribution ρ_∞ are both Gaussian, denoted as $\mathcal{N}(m_0, \Sigma_0)$ and $\mathcal{N}(\bar{x}, Q)$, respectively. This is equivalent to the objective function $f(x)$ being quadratic of the form $f(x) = \frac{1}{2}(x-\bar{x})^\top Q^{-1}(x-\bar{x})$. Therefore, this problem is referred to as the *quadratic Gaussian case*. The following Proposition shows that the mean of the stochastic process (X_t, Y_t) evolves according to the Nesterov ODE (5.5):

Proposition 5.1. *(Quadratic Gaussian case) Consider the variational problem* (5.11) *for the quadratic Gaussian case. Then*

(i) *The stochastic process* (X_t, Y_t) *is a Gaussian process. The Hamilton's equations are given by:*

$$\begin{aligned} \frac{\mathrm{d}X_t}{\mathrm{d}t} &= e^{\alpha_t - \gamma_t} Y_t, \\ \frac{\mathrm{d}Y_t}{\mathrm{d}t} &= -e^{\alpha_t + \beta_t + \gamma_t}(Q^{-1}(X_t - \bar{x}) - \Sigma_t^{-1}(X_t - m_t)) \end{aligned}$$

where m_t *and* Σ_t *are the mean and the covariance of* X_t*.*

(ii) *Upon taking the expectation of both sides, and denoting* $n_t := \mathsf{E}[Y_t]$

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = e^{\alpha_t - \gamma_t} n_t, \quad \frac{\mathrm{d}n_t}{\mathrm{d}t} = -e^{\alpha_t + \beta_t + \gamma_t} \underbrace{Q^{-1}(m_t - \bar{x})}_{\nabla f(m_t)}$$

which is identical to Nesterov ODE (5.5)*.*

Proof sketch. Fix ρ_t. Consider the resulting pair (X_t, Y_t) from (5.19) and let $\tilde{\rho}_t = \mathrm{Law}(X_t)$. The proof follows from showing that a Gaussian ρ_t is a fixed-point of the map $\rho_t \mapsto \tilde{\rho}_t$. □

5.4 Numerical algorithm

The proposed numerical algorithm is based upon an interacting particle implementation of the Hamilton's equation (5.19). Consider a system of N particles $\{(X_t^i, Y_t^i)\}_{i=1}^N$ that evolve according to:

$$\begin{aligned}\frac{\mathrm{d}X_t^i}{\mathrm{d}t} &= e^{\alpha_t-\gamma_t} Y_t^i, \quad X_0^i \sim \rho_0 \\ \frac{\mathrm{d}Y_t^i}{\mathrm{d}t} &= -e^{\alpha_t+\beta_t+\gamma_t}(\nabla f(X_t^i) + \underbrace{I_t^{(N)}(X_t^i)}_{\text{interaction term}}),\end{aligned}$$

with $Y_0^i = \nabla\phi_0(X_0^i)$. The interaction term $I_t^{(N)}$ is an empirical approximation of the $\nabla \log(\rho_t)$ term in (5.19). We propose two types of empirical approximations as follows:

1. Gaussian approximation: Suppose the density is approximated as a Gaussian $\mathcal{N}(m_t, \Sigma_t)$. In this case, $\nabla \log(\rho_t(x)) = -{\Sigma_t}^{-1}(x - m_t)$. This motivates the following empirical approximation of the interaction term:

$$I_t^{(N)}(x) = -{\Sigma_t^{(N)}}^{-1}(x - m_t^{(N)}) \tag{5.23}$$

where $m_t^{(N)} := N^{-1}\sum_{i=1}^N X_t^i$ is the empirical mean and $\Sigma_t^{(N)} := \frac{1}{N-1}\sum_{i=1}^N (X_t^i - m_t^{(N)})(X_t^i - m_t^{(N)})^\top$ is the empirical covariance.

Even though the approximation is asymptotically (as $N \to \infty$) exact only under the Gaussian assumption, it may be used in a more general settings, particularly when the density ρ_t is unimodal. The situation is analogous to the (Bayesian) filtering problem, where an ensemble Kalman filter is used as an approximate solution for non-Gaussian distributions [Evensen, 2003].

2. Diffusion map approximation: This is based upon the diffusion map approximation of the weighted Laplacian operator [Coifman and Lafon, 2006, Hein et al., 2007]. For a C^2 function f, the weighted Laplacian is defined as $\Delta_\rho f := \frac{1}{\rho}\nabla\cdot(\rho\nabla f)$. Denote $e(x) = x$ as the coordinate function on $\mathbb{R}^d$. It is a straightforward calculation to show that $\nabla\log(\rho) = \Delta_\rho e$. This allows one to use the diffusion map approximation of the weighted Laplacian to approximate the interaction term as follows:

$$\text{(DM)}\quad I_t^{(N)}(X_t^i) = \frac{1}{\varepsilon}\frac{\sum_{j=1}^N k_\varepsilon(X_t^i, X_t^j)(X_t^j - X_t^i)}{\sum_{j=1}^N k_\varepsilon(X_t^i, X_t^j)} \tag{5.24}$$

where the kernel $k_\varepsilon(x,y) = \frac{g_\varepsilon(x,y)}{\sqrt{\sum_{i=1}^N g_\varepsilon(y,X^i)}}$ is constructed empirically in terms of the Gaussian kernel $g_\varepsilon(x,y) = \exp(-|x-y|^2/(4\varepsilon))$. The parameter ε is referred to as the kernel bandwidth. The approximation is asymptotically exact as $\varepsilon \downarrow 0$ and $N \uparrow \infty$. The approximation error is of order $O(\varepsilon) + O(\frac{1}{\sqrt{N}\varepsilon^{d/4}})$ where the first term is referred to as the bias error and the second term is referred to as the variance error [Hein et al., 2007]. The variance error is the dominant term in the error for small values of ε, whereas the bias error is the dominant term for large values of ε (see Figure 5.2(d)).

