On the use of transport and optimal control methods for Monte Carlo simulation

Jeremy Heng

St Cross College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2016

Abstract

This thesis explores ideas from transport theory and optimal control to develop novel Monte Carlo methods to perform efficient statistical computation.

The first project considers the problem of constructing a transport map between two given probability measures. In the Bayesian formalism, this approach is natural when one introduces a curve of probability measures connecting the prior to posterior by tempering the likelihood function. The main idea is to move samples from the prior using an ordinary differential equation (ODE), constructed by solving the Liouville partial differential equation (PDE) which governs the time evolution of measures along the curve. In this work, we first study the regularity solutions of Liouville equation should satisfy to guarantee validity of this construction. We place an emphasis on understanding these issues as it explains the difficulties associated with solutions that have been previously reported. After ensuring that the flow transport problem is well-defined, we give a constructive solution. However, this result is only formal as the representation is given in terms of integrals which are intractable. For computational tractability, we proposed a novel approximation of the PDE which yields an ODE whose drift depends on the full conditional distributions of the intermediate distributions. Even when the ODE is time-discretized and the full conditional distributions are approximated numerically, the resulting distribution of mapped samples can be evaluated and used as a proposal within Markov chain Monte Carlo and sequential Monte Carlo (SMC) schemes. We then illustrate experimentally that the resulting algorithm can outperform state-of-the-art SMC methods at a fixed computational complexity.

The second project aims to exploit ideas from optimal control to design more efficient SMC methods. The key idea is to control the proposal distribution induced by a time-discretized Langevin dynamics so as to minimize the Kullback-Leibler divergence of the extended target distribution from the proposal. The optimal value functions of the resulting optimal control problem can then be approximated using algorithms developed in the approximate dynamic programming (ADP) literature. We introduce a novel iterative scheme to perform ADP, provide a theoretical analysis of the proposed algorithm and demonstrate that the latter can provide significant gains over state-of-the-art methods at a fixed computational complexity.

On the use of transport and optimal control methods for Monte Carlo simulation

Jeremy Heng

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2016

Acknowledgements

I am enormously indebted to my supervisor Prof Arnaud Doucet for his unfailing support, guidance and encouragement in the past years, without which, this thesis would not have been possible. Next, sincere thanks are due to my collaborators Dr Adrian Bishop, Dr George Deligiannidis and Dr Yvo Pokern for many stimulating discussions that have benefited my research in many ways. I would like to extend my gratitude to the Clarendon fund for their generous financial support throughout my studies. Thanks also go to my uncle Jeffrey McAskill and aunty Theresa McAskill for supporting all my endeavours over the years. Lastly, I conclude by expressing my appreciation to the two most important women in my life: my mother Cindy Tan for her unconditional love and my girlfriend Si Yun Ng for completing my life.

Abstract

This thesis explores ideas from transport theory and optimal control to develop novel Monte Carlo methods to perform efficient statistical computation.

The first project considers the problem of constructing a transport map between two given probability measures. In the Bayesian formalism, this approach is natural when one introduces a curve of probability measures connecting the prior to posterior by tempering the likelihood function. The main idea is to move samples from the prior using an ordinary differential equation (ODE), constructed by solving the Liouville partial differential equation (PDE) which governs the time evolution of measures along the curve. In this work, we first study the regularity solutions of Liouville equation should satisfy to guarantee validity of this construction. We place an emphasis on understanding these issues as it explains the difficulties associated with solutions that have been previously reported. After ensuring that the flow transport problem is well-defined, we give a constructive solution. However, this result is only formal as the representation is given in terms of integrals which are intractable. For computational tractability, we proposed a novel approximation of the PDE which yields an ODE whose drift depends on the full conditional distributions of the intermediate distributions. Even when the ODE is time-discretized and the full conditional distributions are approximated numerically, the resulting distribution of mapped samples can be evaluated and used as a proposal within Markov chain Monte Carlo and sequential Monte Carlo (SMC) schemes. We then illustrate experimentally that the resulting algorithm can outperform state-of-the-art SMC methods at a fixed computational complexity.

The second project aims to exploit ideas from optimal control to design more efficient SMC methods. The key idea is to control the proposal distribution induced by a time-discretized Langevin dynamics so as to minimize the Kullback-Leibler divergence of the extended target distribution from the proposal. The optimal value functions of the resulting optimal control problem can then be approximated using algorithms developed in the approximate dynamic programming (ADP) literature. We introduce a novel iterative scheme to perform ADP, provide a theoretical analysis of the proposed algorithm and demonstrate that the latter can provide significant gains over state-of-the-art methods at a fixed computational complexity.

Contents

List of Figures

Notation

Most notation except those that are very commonly used will be introduced in the main text. Reference to this list will only be necessary in the case of any ambiguity.

Sets and numbers

The set of natural numbers excluding and including zero are denoted by $\mathbb{N} := \{1, 2, \ldots\}$ and $\mathbb{N}_0 := \{0, 1, \ldots\}$ respectively. The set of real numbers and positive real numbers are written as $\mathbb{R}$ and $\mathbb{R}_+$ respectively. The set of extended real numbers is defined as $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. Let $\lfloor x \rfloor$ denote the largest integer less than or equal to $x \in \mathbb{R}$. We denote the d-dimensional Euclidean space by $\mathbb{R}^d$, equipped with the inner product $\langle x, y \rangle := \sum_{i=1}^d x_i y_i$ for any $x, y \in \mathbb{R}^d$ and its induced norm $|x| := \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^d x_i^2}$ for any $x \in \mathbb{R}^d$. The notation $\cup$ and $\cap$ denote union and intersection of sets respectively. For subsets $A, B \subseteq \Omega$, we denote the relative complement of A in B by $B \setminus A$ and the complement of A by $A^c := \Omega \setminus A$.

Matrices and vectors

For any matrix A, A^T denotes its transpose. For any real square matrix A, we denote its trace by $\mathrm{Tr}(A)$, its determinant by $\det(A)$, its minimum eigenvalue by $\lambda_{\min}(A)$, its spectral matrix norm by $\|A\|_2$, its Frobenius norm by $\|A\|_F := \sqrt{\langle A, A \rangle_F}$, induced by the Frobenius inner product $\langle \cdot, \cdot \rangle_F$ and its inverse (if it exists) by A^{-1}. For any $d \in \mathbb{N}$, we write $0_d := (0, \ldots, 0)^T \in \mathbb{R}^d$ as the vector of zeros, $0_{d \times d}$ as the $d \times d$ matrix of zeros, $1_d := (1, \ldots, 1)^T \in \mathbb{R}^d$ as the vector of ones and I_d as the identity matrix of size d. For any $u, v \in \mathbb{R}^d$, we denote the inner product by $u^T v$ and the outer product by uv^T.

Functions

Given a subset $A \subseteq \Omega$, the indicator function is defined as $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 if $x \in A^c$. For any function $f : \Omega_1 \to \Omega_2$, let $f^{-1}(A) := \{x \in \Omega_1 : f(x) \in A\}$ denote the pre-image of $A \subseteq \Omega_2$ under f. For any function $\varphi : \Omega \to \mathbb{R}$, we define its oscillation as $\mathrm{osc}(\varphi) := \sup_{(x,y) \in \Omega \times \Omega} |\varphi(x) - \varphi(y)|$.

Function spaces

Let $p \in [1, \infty)$, $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\mathcal{L}^p(\mu)$ be the set of measurable functions $\varphi : \Omega \to \mathbb{R}^d$ such that $\|\varphi\|_{L^p(\mu)} := (\int_\Omega |\varphi(x)|^p \mu(\mathrm{d}x))^{1/p} < \infty$. We then define $L^p(\mu)$ as the set of classes of μ-equivalent functions in $\mathcal{L}^p(\mu)$. For any function $\varphi : \Omega \to \mathbb{R}^d$, we define its supremum norm as $\|\varphi\|_\infty := \sup_{x \in \Omega} |\varphi(x)|$ (with $\|\varphi\|_\infty := \infty$ if φ is unbounded) and denote $L^\infty(\Omega)$ as the set of all bounded functions on Ω.

Let $\Omega_1 \subseteq \mathbb{R}^{d_1}$ and $\Omega_2 \subseteq \mathbb{R}^{d_2}$ for some $d_1, d_2 \in \mathbb{N}$. We denote $C(\Omega_1, \Omega_2)$ as the set of continuous functions from Ω_1 to Ω_2 and $C^k(\Omega_1, \Omega_2)$ as the subset of k-times continuously differentiable functions for $k \in \mathbb{N} \cup \{0, \infty\}$, with the case $k = 0$ to mean $C^0(\Omega_1, \Omega_2) := C(\Omega_1, \Omega_2)$ and the case $k = \infty$ defined as $C^\infty(\Omega_1, \Omega_2) := \cap_{k \in \mathbb{N}_0} C^k(\Omega_1, \Omega_2)$.

Measures and σ-algebras

Given a measurable space $(\Omega, \mathcal{F})$, let $\mathcal{P}(\Omega)$ denote the set of all probability measures on this space. If Ω is a topological space, we write $\mathcal{B}(\Omega)$ as its corresponding Borel σ-algebra. Given two probability measures $\mu, \nu \in \mathcal{P}(\Omega)$, we use the notation $\mu \ll \nu$ if μ is absolutely continuous with respect to ν, i.e. $\nu(A) = 0$ implies $\mu(A) = 0$ for all $A \in \mathcal{F}$, and denote $\mathrm{d}\mu/\mathrm{d}\nu$ as the corresponding Radon-Nikodym derivative. For any $x \in \Omega$, δ_x refers to the Dirac measure at x. Given two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, we denote the product σ-algebra on $\Omega_1 \times \Omega_2$ as $\mathcal{F}_1 \otimes \mathcal{F}_2$. For any $\mu \in \mathcal{P}(\Omega)$ and measurable function $f : \Omega \to \Omega$, we define the push-forward of μ as the measure defined by $(f_\# \mu)(A) := \mu(f^{-1}(A))$ for all $A \in \mathcal{F}$.

Probability

All random variables will be written in capital letters and assumed to be supported on some underlying common probability space with probability measure $\mathbb{P}$. We will write $\mathbb{E}$ to denote expectation with respect to $\mathbb{P}$. The variance of a random variable X will be denoted as $\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. Given a probability measure $\mu \in \mathcal{P}(\Omega)$ and a Ω-valued random variable X, we will use the shorthand $X \sim \mu$ to mean $\mathbb{P}(X \in A) = \mu(A)$ for all $A \in \mathcal{F}$. We will denote the Gaussian distribution on $\mathbb{R}^d$ with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ by $\mathcal{N}(\mu, \Sigma)$ and its density (with respect to Lebesgue measure) by $x \mapsto \mathcal{N}(x; \mu, \Sigma)$.

Divergences and probability metrics

Given $\mu, \nu \in \mathcal{P}(\Omega)$, we define the Kullback-Leibler divergence of ν from μ as $\mathrm{KL}(\mu|\nu) = \int_\Omega \log(\mathrm{d}\mu/\mathrm{d}\nu)\, \mathrm{d}\mu$ if $\mu \ll \nu$ and $\int_\Omega \log(\mathrm{d}\mu/\mathrm{d}\nu)\, \mathrm{d}\mu < \infty$, and $\mathrm{KL}(\mu|\nu) = \infty$ otherwise.

For any finite signed measure μ on $(\Omega, \mathcal{F})$, we define its total variation norm as $\|\mu\|_{\mathrm{TV}} := \mu^+(\Omega) + \mu^-(\Omega)$ where the pair (μ^+, μ^-) of positive finite measures refers to the Jordan decomposition of μ. Let d_{TV} be the total variation distance induced by the total variation norm on the vector space of finite signed measures on $(\Omega, \mathcal{F})$, i.e. $d_{\mathrm{TV}}(\mu, \nu) := \frac{1}{2}\|\mu - \nu\|_{\mathrm{TV}}$ for $\mu, \nu \in \mathcal{P}(\Omega)$.

Operators

A Markov transition kernel on $(\Omega, \mathcal{F})$ is a map $K : \Omega \times \mathcal{F} \to [0, 1]$ with the following properties: (i) for any fixed $x \in \Omega$, $K(x, \cdot)$ is a probability measure on $(\Omega, \mathcal{F})$; (ii) for any fixed $A \in \mathcal{F}$, $K(\cdot, A)$ is a measurable function.

Given $\mu \in \mathcal{P}(\Omega)$, Markov transition kernel K on $(\Omega, \mathcal{F})$, measurable functions $\varphi : \Omega \to \mathbb{R}$ and $\xi : \Omega \times \Omega \to \mathbb{R}$, we define the integral $\mu(\varphi) := \int_\Omega \varphi(x)\mu(\mathrm{d}x)$, the probability measure $(\mu K)(\cdot) := \int_\Omega \mu(\mathrm{d}x)K(x, \cdot)$ on $(\Omega, \mathcal{F})$ and the measurable functions $(K\varphi)(\cdot) := \int_\Omega \varphi(u)K(\cdot, \mathrm{d}u)$, $(K\xi)(\cdot) := \int_\Omega \xi(\cdot, u)K(\cdot, \mathrm{d}u)$. Given two Markov kernels M and K on $(\Omega, \mathcal{F})$, the composition is the Markov kernel defined by $(M \circ K)(x, \cdot) := \int_\Omega M(x, \mathrm{d}y)K(y, \cdot)$.

We denote the partial derivative with respect to time variable t and i^{th}-coordinate spatial variable x_i by ∂_t and ∂_{x_i} respectively. For $\varphi \in C^1(\mathbb{R}^d, \mathbb{R})$, we define its gradient $\nabla\varphi : \mathbb{R}^d \to \mathbb{R}^d$ by $\nabla\varphi(x) := (\partial_{x_1}\varphi(x), \ldots, \partial_{x_d}\varphi(x))^T$ and for $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$, the $(i, j)^{th}$ element of its Hessian matrix $\nabla^2\varphi : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ by $(\nabla^2\varphi(x))_{i,j} := \partial_{x_i}\partial_{x_j}\varphi(x)$. For any vector-valued function $\varphi = (\varphi_1, \ldots, \varphi_d)^T$, we define the divergence operator as $\nabla \cdot \varphi(x) := \sum_{i=1}^d \partial_{x_i}\varphi_i(x)$ for $\varphi \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ and the Laplacian operator as $\Delta\varphi(x) := \sum_{i=1}^d \partial^2_{x_i}\varphi_i(x)$ for $\varphi \in C^2(\mathbb{R}^d, \mathbb{R}^d)$. For $\varphi \in C^1(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$, the $(i, j)^{th}$ element of its Jacobian matrix $\nabla\varphi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2\times d_1}$ is defined as $(\nabla\varphi(x))_{i,j} := \partial_{x_j}\varphi_i(x)$.

Asymptotics

We will write $\varphi(n) = O(\xi(n))$ if there exists $c \in \mathbb{R}_+, n_0 \in \mathbb{N}$ such that $|\varphi(n)| \leq c|\xi(n)|$ for all $n \geq n_0$ and $\varphi(n) = o(\xi(n))$ if $\xi(n) \neq 0$ for all $n \in \mathbb{N}$ and $\varphi(n)/\xi(n) \to 0$ as $n \to \infty$. For a sequence of random variables $\{\varphi_n\}_{n\in\mathbb{N}}$ and non-zero constants $\{\xi_n\}_{n\in\mathbb{N}}$, the following are stochastic counterparts of the above statements. We denote $\varphi_n = O_P(\xi_n)$ if for any $\varepsilon > 0$, there exists $c \in \mathbb{R}_+$ such that $\mathbb{P}(|\varphi_n/\xi_n| > c) < \varepsilon$ for all $n \in \mathbb{N}$ and $\varphi_n = o_P(\xi_n)$ if $\varphi_n/\xi_n \to 0$ in probability as $n \to \infty$. Lastly, we will often write $X_n \xrightarrow{d.} \mathcal{N}(\mu, \Sigma)$ as $n \to \infty$ to mean that the sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ converges in distribution to a random variable $X \sim \mathcal{N}(\mu, \Sigma)$.

1 Introduction

Contents

In Section 1.1, we begin by briefly describing Bayesian inference and discuss the necessity for sampling algorithms. In Section 1.2, we then give some preliminary background on state-of-the-art Monte Carlo methods so as to motivate the need for more efficient algorithms.

1.1 Bayesian inference

Let $x \in \mathcal{X} \subseteq \mathbb{R}^d$ denote the unknown parameters of a given statistical model and $\mathcal{B}(\mathcal{X})$ be the corresponding Borel σ-algebra on the parameter space $\mathcal{X}$. Although we will restrict our attention to the Euclidean space, we note that the methods presented in this chapter also apply in more general state spaces. Compared to other paradigms, the distinctive feature of Bayesian inference is the ability to incorporate prior knowledge by prescribing a prior distribution $\pi_0 \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ refers to the set of all probability measures on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. To simplify presentation, we assume that π_0 admits a positive density $\pi_0(x), x \in \mathcal{X}$ with respect to a σ-finite dominating measure $\mathrm{d}x$.[1]

Given a data set $y \in \mathcal{Y}$, the statistical model defines the likelihood function $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. In Bayesian inference, the main object of interest is the posterior distribution, i.e. the conditional distribution of the parameters given the data, obtained using Bayes' theorem

$$\pi(\mathrm{d}x|y) = \frac{\pi_0(\mathrm{d}x)L(x;y)}{Z(y)}, \tag{1.1}$$

where $Z(y) := \int_{\mathcal{X}} \pi_0(\mathrm{d}x)L(x;y) < \infty$ denotes the marginal likelihood of the data y. In this context, the normalizing constant $Z(y)$ is also typically referred to as the model evidence for its role in model selection.

Depending on the application of interest, we note that many desired quantities can be expressed as

$$\int_{\mathcal{X}} \varphi(x)\pi(\mathrm{d}x|y) \tag{1.2}$$

for some test function $\varphi : \mathcal{X} \to \mathbb{R}$. For parameter estimation, we can set $\varphi(x) = x_i$ to obtain the posterior mean, and for posterior probabilities, set $\varphi(x) = \mathbb{1}_A(x)$ where $\mathbb{1}_A$ denotes the indicator function on the subset $A \in \mathcal{B}(\mathcal{X})$ of interest. Lastly, Bayesian prediction can also be recovered by selecting φ as the conditional density of a new observation given the parameters and observed data.

[1] For most of our purposes, it will suffice to take the Lebesgue measure

Apart from simple models, for example when (1.1) admits conjugacy, the integrals in (1.2) and normalizing constant $Z(y)$ would be intractable so one has to resort to numerical approximations. We will see in Section 1.2 that the form of (1.2) renders it amenable to Monte Carlo approximations, which in turn rely on our ability to generate (possibly approximate) samples from the posterior. In this sense, $\pi \in \mathcal{P}(\mathcal{X})$ will be thought of as the target distribution that we would like to simulate from. For notational simplicity, we will henceforth suppress dependency on the observations $y \in \mathcal{Y}$ and write the target distribution as $\pi(\mathrm{d}x) = \gamma(x)Z^{-1}\mathrm{d}x$ where $\gamma(x) := \pi_0(x)L(x)$ denotes the unnormalized density.

1.2 Monte Carlo methods

We first present the basic Monte Carlo method in Section 1.2.1 and move on to more sophisticated methods in subsequent sections.

1.2.1 Basic Monte Carlo

For a probability measure μ on a measurable space $(\Omega, \mathcal{F})$ and a test function $\varphi : \Omega \to \mathbb{R}$, we write $\mu(\varphi) := \int_\Omega \varphi(x)\mu(\mathrm{d}x)$ as the expectation of φ with respect to μ. In this notation, $\pi(\varphi)$ denotes the quantity of interest in (1.2). Assuming access to $N \in \mathbb{N}$ independent and identically distributed (iid) samples $\{X_n\}_{n=1}^N$ from π, a simple Monte Carlo estimator of $\pi(\varphi)$ is given by

$$\pi_N(\varphi) = \frac{1}{N}\sum_{n=1}^{N} \varphi(X_n). \tag{1.3}$$

Notationally, we can view this empirical average as integrating φ with respect to the random probability measure $\pi_N = N^{-1}\sum_{n=1}^N \delta_{X_n}$, where δ_x denotes a Dirac measure at $x \in \mathcal{X}$.

This estimator is unbiased and by the *law of large numbers* (LLN) strongly consistent if $\varphi \in L^1(\pi)$. Additionally, if $\varphi \in L^2(\pi)$, it also satisfies a *central limit theorem* (CLT):

$$\sqrt{N}\left(\pi_N(\varphi) - \pi(\varphi)\right) \xrightarrow{d.} \mathcal{N}\left(0, \sigma^2(\varphi)\right) \tag{1.4}$$

as $N \to \infty$, where $\sigma^2(\varphi) := \pi((\varphi - \pi(\varphi))^2)$ and $\xrightarrow{d.}$ denotes convergence in distribution. When compared to other numerical integration schemes, an interesting feature of the Monte Carlo estimator (1.3) is that its *convergence rate*, measured in terms of *root mean squared error* (RMSE), is $O(N^{-1/2})$ and does not depend on the dimension of the state space. That said, it is entirely possible for the error constant to grow exponentially with dimension.

Although the Monte Carlo approach reduces the integration problem to that of simulation, it is not straightforward to sample from any generic target distribution. In the following sections, we discuss several general strategies to tackle this issue.

1.2.2 Importance sampling

For measures μ and ν on a measurable space $(\Omega, \mathcal{F})$, we will use the notation $\mu \ll \nu$ if μ is absolutely continuous with respect to ν, i.e. $\nu(A) = 0$ implies $\mu(A) = 0$ for any $A \in \mathcal{F}$, and denote $\mathrm{d}\mu/\mathrm{d}\nu$ as the corresponding Radon-Nikodym derivative.

1.2.2.1 Normalized importance sampling

Suppose we can simulate from a distribution $q \in \mathcal{P}(\mathcal{X})$ satisfying $\pi \ll q$. Importance sampling is simply a change of measure which allows us to mitigate the difficulty of sampling from π by sampling instead from a proposal distribution q and correct for this discrepancy with the Radon-Nikodym derivative $\mathrm{d}\pi/\mathrm{d}q$. If q admits a density $q(x), x \in \mathcal{X}$ with respect to $\mathrm{d}x$, then we have

$$\frac{\mathrm{d}\pi}{\mathrm{d}q}(x) = \frac{w(x)}{Z}, \tag{1.5}$$

where $w(x) := \gamma(x)/q(x)$ denotes the (unnormalized) weight function. Using the identity

$$\pi(\varphi) = \frac{q(\varphi w)}{q(w)}, \tag{1.6}$$

which follows from (1.5), and the basic Monte Carlo estimator (1.3), we obtain the *normalized importance sampling* (NIS) estimator:

$$\hat{\pi}_N(\varphi) = \frac{\frac{1}{N}\sum_{n=1}^N \varphi(X_n) w(X_n)}{\frac{1}{N}\sum_{m=1}^N w(X_m)} = \sum_{n=1}^N W_n \varphi(X_n) \tag{1.7}$$

where $\{X_n\}_{n=1}^N$ are iid samples from q and $W_n := w(X_n)/\sum_{m=1}^N w(X_m)$ denotes the normalized weights. Note that the intractable normalizing constant Z is not present in the ratio (1.7). The NIS estimator is strongly consistent and as its bias is $O(N^{-1})$, the RMSE is dominated by its variance that is $O(N^{-1/2})$ (Geweke 1989). Re-writing (1.7) as

$$\sqrt{N}\left(\hat{\pi}_N(\varphi) - \pi(\varphi)\right) = \frac{\frac{1}{\sqrt{N}}\sum_{n=1}^N \left(\varphi(X_n) - \pi(\varphi)\right)\frac{\mathrm{d}\pi}{\mathrm{d}q}(X_n)}{\frac{1}{N}\sum_{m=1}^N \frac{\mathrm{d}\pi}{\mathrm{d}q}(X_m)}, \tag{1.8}$$

and noting that $N^{-1}\sum_{n=1}^N \mathrm{d}\pi/\mathrm{d}q(X_n) \to 1$ in probability and if $\hat{\sigma}^2(\varphi) := q((\varphi - \pi(\varphi))^2(\mathrm{d}\pi/\mathrm{d}q)^2) < \infty$ we have

$$\frac{1}{\sqrt{N}}\sum_{n=1}^N \left(\varphi(X_n) - \pi(\varphi)\right)\frac{\mathrm{d}\pi}{\mathrm{d}q}(X_n) \xrightarrow{d.} \mathcal{N}\left(0, \hat{\sigma}^2(\varphi)\right), \tag{1.9}$$

then an application of Slutsky's lemma gives

$$\sqrt{N}\left(\hat{\pi}_N(\varphi) - \pi(\varphi)\right) \xrightarrow{d.} \mathcal{N}\left(0, \hat{\sigma}^2(\varphi)\right). \tag{1.10}$$

Re-writing $\hat{\sigma}^2(\varphi) = \pi\left((\varphi - \pi(\varphi))^2\mathrm{d}\pi/\mathrm{d}q\right)$ to facilitate comparison with the asymptotic variance in (1.4), interestingly, we see that it is possible to obtain better performance than independent sampling if the proposal is chosen to suit a given test function. In practice, one might be interested in a large class of test functions, so as a guideline it is desirable to select a proposal distribution with heavier tails than the target distribution to ensure that the asymptotic variance $\hat{\sigma}^2(\varphi)$ is finite.

As before, we can view (1.7) as an expectation with respect to the weighted empirical measure $\hat{\pi}_N = \sum_{n=1}^N W_n \delta_{X_n}$, commonly referred to as the *particle approximation* of π. To quantify the difference between these two measures, we could use the metric introduced in Rebeschini and Van Handel (2015):

$$d(\hat{\pi}_N, \pi)^2 := \sup_{\|\varphi\|_\infty \leq 1} \mathbb{E}\left[\left(\hat{\pi}_N(\varphi) - \pi(\varphi)\right)^2\right], \tag{1.11}$$

where the supremum is taken over test functions satisfying $\|\varphi\|_\infty := \sup_{x\in\mathcal{X}} |\varphi(x)| \leq 1$ and the expectation is with respect to the random variables generating the random measure $\hat{\pi}_N$. Using (1.10), we have for sufficiently large enough N that

$$d(\hat{\pi}_N, \pi)^2 \leq \frac{\hat{\sigma}^2(\varphi)}{N} \leq \frac{4q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right)}{N}. \tag{1.12}$$

In fact, it can be shown that this inequality holds for any finite $N \in \mathbb{N}$ (see Agapiou et al. 2015; Theorem 2.1).

1.2.2.2 Normalizing constant estimation

Using the identity $Z = q(w)$, an unbiased estimator of Z is given by

$$\hat{Z}_N = \frac{1}{N} \sum_{n=1}^{N} w(X_n). \tag{1.13}$$

When compared to other normalizing constant estimation methods, *unbiasedness* makes importance sampling particularly appealing as this property can be further exploited to design algorithms to perform inference on models with intractable likelihoods (Andrieu and Roberts 2009, Andrieu et al. 2010). It is easy to see that $\hat{Z}_N$ is strongly consistent and satisfies a CLT

$$\sqrt{N}\left(\frac{\hat{Z}_N}{Z} - 1\right) \xrightarrow{d.} \mathcal{N}\left(0, q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right) - 1\right) \tag{1.14}$$

as $N \to \infty$ if $w \in L^2(q)$. Note also that when $q = \pi$, (1.13) produces a zero variance estimator, i.e. $\hat{Z}_N = Z$, q-almost surely for any $N \in \mathbb{N}$.

1.2.2.3 Computational complexity of importance sampling

It follows from the above discussion that the computational complexity of importance sampling, encompassing both estimation of bounded test functions (1.12) and normalizing constant estimation (1.14), is characterized by the ratio $q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right) N^{-1}$.

The quantity $q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right)$, which is at least one using an application of Jensen's inequality, can be written as $\chi^2(\pi|q) + 1$, where $\chi^2(\pi|q) := q((\mathrm{d}\pi/\mathrm{d}q - 1)^2)$ denotes the χ^2-divergence of the proposal from the target distribution. The relation to a divergence instead of a probability metric is unsurprising since importance sampling is an asymmetric procedure. It is clear that if π and q are close to being mutually singular, then the sample size required for accurate estimation has to be particularly large. Agapiou et al. (2015) studied various limits under which absolute continuity between the measures break down.

The relationship $\chi^2(\pi|q) + 1 \geq \exp(\mathrm{KL}(\pi|q))$ (Gibbs and Su 2002; Theorem 5) suggests that we should scale N exponentially with the *Kullback-Leibler* divergence

of the proposal from the target distribution defined as $\mathrm{KL}(\pi|q) := \pi(\log(\mathrm{d}\pi/\mathrm{d}q))$.[2] Interestingly, Chatterjee and Diaconis (2015) recently showed that this scaling is roughly[3] both sufficient and necessary to control the L^1-error of $\hat{\pi}_N(\varphi)$. The use of Kullback-Leibler divergence is central to this result as it relies on the concentration of $\log(\mathrm{d}\pi/\mathrm{d}q)(\cdot)$, under the measure π, around its expected value. We now illustrate that the latter phenomenon is typical of problems in high dimensions.

Consider the simple case where both the target and proposal distributions factorize into identical marginals, i.e. $\pi(\mathrm{d}x) = \prod_{i=1}^{d} \pi_1(\mathrm{d}x_i)$ and $q(\mathrm{d}x) = \prod_{i=1}^{d} q_1(\mathrm{d}x_i)$, in which case we have $q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right) N^{-1} = q_1\left((\mathrm{d}\pi_1/\mathrm{d}q_1)^2\right)^d N^{-1}$. Of course, as $d \to \infty$, this decline in performance can still happen slowly if $q_1\left((\mathrm{d}\pi_1/\mathrm{d}q_1)^2\right) \approx 1$; hence the need to distinguish between the dimension of the state space d and the intrinsic dimensionality of the importance sampling procedure (Agapiou et al. 2015). Now let $Y \sim \pi$ and Y_i denote the i^{th} component of the vector. Since $\log(\mathrm{d}\pi/\mathrm{d}q)(Y) = \sum_{i=1}^{d} \log(\mathrm{d}\pi_1/\mathrm{d}q_1)(Y_i)$ is given by a sum of iid random variables, assuming that $\sigma^2_{\mathrm{KL}} := \pi_1((\log(\mathrm{d}\pi_1/\mathrm{d}q_1) - \mathrm{KL}(\pi_1|q_1))^2) < \infty$, we have

$$\frac{1}{\sqrt{d}}\left(\log\left(\frac{\mathrm{d}\pi}{\mathrm{d}q}\right)(Y) - \mathrm{KL}(\pi|q)\right) \xrightarrow{d.} \mathcal{N}(0, \sigma^2_{\mathrm{KL}}) \tag{1.15}$$

as $d \to \infty$. This illustrates that in high dimensions, the above-mentioned concentration does take place, and that the fluctuations of $\log(\mathrm{d}\pi/\mathrm{d}q)(Y)$ around $\mathrm{KL}(\pi|q) = d\,\mathrm{KL}(\pi_1|q_1)$ are of order $\sqrt{d}$.

1.2.2.4 Effective sample size

Noting that $q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right)^{-1} N = N$ when $q = \pi$, a natural way to quantify the inefficiency of performing importance sampling relative to independent sampling is to define the notion of *effective sample size* (ESS) based on the ratio $q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right)^{-1} N$. In practice, one approximates the intractable denominator with the identity $q\left((\mathrm{d}\pi/\mathrm{d}q)^2\right) = q(w^2)/q(w)^2$ and the basic Monte Carlo estimator

[2] For completeness, we take $\mathrm{KL}(\pi|q) := \infty$ if π and q are mutually singular or if the integral is not finite.

[3] More precisely, this depends on the order of fluctuations around the expectation.

(1.3) which gives

$$\mathrm{ESS} := \frac{\left(\sum_{n=1}^{N} w(X_n)\right)^2}{\sum_{m=1}^{N} w(X_m)^2} = \left(\sum_{n=1}^{N} W_n^2\right)^{-1}. \tag{1.16}$$

This diagnostic was introduced in Kong et al. (1994) and is routinely used to access the quality of the particle approximation. We will adopt this convention when comparing algorithms whose validity are based on importance sampling. It is easy to see that $1 \leq \mathrm{ESS} \leq N$. The lower bound is attained in a situation typically referred to as *weight degeneracy* where only one sample holds all the normalized weight. On the other hand, the upper bound is achieved when all samples have uniform weights.

1.2.3 Markov Chain Monte Carlo

The main idea behind any *Markov chain Monte Carlo* (MCMC) method for the simulation of a target distribution π is to set up an ergodic Markov chain whose invariant distribution is π. In this section, we describe the celebrated *Metropolis-Hastings* algorithm, discuss its validity and outline several implementation choices.

1.2.3.1 Metropolis-Hastings algorithm

Consider a Markov transition kernel $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ of the form

$$K(x, \mathrm{d}y) = k(x, \mathrm{d}y) + r(x)\delta_x(\mathrm{d}y), \tag{1.17}$$

where $r(x) = 1 - \int_{\mathcal{X}} k(x, \mathrm{d}y)$ is the probability of rejecting a move from the transition kernel k and remaining at $x \in \mathcal{X}$. It is easy to see that if

$$\pi(\mathrm{d}x)k(x, \mathrm{d}y) = \pi(\mathrm{d}y)k(y, \mathrm{d}x) \tag{1.18}$$

as measures on $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X})$, then the detailed balance condition is satisfied and consequently K is π-invariant, i.e. $(\pi K)(\cdot) := \int_{\mathcal{X}} \pi(\mathrm{d}x)K(x, \cdot) = \pi(\cdot)$.

Suppose we are given a proposal transition kernel $q : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ with positive transition density $q(x, y), y \in \mathcal{X}$ with respect to $\mathrm{d}y$ for each $x \in \mathcal{X}$. Although q may not satisfy (1.18), we could enforce equality by introducing a measurable function $\alpha : \mathcal{X} \times \mathcal{X} \to [0, 1]$ and setting $k(x, \mathrm{d}y) = \alpha(x, y)q(x, \mathrm{d}y)$.

Algorithmically, this amounts to accepting or rejecting a transition from q according to the acceptance probability α. Substituting this form of k into (1.18) gives

$$\alpha(x, y) = R(x, y)\alpha(y, x) \tag{1.19}$$

for $(x, y) \in \mathcal{X} \times \mathcal{X}$ where

$$R(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}. \tag{1.20}$$

We now seek a choice of α that satisfies (1.19). First consider the case $R(x, y) < 1$, where moves from x to y occurs too frequently under q. To impose equality, we could always accept moves from y to x, i.e. $\alpha(y, x) = 1$, and limit moves from x to y by setting $\alpha(x, y) = R(x, y)$. Similarly, for the case $R(x, y) \geq 1$, setting $\alpha(x, y) = 1$ and $\alpha(y, x) = R(y, x)$ satisfies (1.19) since $1 = R(x, y)R(y, x)$ due to the form of (1.20). Collapsing these two cases gives the canonical choice $\alpha(x, y) = \min\{1, R(x, y)\}$, which can be shown to be more efficient than other choices that have been proposed (Barker 1965, Hastings 1970) in the sense of *Peskun's ordering* (Peskun 1973).

In summary, the above arguments derive the Metropolis-Hastings (MH) kernel

$$K(x, \mathrm{d}y) = \alpha(x, y)q(x, \mathrm{d}y) + r(x)\delta_x(\mathrm{d}y) \tag{1.21}$$

with $\alpha(x, y) = \min\{1, R(x, y)\}$, $r(x) = 1 - \int_{\mathcal{X}} \alpha(x, y)q(x, \mathrm{d}y)$, introduced in Metropolis et al. (1953) and later generalized by Hastings (1970). As the proposal q can be arbitrary, this provides a general method to construct a reversible Markov transition kernel with π as its specified invariant distribution. Moreover, since the acceptance probability involves a ratio (1.20), it is only necessary to know π and q up to some normalizing constants.

It is easy to see that finite compositions and mixtures of π-invariant Markov transition kernels are also π-invariant. A widely used example is the systematic/random scan *Gibbs sampler* (Geman and Geman 1984), which corresponds to a composition/mixture of kernels given by the full conditional distributions of π that are clearly π-invariant. Whenever sampling from a particular full conditional is intractable, one can also replace this step with a MH kernel targeting this full

conditional distribution – this procedure is commonly known as *Metropolis-within-Gibbs* (Zeger and Karim 1991). Another example of mixing π-invariant kernels is the *hit-and-run* algorithm (Smith 1984).

1.2.3.2 Validity of MCMC methods

We now discuss some theoretical justifications for the use of MCMC methods that are constructed with MH kernels.

Ergodicity. For a MH kernel K to be *ergodic*, it needs to satisfy two important properties: *Harris recurrence* and *aperiodicity*.

By Tierney (1994; Corollary 2), a π-*irreducible* MH kernel has π as its unique invariant distribution and is Harris recurrent, i.e. for all $x \in \mathcal{X}$ and each $A \in \mathcal{B}(\mathcal{X})$ with $\pi(A) > 0$, the Markov chain initialized at x visits A infinitely often with probability one. Irreducibility means that for each $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ with $\pi(A) > 0$, there exists an $N \in \mathbb{N}$ such that the N^{th} iterate of K satisfies $K^N(x, A) > 0$. In the following, we will see that this condition, which ensures that the Markov chain $\{X_n\}$ generated by K eventually visits all regions of interest under π, usually translates to easily verifiable assumptions on the proposal kernel q.

Aperiodicity refers to the absence of cyclic behaviour: more precisely, there is no period $p \geq 2$ and disjoint subsets $A_0, \ldots, A_{p-1} \in \mathcal{B}(\mathcal{X})$ with positive measure under π such that $K(x, A_{(i+1 \mod p)}) = 1$ for all $i = 0, \ldots, p-1$ and $x \in A_i$.

With π-irreducibility and aperiodicity, by Nummelin (1984; Theorem 3.7, Proposition 6.3), we have convergence of the transition probabilities to the target distribution:

$$d_{\mathrm{TV}}\left(K^N(x, \cdot), \pi\right) := \frac{1}{2}\|K^N(x, \cdot) - \pi\|_{\mathrm{TV}} \to 0 \tag{1.22}$$

as $N \to \infty$ for all $x \in \mathcal{X}$. Here $d_{\mathrm{TV}} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty)$ denotes the *total variation* distance induced by the total variation norm $\|\cdot\|_{\mathrm{TV}}$ on the vector space of finite signed measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. It is straightforward to see that this metric coincides with (1.11) when the measures concerned are non-random (Roberts and Rosenthal 2004; Proposition 3b).

Uniform ergodicity. Under appropriate sufficient conditions, it is possible to characterize the rate at which convergence in (1.22) takes place, thus giving rise to stronger forms of ergodicity.

A classical result that dates back to the seminal paper of Doeblin (1938) concerns *uniform ergodicity*:

$$\sup_{x\in\mathcal{X}} d_{\mathrm{TV}}\left(K^N(x,\cdot),\pi\right) \leq C\varrho^N \tag{1.23}$$

for some $C > 0, \varrho \in (0,1)$. This requires existence of $M \in \mathbb{N}, \varepsilon > 0, \nu \in \mathcal{P}(\mathcal{X})$ such that the following *minorization condition* is satisfied

$$K^M(x,A) \geq \varepsilon\nu(A) \tag{1.24}$$

for all $x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$. By taking $A = \mathcal{X}$ in (1.24), observe that $\varepsilon \in (0,1]$. It can be shown that the mapping defined by $\mathcal{P}(\mathcal{X}) \ni \mu \mapsto \mu K^M$ is a contraction, i.e.

$$d_{\mathrm{TV}}\left(\mu K^M, \mu' K^M\right) \leq \Delta_{\mathrm{TV}}\left(K^M\right) d_{\mathrm{TV}}\left(\mu,\mu'\right) \tag{1.25}$$

for all $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ where $\Delta_{\mathrm{TV}}\left(K^M\right) := \sup_{(x,x')\in\mathcal{X}\times\mathcal{X}} d_{\mathrm{TV}}\left(K^M(x,\cdot), K^M(x',\cdot)\right) \leq 1-\varepsilon$ is the *Dobrushin coefficient* of K^M (Dobrushin 1956). Since $(\mathcal{P}(\mathcal{X}), d_{\mathrm{TV}})$ is a complete metric space (Douc et al. 2014b; Proposition 6.16), it follows from the Banach fixed point theorem that (1.23) holds with $C = d_{\mathrm{TV}}\left(\delta_x,\pi\right) = 1, \varrho = (1-\varepsilon)^{\lfloor N/M \rfloor}$.