The resulting interacting particle algorithm is tabulated in Table 5.1. The symplectic method proposed

Algorithm 5.1 Interacting particle implementation of the accelerated gradient flow

Input: ρ_0, ϕ_0, N, t_0, Δt, p, C, K
Output: $\{X_k^i\}_{i=1,k=0}^{N,K}$

Initialize $\{X_0^i\}_{i=1}^N \overset{\text{i.i.d}}{\sim} \rho_0$, $Y_0^i = \nabla\phi_0(X_0^i)$
Compute $I_0^{(N)}(X_0^i)$ with (5.23) or (5.24)
for $k = 0$ to $K-1$ **do**
  $t_{k+\frac{1}{2}} = t_k + \frac{1}{2}\Delta t$
  $Y_{k+\frac{1}{2}}^i = Y_k^i - \frac{1}{2}Cpt_{k+\frac{1}{2}}^{2p-1}(\nabla f(X_k^i) + I_k^{(N)}(X_k^i))\Delta t$
  $X_{k+1}^i = X_k^i + \frac{p}{t_{k+\frac{1}{2}}^{p+1}} Y_k^i \Delta t$
  Compute $I_{k+1}^{(N)}(X_{k+1}^i)$ with (5.23)or (5.24)
  $Y_{k+1}^i = Y_{k+\frac{1}{2}}^i - \frac{1}{2}Cpt_{k+\frac{1}{2}}^{2p-1}(\nabla f(X_{k+1}^i) + I_{k+1}^{(N)}(X_{k+1}^i))\Delta t$
  $t_{k+1} = t_{k+\frac{1}{2}} + \frac{1}{2}\Delta t$
end for

in [Betancourt et al., 2018] is used to carry out the numerical integration. The algorithm is applied to two examples as described in the following sections.

Remark 5.3. *For the case where there is only one particle ($N = 1$), the interaction term is zero and the system (5.22) reduces to the Nesterov ODE (5.5).*

Remark 5.4. *(Comparison with density estimation) The diffusion map approximation algorithm is conceptually different from an explicit density estimation-based approach. A basic density estimation is to approximate $\rho(x) \approx \frac{1}{N}\sum_{i=1}^N g_\varepsilon(x, X_t^i)$ where $g_\varepsilon(x,y)$ is the Gaussian kernel. Using such an approximation, the interaction term is approximated as*

$$(DE)\quad I_t^{(N)}(X_t^i) = \frac{1}{\varepsilon}\frac{\sum_{j=1}^N g_\varepsilon(X_t^i, X_t^j)(X_t^j - X_t^i)}{2\sum_{j=1}^N g_\varepsilon(X_t^i, X_t^j)} \tag{5.25}$$

Despite the apparent similarity of the two formulae, (5.24) for diffusion map approximation and (5.25) for density estimation, the nature of the two approximations is different. The difference arises because the kernel $k_\varepsilon(x,y)$ in (5.24) is data-dependent whereas the kernel in (5.25) is not. While both approximations are exact in the asymptotic limit as $N \uparrow \infty$ and $\varepsilon \downarrow 0$, they exhibit different convergence rates. Numerical experiments presented in Figure 5.2(a)-(d) show that the diffusion map approximation has a much smaller variance for intermediate values of N. Theoretical understanding of the difference is the subject of continuing work.

5.4.1 Gaussian Example

Consider the Gaussian example as described in Sec. 5.3.6. The simulation results for the scalar ($d = 1$) case with initial distribution $\rho_0 = \mathcal{N}(2,4)$ and target distribution $\mathcal{N}(\bar{x}, Q)$ where $\bar{x} = -5.0$ and $Q = 0.25$ is depicted in Figure 5.1-(a)-(b). For this simulation, the numerical parameters are as follows: $N = 100$, $\phi_0(x) = 0.5(x-2)$, $t_0 = 1$, $\Delta t = 0.1$, $p = 2$,$C = 0.625$, and $K = 400$. The result numerically verifies the

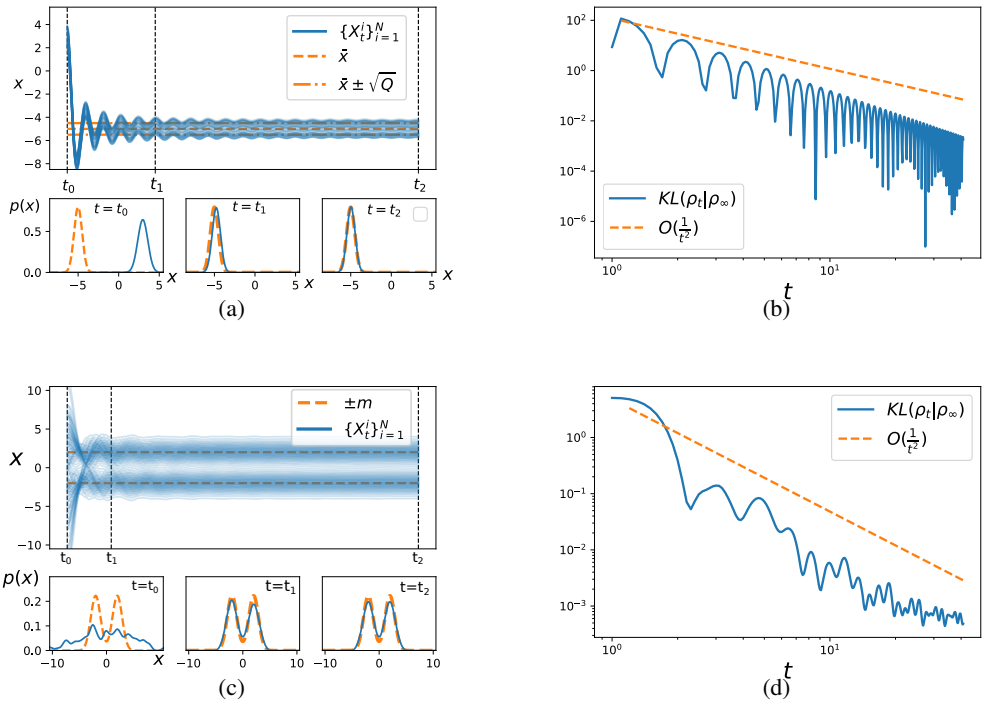


Figure 5.1: Simulation result for the Gaussian case (Example 5.4.1): (a) The time traces of the particles; (b) The KL-divergence as a function of time. Simulation result for the non-Gaussian case (Example 5.4.2): (c) The time traces of the particles; (d) The KL-divergence as a function of time.