Geometric ergodicity. A weaker form of ergodicity is *geometric ergodicity*:

$$d_{\mathrm{TV}}\left(K^N(x,\cdot),\pi\right) \leq C(x)\varrho^N \tag{1.26}$$

for some function C that is π-almost everywhere finite and $\varrho \in (0,1)$. To establish such a result, one typically assumes ergodicity, existence of a *small set*, i.e. the minorization condition (1.24) holds only on a subset $S \subset \mathcal{X}$, and a *drift condition*: there exists $a \in [0,1), b \in \mathbb{R}_+$ and a function $V : \mathcal{X} \to [1,\infty]$ such that

$$K(V) \leq aV + b\mathbb{1}_S \tag{1.27}$$

where $K(V)(x) := \int_{\mathcal{X}} V(y)K(x, \mathrm{d}y)$, and that the small set S is characterized by V. With reference to its connection to the study of stability in dynamical systems, the function V is often called a *Lyapunov function*. Intuitively, if we think of V as describing an energy surface, the condition (1.27) implies that on average the Markov chain drifts towards states of lower energy, and eventually reaching the small set.

There are many different proofs of geometric ergodicity. The traditional way involves controlling the length of excursions outside the small set (Nummelin 1984, Meyn and Tweedie 1993). Another popular approach is to use coupling techniques which has the added benefit of giving explicit expressions of C and ϱ (see Roberts and Rosenthal (2004) and references therein). Such proofs typically use two coupled realizations of a Markov chain that coalesce. Although coalescence is facilitated by the minorization condition when coupled chains both enter the small set, there are many scenarios where coupled chains may come close to each other but fail to coalesce. This has motivated the use of *Wasserstein* distance (introduced in Section 2.2) to analyze convergence of Markov chains (Gibbs 2004, Madras and Sezer 2010). More recent methods establish *V-geometric ergodicity* by exploiting the existence of a spectral gap when we view the transition kernel as an operator on the Banach space of functions with supremum norm weighted by V (Hairer and Mattingly 2008; 2011).

Limit theorems. Lastly, suppose that we have π-irreducibility and aperiodicity which imply that the Markov chain $\{X_n\}$ generated by the MH kernel is ergodic. Under these conditions, we have a strong LLN that allows us to compute expectations under π:

$$\tilde{\pi}_N(\varphi) = \frac{1}{N} \sum_{n=1}^{N} \varphi(X_n) \to \pi(\varphi) \tag{1.28}$$

almost surely as $N \to \infty$ for any $\varphi \in L^1(\pi)$ (Meyn and Tweedie 1993; Theorem 17.0.1). It is worth noting that (1.28) also holds for periodic Markov chains. If we additionally assume $\tilde{\sigma}^2(\varphi) := \lim_{N\to\infty} N\mathbb{E}\left[(\tilde{\pi}_N(\varphi) - \pi(\varphi))^2\right] < \infty$, as MH chains are reversible, it follows from Kipnis and Varadhan (1986) that this estimator

also satisfies a CLT

$$\sqrt{N}\left(\tilde{\pi}_N(\varphi) - \pi(\varphi)\right) \xrightarrow{d.} \mathcal{N}\left(0, \tilde{\sigma}^2(\varphi)\right) \tag{1.29}$$

as $N \to \infty$. The asymptotic variance can be written as $\tilde{\sigma}^2(\varphi) = \sigma^2(\varphi)\tau(\varphi)$, where

$$\tau(\varphi) := 1 + 2\sum_{n=1}^{\infty} \mathrm{Corr}_\pi\left(\varphi(X_0), \varphi(X_n)\right) \tag{1.30}$$

is the *integrated auto-correlation time* (IACT) and Corr_π denotes correlation between random variables generated by the Markov chain at stationarity. Recalling from (1.4) that $\sigma^2(\varphi)$ is the asymptotic variance under independent sampling, we can utilize this relationship to define the notion of ESS for a Markov chain of length N as $\tau(\varphi)^{-1}N$. In practice, the auto-correlation function can be estimated using techniques from time series.

1.2.3.3 Implementation choices

We now describe some implementation strategies that are widely used and discuss some practical considerations. This section only serves to outline some algorithms that will be of interest in this thesis and is by no means an exhaustive review on the vast amount of literature on MCMC that has been developed in the past few decades.

Independent Metropolis-Hastings. In the case where $q(x, \mathrm{d}y) = q(y)\mathrm{d}y$, commonly termed as *independent* MH, the acceptance probability has the form

$$\alpha(x, y) = \min\left\{1, \frac{w(y)}{w(x)}\right\} \tag{1.31}$$

where w is the importance weight function with proposal distribution q. As alluded earlier, the resulting Markov chain is π-irreducible and aperiodic if and only if $q(y) > 0$ $\mathrm{d}y$-almost everywhere. It follows from (1.31) that moves to states with higher weights are always accepted while moves to states with low weights are often rejected.

Observe also that the process could be "stuck" in state x for many iterations if it has a particularly large weight. Such a behaviour can be very pronounced when the tails of the target is heavier than that of the proposal distribution. To mitigate this problem, one should select a proposal so that the weights are bounded – observe

that this also coincides with good importance sampling performance. More precisely, under the boundedness assumption, it is possible to show that the minorization condition (1.24) is satisfied with $M = 1, \varepsilon = \|\mathrm{d}\pi/\mathrm{d}q\|_\infty^{-1}, \nu = \pi$; hence the Markov chain is uniformly ergodic with rate $\varrho = 1 - \|\mathrm{d}\pi/\mathrm{d}q\|_\infty$.

Clearly, optimal performance is attained when $q = \pi$, in which case we recover iid sampling. However, in practice it is difficult to design a proposal that mimics the target distribution closely, and particularly so in high dimensional state spaces.

Random walk Metropolis-Hastings. A more numerically tractable alternative is the use of *local* proposals. The simplest example of this is *random walk* MH (RWMH), where $q(x, y) = f(y-x)$ for some increment density f. For π-irreducibility and aperiodicity, it suffices to have $f > 0$ everywhere. The typical choice of increment distribution is $\mathcal{N}(0_d, \varepsilon^2\Sigma)$, i.e. the Gaussian distribution with mean vector $0_d := (0, \ldots, 0)^T$ and covariance matrix $\varepsilon^2\Sigma$ for some $\varepsilon > 0$ and symmetric positive definite matrix Σ. Under this choice, the acceptance probability takes the form

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}, \tag{1.32}$$

from which it is apparent that moves to states with higher target density are always accepted while moves to states of lower target density are sometimes accepted. The former behaviour reveals that the RWMH chain performs an exploration of the state space guided by only *local* information, i.e. target density values, while the latter behaviour distinguishes the sampling algorithm from an optimization routine. Sufficient conditions for geometric ergodicity of the resulting algorithm are given in Roberts and Tweedie (1996b) and Jarner and Hansen (2000).

Metropolis adjusted Langevin algorithm. The above methodology can be improved upon by using also *gradient* information when this is available. We denote the partial derivative with respect to time variable t and i^{th}-coordinate spatial variable x_i by ∂_t and ∂_{x_i} respectively. For any scalar-valued function $\varphi \in C^1(\mathbb{R}^d, \mathbb{R})$, its gradient is defined as $\nabla\varphi(x) := (\partial_{x_1}\varphi(x), \ldots, \partial_{x_d}\varphi(x))^T$. For any vector-valued function $\varphi = (\varphi_1, \ldots, \varphi_d)^T$, we define the divergence operator

as $\nabla \cdot \varphi(x) := \sum_{i=1}^{d} \partial_{x_i} \varphi_i(x)$ for $\varphi \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ and the Laplacian operator as $\Delta\varphi(x) := \sum_{i=1}^{d} \partial_{x_i}^2 \varphi_i(x)$ for $\varphi \in C^2(\mathbb{R}^d, \mathbb{R}^d)$.

Consider the case where Σ is the identity matrix I_d of size d. If we view the above random walk as a discretization of a d-dimensional standard *Brownian motion* $\{B_t\}_{t\geq 0}$, then a better proposal can be derived from the *Langevin diffusion* $\{L_t\}_{t\geq 0}$, defined as the solution of the following stochastic differential equation (SDE)

$$\mathrm{d}L_t = \frac{1}{2}\nabla \log \pi(L_t)\mathrm{d}t + \mathrm{d}B_t. \tag{1.33}$$

The time evolution of the marginal distribution of L_t, denoted as $\tilde{\pi}_t$, is governed by the *Fokker-Planck* partial differential equation (PDE)

$$\partial_t \tilde{\pi}_t = -\frac{1}{2}\nabla \cdot (\tilde{\pi}_t \nabla \log \pi) + \frac{1}{2}\Delta \tilde{\pi}_t. \tag{1.34}$$

It is straightforward to check that π is the stationary solution of (1.34), hence the Langevin diffusion admits π as its invariant measure. It can also be shown that the solution $\{\tilde{\pi}_t\}_{t\geq 0}$ of (1.34) is such that the Kullback-Leibler divergence of π from $\tilde{\pi}_t$ decreases with time and moreover follows the direction of steepest descent of this functional with respect to the Wasserstein distance (Jordan et al. 1998).

Now consider an *Euler-Maruyama* discretization of (1.33) with step size ε^2:

$$Y_n = Y_{n-1} + \frac{\varepsilon^2}{2}\nabla \log \pi(Y_{n-1}) + \varepsilon Z_n \tag{1.35}$$

where $\{Z_n\}$ is a sequence of iid samples from $\mathcal{N}(0_d, I_d)$. The use of the time discretized process $\{Y_n\}$ for sampling was proposed in earlier papers by Parisi (1981) and Grenander and Miller (1994). However, due to the time discretization, π-invariance is only approximately preserved for sufficiently small ε^2 (Mattingly et al. 2002). Moreover, Roberts and Tweedie (1996a) showed that the discretized process might even be *transient* for well-behaved target distributions. As noted by Besag (1994), using (1.35) as a proposal move within a MH kernel, i.e. setting

$$q(x, y) = \mathcal{N}\left(y; x + \frac{\varepsilon^2}{2}\nabla \log \pi(x), \varepsilon^2 I_d\right), \tag{1.36}$$

allows us to preserve π-invariance. Here $x \mapsto \mathcal{N}(x; \mu, \Sigma)$ denotes the Gaussian density with mean μ and covariance Σ. Ergodicity properties of the resulting

Markov chain, now commonly known as the *Metropolis adjusted Langevin algorithm* (MALA), can be found in Roberts and Tweedie (1996a).

Choice of algorithmic parameters. We now turn our attention to the choice of algorithmic parameters ε^2 and Σ. It is possible to let these parameters depend on an initial portion of the sequence $\{X_n\}$ up to some time $N_0 \in \mathbb{N}$, in which case we obtain *adaptive MCMC* methods (see Andrieu and Thoms 2008). Too large values of the scale parameter ε^2 would lead to many rejected moves while too small values would result in a very slow moving Markov chain – therefore there should be a notion of *optimal scaling*.

Consider a target distribution that factorizes into identical marginals, i.e. $\pi(\mathrm{d}x) = \prod_{i=1}^{d} \pi_1(\mathrm{d}x_i)$. In Roberts et al. (1997) and Roberts and Rosenthal (1998), under some regularity conditions on π_1, the authors proved that appropriately rescaled RWMH and MALA chains converge weakly to diffusion processes as $d \to \infty$. They then maximized the speed measure of these limiting diffusions, which can be shown to be equivalent to minimizing the auto-correlation (for any test function) with respect to the scale parameter. Optimality is achieved in RWMH if $\varepsilon^2 = O(d^{-1})$ with optimal acceptance rate of 0.234 and for MALA if $\varepsilon^2 = (d^{-1/3})$ with optimal acceptance rate of 0.574. Such optimal scaling results partially justifies the standard practice of selecting ε^2 as large as possible to achieve acceptance probabilities close to these guidelines. Moreover, they also quantify the asymptotic inefficiency of RWMH compared to MALA, if the additional cost of computing gradients is also taken into account. In practice, it is important to treat these guidelines as only heuristics as these analysis were done in the overly simplified case of target distributions that factorize and in the large dimension asymptotic.

The covariance matrix Σ should be chosen to reflect the geometry of the target distribution. Such curvature information could be obtained for example by computing the Hessian of the target density or estimated using samples from a pilot run of the MCMC chain. It is also possible to design more sophisticated

MCMC methods that exploit ideas from differential geometry – see Girolami and Calderhead (2011) for more details.

1.2.4 Sequential Monte Carlo samplers

In this section, we describe the *sequential Monte Carlo* (SMC) sampler introduced in Del Moral et al. (2006), which is a methodology to sample sequentially from a sequence of probability measures defined on a common measurable space. We will see that this sampler falls under a wider class of algorithms known as *SMC methods*, or *particle filters* in the case of inference for hidden Markov models, that deals with the setup where the dimension of target distributions increase with time. These methods have been the topic of intensive research in the last two decades (Doucet et al. 2001, Cappé et al. 2006, Doucet and Johansen 2009) and its theoretical properties are now well understood (Chopin 2004, Del Moral 2004, Künsch 2005, Del Moral 2013, Whiteley 2013, Bérard et al. 2014, Douc et al. 2014a).

1.2.4.1 Constructing a sequence of probability measures

We begin by addressing the issue of defining a sensible sequence of probability measures when only a single target distribution π is of interest.

If π is the posterior distribution associated to a data set $y = (y_1, \ldots, y_T)^T$ with $T \in \mathbb{N}$ observations, then we could perform *sequential Bayesian inference* by exploring the partial posteriors $\pi_t(\mathrm{d}x) := \pi(\mathrm{d}x|y_{1:t})$, where the shorthand $y_{1:t}$ refers to the sequence $y_1, \ldots, y_t$. As noted by Chopin (2002), such a procedure might provide a beneficial tempering effect and potential reduction in computational complexity.

Instead of assuming that the likelihood admits such a structure, an alternative way to induce this desired tempering behaviour is to exponentiate the likelihood with an increasing sequence $\{\lambda_t\}_{t=0}^T \subseteq [0,1]$ in the Bayes' update (1.1):

$$\pi_t(\mathrm{d}x) := \frac{\pi_0(\mathrm{d}x)L(x)^{\lambda_t}}{Z_t} \tag{1.37}$$

where $Z_t := \int_{\mathcal{X}} \pi_0(\mathrm{d}x)L(x)^{\lambda_t}$ (note that this is finite since Z is). If the prior and posterior distributions are distant, or equivalently if the likelihood is particularly

informative or induces complex structures in the posterior, then importance sampling with π_0 as proposal and MCMC methods targeting π would be ineffective. The rationale behind introducing a sequence of *bridging* distributions with a tempering schedule is to gradually evolve the prior π_0 (assume $\lambda_0 = 0$) to the posterior $\pi_T = \pi$ (assume $\lambda_T = 1$) and to apply Monte Carlo methods in a manner that exploits the proximity between successive distributions. Such an approach would be beneficial as long as the errors involved in solving T simpler problems sequentially do not accumulate too badly. It is also apparent that the benefits of likelihood tempering is closely related to the ideas behind *simulated annealing* in the context of optimization (Kirkpatrick et al. 1983).

It is worth noting that we can also construct other sequence of distributions to suit other applications (see Section 3.5.2). For example, when one is interested in estimating the probability of a rare event $A \in \mathcal{B}(\mathcal{X})$ under a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ which we can simulate from, i.e. $\mu(A) \approx 0$, then it is natural to define the sequence $\pi_t(\mathrm{d}x) := \mu(\mathrm{d}x)\mathbb{1}_{A_t}(x)/\mu(\mathbb{1}_{A_t})$ for some $A_t \in \mathcal{B}(\mathcal{X})$ such that $\mathcal{X} =: A_0 \supset \ldots \supset A_T := A$ (Johansen et al. 2006).

1.2.4.2 Sequential importance sampling

Suppose that we can initialize particles by sampling from the prior distribution π_0. To move to the next intermediate distribution π_1, particles are then propagated according to a Markov kernel K_1. As alluded earlier, since π_0 and π_1 are not too distant, it should be possible to construct a kernel to move particles to regions of the state space with high probability mass under π_1. For example, if K_1 is a π_1-invariant MCMC kernel that converges to its equilibrium distribution reasonably quickly, i.e. fast mixing, we expect the marginal distribution of the particles at time $t = 1$ to be close to π_1. For subsequent times, we continue moving particles with Markov kernels $\{K_t\}_{t=2}^T$ until the terminal time T. Hence in summary, for all $n = 1, \ldots, N$ we initialize $X_0^n \sim \pi_0$, and iterate $X_t^n \sim K_t(X_{t-1}^n, \cdot)$ for $t = 1, \ldots, T$.

For notational convenience, we write the t^{th}-step transition kernel as $K_{1:t} := K_1 \circ \cdots \circ K_t$, where $\circ$ denotes composition of Markov kernels. For each time

$t = 1, \ldots, T$, we will see in Section 1.2.4.4 that ideally we would like to perform importance sampling on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with π_t as the target distribution and the marginal distribution $q_t := \pi_0 K_{1:t}$ as the proposal, but the latter is typically intractable. The key idea behind Del Moral et al. (2006) is to mitigate this difficulty by introducing a sequence of *backward* Markov kernels $\{L_t\}_{t=0}^{T-1}$ and perform importance sampling on path space $(\mathcal{X}^{t+1}, \mathcal{B}(\mathcal{X})^{\otimes t+1})$ between an artificially constructed target distribution

$$P_t(\mathrm{d}x_{0:t}) := \pi_t(\mathrm{d}x_t) \prod_{k=1}^{t} L_{k-1}(x_k, \mathrm{d}x_{k-1}) \tag{1.38}$$

and the joint distribution of the particle dynamics

$$Q_t(\mathrm{d}x_{0:t}) := \pi_0(\mathrm{d}x_0) \prod_{k=1}^{t} K_k(x_{k-1}, \mathrm{d}x_k) \tag{1.39}$$

as the proposal. If the backward kernels are selected so that $P_t \ll Q_t$ then

$$\frac{\mathrm{d}P_t}{\mathrm{d}Q_t}(x_{0:t}) = \frac{W_t(x_{0:t})}{Z_t}, \tag{1.40}$$

and the (unnormalized) weight function $W_t : \mathcal{X}^{t+1} \to [0, \infty)$ can be written as

$$W_t(x_{0:t}) = \prod_{k=1}^{t} w_k(x_{k-1}, x_k) \tag{1.41}$$

where the incremental weight function $w_t : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is the unnormalized Radon-Nikodym derivative of $\pi_t(\mathrm{d}x_t) L_{t-1}(x_t, \mathrm{d}x_{t-1}) \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ with respect to $\pi_{t-1}(\mathrm{d}x_{t-1}) K_t(x_{t-1}, \mathrm{d}x_t) \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, i.e.

$$w_t(x_{t-1}, x_t) := \frac{\gamma_t(\mathrm{d}x_t) L_{t-1}(x_t, \mathrm{d}x_{t-1})}{\gamma_{t-1}(\mathrm{d}x_{t-1}) K_t(x_{t-1}, \mathrm{d}x_t)} \tag{1.42}$$

for $t = 1, \ldots, T$. Hence a particle approximation of π_t is given by

$$\hat{\pi}_t^N = \sum_{n=1}^{N} W_t^n \delta_{X_t^n}, \quad W_t^n := \frac{W_t(X_{0:t}^n)}{\sum_{m=1}^{N} W_t(X_{0:t}^m)}. \tag{1.43}$$

Using the identity $Z_t = Q_t(W_t)$, an unbiased estimator of Z_t is given by

$$\hat{Z}_t^N = \frac{1}{N} \sum_{n=1}^{N} W_t(X_{0:t}^n). \tag{1.44}$$

Iterating this procedure from $t = 1$ to $t = T$ is typically known as *sequential importance sampling* (SIS). When only the terminal time is of interest, we will write $P := P_T, Q := Q_T$ and $W := W_T$.

1.2.4.3 Sequential importance sampling resampling

Although for each $t = 1, \ldots, T$, P_t admits π_t as its marginal distribution by construction, this comes at the cost of having to perform importance sampling on an extended space that increases in dimension over time. Hence from the discussion in Section 1.2.2.3, we expect the quality of the particle approximation to degrade with time. In practice, this manifests as *weight degeneracy* which is a well-known phenomenon in the SMC literature. To mitigate such degeneracy, one usually employs a *resampling* step which serves to prune particles with low weights and multiply particles with high weights.

Intuitively, this procedure is beneficial as it allows the algorithm to focus its computational effort on promising regions of the state space. For a more theoretical justification, under mixing assumptions on $\{K_t\}_{t=1}^T$, it can be shown that if successive distributions $\{\pi_t\}_{t=0}^T$ are very distant, then the asymptotic variance (given in Section 1.2.4.5) of SIS (1.63) diverges with time while that of SIS with resampling (1.65) remains *uniformly* bounded (Chopin (2004; Theorem 5), Del Moral et al. (2006; Remark 4)).

Let $\mathcal{G}_t$ be the σ-algebra generated by $\{W_t^n, X_{0:t}^n\}_{n=1}^N$. More precisely, a resampling step is a random mapping that takes (1.43) to another empirical measure

$$\bar{\pi}_t^N = \sum_{n=1}^N \frac{N_t^n}{N} \delta_{X_t^n} = \frac{1}{N} \sum_{n=1}^N \delta_{\bar{X}_t^n}, \tag{1.45}$$

where the number of offsprings of particle X_t^n, i.e. $N_t^n := \#\{m = 1, \ldots, N : \bar{X}_t^m = X_t^n\}$, is constrained to satisfy $\mathbb{E}[N_t^n | \mathcal{G}_t] = NW_t^n$ for $n = 1, \ldots, N$. The simplest resampling scheme is *multinomial resampling* (Gordon et al. 1993), where $(N_t^1, \ldots, N_t^N)$ is distributed according to a multinomial distribution with N trials and probabilities $\{W_t^n\}_{n=1}^N$. Other schemes with lower variance such as *residual resampling* (Whitley 1994, Liu and Chen 1998), *stratified resampling* and *systematic resampling* (Kitagawa 1996, Fearnhead and Clifford 2003) are more commonly used in practice – see Douc and Cappé (2005), Hol et al. (2006) for comparisons of these schemes.

It should be stressed that resampling whenever the quality of the particle approximation is adequate is detrimental as it increases the variance of Monte Carlo estimates. Hence in practice, one usually employs resampling adaptively, i.e. whenever some measure of weight degeneracy (for example ESS defined in Section 1.2.2.4) falls below a pre-specified threshold. We shall henceforth refer to an algorithm which combines SIS and (possibly) resampling as the *SMC sampler* and outline its steps in Algorithm 1.

From (1.42), we have the identity

$$\frac{Z_t}{Z_{t-1}} = \int_{\mathcal{X}^2} w_t(x_{t-1}, x_t)\pi_{t-1}(\mathrm{d}x_{t-1})K_t(x_{t-1}, \mathrm{d}x_t). \tag{1.46}$$

Hence given a (possibly resampled) particle approximation $\hat{\pi}_{t-1}^N = \sum_{n=1}^N W_{t-1}^n \delta_{X_{t-1}^n}$ of π_{t-1}, we can obtain a consistent approximation of $R_t := Z_t/Z_{t-1}$ with

$$\hat{R}_t^N = \sum_{n=1}^N W_{t-1}^n w_t(X_{t-1}^n, X_t^n). \tag{1.47}$$

Since $Z_t = \prod_{k=1}^t R_k$, it is natural to form an estimator of Z_t using the product

$$\hat{Z}_t^N := \prod_{k=1}^t \hat{R}_k^N. \tag{1.48}$$

While consistency of $\hat{Z}_t^N$ follows straightforwardly from that of (1.47), a remarkable property of this estimator is its *unbiasedness*, when averaged over all random variables generated by the SMC sampler (Del Moral 2004; Theorem 7.4.2). Unlike the corresponding SIS estimator (1.44), such a result is non-trivial since the particles interact through the resampling steps.

1.2.4.4 Choice of backward kernels

Introducing the artificial backward Markov kernels $\{L_t\}_{t=0}^{T-1}$ to circumvent the difficulty of evaluating the marginal distributions $q_t = \pi_0 K_{1:t}$ leaves a degree of freedom over the choice of these backward kernels. We first identify the optimal choice which is intractable before discussing approximations that yield various suboptimal but tractable choices.

Algorithm 1 Sequential Monte Carlo sampler

Input: particles N, time steps T, probability measures $\{\pi_t\}_{t=0}^T$, Markov kernels $\{K_t\}_{t=1}^T$, backward kernels $\{L_t\}_{t=0}^{T-1}$, resampling threshold $\theta \in (0,1)$.

1. Initialization: sample $X_0^n \sim \pi_0$ and set $W_0^n = N^{-1}$ for $n = 1, \dots, N$.
2. For $t = 1, \dots, T$,
   (a) sample $X_t^n \sim K_t(X_{t-1}^n, \cdot)$ for $n = 1, \dots, N$;
   (b) evaluate incremental weights $\left\{w_t(X_{t-1}^n, X_t^n)\right\}_{n=1}^N$ using (1.42);
   (c) update and normalize weights for $n = 1, \dots, N$

$$W_t^n = W_{t-1}^n w_t(X_{t-1}^n, X_t^n) \Big/ \sum_{m=1}^N W_{t-1}^m w_t(X_{t-1}^m, X_t^m);$$

   (d) if $\mathrm{ESS}_t < \theta N$, resample particles and set $W_t^n = N^{-1}$ for $n = 1, \dots, N$;
   (e) compute ratio of normalizing constants estimate $\hat{R}_t^N$ using (1.47).

Output: particles $\{X_T^n\}_{n=1}^N$, normalized weights $\{W_T^n\}_{n=1}^N$ and normalizing constant estimate $\hat{Z}_T^N = \prod_{t=1}^T \hat{R}_t^N$.

Optimal backward kernels. By Del Moral et al. (2006; Proposition 1), the optimal choice of backward kernels $\{L_t^*\}_{t=0}^{T-1}$ in terms of minimizing the variance of the importance weight W_t satisfies

$$q_t(\mathrm{d}x_t) L_{t-1}^*(x_t, \mathrm{d}x_{t-1}) = q_{t-1}(\mathrm{d}x_{t-1}) K_t(x_{t-1}, \mathrm{d}x_t) \tag{1.49}$$

as measures on $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X})$ for $t = 1, \dots, T$. To see this, note that under the choice (1.49) we have $W_t(x_{0:t}) = \mathrm{d}\gamma_t / \mathrm{d}q_t(x_t)$, hence by the variance decomposition formula

$$\mathbb{V}\left[W_t(X_{0:t})\right] = \mathbb{E}\left[\mathbb{V}\left[W_t(X_{0:t}) | X_t\right]\right] + \mathbb{V}\left[\mathbb{E}\left[W_t(X_{0:t}) | X_t\right]\right] = \mathbb{V}\left[\frac{\mathrm{d}\gamma_t}{\mathrm{d}q_t}(X_t)\right]. \tag{1.50}$$

This result is intuitive: under the optimal choice, we simply perform importance sampling on the marginal space $\mathcal{X}$ instead of the extended space $\mathcal{X}^{t+1}$. In terms of marginal distributions, observe that integrating over the x_t variable in (1.49) gives $q_t L_{t-1}^* = q_{t-1}$. To understand these optimal backward kernels in terms of the joint distribution Q_t, note the following forward-backward decomposition

$$Q_t(\mathrm{d}x_{0:t}) = \pi_0(\mathrm{d}x_0) \prod_{k=1}^t K_k(x_{k-1}, \mathrm{d}x_k) = q_t(\mathrm{d}x_t) \prod_{k=1}^t L_{k-1}^*(x_k, \mathrm{d}x_{k-1}). \tag{1.51}$$

Although employing these optimal backward kernels is not feasible, as this requires computing weights that are intractable, knowing the form of (1.49) allows us to construct approximations. While the validity of SMC sampler holds for any choice of backward kernels, algorithmic performance will depend on the quality of such approximations.

Locally optimal backward kernels. We begin with a local argument by assuming that q_{t-1} is sufficiently close to π_{t-1}. Using this approximation in (1.49) gives a suboptimal choice $\{\tilde{L}_t\}_{t=0}^{T-1}$ that satisfies

$$\pi_{t-1}K_t(\mathrm{d}x_t)\tilde{L}_{t-1}(x_t, \mathrm{d}x_{t-1}) = \pi_{t-1}(\mathrm{d}x_{t-1})K_t(x_{t-1}, \mathrm{d}x_t), \tag{1.52}$$

with corresponding incremental weight of the form

$$w_t(x_{t-1}, x_t) = \frac{\mathrm{d}\gamma_t}{\mathrm{d}\gamma_{t-1}K_t}(x_t) \tag{1.53}$$

for $t = 1, \ldots, T$. In terms of marginal distributions, observe that (1.52) implies $(\pi_{t-1}K_t)\tilde{L}_{t-1} = \pi_{t-1}$, which suggests that $\tilde{L}_{t-1}$ could be thought of as a *locally optimal* backward kernel at time t. This choice of backward kernels is not generally applicable as (1.53) involves a possibly intractable integral – see Del Moral et al. (2006; Section 5) and Griffin et al. (2016) for applications where this is used.

Time reversed backward kernels. Consider the case where K_t is chosen to be π_t-invariant. If successive distributions are sufficiently close, approximating π_{t-1} with π_t in (1.52) gives another suboptimal choice $\{\hat{L}_t\}_{t=0}^{T-1}$ that satisfies

$$\pi_t(\mathrm{d}x_t)\hat{L}_{t-1}(x_t, \mathrm{d}x_{t-1}) = \pi_t(\mathrm{d}x_{t-1})K_t(x_{t-1}, \mathrm{d}x_t), \tag{1.54}$$

with corresponding incremental weight of the form

$$w_t(x_{t-1}, x_t) = \frac{\mathrm{d}\gamma_t}{\mathrm{d}\gamma_{t-1}}(x_{t-1}) \tag{1.55}$$

for $t = 1, \ldots, T$. Observe that $\hat{L}_{t-1}$ is simply the *time reversal* of the π_t-invariant Markov kernel K_t. When using the tempering schedule in (1.37), under this choice of backward kernels, the SIS algorithm (i.e. no resampling) coincides with work

by Crooks (1998) in thermodynamic integration, the *annealed importance sampler* (AIS) introduced by Neal (2001) and work in Jarzynski (1997) when the sampler is formulated in continuous time. For sequential Bayesian inference problems, the SMC sampler recovers the algorithm proposed by Chopin (2002).

Note from (1.55) that a particle's weight W_t^n will not depend on its location X_t^n at time t. Therefore, to promote sample diversity, the Markov kernel K_t should only be applied after the particle approximation $\hat{\pi}_t^N = \sum_{n=1}^N W_t^n \delta_{X_{t-1}^n}$ has (possibly) been resampled. Interestingly, as noted by Del Moral et al. (2012a) and Schäfer and Chopin (2013), it is possible to exploit this independence between particle weights and locations at time t to adaptively specify the tempering schedule $\{\lambda_t\}_{t=0}^T$. The main idea is to observe that, owing to the form of (1.55), the effective sample size at time t, denoted by ESS_t, is a function of λ_t. Hence we can solve for a value such that

$$\mathrm{ESS}_t(\lambda_t) = \varrho \mathrm{ESS}_{t-1} \tag{1.56}$$

for some pre-specified proportion $\varrho \in (0, 1)$. If one performs adaptive resampling, it is not clear that ESS is the most suitable quantity in (1.56) as it measures the accumulated discrepancy between P_t and Q_t from the last resampling time. To induce more uniformity in the distance between successive distributions $\{\pi_t\}_{t=0}^T$, Zhou et al. (2016) argue convincingly that it is more appropriate to perform adaptation using a quantity they termed as *conditional* ESS

$$\mathrm{cESS}_t := \frac{N\left(\sum_{n=1}^N W_{t-1}^n w_t(X_{t-1}^n)\right)^2}{\sum_{m=1}^N W_{t-1}^m w_t(X_{t-1}^m)^2}, \tag{1.57}$$

which can be thought of as a measure of how good π_{t-1} is as an importance sampling proposal for π_t. If we resample at every time iteration, then $W_{t-1}^n = N^{-1}$ for all $n = 1, \ldots, N$, so (1.57) coincides with the usual definition of ESS and we recover the procedure in (1.56).

Despite the benefits of adaptation, the independence between a particle's weight and location is counter-intuitive: even if K_t is fast mixing (consider the extreme case where $K_t(x_{t-1}, \mathrm{d}x_t) = \pi_t(\mathrm{d}x_t)$), the variance of the incremental weight (1.55) would still be particularly high if π_{t-1} and π_t are distant. In contrast, the backward

kernels $\{\tilde{L}_t\}_{t=0}^{T-1}$ are more sensible as the incremental weight (1.53) reflects the mixing properties of K_t and depends on the particle's location after the move.

1.2.4.5 Validity of SMC samplers

We now give some theoretical justifications for the use of SMC samplers. While the validity of SIS follow from standard iid asymptotics (Section 1.2.2), establishing convergence of SMC algorithms is much more involved due to the interaction of particles in the resampling steps. Many results on general SMC methods exist and can be found for example in the monograph by (Del Moral 2004; 2013) and specialized to the context of SMC samplers as done in Del Moral et al. (2006; Section 3.4). Moreover, as a consequence of identifying algorithms such as Chopin (2002) in the SMC framework, many convergence results apply straightforwardly.

Consider the likelihood tempering case (1.37) with $\lambda_t = t\Delta\lambda$ for $t = 0, \ldots, T$ and the use of π_t-invariant Markov kernel K_t and its time reversal (1.54) as backward kernel. For notational simplicity, given $\mu \in \mathcal{P}(\mathcal{X})$ and $N \in \mathbb{N}$, we define the *Bayes'* operator $(\mathsf{B}\mu)(\mathrm{d}x) := \mu(\mathrm{d}x)L(x)^{\Delta\lambda}/\mu(L^{\Delta\lambda})$ and the *sampling* operator $\mathsf{S}^N\mu := N^{-1}\sum_{n=1}^N \delta_{X_n}$, where $\{X_n\}_{n=1}^N$ are iid samples from μ. In these notation, we can write a SMC sampler with multinomial resampling at every time step succinctly as

$$\hat{\pi}_t^N = \mathsf{B}\mathsf{S}^N(\hat{\pi}_{t-1}^N K_{t-1}) \tag{1.58}$$

for $t = 1, \ldots, T$ with initialization at $\hat{\pi}_0^N = \pi_0$. For $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and the metric d defined in (1.11), using the fact that $d(\mathsf{B}\mu, \mathsf{B}\nu) \leq 2\kappa^{-2}d(\mu, \nu)$ holds under the assumption $\kappa \leq L(x)^{\Delta\lambda} \leq \kappa^{-1}$ for all $x \in \mathcal{X}$ and some $\kappa \in (0, 1)$, $d(\mathsf{S}^N\mu, \mu) \leq N^{-1/2}$ and $d(\mu K, \nu K) \leq d(\mu, \nu)$ for any Markov kernel K on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we have by triangle inequality that

$$\begin{aligned} d(\hat{\pi}_t^N, \pi_t) &= d(\mathsf{B}\mathsf{S}^N(\hat{\pi}_{t-1}^N K_{t-1}), \mathsf{B}(\pi_{t-1}K_{t-1})) \\ &\leq 2\kappa^{-2} d(\mathsf{S}^N(\hat{\pi}_{t-1}^N K_{t-1}), \pi_{t-1}K_{t-1}) \\ &\leq 2\kappa^{-2}\left\{d(\mathsf{S}^N(\hat{\pi}_{t-1}^N K_{t-1}), \hat{\pi}_{t-1}^N K_{t-1}) + d(\hat{\pi}_{t-1}^N K_{t-1}, \pi_{t-1}K_{t-1})\right\} \\ &\leq 2\kappa^{-2}\left\{\frac{1}{\sqrt{N}} + d(\hat{\pi}_{t-1}^N, \pi_{t-1})\right\}. \end{aligned} \tag{1.59}$$

By induction, we obtain

$$d(\hat{\pi}_t^N, \pi_t) \leq \frac{1}{\sqrt{N}} \sum_{k=1}^{t} (2\kappa^{-2})^k \tag{1.60}$$

for $t = 1, \ldots, T$, which establishes consistency of the SMC sampler at the usual Monte Carlo rate of $N^{-1/2}$ (Rebeschini and Van Handel 2015). Clearly, this upper bound is crude as it suggests that the errors accumulate exponentially fast in the number of time steps which is definitely not the case in practice. To obtain a bound that is *uniform* in T, a more refined analysis which exploits ergodicity properties of the Markov kernels $\{K_t\}_{t=1}^T$ is needed (Del Moral and Guionnet 2001).

Under appropriate integrability conditions, we also have CLT for estimates of expectation of test functions $\varphi : \mathcal{X} \to \mathbb{R}$

$$\sqrt{N}\left(\hat{\pi}_t^N(\varphi) - \pi_t(\varphi)\right) \xrightarrow{d.} \mathcal{N}(0, \sigma_t^2(\varphi)) \tag{1.61}$$

and normalizing constants

$$\sqrt{N}\left(\frac{\hat{Z}_t^N}{Z_t} - 1\right) \xrightarrow{d.} \mathcal{N}(0, \sigma_t^2) \tag{1.62}$$

for both the SIS algorithm and the SMC sampler. As mentioned earlier, for SIS it follows directly from (1.10) that

$$\sigma_t^2(\varphi) = Q_t\left((\varphi - \pi_t(\varphi))^2 (\mathrm{d}P_t/\mathrm{d}Q_t)^2\right) \tag{1.63}$$

and from (1.14) that

$$\sigma_t^2 = Q_t\left((\mathrm{d}P_t/\mathrm{d}Q_t)^2\right) - 1. \tag{1.64}$$

For general SMC methods with multinomial resampling performed at every time step, these results were established by Chopin (2004; Theorem 1), Del Moral (2004; Proposition 9.4.2) and tailored to the SMC sampler setup in Del Moral et al. (2006; Proposition 2).

We introduce some notation to write down expressions of these asymptotic variances compactly. For each $t = 1, \ldots, T$, we denote the time $k = 0, \ldots, t$ marginal distribution of $P_t(\mathrm{d}x_{0:t})$ as $P_{t,k}(\mathrm{d}x_k)$ and the conditional distribution of x_t given x_k

as $P_{t|k}(\mathrm{d}x_t|x_k)$. Note that the case $k = t$ is to be understood as $P_{t|t} := P_{t,t} = \pi_t$ and we use the notation $P_{t|k}(\varphi)(x_k) := \int_{\mathcal{X}} \varphi(x_t) P_{t|k}(\mathrm{d}x_t|x_k)$. We define the measures $\widetilde{P}_{t,k-1:k}(\mathrm{d}x_{k-1}, \mathrm{d}x_k) := P_{t,k}(\mathrm{d}x_k) L_{k-1}(x_k, \mathrm{d}x_{k-1})$ and $\widetilde{Q}_{k-1:k}(\mathrm{d}x_{k-1}, \mathrm{d}x_k) := \pi_{k-1}(\mathrm{d}x_{k-1}) K_k(x_{k-1}, \mathrm{d}x_k)$ on $(\mathcal{X} \times \mathcal{X}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X}))$ for $k = 1, \ldots, t$. In these notation, we have

$$\begin{aligned} \sigma_t^2(\varphi) = \pi_0 \left((P_{t|0}(\varphi) - \pi_t(\varphi))^2 (\mathrm{d}P_{t,0}/\mathrm{d}\pi_0)^2 \right) \\ + \sum_{k=1}^{t} \widetilde{Q}_{k-1:k} \left((P_{t|k}(\varphi) - \pi_t(\varphi))^2 (\mathrm{d}\widetilde{P}_{t,k-1:k}/\mathrm{d}\widetilde{Q}_{k-1:k})^2 \right) \end{aligned} \tag{1.65}$$

and

$$\sigma_t^2 = \pi_0 \left((\mathrm{d}P_{t,0}/\mathrm{d}\pi_0)^2 \right) - 1 + \sum_{k=1}^{t} \left\{ \widetilde{Q}_{k-1:k} \left((\mathrm{d}\widetilde{P}_{t,k-1:k}/\mathrm{d}\widetilde{Q}_{k-1:k})^2 \right) - 1 \right\}. \tag{1.66}$$

Observe that in these expressions, the form of the summands are similar to those in (1.63) and (1.64) with the proposal distribution Q_t in SIS replaced by $\widetilde{Q}_{k-1:k}$. This reveals that the resampling step has the effect of truncating the importance sampling error locally in time. Similar expressions of these asymptotic variances in the case of adaptive resampling and other resampling schemes can be found in Del Moral et al. (2012b) and Beskos et al. (2016).

1.3 Summary and looking ahead

We started this chapter by showing that a practical implementation of Bayesian inference relies on our ability to evaluate integrals. The use of Monte Carlo methods allows us to tackle the integration problem with (approximate) simulation from a target distribution.