$O(e^{-\beta_t}) = O(\frac{1}{t^2})$ convergence rate derived in Theorem 5.1 for the case where the target distribution is Gaussian.

5.4.2 Non-Gaussian example

This example involves a non-Gaussian target distribution $\rho_\infty = \frac{1}{2}\mathcal{N}(-m,\sigma^2) + \frac{1}{2}\mathcal{N}(m,\sigma^2)$ which is a mixture of two one-dimensional Gaussians with $m = 2.0$ and $\sigma^2 = 0.8$. The simulation results are depicted in Figure 5.1-(c)-(d). The numerical parameters are same as in the Example 5.4.1. The interaction term is approximated using the diffusion map approximation with $\varepsilon = 0.01$. The numerical result depicted in Figure 5.1-(c) show that the diffusion map algorithm converges to the mixture of Gaussian target distribution. The result depicted in Figure 5.1-(d) suggests that the convergence rate $O(e^{-\beta_t})$ also appears to hold for this non-log-concave target distribution. Theoretical justification of this is subject of continuing work.

5.4.3 Comparison with MCMC and HMCMC

This section contains numerical experiment comparing the performance of the accelerated algorithm 5.1 using the diffusion map (DM) approximation (5.24) and the density estimation (DE)-based approximation (5.25) with the Markov chain Monte-Carlo (MCMC) algorithm studied in [Durmus and Moulines, 2016] and the Hamiltonian MCMC algorithm studied in [Cheng et al., 2017].

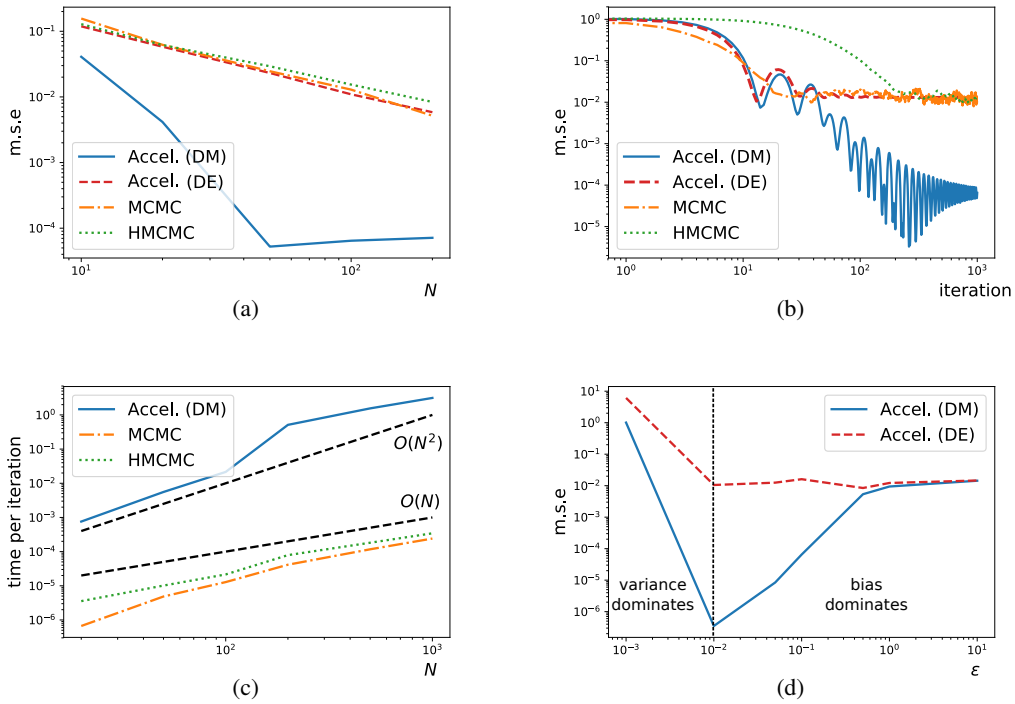


Figure 5.2: Simulation-based comparison of the performance of the accelerated algorithm 5.1 using the diffusion map (DM) approximation (5.24), the density estimation (DE)-based approximation (5.25) with the MCMC and HMCMC algorithms: (a) the mean-squared error (m.s.e) (5.26) as a function of the number of samples N; (b) the m.s.e as a function of the number of iterations; (c) the average computational time per iteration as a function of the number of samples; (d) m.s.e comparison between the diffusion map and the density estimation-based approaches as a function of the kernel bandwidth ε.

We consider the problem setting of the mixture of Gaussians as in example 5.4.2. All algorithms are simulated with a fixed step-size of $\Delta t = 0.1$ for $K = 1000$ iterations. The performance is measured by computing the mean-squared error in estimating the expectation of the function $\psi(x) = x1_{x\geq 0}$ denoted as $\hat{\psi} := \int \psi(x)\rho_\infty(x)\mathrm{d}x$. The mean-square error at the k-th iteration is computed by averaging the error over $M = 100$ runs:

$$\text{m.s.e}_k = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N}\psi(X_{t_k}^{i,m}) - \hat{\psi}\right)^2 \tag{5.26}$$

The numerical results are depicted in Figure 5.2. Figure 5.2(a) depicts the m.s.e as a function of N. It is observed that the accelerated algorithm 5.1 with the diffusion map approximation admits an order of magnitude better m.s.e for the same number of particles. It is also observed that the m.s.e decreases rapidly for intermediate values of N before saturating for large values of N, where the bias term dominates (see discussion following Eq. 5.24).

Figure 5.2(b) depicts the m.s.e as a function of the number of iterations for a fixed number of particles $N = 100$. It is observed that the accelerated algorithm 5.1 displays the quickest convergence amongst the algorithms tested.

Figure 5.2(c) depicts the average computational time per iteration as a function of the number of samples N. The computational time of the diffusion map approximation scales as $O(N^2)$ because it involves computing a $N \times N$ matrix $[k_\varepsilon(X^i, X^j)]_{i,j=1}^N$, while the computational cost of the MCMC and HMCMC algorithms scale as $O(N)$. The computational complexity may be improved by (i) exploiting the sparsity structure of the $N \times N$ matrix ; (ii) sub-sampling the particles in computing the empirical averages; (iii) adaptively updating the $N \times N$ matrix according to a certain error criteria.