We have seen that the efficiency of algorithms such as importance sampling (Section 1.2.2) and independent MH (Section 1.2.3.3) is very dependent on the construction of a proposal distribution that mimics the target distribution closely. Designing such *global* approximations is difficult in practice and particularly so for high dimensional problems.

On the other hand, although *local* strategies like RWMH and MALA (Section 1.2.3.3) are numerically more tractable and are backed by theoretical guarantees,

convergence of these methods can be excruciatingly slow in practice. For example, for highly *multi-modal* target distributions that are induced when performing Bayesian inference in mixture models (Section 3.5.1), these local algorithms typically get trapped in a local mode and therefore fail to characterize all posterior modes for any practical amount of compute time (Celeux et al. 2000).

It is well-known that combining such local schemes with *population* based approaches such as SMC samplers (Section 1.2.4) and population MCMC (Geyer 1991, Liang and Wong 2001), that exploit tempering to gradually introduce the complex structures of the likelihood, result in algorithms that can tackle challenging sampling problems in reasonable compute time (Jasra et al. 2007). We will therefore treat SMC sampler with local MCMC moves and the time reversed backward kernels (1.54) as *state-of-the-art* and compare it against sampling algorithms developed in this thesis. Following our earlier discussion, we will often refer to this approach as AIS and stress the additional resampling operation when necessary.

In this thesis, we explore two novel directions with the objective of developing more efficient SMC samplers. Firstly, we consider the use of *transport theory* to construct good global proposal distributions. Chapter 2 contains a review of existing methods and Chapter 3 focuses on the problem of constructing transport maps using flows. Secondly, in Chapter 4 we exploit ideas from *optimal control* to control a specific SMC sampler, with proposal distribution induced by a time-discretized Langevin dynamics, so as to minimize the Kullback-Leibler divergence of the extended target distribution from the proposal.

2

The transport problem

Contents

The purpose of this chapter is twofold. First, we introduce some background necessary to formulate the transport problem (Section 2.1-2.2). Second, we describe well-known ways to solve the transport problem (Section 2.3) and discuss existing

methods to compute them (Section 2.4).

2.1 Introduction

We begin with the notion of *coupling* the prior distribution π_0 with the posterior distribution π.

Definition 2.1. *Coupling $\pi_0, \pi \in \mathcal{P}(\mathcal{X})$ refers to the construction of random variables $(X_0, X_1) \sim Q \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ such that $Q(\cdot, \mathcal{X}) = \pi_0$ and $Q(\mathcal{X}, \cdot) = \pi_1$ as measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.*

The joint distribution Q is usually known as the *transport plan* and the couple (X_0, X_1) the *coupling* of π_0 and π. Clearly, couplings always exist since there is the *trivial* coupling where X_0 and X_1 are independent, i.e. $Q(\mathrm{d}x_0, \mathrm{d}x_1) = \pi_0(\mathrm{d}x_0)\pi(\mathrm{d}x_1)$. In contrast, we have a *deterministic* coupling if X_1 is a function of X_0.

Definition 2.2. *A coupling (X_0, X_1) of π_0 with π is said to be deterministic if there exists a function $T : \mathcal{X} \to \mathcal{X}$ such that $X_1 = T(X_0)$. We will refer to T as a transport map if it is additionally a C^1-diffeomorphism.*

The associated transport plan is given by $Q(\mathrm{d}x_0, \mathrm{d}x_1) = \pi_0(\mathrm{d}x_0)\delta_{T(x_0)}(\mathrm{d}x_1)$. For any $\mu \in \mathcal{P}(\mathcal{X})$ and measurable function $f : \mathcal{X} \to \mathcal{X}$, we define the push-forward of μ as the measure defined by $(f_{\#}\mu)(A) := \mu(f^{-1}(A))$ for all $A \in \mathcal{B}(\mathcal{X})$, where $f^{-1}(A) := \{x \in \mathcal{X} : f(x) \in A\}$ is the pre-image of A under f. In this notation, Definition 2.2 relates the measures via $\pi = T_{\#}\pi_0$.

Note that for any transport map T, we have the *change of variables formula*

$$\int_{\mathcal{X}} \varphi(x_1)\pi(\mathrm{d}x_1) = \int_{\mathcal{X}} \varphi(T(x_0))\, \pi_0(\mathrm{d}x_0), \tag{2.1}$$

for all $\varphi \in L^1(\pi)$. Setting $\varphi = \mathbb{1}_A$ in (2.1) for some $A \in \mathcal{B}(\mathcal{X})$ and applying change of variables in the integral gives

$$\int_A \pi(x_1)\,\mathrm{d}x_1 = \int_{T^{-1}(A)} \pi_0(x_0)\,\mathrm{d}x_0 = \int_A \pi_0(T^{-1}(x_1)) \left|\det\left(\nabla T^{-1}(x_1)\right)\right| \mathrm{d}x_1, \tag{2.2}$$

where $\nabla T^{-1} : \mathcal{X} \to \mathbb{R}^{d \times d}$ denotes the Jacobian matrix of T^{-1}. It follows that the densities are related via

$$\pi(x) = \pi_0(T^{-1}(x)) \left| \det \left(\nabla T^{-1}(x) \right) \right| \tag{2.3}$$

$\mathrm{d}x$-almost everywhere. It is often more convenient to invoke the inverse function theorem and compute the Jacobian determinant as $\det\left(\nabla T^{-1}(x)\right) = \det\left(\nabla T(T^{-1}(x))\right)^{-1}$. Equation (2.3) still holds even if the C^1-diffeomorphic assumption on T is reduced to an injective (or take multiplicity into account), locally Lipschitz map (see Evans and Gariepy 2015; Chapter 3).

It should be stressed that simulation from the posterior distribution would be straightforward if we had access to a transport map T: we simply need to sample $X_0 \sim \pi_0$, which we will assume to be possible since prior distributions are usually tractable, and then evaluate the transport map at this sample, i.e. $X_1 = T(X_0)$. Given π_0 and π, the problem of studying existence and regularity of diffeomorphisms T satisfying (2.3) is known as the *Jacobian problem* (Moser 1965, Dacorogna and Moser 1990). For such results to be computationally useful, they should also be *constructive*. The problem of constructing a transport map shall henceforth be referred to as the *transport problem*.

2.2 Optimal transport

The transport problem is typically underdetermined. To see this, consider the case $\pi_0 = \pi = \mathcal{N}(0_2, I_2) \in \mathcal{P}(\mathbb{R}^2)$. Clearly $T_\theta(x) := R(\theta)x$ is a transport map for any $\theta \in [0, 2\pi)$, where $R(\theta) := \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ is the clockwise rotation matrix of angle θ.

2.2.1 Monge's formulation

In the above example, intuitively, we would like to favour T_0 and penalize rotational solutions. This can be done by introducing a transportation cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and defining the functional

$$\mathcal{M}(T) := \int_{\mathcal{X}} c(x, T(x))\, \pi_0(\mathrm{d}x) \tag{2.4}$$

which associates each transport map T with a cost. We then consider the following minimization problem

$$\inf_{\{T\,:\,\pi=T_{\#}\pi_0\}} \mathcal{M}(T) \tag{2.5}$$

and define the *optimal* transport map as the map T^* that attains the minimum (if it exists). To guarantee existence of solutions to (2.5), the cost function c needs to be sufficiently varying to distinguish between transport maps.

The case $c(x, y) = |x - y|$ was first studied by French geometer *Gaspard Monge* in 1781 as a civil engineering problem, where the cost function represents the work needed to transport an amount of material extracted from a mine located at x to a construction site located at y. A rigorous treatment was given much later in Sudakov (1979) followed by Evans and Gangbo (1999), Trudinger and Wang (2001), Caffarelli et al. (2002), Ambrosio (2003).

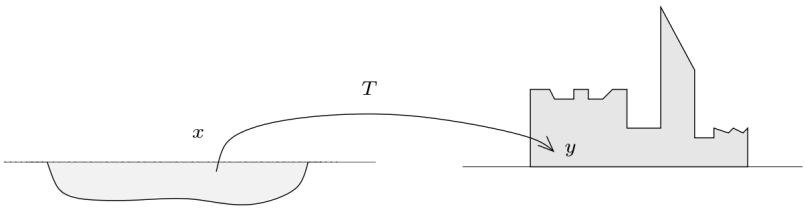

Figure 2.1: Illustration of Monge's civil engineering problem (figure taken from Villani (2008)).

2.2.2 Kantorovich's formulation

Years later, Russian mathematician *Leonid Kantorovich* proposed a relaxation of Monge's problem by allowing the mass leaving mine x to be split among several sites y and conversely mass arriving site y can also come from several possible mines x. In the Kantorovich (2006) formulation, we consider

$$\inf_{Q\in\mathcal{C}(\pi_0,\pi)} Q(c), \tag{2.6}$$

where $\mathcal{C}(\pi_0, \pi)$ is the set of all possible transport plans, and define the *optimal transport plan* as the distribution $Q^* \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ that attains the minimum (see

Villani (2008; Theorem 4.1) for conditions to guarantee existence). Note that in (2.6), both the cost functional and the marginal constraints are linear in the transport plan Q and in this sense may be thought of as the continuous analogue of *linear programming* for discrete state space problems. We will describe a method to compute Q^* which exploits this connection in Section 2.4.1.

2.2.3 Equivalence between Monge's and Kantorovich's formulation

Consider the case $c(x, y) = |x - y|^2$ and assume that π_0, π have finite second moments with dominating measure $\mathrm{d}x$ taken to be the Lebesgue measure. A classical result in the optimal transport literature (Villani 2008; Chapter 10) is that Monge's and Kantorovich's formulations admit the same unique solution in the sense that $\mathcal{M}(T^*) = Q^*(c)$ and $Q^*(A) = \pi_0(\{x \in \mathcal{X} : (x, T^*(x)) \in A\})$ for any $A \in \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X})$. Moreover, the optimal transport map is monotone and characterized by $T^* = \nabla\phi$ for some convex function ϕ.

It follows from (2.3) that ϕ should satisfy (in a suitable weak sense) the *Monge-Ampère* equation:

$$\pi\left(\nabla\phi(x)\right) = \pi_0(x)\left|\det\left(\nabla^2\phi(x)\right)\right|^{-1}, \tag{2.7}$$

where $\nabla^2\phi : \mathcal{X} \to \mathbb{R}^{d\times d}$ denotes the Hessian matrix of ϕ. Although numerical resolution of (2.7) has received a great deal of attention in recent years, developing accurate and efficient methods to solve this non-linear second-order elliptic PDE remains a challenge even in low dimensions – see Benamou and Brenier (2000) for a variational approach, Loeper and Rapetti (2005) for an algorithm based on Newton's method and Froese and Oberman (2011) for a finite difference solver.

Although efficient implementation still remains rather limited, the theoretical implications of the Monge-Kantorovich problem have been far-reaching; we refer to Villani (2003; 2008) for a modern account of the theory. An example of this is the metric $W_1(\pi_0, \pi) := Q^*(c)$ it defines on the subspace $\{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} c(x_0, x)\,\mu(\mathrm{d}x) < \infty\}$ for some arbitrary $x_0 \in \mathcal{X}$, commonly known as the

Wasserstein distance (of order 1), whenever c is a distance on $\mathcal{X}$ (Villani 2008; Chapter 6). Noting that we recover the total variation distance with cost function $c(x, y) = \mathbb{1}_{x \neq y}$ as the trivial metric on $\mathcal{X}$, we see that the Wasserstein distance with $c(x, y) = |x - y|^2$ can be an appealing alternative to the total variation distance as it reflects the geometry of the state space.

2.3 Knothe-Rosenblatt transport

From the above discussion, we see that imposing unicity in the transport problem comes at the price of tractability, i.e. absence of an explicit expression of the optimal transport map, and computability as solving the Monge-Ampère PDE numerically is challenging even in low dimensions. We now describe the *Knothe-Rosenblatt rearrangement*, better known to statisticians as the *conditional quantile transform*, which is a transport map that is explicit – although not necessarily optimal. This was introduced by Rosenblatt (1952) within a statistical context and rediscovered by Knothe (1957) when extending the Brunn–Minkowski inequalities. Throughout Section 2.3, we shall assume that $\mathcal{X} = \mathbb{R}^d$ and the dominating measure $\mathrm{d}x$ is the Lebesgue measure.

2.3.1 Optimal transport on $\mathbb{R}$

We shall first consider the one-dimensional case $d = 1$. Define the cumulative distribution function (CDF) of π_0 and π by

$$F(x) := \pi_0((-\infty, x]), \quad G(x) := \pi((-\infty, x]) \tag{2.8}$$

for $x \in \mathbb{R}$ and their generalized inverses

$$F^{-1}(u) := \inf\{x \in \mathbb{R} : F(x) \geq u\}, \quad G^{-1}(u) := \inf\{x \in \mathbb{R} : G(x) \geq u\} \tag{2.9}$$

for $u \in [0, 1]$, respectively. If π_0, π have finite second moments, by Villani (2003; Theorem 2.18), $T = G^{-1} \circ F$ is the unique monotone optimal transport map of the Monge-Kantorovich problem (since they coincide) with quadratic cost $c(x, y) = |x - y|^2$. Under some regularity assumption on the prior density and likelihood

function, more precisely that $\pi_0, L \in C^1(\mathbb{R}, \mathbb{R}_+)$, T is a C^1-diffeomorphism with derivative $\mathrm{d}T(x)/\mathrm{d}x = \pi_0(x)/\pi(T(x)) > 0$.

Moreover, using the facts that $x = F^{-1} \circ F(x)$ for π_0-almost everywhere x and the push-forward $F_\# \pi_0$ is the uniform distribution on $[0, 1]$, by the change of variables formula (2.1)

$$\begin{aligned} \inf_{Q \in \mathcal{C}(\pi_0, \pi)} Q(c) &= \inf_{\{T \,:\, \pi = T_\# \pi_0\}} \mathcal{M}(T) \\ &= \int_{\mathbb{R}} \left(x - G^{-1} \circ F(x) \right)^2 \pi_0(\mathrm{d}x) \\ &= \int_{\mathbb{R}} \left(F^{-1} \circ F(x) - G^{-1} \circ F(x) \right)^2 \pi_0(\mathrm{d}x) \\ &= \int_0^1 \left(F^{-1}(x) - G^{-1}(x) \right)^2 \mathrm{d}x, \end{aligned} \tag{2.10}$$

so we have an explicit expression of the optimal transportation cost. To gain some intuition, writing

$$\int_{-\infty}^{x} \pi_0(\mathrm{d}y) = \int_{-\infty}^{T(x)} \pi(\mathrm{d}y) \tag{2.11}$$

reveals that the transport is achieved by an increasing arrangement of the corresponding probability mass.

2.3.2 Increasing rearrangement in $\mathbb{R}^d$ for $d \geq 2$.

For any $\mu \in \mathcal{P}(\mathbb{R}^d)$, we write $\mu(\mathrm{d}x_1) \in \mathcal{P}(\mathbb{R})$ as the marginal distribution of the first component and $\mu(\mathrm{d}x_i | x_{1:i-1}) \in \mathcal{P}(\mathbb{R})$ as the increasing conditional distribution of x_i given $x_1, \ldots, x_{i-1}$ for $i = 2, \ldots, d$. We now consider generalizing the above transport map to dimensions $d \geq 2$. The construction is based on the following decompositions:

$$\pi_0(\mathrm{d}x) = \pi_0(\mathrm{d}x_1) \prod_{i=2}^{d} \pi_0(\mathrm{d}x_i | x_{1:i-1}), \quad \pi(\mathrm{d}x) = \pi(\mathrm{d}x_1) \prod_{i=2}^{d} \pi(\mathrm{d}x_i | x_{1:i-1}). \tag{2.12}$$

We shall assume that we have access to the first marginals and the increasing conditionals in (2.12) and define the respective CDFs

$$\begin{aligned} F_1(x_1) &:= \pi_0((-\infty, x_1] \times \mathbb{R}^{d-1}), \quad G_1(x_1) := \pi((-\infty, x_1] \times \mathbb{R}^{d-1}), \\ F_i(x_i | x_{1:i-1}) &:= \pi_0((-\infty, x_i] | x_{1:i-1}), \quad G_i(x_i | x_{1:i-1}) := \pi((-\infty, x_i] | x_{1:i-1}) \end{aligned} \tag{2.13}$$

for $i = 2, \ldots, d, x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ and their right-continuous inverses

$$\begin{aligned} F_1^{-1}(u_1) &:= \inf\{x_1 \in \mathbb{R} : F_1(x_1) \geq u_1\}, \\ G_1^{-1}(u_1) &:= \inf\{x_1 \in \mathbb{R} : G_1(x_1) \geq u_1\}, \\ F_i^{-1}(u_i|x_{1:i-1}) &:= \inf\{x_i \in \mathbb{R} : F_i(x_i|x_{1:i-1}) \geq u_i\}, \\ G_i^{-1}(u_i|x_{1:i-1}) &:= \inf\{x_i \in \mathbb{R} : G_i(x_i|x_{1:i-1}) \geq u_i\}, \end{aligned} \tag{2.14}$$

for $i = 2, \ldots, d, u = (u_1, \ldots, u_d)^T \in [0,1]^d$. The main idea is to apply the optimal transport in the one dimensional case described earlier to transport the marginal distributions in the first component and the corresponding increasing conditional distributions. This gives the *Knothe-Rosenblatt* transport map $x \mapsto T(x) = (T_1(x_1), \ldots, T_d(x_{1:d}))^T$ defined by

$$\begin{aligned} y_1 &:= T_1(x_1) := G_1^{-1}(F_1(x_1)), \\ y_i &:= T_i(x_{1:i}) := G_i^{-1}(F_i(x_i|x_{1:i-1})|y_{1:i-1}), \end{aligned} \tag{2.15}$$

for $i = 2, \ldots, d$. Note that T is monotone and *lower triangular*, i.e. the i^{th} component T_i depends only on the first i variables $x_1, \ldots, x_i$. As before, if we assume that $\pi_0, L \in C^1(\mathbb{R}, \mathbb{R}_+)$, then T is a C^1-diffeomorphism with a lower triangular Jacobian matrix that has positive entries on its diagonal.

While the Knothe-Rosenblatt transport is not optimal, interestingly, it is possible to represent it as the limit of optimal transports. More precisely, assume π_0, π have finite second moments and let T_ε be the optimal transport map of the Monge-Kantorovich problem with weighted quadratic cost

$$c_\varepsilon(x, y) = \sum_{i=1}^{d} \alpha_i(\varepsilon)(x_i - y_i)^2 \tag{2.16}$$

for some positive weights $\{\alpha_i\}_{i=1}^d$ that depend on the parameter $\varepsilon > 0$. By Carlier et al. (2010; Theorem 2.1), under the constraint that

$$\lim_{\varepsilon \to 0} \frac{\alpha_i(\varepsilon)}{\alpha_{i-1}(\varepsilon)} = 0 \tag{2.17}$$

for $i = 2, \ldots, d$, the map T_ε converges in $L^2(\pi_0)$ to the Knothe-Rosenblatt transport map T defined in (2.15).

Although the Knothe-Rosenblatt transport can be useful in certain scenarios, its broad application is very much limited by the requirement that the marginal and increasing conditional distributions of π_0 and π are known. Moreover, for the sole purpose of simulating from the target distribution π, the way in which the transport is constructed is superfluous since the push-forward of π_0 by the map $x \mapsto F(x) = (F_1(x_1), \dots, F_d(x_d|x_{1:d-1}))^T$ is simply the uniform distribution on the unit hypercube $[0,1]^d$, which is straightforward to simulate from. Lastly, it is also worth noting that the Knothe-Rosenblatt transport has the undesirable property of not being invariant under permutation of the coordinates.

2.4 Computing a transport

We now turn our attention to the computation of a transport map. The objective of this section is to discuss a few strategies that have been proposed in various literatures.

2.4.1 A discrete approach

The methods described in Section 2.4.1 can be thought of a discrete approach to the transport problem. We begin by re-formulating resampling in the framework of coupling empirical measures. Although the optimal coupling exhibits attractive properties, exact computation is costly. Therefore we discuss an entropic regularization which offers reduced complexity. Lastly, we describe a transformation of the optimal coupling that converges to the optimal Monge-Kantorovich transport between π_0 and π.

2.4.1.1 Resampling as coupling

In applications such as *data assimilation* and *non-linear filtering*, although the prior distribution is intractable, we have access to $N \in \mathbb{N}$ iid samples $\{X_0^n\}_{n=1}^N$ from π_0. The corresponding empirical measure is given by $\hat{\pi}_0^N = N^{-1} \sum_{n=1}^N \delta_{X_0^n}$. In this scenario, to perform Bayes' update, one usually uses the prior as the proposal

distribution in an importance sampling approximation (Gordon et al. 1993). This gives the following particle approximation of π

$$\hat{\pi}^N = \sum_{n=1}^{N} W^n \delta_{X_0^n}, \quad W^n := \frac{L(X_0^n)}{\sum_{m=1}^{N} L(X_0^m)}. \tag{2.18}$$

For notational convenience, we write $W := (W^1, \ldots, W^N)^T \in \mathbb{R}_+^N$ as the vector of normalized weights, $1_N := (1, \ldots, 1)^T \in \mathbb{R}^N$ as the vector of ones, $\Omega_N := \{X_0^1, \ldots, X_0^N\} \in \mathcal{X}^N$ as the discrete state space of interest and $U := N^{-1}1_N$ as the uniform distribution on Ω_N. In this setup, $\mathcal{P}(\Omega_N) = \{q \in \mathbb{R}_+^N : q^T 1_N = 1\}$ is the open $(N-1)$-dimensional simplex and

$$\mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N) = \left\{Q \in \mathbb{R}_+^{N \times N} : Q1_N = U, Q^T 1_N = W\right\} \subset \mathcal{P}(\Omega_N \times \Omega_N) \tag{2.19}$$

is the transport polytope. Observe that for any $Q \in \mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N)$, sampling X_1^n from the transition probability associated to the n^{th} particle

$$N(Q_{n,1}, \ldots, Q_{n,N})^T \in \mathcal{P}(\Omega_N) \tag{2.20}$$

for all $n = 1, \ldots, N$ and returning the empirical measure $\bar{\pi}^N = N^{-1} \sum_{n=1}^{N} \delta_{X_1^n}$ yields a *valid* resampling scheme (Section 1.2.4.3) due to the marginal constraint $Q^T 1_N = W$. Hence resampling can be viewed as constructing a coupling and many popular resampling schemes can be casted into this framework (Reich 2013); for example multinomial resampling corresponds to the independent coupling with $Q = UW^T$.

Let $C \in \mathbb{R}^{N \times N}$ denote the matrix of pairwise squared Euclidean distances, i.e. $C_{n,m} = |X_0^n - X_0^m|^2$ for $n, m = 1, \ldots, N$, and $\langle \cdot, \cdot \rangle_F$ the Frobenius inner product on $\mathbb{R}^{N \times N}$. In these notation, the Monge-Kantorovich problem (2.6) between $\hat{\pi}_0^N$ and $\hat{\pi}^N$ can be written as

$$Q^* = \text{argmin}_{Q \in \mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N)} \langle C, Q \rangle_F, \tag{2.21}$$

assuming a unique optimum exists. The minimization (2.21) is a *linear program* which can be solved exactly using *network simplex* or *interior point* methods. In contrast to resampling schemes that are commonly used, Q^* also takes *particle locations* into account and this feature can be exploited to couple particle filters (Jacob et al. 2016). Between particles of similar weights, Q^* also has the interesting property of favouring particles that are more distant.

2.4.1.2 Entropic regularization

Despite its nice features, computing Q^* exactly using any of the above-mentioned methods comes with a huge computational cost of at least $O(N^3 \log N)$ operations (see Pele and Werman 2009; Section 2.1). Cuturi (2013) considered the following regularization

$$Q^\alpha := \text{argmin}_{Q \in \mathcal{C}_\alpha(\hat{\pi}_0^N, \hat{\pi}^N)} \langle C, Q \rangle_F \tag{2.22}$$

for some $\alpha \in [0, \infty]$, where $\mathcal{C}_\alpha(\hat{\pi}_0^N, \hat{\pi}^N) := \left\{ Q \in \mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N) : \text{KL}\left(Q | UW^T\right) \leq \alpha \right\}$. This reduces the computational complexity to $O(N^2)$ which, in the absence of additional structure, is the best one can ask for as computing the cost matrix C and sampling from any joint distribution $Q \in \mathcal{P}(\Omega_N \times \Omega_N)$ already require $O(N^2)$ operations. The convex optimization problem (2.22) can be solved efficiently using Sinkhorn (1967) fixed point iteration which is known to converge linearly at a rate that depends on the value of the regularization parameter α. Truncating at finite iterations yields an approximation $\hat{Q}$ of Q^α that might not be an element of $\mathcal{C}_\alpha(\hat{\pi}_0^N, \hat{\pi}^N)$ and thus leads to a possibly invalid resampling scheme. However, as noted by Jacob et al. (2016; Section 3.2), it is possible to construct a transport plan $Q \in \mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N)$ that is close to $\hat{Q}$.

This regularization is best understood in entropic terms. For any $q \in \mathcal{P}(\Omega_N)$ and $Q \in \mathcal{P}(\Omega_N \times \Omega_N)$, we define their entropies by $h(q) := -\sum_{n=1}^N q_n \log q_n$ and $h(Q) := -\sum_{n,m=1}^N Q_{n,m} \log Q_{n,m}$, respectively. It can be shown that

$$h(Q) \leq h(U) + h(W) \tag{2.23}$$

for all $Q \in \mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N)$, with equality attained at the independent coupling $Q = UW^T$ (Cover and Thomas 2012; Chapter 2). In terms of entropy, the constraint in (2.22) is equivalent to

$$h(Q) \geq h(U) + h(W) - \alpha. \tag{2.24}$$

At $\alpha = 0$, we have the independent coupling $Q^0 = UW^T$ which maximizes entropy, and as α increases, solutions that have less entropy become admissible. It is well-known in linear programming that the optimal transport plan Q^* always occurs

at an extreme point of the transport polytope and has low entropy in the sense of having at most $2N-1$ non-zero elements (Brualdi 2006; Corollary 8.1.3). Therefore when α is large enough, we have $Q^\alpha = Q^*$ as the Kullback-Leibler ball $\mathcal{C}_\alpha(\hat{\pi}_0^N, \hat{\pi}^N)$ overlaps with $\mathcal{C}(\hat{\pi}_0^N, \hat{\pi}^N)$ in a neighbourhood of Q^* (Cuturi 2013; Property 1).

While it is tempting to solve the regularized problem with a large value of α in the hope of recovering Q^*, as discussed in Cuturi (2013), this causes numerical instabilities and convergence of the Sinkhorn's iteration will become excruciatingly slow with α large.

2.4.1.3 Convergence to optimal transport

After computing the optimal transport plan Q^*, instead of sampling from the transition probabilities (2.20) to obtain the coupling as described above, Reich (2013) considered the following linear transformation

$$X_1^n := T_N(X_0^n) := \sum_{m=1}^N P_{n,m}^* X_0^m \tag{2.25}$$

with Markov transition matrix $P^* := NQ^*$. The rationale here is to exploit the low entropic behaviour of Q^*: the transition probabilities in (2.20) are quasi-deterministic, i.e. has positive probability mass on only a few points in Ω_N.

Noting that we have

$$\frac{1}{N}\sum_{n=1}^N X_1^n = \sum_{m=1}^N \sum_{n=1}^N Q_{n,m}^* X_0^m = \sum_{m=1}^N W^m X_0^m \tag{2.26}$$

due to the marginal constraint $Q^{*T} 1_N = W$, by consistency of normalized importance sampling (Section 1.2.2.1), it follows that this procedure maintains consistency of the posterior mean. Moreover, as the empirical measures $\hat{\pi}_0^N$ and $\hat{\pi}^N$ are consistent approximations of π_0 and π, respectively, as $N \to \infty$ we might expect T_N to converge in a suitable sense to the optimal transport of the Monge-Kantorovich problem between π_0 and π with quadratic cost. This is the main result in Reich (2013; Theorem 3.2).

Returning to the regularized transport plan Q^α in (2.22), we note that its $N \to \infty$ limit will not be a transport due to the entropic regularization. We

conjecture that this limit will be the minimizer

$$Q(\sigma^2) := \text{argmin}_{Q \in \mathcal{C}(\pi_0, \pi)} \text{KL}(Q|P), \tag{2.27}$$

where $P(\mathrm{d}x_0, \mathrm{d}x_1) := \mu(\mathrm{d}x_0)\mathcal{N}(x_1; x_0, \sigma^2 I_d)\mathrm{d}x_1 \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ for some arbitrary $\mu \in \mathcal{P}(\mathcal{X})$ and noise parameter $\sigma^2 > 0$ that depends on the value of the regularization parameter α. The minimization (2.27), known as the *Schrödinger bridge* problem (Schrödinger 1931, Jamison 1975), can be thought of as a large deviation principle with marginal constraints. Existence and uniqueness of this minimizer was studied in Beurling (1960) and Jamison (1974). We note that Gaussianity of the transition kernel associated to P is inherited from the use of squared Euclidean distance in (2.22). Mikami (2004) established that as $\sigma^2 \to 0$, $Q(\sigma^2)$ converges to the optimal Monge-Kantorovich transport between π_0 and π; this zero noise limit is analogous to the above discussion on the $\alpha \to \infty$ limit in discrete state spaces.

2.4.2 A global approach

We now return to continuous state spaces and outline the main ideas behind a series of papers by Moselhy and Marzouk (2012), Parno and Marzouk (2014), Marzouk et al. (2016), where the task of constructing a transport map was formulated as a variational problem.

2.4.2.1 Variational formulation

Consider the minimization problem

$$\inf_{T \in \mathsf{T}} \text{KL}(T_{\#}\pi_0|\pi) \tag{2.28}$$

where T is a set of mappings whose properties are to be specified. Although the functional $T \mapsto \text{KL}(T_{\#}\pi_0|\pi)$ is not amenable to Monte Carlo approximation, if we additionally impose that the maps in T are C^1-diffeomorphisms, then using the change of variables formula (2.1), (2.28) can be written as

$$\inf_{T \in \mathsf{T}} \text{KL}(\pi_0|T^{-1}_{\#}\pi). \tag{2.29}$$

As $T \mapsto \mathrm{KL}(\pi_0 | T_\#^{-1}\pi)$ is an expectation with respect to the prior distribution, we can use the basic Monte Carlo estimator (1.3):

$$\inf_{T \in \mathsf{T}} \frac{1}{N} \sum_{n=1}^{N} \log \left(\frac{\mathrm{d}\pi_0}{\mathrm{d}T_\#^{-1}\pi} \right) (X^n) \tag{2.30}$$

where $\{X^n\}_{n=1}^N$ are $N \in \mathbb{N}$ iid samples from π_0. We note that the variance of this estimator will depend on how close the push-forward measure $T_\#^{-1}\pi$ is from the prior distribution and that it is possible to use results from statistical learning (Vapnik 1998) to get bounds on the expected difference between (2.29) and its empirical counterpart (2.30) as a function of N and properties of T (see also Appendix B).

2.4.2.2 Approximating Knothe-Rosenblatt

With the aim of approximating the Knothe-Rosenblatt transport described in Section 2.3, the authors proposed to select T as the set $\mathsf{T}_+^\triangle$ of lower triangular maps $x \mapsto T(x) = (T_1(x_1), \ldots, T_d(x_{1:d}))^T$ satisfying the following positivity condition

$$\partial_{x_i} T_i(x_{1:i}) > 0, \quad \text{for } i = 1, \ldots, d \text{ and } x = (x_1, \ldots, x_d)^T \in \mathcal{X}, \tag{2.31}$$

which enforces invertibility of an admissible triangular map. Applying (2.3) to obtain the density of $T_\#^{-1}\pi$ and dropping terms that are irrelevant for the optimization gives

$$\inf_{T \in \mathsf{T}_+^\triangle} \sum_{n=1}^{N} - \log \gamma(T(X^n)) - \log \det \left(\nabla T(X^n) \right), \tag{2.32}$$

where $\gamma : \mathcal{X} \to \mathbb{R}_+$ denotes the unnormalized density of π. For the minimization in (2.32) to be numerically tractable, one has to approximate $\mathsf{T}_+^\triangle$ with a finite dimensional subspace. To do so, the authors considered a parameterization given by a multivariate polynomial basis expansion:

$$T_i(x_{1:i}; \beta_i) = \sum_{m=1}^{M} \beta_{i,m} \phi_{i,m}(x_{1:i}) \tag{2.33}$$

for $i = 1, \ldots, d$ and $M \in \mathbb{N}$, where $\beta = (\beta_{i,j}) \in \mathbb{R}^{d \times M}$ are the parameters to be inferred and each multivariate polynomial has a tensor product form $\phi_{i,m}(x_{1:i}) = \prod_{j=1}^{i} \varphi_{\alpha(i,m)_j}(x_j)$ with the index function $(i, m) \mapsto \alpha(i, m) \in \mathbb{N}_0^i$ prescribing the degree of the univariate polynomial φ. The authors advocate designing α to exploit

any *conditional independence* structure in the target distribution (Parno et al. 2015) and using the *polynomial chaos* expansion (Xiu and Karniadakis 2002) whenever polynomials that are orthogonal to the prior distribution can be easily evaluated.

Observe that while the parameterization (2.33) has the desired lower triangular structure by construction, the positivity condition (2.31) might not be satisfied for all values of $\beta \in \mathbb{R}^{d\times M}$ and $x \in \mathcal{X}$. As a partial fix, the authors proposed to enforce this condition at the sampled points:

$$\partial_{x_i} T_i(X^n_{1:i}; \beta_i) = \sum_{m=1}^{M} \beta_{i,m} \partial_{x_i} \phi_{i,m}(X^n_{1:i}) > 0, \quad \text{for } i = 1, \dots, d \text{ and } n = 1, \dots, N. \tag{2.34}$$

This imposes a finite set of linear constraints on the minimization problem (2.32). Hence in terms of the parameters, we seek

$$\inf_{\{\beta\in\mathbb{R}^{d\times M}\,:\,(2.34)\text{ holds}\}} \sum_{n=1}^{N} -\log \gamma(T(X^n;\beta)) - \log\det\left(\nabla T(X^n;\beta)\right), \tag{2.35}$$

which is a non-convex (constrained) optimization problem in general. As noted by Kim et al. (2013), since the parameterized map $\beta \mapsto T(x;\beta)$ is linear in the parameters, under the assumption of log-concavity of the prior density and likelihood function, (2.35) is a convex optimization problem.

2.4.2.3 Composite maps

As the prior and posterior distributions become more distant, we expect the above procedure to require polynomials of higher degree in order to obtain a map that is reasonably close to being a transport. Therefore the computational effort of the resulting algorithm would increase as this leads to a larger number of parameters in the optimization problem (2.35). In this situation, using likelihood tempering (Section 1.2.4.1) to introduce a sequence of bridging distributions $\{\pi_t\}_{t=0}^{T}$ can potentially reduce the overall computational complexity.

One possibility here is to exploit the fact that successive distributions are closer, so this allows the algorithm to move between distributions with polynomials of lower degree. More precisely, for each $t = 1, \dots, T$, we construct the map S_t that

aims to transport π_{t-1} to π_t, i.e. $\pi_t \approx S_{t\#}\pi_{t-1}$, and define the composite map $S := S_T \circ \cdots \circ S_1$ to approximate the desired transport.

In Moselhy and Marzouk (2012), the authors considered another possibility which uses the structure of the optimization problem. The difference is to build the t^{th} composite map $S^t := S_t \circ \cdots \circ S_1$ by minimizing $\mathrm{KL}(S^t_\# \pi_0 | \pi_t)$, whilst having fixed all coefficients associated with preceding maps $S_1, \ldots, S_{t-1}$ that have already been determined. Their argument is that unlike the previous approach, where errors might accumulate if intermediate transports are not computed exactly, building an approximation of the desired transport $S := S^T$ in this manner allows preceding maps $S_1, \ldots, S_{T-1}$ to be constructed with looser tolerances while still maintaining accuracy of S if the final map S_T is computed with tighter error tolerance.

2.4.3 An iterative approach

We have seen in the previous section that seeking a transport map directly is a difficult task. In the following, we present an idea from Liu and Wang (2016) which can be seen as a way to build a transport iteratively. Consider the following re-formulation of the minimization problem in (2.28):

$$\inf_{f \in \mathsf{F}} \mathrm{KL}((\mathrm{Id} + f)_\# \pi_0 | \pi) \tag{2.36}$$

where $\mathrm{Id}(x) := x$ is the identity map and F is a set of possible perturbations to be specified. The main idea is to work with a tractable class F and select a perturbation direction $f \in \mathsf{F}$ that decreases the Kullback-Leibler divergence.

2.4.3.1 Gradient descent in a reproducing kernel Hilbert space

Suppose we have a positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and let H denote the *reproducing kernel Hilbert space* (RKHS) with inner product $\langle \cdot, \cdot \rangle_{\mathsf{H}}$ for which k is the reproducing kernel, i.e. $\langle \varphi, k(\cdot, x) \rangle_{\mathsf{H}} = \varphi(x)$ for all $\varphi \in \mathsf{H}$. We note that existence of a unique H is guaranteed by the *Moore-Aronszajn* theorem. We define $\mathsf{F} := \mathsf{H}^d$ as the Cartesian product of vector-valued functions $f = (f_1, \ldots, f_d)^T$ with $f_i \in \mathsf{H}$ for all $i = 1, \ldots, d$, and note that this is also a Hilbert space equipped

with the inner product $\langle f, g\rangle_{\mathsf{F}} := \sum_{i=1}^{d}\langle f_i, g_i\rangle_{\mathsf{H}}$ for $f, g \in \mathsf{F}$. For any functional $F : \mathsf{F} \to \mathbb{R}$, we write $\nabla F(f)$ as the functional gradient of F at f, defined by $F(f + \varepsilon g) = F(f) + \varepsilon\langle \nabla F(f), g\rangle_{\mathsf{F}} + O(\varepsilon^2)$ for all $g \in \mathsf{F}, \varepsilon \in \mathbb{R}$.

By Liu and Wang (2016; Theorem 3.3), $f^{\pi_0} = (f_1^{\pi_0}, \ldots, f_d^{\pi_0})^T$ defined as $f_i^{\pi_0}(x) := \pi_0\left(k(x, \cdot)\partial_{x_i} \log \pi(\cdot) + \partial_{x_i} k(x, \cdot)\right)$ for $i = 1, \ldots, d$ satisfies

$$f^{\pi_0} = -\nabla \mathrm{KL}((\mathrm{Id} + f)_{\#}\pi_0|\pi)\Big|_{f=0}. \tag{2.37}$$

Moreover, the squared norm of $f^{\pi_0} \in \mathsf{F}$ equals to the *Stein discrepancy* between π_0 and π that is induced by the RKHS (see Oates et al. 2017). This result prompts the following iterative *gradient descent* algorithm

$$\pi_t = (\mathrm{Id} + \Delta_t f^{\pi_{t-1}})_{\#}\, \pi_{t-1} \tag{2.38}$$

with a sequence of sufficiently small step sizes $\{\Delta_t\}$ and initialization at the prior distribution. Note that the iteration (2.38) is typically intractable as an evaluation of the functional gradient $f^{\pi_{t-1}}$ requires computing an expectation with respect to π_{t-1}. We can form a particle approximation of (2.38) by iterating

$$\hat{\pi}_t^N = \left(\mathrm{Id} + \Delta_t f^{\hat{\pi}_{t-1}^N}\right)_{\#} \hat{\pi}_{t-1}^N \tag{2.39}$$

with initialization at $\hat{\pi}_0^N = N^{-1}\sum_{n=1}^{N}\delta_{X_0^n}$, where $\{X_0^n\}_{n=1}^N$ are $N \in \mathbb{N}$ iid samples from π_0. This gives $\hat{\pi}_t^N = N^{-1}\sum_{n=1}^{N}\delta_{X_t^n}$ with the following particle update

$$X_t^n = X_{t-1}^n + \frac{\Delta_t}{N}\sum_{m=1}^{N}\left(k(X_{t-1}^n, X_{t-1}^m)\nabla \log \pi(X_{t-1}^m) + \nabla k(X_{t-1}^n, X_{t-1}^m)\right) \tag{2.40}$$

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

To build some intuition for (2.40), we consider the commonly used *radial basis function* $k(x, y) = \exp\left(-|x - y|^2/h\right)$ for some bandwidth $h > 0$. The first term in the update can be thought of as a *smoothed* version of the gradient of the log-posterior density, obtained by a weighted average of the particles' gradients against the kernel function. Noting that $\nabla k(x, y) = 2(x - y)k(x, y)/h$, we observe that the second term acts as a *repulsive force* between particles with the bandwidth h controlling the strength of this repulsion. Since (2.40) reduces to gradient descent

on the log-posterior density when $N = 1$, it follows that the repulsive interaction between particles is crucial in preventing the particle approximation from collapsing to the local modes of π. In summary, we see that the qualitative behaviour of the resulting approximation will be particularly sensitive to both the bandwidth and the number of particles used.