Finally, we provide comparison between diffusion map approximation (5.25) and the density-based approximation (5.25): Figure 5.2(d) depicts the m.s.e for these two approximations as a function of the kernel-bandwidth ε for a fixed number of particles $N = 100$. For very large and for very small values of ε, where bias and variance dominates the error, respectively, the two algorithms have similar m.s.e. However, for intermediate values of ε, the diffusion map approximation has smaller variance, and thus lower m.s.e.

5.5 Supplementary information

5.5.1 Proof of Thm. 5.1

Proof. (i) The Hamiltonian function defined in (5.12) is equal to

$$\mathsf{H}(t,x,\rho,y,u) = y \cdot u - e^{\gamma_t - \alpha_t}\frac{1}{2}|u|^2 + e^{\alpha_t+\gamma_t\beta_t}\tilde{F}(\rho,x)$$

after inserting the formula for the Lagrangian. According to the maximum principle in probabilistic form for (mean-field) optimal control problems (see [Carmona and Delarue, 2017, Sec. 6.2.3]), the optimal control law $U_t^* = \arg\min_v \mathsf{H}(t,X_t,\rho_t,Y_t,v) = e^{\alpha_t-\gamma_t}Y_t$ and the Hamilton's equations are

$$\begin{aligned}\frac{\mathrm{d}X_t}{\mathrm{d}t} &= +\nabla_y \mathsf{H}(t,X_t,\rho_t,Y_t,U_t^*) = U_t^* = e^{\alpha_t-\gamma_t}Y_t \\ \frac{\mathrm{d}Y_t}{\mathrm{d}t} &= -\nabla_x \mathsf{H}(t,X_t,\rho_t,Y_t,U_t^*) - \tilde{\mathsf{E}}[\nabla_\rho \mathsf{H}(t,\tilde{X}_t,\rho_t,\tilde{Y}_t,\tilde{U}_t^*)(X_t)]\end{aligned}$$

where $\tilde{X}_t, \tilde{Y}_t, \tilde{U}_t^*$ are independent copies of X_t, Y_t, U_t^*. The derivatives

$$\begin{aligned}\nabla_x \mathsf{H}(t,x,\rho,y,u) &= e^{\alpha_t+\beta_t+\gamma_t}\nabla_x\tilde{F}(\rho,x) \\ \nabla_\rho \mathsf{H}(t,x,\rho,y,u) &= e^{\alpha_t+\beta_t+\gamma_t}\nabla_\rho\tilde{F}(\rho,x)\end{aligned}$$

It follows that

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -e^{\alpha_t+\beta_t+\gamma_t}\left(\nabla_x\tilde{F}(\rho_t,X_t) + \tilde{\mathsf{E}}[\nabla_\rho\tilde{F}(\rho_t,\tilde{X}_t)(X_t)]\right) = -e^{\alpha_t+\beta_t+\gamma_t}\nabla_\rho \mathsf{F}(\rho)(X_t)$$

where we used the definition $\mathsf{F}(\rho) = \int \tilde{F}(x,\rho)\rho(x)\mathrm{d}x$ and the identity [Carmona and Delarue, 2017, Sec. 5.2.2 Example 3]

$$\nabla_\rho \mathsf{F}(\rho)(x) = \nabla_x\tilde{F}(\rho,x) + \int \nabla_\rho\tilde{F}(\rho,\tilde{x})(x)\rho(\tilde{x})\mathrm{d}\tilde{x}$$

(ii) The Hamiltonian function defined in (5.15) is equal to

$$\mathscr{H}(t,\rho,\phi,u) = \int \left[\nabla\phi(x)\cdot u(x) - \frac{1}{2}e^{\gamma_t-\alpha_t}|u(x)|^2\right]\rho(x)\mathrm{d}x + e^{\alpha_t+\gamma_t+\beta_t}\mathsf{F}(\rho)$$

after inserting the formula for the Lagrangian. According to the maximum principle for pde formulation of mean-field optimal control problems (see [Carmona and Delarue, 2017, Sec. 6.2.4]) the optimal control vector field is $u_t^* = \arg\min_v \mathscr{H}(t,\rho_t,\phi_t,v) = e^{\alpha_t-\gamma_t}\nabla\phi_t$ and the Hamilton's equations are:

$$\begin{aligned}
\frac{\partial\rho_t}{\partial t} &= +\frac{\partial\mathscr{H}}{\partial\phi}(t,\rho_t,\phi_t,u_t) = -\nabla\cdot(\rho_t\nabla u_t^*)\\
\frac{\partial\phi_t}{\partial t} &= -\frac{\partial\mathscr{H}}{\partial\rho}(t,\rho_t,\phi_t,u_t) = -(\nabla\phi\cdot u^* - e^{\gamma_t-\alpha_t}\frac{1}{2}|u_t^*|^2 + e^{\alpha_t+\gamma_t+\beta_t}\frac{\partial\mathsf{F}}{\partial\rho}(\rho_t))
\end{aligned}$$

inserting the formula $u_t^* = e^{\alpha_t-\gamma_t}\nabla\phi_t$ concludes the result.

(iii) Consider the (ρ_t,ϕ_t) defined from (5.17). The distribution ρ_t is identified with a stochastic process $\tilde{X}_t$ such that $\frac{\mathrm{d}\tilde{X}_t}{\mathrm{d}t} = e^{\alpha_t-\gamma_t}\nabla\phi_t(\tilde{X}_t)$ and $\mathrm{Law}(\tilde{X}_t) = \rho_t$. Then define $\tilde{Y}_t = \nabla\phi_t(\tilde{X}_t)$. Taking the time derivative shows that