3

The flow transport problem

Contents

Throughout this chapter, we will consider the case $\mathcal{X} = \mathbb{R}^d$ for ease of presentation and assume that $\pi_0, L \in C^1(\mathbb{R}^d, \mathbb{R}_+)$. In Section 3.1, we begin by introducing a continuous curve of probability measures connecting the prior and posterior distributions by tempering the likelihood. This leads naturally to the question of designing particle dynamics to track changes in the underlying distributions along the curve. We formulate the problem of constructing transport maps using flows in Section 3.2 and survey existing methods to solve this problem before giving our main results in Section 3.3. We then discuss implementation issues in Section 3.4 and conclude with some applications in Section 3.5.

3.1 A curve from prior to posterior

We return to the discussion in Section 1.2.4.1 on likelihood tempering and consider the limit as the number of bridging distributions in (1.37) goes to infinity. This amounts to defining a curve of probability measures $\mathcal{C}_\pi = \{\pi_t\}_{t\in[0,1]}$, smoothly bridging the prior π_0 to the posterior $\pi_1 := \pi$, by gradually introducing the likelihood using a strictly increasing C^1-function $\lambda : [0,1] \to [0,1]$ such that $\lambda(0) = 0$ and $\lambda(1) = 1$:

$$\pi_t(\mathrm{d}x) := \frac{\pi_0(\mathrm{d}x)L(x)^{\lambda(t)}}{Z(t)} \tag{3.1}$$

with

$$Z(t) := \int_{\mathbb{R}^d} \pi_0(\mathrm{d}x)L(x)^{\lambda(t)}. \tag{3.2}$$

We will denote the time derivative of λ by $\lambda' : [0,1] \to \mathbb{R}_+$ and henceforth assume that $\log L \in L^1(\pi_t)$ for all $t \in [0,1]$. The rationale of moving to the continuum is that this allows us to study the time evolution of π_t along the curve $\mathcal{C}_\pi$.

3.1.1 Properties

We first establish some simple properties of the normalization function $t \mapsto Z(t)$.

Lemma 3.1. *The normalization function $Z : [0,1] \to \mathbb{R}_+$ defined in (3.2) lies in $C^1([0,1], \mathbb{R}_+)$ and*

$$\frac{\mathrm{d}}{\mathrm{d}t} Z(t) = \int_{\mathbb{R}^d} \pi_0(\mathrm{d}x) \partial_t L(x)^{\lambda(t)} = \lambda'(t) \int_{\mathbb{R}^d} \pi_0(\mathrm{d}x) L(x)^{\lambda(t)} \log L(x). \tag{3.3}$$

Proof. Let $\{t_m\}_{m\in\mathbb{N}} \subset [0,1]$ be a sequence such that $t_m \to t_*$. Note that $\pi_0(x)L(x)^{\lambda(t_m)}$ converges pointwise to $\pi_0(x)L(x)^{\lambda(t_*)}$ and

$$\pi_0(x)L(x)^{\lambda(t)} \leq \pi_0(x) \sup_{t\in[0,1]} L(x)^{\lambda(t)} \leq \pi_0(x)\,(1 + L(x)) \tag{3.4}$$

holds for all $(t,x) \in [0,1] \times \mathbb{R}^d$. Hence applying dominated convergence theorem with dominating function $x \mapsto \pi_0(x)\,(1 + L(x))$, which is integrable since $Z < \infty$, establishes continuity.

For differentiability, let $\{t_m\}_{m\in\mathbb{N}}$ be a sequence such that $t_m \to t_*$ and $t_m \in [0,1] \setminus \{t_*\}$ for each $m \in \mathbb{N}$. We write

$$\frac{Z(t_m) - Z(t_*)}{t_m - t_*} = \int_{\mathbb{R}^d} f_m(x)\,\mathrm{d}x \tag{3.5}$$

with

$$f_m(x) := \frac{\pi_0(x)\left(L(x)^{\lambda(t_m)} - L(x)^{\lambda(t_*)}\right)}{t_m - t_*}. \tag{3.6}$$

Since $\lambda \in C^1([0,1],[0,1])$, $f_m(x)$ converges pointwise to $\pi_0(x)\partial_t L(x)^{\lambda(t)}|_{t=t_*} = \pi_0(x)\lambda'(t_*)L(x)^{\lambda(t_*)} \log L(x)$ for each $x \in \mathbb{R}^d$. Applying the mean value theorem to $t \mapsto L(x)^{\lambda(t)}$ gives $f_m(x) = \pi_0(x)\lambda'(c_m)L(x)^{\lambda(c_m)} \log L(x)$ for some c_m between t_m and t_*. Note that

$$|f_m(x)| \leq \pi_0(x)|\log L(x)| \sup_{t\in[0,1]} \lambda'(t)L(x)^{\lambda(t)} \leq \pi_0(x)|\log L(x)|\|\lambda'\|_\infty(1 + L(x)) \tag{3.7}$$

holds for all $(m,x) \in \mathbb{N} \times \mathbb{R}^d$. Differentiability and (3.3) follows from another application of dominated convergence theorem with dominating function $x \mapsto \pi_0(x)|\log L(x)|\|\lambda'\|_\infty\,(1+L(x))$, which is integrable since $\log L \in L^1(\pi_t)$ for $t = 0, 1$. Continuity of the derivative of $Z(t)$ follows using similar arguments. □

Intuitively, we expect the curve of probability measures $\mathcal{C}_\pi$ to be "continuous" in some sense. This is made precise in the following.

Lemma 3.2. *The curve of probability measures $\mathcal{C}_\pi = \{\pi_t\}_{t\in[0,1]}$ defined in (3.1) is narrowly continuous, i.e. for any bounded function $\varphi \in C(\mathbb{R}^d, \mathbb{R})$ and any sequence $\{t_m\}_{m\in\mathbb{N}} \subset [0,1]$ such that $t_m \to t_*$ we have*

$$\lim_{m\to\infty} \pi_{t_m}(\varphi) = \pi_{t_*}(\varphi). \tag{3.8}$$

Proof. Continuity of $\lambda(t)$ and $Z(t)$ implies that $t \mapsto \pi_t(x) \in C([0,1], \mathbb{R}_+)$ for each $x \in \mathbb{R}^d$. Hence for any bounded function $\varphi \in C(\mathbb{R}^d, \mathbb{R})$ and any sequence $\{t_m\}_{m\in\mathbb{N}} \subset [0,1]$ such that $t_m \to t_*$, we have $\varphi(x)\pi_{t_m}(x) \to \varphi(x)\pi_{t_*}(x)$ pointwise. Note that

$$|\varphi(x)\pi_{t_m}(x)| \le \|\varphi\|_\infty \pi_0(x) \frac{\sup_{t\in[0,1]} L(x)^{\lambda(t)}}{\inf_{t\in[0,1]} Z(t)} \tag{3.9}$$

holds for all $(m, x) \in \mathbb{N} \times \mathbb{R}^d$. Since $Z(t)$ is continuous on $[0,1]$ by Lemma 3.1, the infimum in (3.9) is attained and is strictly positive under positivity assumptions made on the prior density and likelihood function. Hence the upper bound in (3.9) is integrable and (3.8) follows from an application of the dominated convergence theorem. □

3.1.2 Time evolution along curve

By differentiating the density of (3.1) with respect to the pseudo-time variable t, we obtain

$$\partial_t \pi_t(x) = \lambda'(t) \left(\log L(x) - I_t\right) \pi_t(x), \tag{3.10}$$

for each $x \in \mathbb{R}^d$ where

$$I_t := \frac{1}{\lambda'(t)} \frac{\mathrm{d}}{\mathrm{d}t} \log Z(t) = \frac{\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} \pi_0(x) L(x)^{\lambda(t)} \,\mathrm{d}x}{\lambda'(t) Z(t)} = \pi_t(\log L). \tag{3.11}$$

The last equality in (3.11) requires the validity of interchanging the order of differentiation with respect to the time variable and integration with respect to the spatial variable which is justified by Lemma 3.1. From a statistical perspective,

this assumes that the family of models $\{\pi_t\}_{t\in[0,1]}$ is regular so (3.11) is simply a consequence of the fact that the score function has zero expectation for each $t \in [0,1]$. Since $Z(0) = 1$, integrating (3.11) on $[0,1]$ recovers the identity

$$\log Z = \int_0^1 \lambda'(t)\, I_t \,\mathrm{d}t \tag{3.12}$$

commonly known as *path sampling* in statistics (Gelman and Meng 1998) and *thermodynamic integration* in molecular dynamics where $\log Z$ represents the free energy difference between two macroscopic states of a system (Frenkel and Smit 2001).

Equation (3.10) reveals that the expected log-likelihood I_t plays the role of a *reference value* which controls the time evolution of the density $\pi_t(x)$, i.e. in logarithmic scale, the local behaviour around a point $x \in \mathbb{R}^d$ is such that there is an increase or decrease in density if $\log L(x) > I_t$ or $\log L(x) < I_t$ respectively. In what follows, we will see that this difference, when integrated with respect to π_t, provides us with the right direction to move particles at time t. It is also intuitive that the factors $\pi_t(x)$ and $\lambda'(t)$ are present in (3.10), as the change in density must be proportional to how much probability mass there is locally and how quickly we introduce the likelihood. It will be apparent later that these factors dictate the speed of particles.

In this chapter, we focus on a fluid dynamics interpretation of the transport problem. If we perceive probability mass as an infinite ensemble of fluid particles, we could attempt to prescribe an appropriate velocity field to move these particles deterministically so as to mimic the time evolution of π_t over the pseudo-time interval $t \in [0,1]$. Loosely speaking, we may think of the action of particles under such a velocity field as implicitly defining flow transport maps $\{T_t\}_{t\in[0,1]}$ satisfying $\pi_t = T_{t\#}\pi_0$ for each $t \in [0,1]$.

3.1.3 Path sampling

We now give an alternative derivation of (3.12) from a different perspective. Consider the situation where one performs importance sampling with a proposal that is distant from the target distribution and consequently the estimator (1.13) of Z exhibits

large variance. In such a scenario, we expect the use of likelihood tempering to introduce a sequence of bridging distributions to be beneficial.

Let $\{\pi_{t_m}\}_{m=0}^M$ be a finite collection of probability measures along the curve (3.1) corresponding to a uniform grid of times $t_m = m\Delta t$ for $m = 0, \ldots, M$ with step size $\Delta t = M^{-1}$. Noting that $Z(0) = 1, Z(1) = Z$ and using the identity $Z(t_m)/Z(t_{m-1}) = \pi_{t_{m-1}}\left(L^{\lambda(t_m)-\lambda(t_{m-1})}\right)$, we have the telescopic product

$$Z = \prod_{m=1}^{M} \frac{Z(t_m)}{Z(t_{m-1})} = \prod_{m=1}^{M} \pi_{t_{m-1}}\left(L^{\lambda(t_m)-\lambda(t_{m-1})}\right). \tag{3.13}$$

Assuming that simulation from the intermediate distributions is feasible, for Δt sufficiently small or equivalently M large, the ratio $Z(t_m)/Z(t_{m-1})$ can be well approximated by importance sampling. This prompts a limiting argument as the number of bridging distributions go to infinity.

Taking logarithm of (3.13) gives

$$\log Z = \sum_{m=1}^{M} \log \pi_{t_{m-1}}\left(L^{\lambda(t_m)-\lambda(t_{m-1})}\right). \tag{3.14}$$

For each $t \in [0,1]$, we define the function $f_t(s) := \pi_t\left(L^{\lambda(t+s)-\lambda(t)}\right)$ for $s \in [0, 1-t]$. Noting that we can write $f_t(s) = Z(t+s)/Z(t)$, it follows from Lemma 3.1 that $f_t \in C^1([0, 1-t], \mathbb{R}_+)$ and in particular at $s = 0$, we have $f_t(0) = 1$. Therefore using the mean value theorem for the function $\log f_t$ on the interval $[0, s]$ gives

$$\frac{\log f_t(s)}{s} = \frac{\mathrm{d}}{\mathrm{d}s} \log f_t(s)\Big|_{s=c} = \frac{\mathrm{d}}{\mathrm{d}s} \log Z(t+s)\Big|_{s=c} = \lambda'(t+c) I_{t+c} \tag{3.15}$$

for some $c \in (0, s)$. The last equality in (3.15) follows from Equation (3.3). Now apply (3.15) to each summand in (3.14) on the subinterval $[t_{m-1}, t_m]$ to obtain

$$\log Z = \sum_{m=1}^{M} \Delta t \lambda'(c_m) I_{c_m} \tag{3.16}$$

for some $c_m \in (t_{m-1}, t_m), m = 1, \ldots, M$. Since $t \mapsto \lambda'(t) I_t = (\mathrm{d} \log Z/\mathrm{d}t)(t)$ is continuous on $[0,1]$, convergence of the Riemann sum (3.16) yields the path sampling identity (3.12).

3.2 Particle dynamics and Liouville's equation

Consider a particle in $\mathbb{R}^d$, initialized at time $t = 0$ with a random draw $X_0 \sim \pi_0$, and evolved deterministically according to the following ordinary differential equation (ODE)

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x), \tag{3.17}$$

with drift function $f : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$. Under appropriate regularity conditions on f which will be detailed later, this ODE admits a unique solution $x(t; X_0)$ on a unique maximal time interval $i(X_0)$ and we have $i(X_0) = [0, 1]$ for π_0-almost everywhere X_0. Therefore the map $X_t := T_t(X_0) := x(t; X_0)$ is well-defined and is a C^1-diffeomorphism for each $t \in [0, 1]$.

Additionally, if we denote the marginal distribution of X_t by $\tilde{\pi}_t$, the curve of distributions $\mathcal{C}_{\tilde{\pi}} = \{\tilde{\pi}_t\}_{t\in[0,1]}$ satisfies, under appropriate regularity conditions, the *Liouville PDE* (Gardiner 1985; eq. (3.5.13), p. 54) also known as the *continuity equation* (Ambrosio et al. 2005; eq. (8.1.1), p. 169):

$$\partial_t \tilde{\pi}_t(x) = -\sum_{i=1}^{d} \partial_{x_i}(\tilde{\pi}_t(x) f_i(t, x)) \tag{3.18}$$

for $(t, x) \in (0, 1) \times \mathbb{R}^d$. We will write (3.18) more succinctly as

$$\partial_t \tilde{\pi}_t = -\nabla \cdot (\tilde{\pi}_t f) \tag{3.19}$$

and note that this corresponds to the Fokker-Planck PDE (1.34) when the diffusivity term from Brownian motion is set to zero. An informal but intuitive derivation of Liouville's PDE will be given in the following section. We will call $\mathcal{C}_{\tilde{\pi}} = \{\tilde{\pi}_t\}_{t\in[0,1]}$ a weak solution of (3.19) if

$$\int_0^1 \int_{\mathbb{R}^d} \left(\partial_t \varphi + \langle f, \nabla \varphi \rangle\right) \tilde{\pi}_t(\mathrm{d}x)\, \mathrm{d}t = 0 \tag{3.20}$$

for all compactly supported $\varphi \in C^\infty((0, 1) \times \mathbb{R}^d, \mathbb{R})$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^d$.

3.2.1 An informal derivation of Liouville's equation

Consider a d-dimensional hyper-rectangle $\Delta V(x)$, defined formally as the Cartesian product of intervals $(x_i, x_i+\Delta_i)$ for $i = 1, \ldots, d$ and some small $\Delta = (\Delta_1, \ldots, \Delta_d)^T \in \mathbb{R}^d_+$, to be thought of as an infinitesimal control volume at a point $x \in \mathbb{R}^d$ as depicted in Figure 3.1.

If we perceive particles as constituents of a fluid representing probability mass, then the fluid flow driven by a velocity field f will cause the probability mass in $\Delta V(x)$ to change. Along the i^{th} axis, for sufficiently small $|\Delta|_\infty := \max_{i=1,\ldots,d} \Delta_i$, this change is given by the difference between the rate at which mass flows into $\Delta V(x)$

$$\tilde{\pi}_t(x) f_i(t,x) \prod_{j \neq i} \Delta_j + o(|\Delta|^d_\infty) \tag{3.21}$$

and the rate at which mass flows out of $\Delta V(x)$

$$\tilde{\pi}_t(x + \Delta_i e_i) f_i(t, x + \Delta_i e_i) \prod_{j \neq i} \Delta_j + o(|\Delta|^d_\infty), \tag{3.22}$$

where $\{e_i\}_{i=1}^d$ denote the canonical basis vectors in $\mathbb{R}^d$. In fluid dynamics terminology, the leading terms in (3.21) and (3.22) are simply the density multiplied by the volume metric flow rate in and out of the control volume respectively.

Summing over all axes yields the net rate at which probability mass is accumulating in $\Delta V(x)$:

$$\sum_{i=1}^{d} \left(\tilde{\pi}_t(x) f_i(t,x) \prod_{j \neq i} \Delta_j - \tilde{\pi}_t(x + \Delta_i e_i) f_i(t, x + \Delta_i e_i) \prod_{j \neq i} \Delta_j \right) + o(|\Delta|^d_\infty). \tag{3.23}$$

For probability mass to be conserved, (3.23) has to be equal to

$$\partial_t \tilde{\pi}_t(x) \prod_{i=1}^{d} \Delta_i + o(|\Delta|^d_\infty). \tag{3.24}$$

Equating (3.23) and (3.24) and dividing by the volume $\prod_{i=1}^d \Delta_i$ of $\Delta V(x)$ gives

$$\partial_t \tilde{\pi}_t(x) = \sum_{i=1}^{d} \frac{\tilde{\pi}_t(x) f_i(t,x) - \tilde{\pi}_t(x + \Delta_i e_i) f_i(t, x + \Delta_i e_i)}{\Delta_i} + o(1). \tag{3.25}$$

Finally, taking the limit of $|\Delta|_\infty \to 0$ gives (3.18).

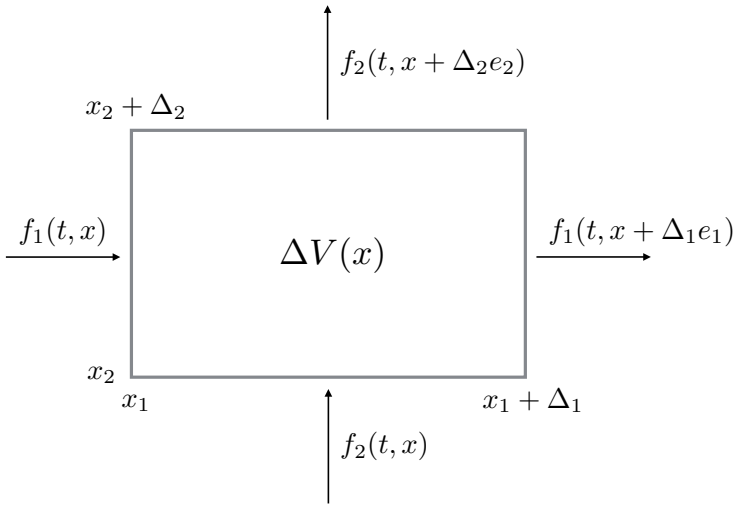

Figure 3.1: Illustrating the conservation of mass argument in $\mathbb{R}^2$.

3.2.2 Equivalence between Lagrangian and Eulerian perspectives

At this point, while it is clear that we should set up the flow transport problem as identifying a velocity field f such that the curve of target distributions $\mathcal{C}_\pi = \{\pi_t\}_{t\in[0,1]}$ in (3.1) is a weak solution of Liouville equation (3.19), we have to proceed with caution as f has to be regular enough for the resulting ODE (3.17) to admit a unique solution globally defined on $[0, 1]$. We now discuss sufficient conditions to ensure equivalence between the *Eulerian* perspective characterized by Liouville PDE (3.19) and the *Lagrangian* perspective described in terms of particle trajectories governed by the ODE (3.17).

We first introduce some necessary definitions before presenting the main theorem.