$$\begin{aligned}
\frac{\mathrm{d}\tilde{Y}_t}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\nabla\phi_t(\tilde{X}_t) = \nabla^2\phi_t(\tilde{X}_t)\frac{\mathrm{d}\tilde{X}_t}{\mathrm{d}t} + \nabla\frac{\partial\phi_t}{\partial t}(X_t)\\
&= e^{\alpha_t-\gamma_t}\nabla^2\phi_t(\tilde{X}_t)\nabla\phi_t(\tilde{X}_t) - e^{\alpha_t-\gamma_t}\nabla^2\phi_t(\tilde{X}_t)\nabla\phi_t(X_t) - e^{\alpha_t+\beta_t+\gamma_t}\nabla\frac{\partial\mathsf{F}}{\partial\rho}(\rho_t)(\tilde{X}_t)\\
&= -e^{\alpha_t+\beta_t+\gamma_t}\nabla\frac{\partial\mathsf{F}}{\partial\rho}(\rho_t)(\tilde{X}_t)\\
&= -e^{\alpha_t+\beta_t+\gamma_t}\nabla_\rho\mathsf{F}(\rho_t)(\tilde{X}_t)
\end{aligned}$$

with the initial condition $\tilde{Y}_0 = \nabla\phi_0(\tilde{X}_0)$, where we used the identity $\nabla_x\frac{\partial\mathsf{F}}{\partial\rho}(\rho) = \nabla_\rho\mathsf{F}(\rho)$ [Carmona and Delarue, 2017, Prop. 5.48]. Therefore the equations for $\tilde{X}_t$ and $\tilde{Y}_t$ are identical. Hence one can identify (X_t,Y_t) with $(\tilde{X}_t,\tilde{Y}_t)$.

(iv) The energy functional

$$V(t) = \underbrace{\frac{1}{2}\mathsf{E}\left[|X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t)|^2\right]}_{\text{first term}} + \underbrace{e^{\beta_t}(\mathsf{F}(\rho) - \mathsf{F}(\rho_\infty))}_{\text{second term}}$$

Then the derivative of the first term is

$$\mathsf{E}\left[(X_t + e^{-\gamma_t}Y_t - T_{\rho_t}^{\rho_\infty}(X_t))\cdot(e^{\alpha_t-\gamma_t}Y_t - \dot{\gamma}_t e^{-\gamma_t}Y_t - e^{\alpha_t+\beta_t}\nabla_\rho\mathsf{F}(\rho_t)(X_t) + \xi(T_{\rho_t}^{\rho_\infty}(X_t)))\right]$$

where $\xi(T_{\rho_t}^{\rho_\infty}(X_t)) := \frac{\mathrm{d}}{\mathrm{d}t}T_{\rho_t}^{\rho_\infty}(X_t)$. Using the scaling condition $\dot{\gamma}_t = e^{\alpha_t}$ the derivative of the first

term simplifies to

$$\mathsf{E}\left[(X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot (-e^{\alpha_t+\beta_t} \nabla_\rho \mathsf{F}(\rho_t)(X_t) + \xi(T_{\rho_t}^{\rho_\infty}(X_t)))\right]$$

We claim that when the dimension $d = 1$, the expectation

$$\mathsf{E}[(X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot \xi(T_{\rho_t}^{\rho_\infty}(X_t))] = 0 \tag{5.27}$$

We present the proof for the claim at the end. Assuming that the claim is true, the derivative of the first term simplifies to

$$\mathsf{E}\left[(X_t + e^{-\gamma_t} Y_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot (-e^{\alpha_t+\beta_t} \nabla_\rho F(\rho_t)(X_t))\right]$$

The derivative of the second term is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\text{second term}) &= \dot{\beta}_t e^{\beta_t}(\mathsf{F}(\rho_t) - \mathsf{F}(\rho_\infty)) + e^{\beta_t}\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{F}(\rho_t) \\ &= e^{\alpha_t+\beta_t}(\mathsf{F}(\rho_t) - \mathsf{F}(\rho_\infty)) + e^{\beta_t}\mathsf{E}[\nabla_\rho \mathsf{F}(\rho_t)(X_t) e^{\alpha_t-\gamma_t} Y_t]\end{aligned}$$

where we used the scaling condition $\dot{\beta}_t = e^{\alpha_t}$ and the chain-rule for the Wasserstein gradient [Ambrosio et al., 2008, Ch. 10, E. Chain rule]. Adding the derivative of the first and second term yields:

$$\frac{\mathrm{d}V}{\mathrm{d}t}(t) = e^{\alpha_t+\beta_t}\left(\mathsf{F}(\rho_t) - \mathsf{F}(\rho_\infty) - \mathsf{E}\left[(X_t - T_{\rho_t}^{\rho_\infty}(X_t)) \cdot \nabla_\rho \mathsf{F}(\rho_t)(X_t)\right]\right)$$

which is negative by variational inequality characterization of the displacement convex function $\mathsf{F}(\rho)$ [Ambrosio et al., 2008, Eq. 10.1.7].

We now present the proof of the claim (5.27) under the assumption that $d = 1$. According to Brenier theorem [Villani, 2003], there exists a convex function ψ_t such that $T_{\rho_t}^{\rho_\infty}(x) = \nabla\psi_t(x)$ and $T_{\rho_\infty}^{\rho_t}(x) = \nabla\psi_t^\star(x)$ where $\psi_t^\star$ is the convex conjugate of ψ_t. Because ρ_∞ is the push-forward of ρ_t under the map $\nabla\psi_t$, we have

$$\mathsf{E}[g(\nabla\psi_t(X_t))] = \int g(x)\rho_\infty(x)\mathrm{d}x,$$

for all measurable functions g. Upon taking the derivative with respect to time,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}[g(\nabla\psi_t(X_t))] = \frac{\mathrm{d}}{\mathrm{d}t}\int g(x)\rho_\infty(x)\mathrm{d}x = 0$$

Hence by application of the dominated convergence theorem (DCT) and interchanging the expectation and the derivative,

$$\mathsf{E}[\frac{\mathrm{d}}{\mathrm{d}t}g(\nabla\psi_t(X_t))] = \mathsf{E}[\nabla g(\nabla\psi_t(X_t)) \cdot \xi(\nabla\psi_t(X_t))] = 0 \tag{5.28}$$

Letting $g(x) = \psi^*(x) - e^{-\gamma_t} \int_{-\infty}^{x} \nabla\phi_t(\nabla\psi^\star(z))\mathrm{d}z - \frac{1}{2}|x|^2$ where ϕ_t is defined in part-(ii) of the theorem 5.1 concludes

$$\begin{aligned} 0 = \mathsf{E}[\nabla g(\nabla\psi_t(X_t)) \cdot \xi(\nabla\psi_t(X_t)))] &= \mathsf{E}[X_t - e^{-\gamma_t}\nabla\phi_t(X_t) - \nabla\psi_t(X_t)) \cdot \xi(\nabla\psi_t(X_t)))] \\ &= \mathsf{E}[X_t - e^{-\gamma_t} Y_t - \nabla\psi_t(X_t)) \cdot \xi(\nabla\psi_t(X_t)))] \end{aligned}$$

where we used $Y_t = \nabla\phi_t(X_t)$ from part-(iii) of Theorem 5.1. This concludes the proof of the claim. Note that the application of DCT in (5.28) follows from smoothness of $g(x)$ and assuming $T_{\rho_t}^{\rho_\infty}(x)$ is differentiable with respect to time. Showing $T_{\rho_t}^{\rho_\infty}(x)$ is differentiable with respect to time is technical out of the scope of this work.

□

5.5.2 Wasserstein gradient and Gâteaux derivative

This section contains definitions of the Wasserstein gradient and Gâteaux derivative [Ambrosio et al., 2008, Carmona and Delarue, 2017].

Let $\mathsf{F} : \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d) \to \mathbb{R}$ be a (smooth) functional on the space of probability distributions.

Gâteaux derivative: The Gâteaux derivative of F at $\rho \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ is a real-valued function on $\mathbb{R}^d$ denoted as $\frac{\partial \mathsf{F}}{\partial \rho}(\rho) : \mathbb{R}^d \to \mathbb{R}$. It is defined as a function that satisfies the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{F}(\rho_t)\bigg|_{t=0} = \int_{\mathbb{R}^d} \frac{\partial \mathsf{F}}{\partial \rho}(\rho)(x)(-\nabla\cdot(\rho(x)u(x)))\mathrm{d}x$$

for all path ρ_t in $\mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ such that $\frac{\partial \rho_t}{\partial t} = -\nabla\cdot(\rho_t u)$ with $\rho_0 = \rho \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$.

Wasserstein gradient: The Wasserstein gradient of F at ρ is a vector-field on $\mathbb{R}^d$ denoted as $\nabla_\rho \mathsf{F}(\rho) : \mathbb{R}^d \to \mathbb{R}^d$. It is defined as a vector-field that satisfies the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{F}(\rho_t)\bigg|_{t=0} = \int_{\mathbb{R}^d} \nabla_\rho \mathsf{F}(\rho)(x)\cdot u(x)\ \rho(x)\mathrm{d}x$$

for all path ρ_t in $\mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$ such that $\frac{\partial \rho_t}{\partial t} = -\nabla\cdot(\rho_t u)$ with $\rho_0 = \rho \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$.

The two definitions imply the following relationship [Carmona and Delarue, 2017, Prop. 5.48]:

$$\nabla_\rho \mathsf{F}(\rho)(\cdot) = \nabla_x \frac{\partial \mathsf{F}}{\partial \rho}(\rho)(\cdot)$$

Example: Let $\mathsf{F}(\rho) = \int \log(\frac{\rho(x)}{\rho_\infty(x)})\rho(x)\mathrm{d}x$ be the relative entropy functional. Consider a path ρ_t in $\mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$

such that $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t u)$ with $\rho_0 = \rho \in \mathscr{P}_{\mathrm{ac},2}(\mathbb{R}^d)$. Then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{F}(\rho_t) &= \int \log(\frac{\rho_t(x)}{\rho_\infty(x)})\frac{\partial \rho_t}{\partial t}(x)\mathrm{d}x + \int \frac{\partial \rho_t}{\partial t}(x)\mathrm{d}x \\
&= -\int \log(\frac{\rho_t(x)}{\rho_\infty(x)})\nabla \cdot (\rho_t(x)u(x))\mathrm{d}x \\
&= \int \nabla_x \log(\frac{\rho_t(x)}{\rho_\infty(x)}) \cdot u(x)\ \rho_t(x)\mathrm{d}x
\end{aligned}$$

where the divergence theorem is used in the last step. The definitions of the Gâteaux derivative and Wasserstein gradient imply

$$\begin{aligned}
\frac{\partial \mathsf{F}}{\partial \rho}(\rho)(x) &= \log(\frac{\rho(x)}{\rho_\infty(x)}) \\
\nabla_\rho \mathsf{F}(\rho)(x) &= \nabla_x \log(\frac{\rho(x)}{\rho_\infty(x)})
\end{aligned}$$

References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

P. M. Anselone. *Collectively compact operator approximation theory and applications to integral equations*. Prentice Hall, 1971.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

K. Atkinson. A survey of numerical methods for the solution of fredholm integral equations of the second kind. 1976.

A. Bain and D. Crisan. *Fundamentals of stochastic filtering*, volume 3. Springer, 2009.

C. T. Baker. The numerical treatment of integral equations. 1977.

D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab*, 13:60–66, 2008.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

M. Belkin. Problems of learning on manifolds. 2003.

M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems*, pages 129–136, 2007.

T. Bengtsson, P. Bickel, and B. Li. Curse of dimensionality revisited: Collapse of the particle filter in very large scale systems. In *IMS Lecture Notes - Monograph Series in Probability and Statistics: Essays in Honor of David F. Freedman*, volume 2, pages 316–334. Institute of Mathematical Sciences, 2008.

A. Bensoussan, J. Frehse, P. Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

K. Bergemann and S. Reich. An ensemble Kalman-Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 21(3):213–219, 2012.

K. Berntorp. Feedback particle filter: Application and evaluation. In *18th Int. Conf. Information Fusion, Washington, DC*, 2015.

K. Berntorp and P. Grover. Data-driven gain computation in the feedback particle filter. In *2016 American Control Conference (ACC)*, pages 2711–2716, 2016.

A. Beskos, D. Crisan, A. Jasra, and N. Whiteley. Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions. *Advances in Applied Probability*, 46(1):279–306, 2014.

M. Betancourt, M. I. Jordan, and A. C. Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

A. N. Bishop and P. Del Moral. On the stability of Kalman–Bucy diffusion processes. *SIAM Journal on Control and Optimization*, 55(6):4015–4047, 2017.

A. N. Bishop and P. Del Moral. On the stability of matrix-valued Riccati diffusions. *arXiv preprint arXiv:1808.00235*, 2018.

A. N. Bishop, P. Del Moral, K. Kamatani, and B. Remillard. On one-dimensional Riccati diffusions. *arXiv preprint arXiv:1711.10065*, 2017.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

A. Budhiraja, L. Chen, and C. Lee. A survey of numerical methods for nonlinear filtering problems. *Physica D: Nonlinear Phenomena*, 230(1):27–36, 2007.

O. Cappé, E. Moulines, and T. Rydén. Inference in hidden Markov models. In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.

R. Carmona and F. Delarue. Probabilistic theory of mean field games with applications, 2017.

A. Chaudhuri, D. Kakde, C. Sadek, L. Gonzalez, and S. Kong. The mean and median criteria for kernel bandwidth selection for support vector data description. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 842–849. IEEE, 2017.

C. Chen, R. Zhang, W. Wang, B. Li, and L. Chen. A unified particle-optimization framework for scalable bayesian sampling. *arXiv preprint arXiv:1805.11659*, 2018.

N. Chen and A. Majda. Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. *Entropy*, 20(7):509, 2018.

Y. Chen, T. Georgiou, and M. Pavon. Optimal steering of a linear stochastic system to a final probability distribution, part I. *IEEE Trans. Autom. Control*, 61(5):1158–1169, 2016.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

Y. Cheng and S. Reich. A McKean optimal transportation perspective on Feynman-Kac formulae with application to data assimilation. *arXiv preprint arXiv:1311.6300*, 2013.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

D. Crisan and J. Xiong. Numerical solutions for a class of SPDEs over bounded domains. *ESAIM: Proc.*, 19:121–125, 2007.

D. Crisan and J. Xiong. Approximate McKean-Vlasov representations for a class of SPDEs. *Stochastics*, 82(1):53–68, 2010.

F. Daum, J. Huang, and A. Noushin. Exact particle flow for nonlinear filters. In *SPIE Defense, Security, and Sensing*, pages 769704–769704, 2010.

F. Daum, J. Huang, and A. Noushin. Generalized Gromov method for stochastic particle flow filters. In *SPIE Defense+ Security*, pages 102000I–102000I. International Society for Optics and Photonics, 2017.

J. de Wiljes, S. Reich, and W. Stannat. Long-time stability and accuracy of the ensemble Kalman-Bucy filter for fully observed processes and small measurement noise. *arXiv preprint arXiv:1612.06065*, 2016.

P. Del Moral. Feynman-Kac formulae. In *Feynman-Kac Formulae*, pages 47–93. Springer, 2004.

P. Del Moral and A. Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. In *Annales de l'IHP Probabilités et statistiques*, volume 37, pages 155–194, 2001.

P. Del Moral and J. Tugaut. On the stability and the uniform propagation of chaos properties of ensemble Kalman-Bucy filters. *arXiv preprint arXiv:1605.09329*, 2016.

P. Del Moral, A. Kurtzmann, and J. Tugaut. On the stability and the uniform propagation of chaos of a class of extended ensemble Kalman–Bucy filters. *SIAM Journal on Control and Optimization*, 55(1):119–155, 2017.

P. Del Moral et al. Measure-valued processes and interacting particle systems. application to nonlinear filtering problems. *The Annals of Applied Probability*, 8(2):438–495, 1998.

A. M. Doucet, A.and Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704, 2009.

A. Durmus and E. Moulines. High-dimensional bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

L. C. Evans. Partial differential equations and Monge-Kantorovich mass transfer. *Current developments in mathematics*, pages 65–126, 1997.

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.

G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

C. Frogner and T. Poggio. Approximate inference with wasserstein gradient flows. *arXiv preprint arXiv:1806.04542*, 2018.

E. Giné, V. Koltchinskii, et al. Empirical graph Laplacian approximation of Laplace–Beltrami operators: Large sample results. In *High dimensional probability*, pages 238–259. Institute of Mathematical Statistics, 2006.

C. R. Givens, R. M. Shortt, et al. A class of Wasserstein metrics for probability distributions. *Michigan Math. J*, 31(2), 1984.

P. W. Glynn and S. P. Meyn. A Liapunov bound for solutions of the Poisson equation. *The Annals of Probability*, pages 916–931, 1996.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

N. Gordon, B. Ristic, and S. Arulampalam. Beyond the Kalman filter: Particle filters for tracking applications. *Artech House, London*, 830:5, 2004.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113, 1993.

F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund. Particle filters for positioning, navigation, and tracking. *IEEE Transactions on signal processing*, 50(2):425–437, 2002.

M. Hein, J. Audibert, and U. Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. In *Learning theory*, pages 470–485. Springer, 2005.

M. Hein, J. Audibert, and U. Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(Jun):1325–1368, 2007.

J. Heng, A. Doucet, and Y. Pokern. Gibbs flow for approximate transport with applications to Bayesian computation. *arXiv preprint arXiv:1509.08787*, 2015.

V. Hutson, J. Pym, and M. Cloud. *Applications of functional analysis and operator theory*, volume 200. Elsevier, 2005.

M. Isard and A. Blake. Condensation—conditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28, 1998.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent. *arXiv preprint arXiv:1704.08227*, 2017.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1):95–108, 1961.

E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2002.

D. Kelly, K. J. Law, and A. M. Stuart. Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27(10):2579, 2014.

S. Kim, R. Ma, D. Mesa, and T. P. Coleman. Efficient Bayesian inference methods via convex optimization and optimal transport. In *2013 IEEE International Symposium on Information Theory*, pages 2259–2263. IEEE, 2013.

I. Kontoyiannis, S. P. Meyn, et al. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab*, 10(3):61–123, 2005.

E. Kwiatkowski and J. Mandel. Convergence of the square root ensemble Kalman filter in the large ensemble limit. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1–17, 2015.

R. S. Laugesen, P. G. Mehta, S. P. Meyn, and M. Raginsky. Poisson's equation in nonlinear filtering. *SIAM Journal on Control and Optimization*, 53(1):501–525, 2015.

K. Law, A. Stuart, and K. Zygalakis. *Data assimilation: a mathematical introduction*, volume 62. Springer, 2015.

F. Le Gland, V. Monbet, and V. Tran. *Large sample asymptotics for the ensemble Kalman filter*. PhD thesis, INRIA, 2009.

C. Liu, J. Zhuo, P. Cheng, R. Zhang, J. Zhu, and L. Carin. Accelerated first-order methods on the Wasserstein space for Bayesian inference. *arXiv preprint arXiv:1807.01750*, 2018.

Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

Y. Liu, F. Shang, J. Cheng, H. Cheng, and L. Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4868–4877, 2017.

R. Ma and T. P. Coleman. Generalizing the posterior matching scheme to higher dimensions via optimal transportation. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 96–102. IEEE, 2011.

J. Mandel, L. Cobb, and J. D. Beezley. On the convergence of the ensemble Kalman filter. *Applications of Mathematics*, 56(6):533–541, 2011.

Y. Matsuura, R. Ohata, K. Nakakuki, and R. Hirokawa. Suboptimal gain functions of feedback particle filter derived from continuation method. In *AIAA Guidance, Navigation, and Control Conference*, page 1620, 2016.

R. J. McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

H. P. McKean. A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966.

S. Meyn. *Control techniques for complex networks*. Cambridge University Press, 2008.

S. Meyn and R. Tweedie. Markov chains and stochastic stability, cambridge, 2009.

M. Montemerlo, S. Thrun, D. Koller, B. Wegbreit, et al. Fastslam: A factored solution to the simultaneous localization and mapping problem. *Aaai/iaai*, 593598, 2002.

P. Moral. Feynman-kac formulae: Genealogical and interacting particle systems with applications, probability and its applications, 2004.

R. M. Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

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.

D. Oliver, A. Reynolds, and N. Liu. *Inverse theory for petroleum reservoir characterization and history matching*. Cambridge University Press, Cambridge, 2008.

A. Radhakrishnan, A. Devraj, and S. Meyn. Learning techniques for feedback particle filter design. In *Conference on Decision and COntrol (CDC), 2016*, pages 648–653. IEEE, 2014.

P. Rebeschini and R. Van Handel. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015.

S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 2011.

S. Reich. Data assimilation-the Schrödinger perspective. *arXiv preprint arXiv:1807.08351*, 2018.

S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

P. H. Richemond and B. Maginnis. On Wasserstein reinforcement learning and the fokker-planck equation. *arXiv preprint arXiv:1712.07185*, 2017.

B. Ristic, S. Arulampalam, and N. Gordon. Beyond the Kalman filter. *IEEE Aerospace and Electronic Systems Magazine*, 19(7):37–38, 2004.

G. Roberts and J. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

P. M. Stano. *Nonlinear State and Parameter Estimation for Hopper Dredgers*. PhD thesis, Ph. D. dissertation). Delft University of Technology, 2013.

P. M. Stano, A. K. Tilton, and R. Babuska. Estimation of the soil-dependent time-varying parameters of the hopper sedimentation model: The FPF versus the BPF. *Control Engineering Practice*, 24:67–78, 2014.

D. W. Stroock. *An introduction to Markov processes*, volume 230. Springer Science & Business Media, 2013.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

S. C. Surace, A. Kutschireiter, and J.-P. Pfister. How to avoid the curse of dimensionality: scalability of particle filters with and without importance weights. *ArXiv e-prints*, Mar. 2017.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

A.-S. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

A. Taghvaei and P. Mehta. Accelerated flow for probability distributions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6076–6085. PMLR, 2019. URL http://proceedings.mlr.press/v97/taghvaei19a.html.

A. Taghvaei and P. G. Mehta. An optimal transport formulation of the linear feedback particle filter. In *American Control Conference (ACC), 2016*, pages 3614–3619. IEEE, 2016a.

A. Taghvaei and P. G. Mehta. Gain function approximation in the feedback particle filter. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 5446–5452. IEEE, 2016b.

A. Taghvaei and P. G. Mehta. Error analysis for the linear feedback particle filter. In *2018 Annual American Control Conference (ACC)*, pages 4261–4266. IEEE, 2018a.

A. Taghvaei and P. G. Mehta. Error analysis of the stochastic linear feedback particle filter. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 7194–7199. IEEE, 2018b.

A. Taghvaei, P. G. Mehta, and S. P. Meyn. Error estimates for the kernel gain function approximation in the feedback particle filter. In *American Control Conference (ACC), 2017*, pages 4576–4582. IEEE, 2017.

A. Taghvaei, J. De Wiljes, P. G. Mehta, and S. Reich. Kalman filter and its modern extensions for the continuous-time nonlinear filtering problem. *Journal of Dynamic Systems, Measurement, and Control*, 140(3):030904, 2018.

A. K. Tilton, S. Ghiotto, and P. G. Mehta. A comparative study of nonlinear filtering techniques. In *Proc. 16th Int. Conf. on Inf. Fusion*, pages 1827–1834, Istanbul, Turkey, July 2013.

D. Ting, L. Huang, and M. Jordan. An analysis of the convergence of graph Laplacians. *arXiv preprint arXiv:1101.5435*, 2011.

X. T. Tong, A. J. Majda, and D. Kelly. Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity*, 29(2):657, 2016.

C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

U. Von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

J. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, 2002.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, page 201614734, 2016.

J. Xiong. *An introduction to stochastic filtering theory*, volume 18 of *Oxford Graduate Texts in Mathematics*. Oxford University Press, 2008.

T. Yang, P. G. Mehta, and S. P. Meyn. Feedback particle filter. *IEEE transactions on Automatic control*, 58 (10):2465–2480, 2013.

T. Yang, R. S. Laugesen, P. G. Mehta, and S. P. Meyn. Multivariable feedback particle filter. *Automatica*, 71:10–23, 2016.

C. Zhang, A. Taghvaei, and P. G. Mehta. A mean-field optimal control formulation for global optimization. *IEEE Transactions on Automatic Control*, 64(1):279–286, 2019.

R. Zhang, C. Chen, C. Li, and L. Carin. Policy optimization as Wasserstein gradient flows. *arXiv preprint arXiv:1808.03030*, 2018.