Definition 3.3. *Let $\mathcal{L}(\mathcal{C}_\pi)$ denote the set of all velocity fields such that $\mathcal{C}_\pi = \{\pi_t\}_{t\in[0,1]}$ is a weak solution of Liouville equation (3.19) on $(0, 1) \times \mathbb{R}^d$.*

Definition 3.4. *Let $\mathcal{E}(\mathcal{C}_\pi)$ be the set of all velocity fields $f : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ satisfying the following conditions:*

- ***A1**. (local Lipschitz continuity) f is a Borel function and is locally Lipschitz in the spatial variable, i.e. for every compact set $A \subset \mathbb{R}^d$, there exists*

a constant $C(A) > 0$ *such that* $|f(t,x) - f(t,y)| \leq C(A)|x-y|$ *for all* $(t,x),(t,y) \in [0,1] \times A$*;*

- ***A2****. (space-time integrability)* $\int_0^1 \int_{\mathbb{R}^d} |f(t,x)| \, \pi_t(\mathrm{d}x) \, \mathrm{d}t < \infty$.

Theorem 3.5. *(Ambrosio et al. (2005; Lemma 8.1.6, Proposition 8.1.8)) If* f *is a velocity field in* $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$*, then the following two statements are equivalent:*

(Eulerian perspective) *the narrowly continuous curve* $\mathcal{C}_\pi = \{\pi_t\}_{t\in[0,1]}$ *is a weak solution of Liouville equation (3.19) on* $(0,1) \times \mathbb{R}^d$*;*

(Lagrangian perspective) *for* π_0*-almost everywhere* $x_0 \in \mathbb{R}^d$*, there exists a unique solution* $x(t;x_0)$ *to the ODE (3.17) globally defined on* $[0,1]$*, therefore the flow maps* $T_t(x_0) := x(t;x_0)$ *are well-defined and satisfy the flow transport property, i.e.* $\pi_t = T_{t\#}\pi_0$ *for each* $t \in [0,1]$*.*

Condition A1 is an assumption which provides existence of a unique solution to (3.17) on a unique maximal time interval $i(x_0)$ for each initial condition $x_0 \in \mathbb{R}^d$ (Ambrosio et al. 2005; Lemma 8.1.4); we note that it only requires f to be a Borel function whereas standard results require $f \in C([0,1] \times \mathbb{R}^d, \mathbb{R}^d)$ (Walter 1998; Theorem III.3.VII). However, this time interval $i(x_0)$ may not contain $[0,1]$, hence A2 acts as a supplementary condition to ensure that $i(x_0) = [0,1]$.

It is possible to weaken condition A1; see DiPerna and Lions (1989) for earlier work and Ambrosio (2004), Ambrosio et al. (2005; Theorem 8.2.1) for recent advances. In fact, it is possible to completely remove assumption A1, in which case, the flow map $x_0 \mapsto T_t(x_0)$ is not uniquely defined as the ODE (3.17) could have several solutions corresponding to an initial condition x_0. Interestingly, it is still possible to give a probabilistic representation of these solutions by considering suitable probability measures on $C([0,1], \mathbb{R}^d)$ and equivalence between the Lagrangian and Eulerian perspectives can be understood in a more subtle sense. This generality is not pursued here but details can be found in Ambrosio et al. (2005; Section 8.2).

Observe that the integrability condition A2 implies

$$|f(t,x)| \, \pi_t(x) \to 0 \text{ as } |x| \to \infty \tag{3.26}$$

for each $t \in [0,1]$. We shall refer to velocity fields that satisfy (3.26) as having the *vanishing property*. With Theorem 3.5 in place, we can now formally define the *flow transport problem* as identifying a velocity field in $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$. It should be stressed that the regularity conditions in Theorem 3.5 are not mere mathematical subtleties one can ignore. In Section 3.3.6, we illustrate this by constructing a velocity field in $\mathcal{L}(\mathcal{C}_\pi)$ and prove that it yields divergent particle trajectories.

3.2.3 Connection to optimal transport

We now digress slightly by drawing a connection to the optimal transport problem discussed in Section 2.2. Like the approach taken in this chapter, it was shown in Benamou and Brenier (2000) that the Monge-Kantorovich problem also admits a fluid dynamics interpretation.

Suppose that π_0 and π have finite second moments and densities with respect to Lebesgue measure, in which case, we know from Section 2.2.3 that the optimal transport is given by $T^* = \nabla\phi$ for some convex function ϕ. In the Benamou and Brenier (2000) formulation, we seek a curve of probability measures connecting π_0 and π (which ensures the transport) using a flow which minimizes a kinetic energy functional (to obtain uniqueness). In particular, the squared Wasserstein distance with quadratic cost equals

$$\inf \int_0^1 \int_{\mathbb{R}^d} |f(t,x)|^2 \, \tilde{\pi}_t(\mathrm{d}x) \, \mathrm{d}t \tag{3.27}$$

where the infimum is taken over (sufficiently smooth) velocity fields f and curves of probability measures (which are absolutely continuous with respect to Lebesgue measure) $\mathcal{C}_{\tilde{\pi}} = \{\tilde{\pi}_t\}_{t\in[0,1]}$ such that Liouville equation (3.19) holds with boundary conditions

$$\tilde{\pi}_0 = \pi_0, \quad \tilde{\pi}_1 = \pi. \tag{3.28}$$

To see this, let f and $\mathcal{C}_{\tilde{\pi}} = \{\tilde{\pi}_t\}_{t\in[0,1]}$ be a pair from the admissible set. Denote the corresponding flow maps by $\{T_t\}_{t\in[0,1]}$, i.e. $\tilde{\pi}_t = T_{t\#}\tilde{\pi}_0$ for each $t \in [0,1]$. By the

change of variables formula (2.1), initial condition in (3.28), the Eulerian-Lagrangian relationship $\partial_t T_t(x) = f(t, T_t(x))$ and Jensen's inequality, we have

$$\begin{aligned}\int_0^1 \int_{\mathbb{R}^d} |f(t,x)|^2 \, \tilde{\pi}_t(\mathrm{d}x)\, \mathrm{d}t &= \int_0^1 \int_{\mathbb{R}^d} |f(t, T_t(x))|^2 \, \pi_0(\mathrm{d}x)\, \mathrm{d}t \\ &= \int_0^1 \int_{\mathbb{R}^d} |\partial_t T_t(x)|^2 \, \pi_0(\mathrm{d}x)\, \mathrm{d}t \\ &\geq \int_{\mathbb{R}^d} \left| \int_0^1 \partial_t T_t(x)\, \mathrm{d}t \right|^2 \pi_0(\mathrm{d}x).\end{aligned} \tag{3.29}$$

Apply the second fundamental theorem of calculus and the initial condition $T_0(x) = x$ to obtain

$$\int_0^1 \partial_t T_t(x)\, \mathrm{d}t = T_1(x) - x. \tag{3.30}$$

Combining (3.29) and (3.30) gives

$$\int_0^1 \int_{\mathbb{R}^d} |f(t,x)|^2 \, \tilde{\pi}_t(\mathrm{d}x)\, \mathrm{d}t \geq \int_{\mathbb{R}^d} |T_1(x) - x|^2 \, \pi_0(\mathrm{d}x) \geq \int_{\mathbb{R}^d} |\nabla\phi(x) - x|^2 \, \pi_0(\mathrm{d}x), \tag{3.31}$$

since $\pi = T_{1\#}\pi_0$ by the terminal condition in (3.28) and $T^* = \nabla\phi$ is the optimal transport. It follows that the infimum in (3.27) is attained by the flow maps $\{T_t^*\}_{t\in[0,1]}$ satisfying

$$\partial_t T_t^*(x) = \nabla\phi(x) - x \tag{3.32}$$

with initial condition $T_0^*(x) = x$. Hence the optimal flow maps are given by the interpolation

$$T_t^*(x) = (1-t)x + t\nabla\phi(x) \tag{3.33}$$

for $t \in [0,1]$.

3.3 Solving the flow transport problem

In this section, we first survey existing methods to solve the flow transport problem before describing our main results. We begin by noting that the flow transport problem defined above is typically underdetermined as illustrated in the following example.

Example 3.6. *Consider the trivial curve $\mathcal{C}_\pi$ given by $\pi_t = \mathcal{N}(0_2, I_2)$ for $t \in [0,1]$. For this curve, the two time-invariant velocity fields $f(x_1, x_2) = 0_2$ and $f(x_1, x_2) = (-x_2, x_1)^T$ both lie in $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$.*

We now consider a way to obtain unicity by seeking the minimal kinetic energy velocity field.

3.3.1 Minimizing kinetic energy

Following Reich (2011; 2012), we seek the velocity field which minimizes the kinetic energy functional at each time $t \in [0,1]$, i.e.

$$f(t,\cdot) := \text{argmin}_{u \in L^2(\pi_t) \cap \mathcal{L}(\pi_t)} \pi_t \left(u^T M_t u \right) \tag{3.34}$$

for some positive definite mass matrix $M_t \in \mathbb{R}^{d \times d}$. The set of admissible functions in (3.34) are $L^2(\pi_t)$-integrable functions that are classical solutions to Liouville equation (3.19) at time t given $\mathcal{C}_{\tilde{\pi}} = \mathcal{C}_\pi$. This leads to minimization of the following Lagrangian

$$E_t(u, \varphi) = \pi_t \left(u^T M_t u \right) + \int_{\mathbb{R}^d} \varphi \left(\partial_t \pi_t + \nabla \cdot (\pi_t u) \right) \mathrm{d}x \tag{3.35}$$

over $u \in L^2(\pi_t)$, where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a Lagrange multiplier enforcing the flow transport. It follows from the Euler-Lagrange equations that the desired velocity field is of the form $f(t,x) = M_t^{-1} \nabla \phi(t,x)$ with $\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ given by the solution of the following linear second-order elliptic PDE

$$\partial_t \pi_t = -\nabla \cdot (\pi_t M_t^{-1} \nabla \phi). \tag{3.36}$$

The regularization adopted here is very similar to Benamou and Brenier (2000) fluid dynamic formulation of the Monge-Kantorovich problem – see Equation (3.27). The main difference is that the curve of probability measures connecting π_0 and π is fixed by tempering the likelihood in (3.1). Moreover, for the purpose of solving the transport problem, we see that forsaking optimality allows us to reduce the non-linear Monge-Ampère PDE (2.7) to the linear PDE (3.36), which also arises in the context of optimal non-linear filtering and can be adequately approximated by

finite element methods for low dimensional problems (Yang et al. 2013, Laugesen et al. 2015, Yang et al. 2016).

We now examine a Gaussian example where both the elliptic PDE (3.36) and the corresponding minimal kinetic energy velocity field are analytically tractable (Bergemann and Reich 2012). The resulting flow is equivalent to the *Kalman-Bucy filter*. By exploiting linearity of (3.36), it is also possible to extend to the case where $\mathcal{C}_\pi$ is in the Gaussian mixture family (Reich 2012).

Example 3.7. *Consider the prior distribution* $\pi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ *and likelihood function*

$$L(x; y) = \exp\left(-\frac{1}{2}(Hx - y)^T R^{-1}(Hx - y)\right) \tag{3.37}$$

with $H \in \mathbb{R}^{p\times d}$ *for some* $p \in \mathbb{N}$*, symmetric positive definite* $R \in \mathbb{R}^{p\times p}$ *and observation* $y \in \mathbb{R}^p$*. By conjugacy, the curve* $\mathcal{C}_\pi$ *lies in the Gaussian family, i.e.* $\pi_t = \mathcal{N}(\mu_t, \Sigma_t)$ *for* $t \in [0, 1]$ *with*

$$\Sigma_t^{-1} = \Sigma_0^{-1} + \lambda(t)H^T R^{-1} H, \quad \mu_t = \Sigma_t\left(\Sigma_0^{-1}\mu_0 + \lambda(t)H^T R^{-1} y\right), \tag{3.38}$$

and the expected log-likelihood is

$$I_t = -\frac{1}{2}\left(\mathrm{Tr}(H^T R^{-1} H\Sigma_t) + (H\mu_t - y)^T R^{-1}(H\mu_t - y)\right), \tag{3.39}$$

where $\mathrm{Tr}(A)$ *denotes the trace of a square matrix* A*. In this Gaussian setting, the analytical solution to the elliptic PDE (3.36) is*

$$\phi(t, x) = -\frac{\lambda'(t)}{4}(Hx + H\mu_t - 2y)^T R^{-1}(Hx + H\mu_t - 2y) \tag{3.40}$$

with $M_t = \Sigma_t^{-1}$ *and the minimal kinetic energy velocity field (3.34) is given by (Bergemann and Reich 2012)*

$$f(t, x) = -\frac{\lambda'(t)}{2}\Sigma_t H^T R^{-1}(Hx + H\mu_t - 2y) \tag{3.41}$$

which is clearly an element of $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$*.*

3.3.2 Extended Kalman-Bucy filter

Consider a modification of Example 3.7, where we have $\pi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ as before but the likelihood function

$$L(x; y) = \exp\left(-\frac{1}{2}(H(x) - y)^T R^{-1}(H(x) - y)\right) \tag{3.42}$$

now involves a non-linear function $H : \mathbb{R}^d \to \mathbb{R}^p$. Following Bunch and Godsill (2016), we can construct an approximate flow transport based on the following truncated Taylor expansion

$$H(x) \approx H(x_*) + \nabla H(x_*)(x - x_*). \tag{3.43}$$

for some $x_* \in \mathbb{R}^d$ to be specified. This gives a linearized likelihood

$$\hat{L}(x; x_*) = \exp\left(-\frac{1}{2}\left(\hat{H}(x_*)x - \hat{y}(x_*)\right)^T R^{-1}\left(\hat{H}(x_*)x - \hat{y}(x_*)\right)\right) \tag{3.44}$$

with $\hat{H}(x_*) = \nabla H(x_*)$, $\hat{y}(x_*) = y - H(x_*) + \nabla H(x_*)x_*$ and the approximate curve of probability measures

$$\hat{\pi}_t(\mathrm{d}x; x_*) := \frac{\pi_0(\mathrm{d}x)\hat{L}(x; x_*)^{\lambda(t)}}{\pi_0(\hat{L}^{\lambda(t)})} = \mathcal{N}\left(x; \hat{\mu}_t(x_*), \hat{\Sigma}_t(x_*)\right)\mathrm{d}x \tag{3.45}$$

where

$$\begin{aligned} \hat{\Sigma}_t^{-1}(x_*) &= \Sigma_0^{-1} + \lambda(t)\hat{H}(x_*)^T R^{-1}\hat{H}(x_*), \\ \hat{\mu}_t(x_*) &= \hat{\Sigma}_t(x_*)\left(\Sigma_0^{-1}\mu_0 + \lambda(t)\hat{H}(x_*)^T R^{-1}\hat{y}(x_*)\right). \end{aligned} \tag{3.46}$$

The main idea behind Bunch and Godsill (2016) is to "continuously refresh" the linearization point x_* and in this sense may be thought of as a continuous time limit of an *extended Kalman filter* update. More precisely, using (3.41) with linearization at the current state, the authors constructed an approximate flow based on

$$\hat{f}(t, x) = -\frac{\lambda'(t)}{2}\hat{\Sigma}_t(x)\hat{H}(x)^T R^{-1}\left(\hat{H}(x)x + \hat{H}(x)\hat{\mu}_t(x) - 2\hat{y}(x)\right). \tag{3.47}$$

Although Bunch and Godsill (2016) did not provide an error analysis, good experimental performance was reported for highly non-linear filtering problems in dimensions up to $d = 6$.

3.3.3 Poisson equation

The following is based on Moser (1965) constructive proof to solve the Jacobian problem mentioned in Section 2.1. The main difference from Moser's original construction is that we employ the curve (3.1) instead of the linear interpolation $t \mapsto (1-t)\pi_0 + t\pi$, which is not suitable in our context as the normalizing constant Z is intractable.

The key idea is to seek a velocity field of the form $f(t,x) = \nabla\phi(t,x)/\pi_t(x)$ for some sufficiently smooth $\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}$. The requirement that $f \in \mathcal{L}(\mathcal{C}_\pi)$ reduces Liouville equation (3.19) to the *Poisson equation*

$$\Delta\phi = -\partial_t \pi_t. \tag{3.48}$$

We impose the tail condition $|\phi(t,x)| \to 0$ as $|x| \to \infty$ for each $t \in [0,1]$ which implies the vanishing property (3.26). As noted by Barron and Luo (2007), Daum et al. (2011), the solution of (3.48) can be written as the convolution of $-\partial_t\pi_t$ and *Green's function* $G : \mathbb{R}^d \to \mathbb{R}$ associated to the Laplacian on $\mathbb{R}^d$

$$\phi(t,x) = -\int_{\mathbb{R}^d} G(x-y)\partial_t\pi_t(y)\,\mathrm{d}y. \tag{3.49}$$

For $d \geq 3$, it can be shown that $G(x) = |x|^{2-d}/\left((2-d)S_{d-1}\right)$ where S_{d-1} denotes the surface area of a d-dimensional hypersphere. Using (3.10) and assuming validity of differentiating under the integral[1], we obtain

$$f(t,x) = \frac{\lambda'(t)}{S_{d-1}} \int_{\mathbb{R}^d} \frac{(y-x)}{|x-y|^d} \left(\log L(y) - I_t\right) \pi_t(\mathrm{d}y) \Big/ \pi_t(x). \tag{3.50}$$

If we have access to a particle approximation $\hat{\pi}_t^N = N^{-1}\sum_{n=1}^N \delta_{X_t^n}$ of π_t and an estimator $\hat{Z}_t^N$ of Z_t, (3.50) can be approximated by

$$\hat{f}_N(t,x) = \frac{\lambda'(t)}{S_{d-1}} \sum_{n=1}^N \frac{(X_t^n - x)}{|x - X_t^n|^d} \left(\log L(X_t^n) - \hat{\pi}_t^N(\log L)\right) \Big/ (\gamma_t(x)/\hat{Z}_t^N). \tag{3.51}$$

Preliminary experiments on simple examples reveal that numerical integration of an ODE with drift function (3.51) is very unstable as $|\hat{f}_N(t,x)| \to \infty$ whenever $|x - X_t^n| \to 0$ for any $n = 1, \ldots, N$.

[1] Justifying this operation is not straightforward; see Crisan and Xiong (2010; Proposition 4.1) for the case where $x \mapsto \log L(x)$ is assumed to be bounded.

3.3.4 Incompressible flow

We now describe an method proposed by Daum and Huang (2008; 2009) and discuss its limitations. Consider a velocity field $f \in \mathcal{L}(\mathcal{C}_\pi)$ and apply chain rule in Liouville equation (3.19)

$$\pi_t \left(\nabla \cdot f\right) + \langle \nabla \pi_t, f \rangle = -\partial_t \pi_t. \tag{3.52}$$

The authors considered using *incompressible* flows, i.e. restricting f to the class of velocity fields that have zero divergence. As

$$\langle \nabla \pi_t, f \rangle = -\partial_t \pi_t \tag{3.53}$$

is an underdetermined equation in the unknown f, an application of Cauchy-Schwarz inequality $|\partial_t \pi_t| = |\langle \nabla \pi_t, f \rangle| \leq |\nabla \pi_t||f|$ shows that the minimal kinetic energy solution is

$$f(t,x) = \frac{-\partial_t \pi_t(x) \nabla \pi_t(x)}{|\nabla \pi_t(x)|^2} = \frac{\lambda'(t)\left(I_t - \log L(x)\right) \nabla \log \pi_t(x)}{|\nabla \log \pi_t(x)|^2}. \tag{3.54}$$

As incompressible flows preserve Lebesgue measure (Leimkuhler and Matthews 2015; p. 72), we expect (3.54) to be sensible only when the flow transport can be achieved by such flow maps. In the following example, we illustrate that the equivalence between the Eulerian and Lagrangian perspectives, necessary to guarantee validity of the above construction, might break down for certain curves of probability measures.

Example 3.8. *Consider the one-dimensional case $d = 1$ with prior distribution $\pi_0 = \mathcal{N}(\mu_0, \sigma_0^2)$ and likelihood function $L(x) = \exp(ax^2 + bx)$ for some $(a, b) \in (-\infty, 1/(2\sigma_0^2)) \times \mathbb{R}$. By conjugacy, $\mathcal{C}_\pi$ lies in the Gaussian family, i.e. $\pi_t = \mathcal{N}(\mu_t, \sigma_t^2)$ for $t \in [0,1]$ with*

$$\sigma_t^2 = \frac{\sigma_0^2}{1 - 2a\sigma_0^2 \lambda(t)}, \quad \mu_t = \frac{\sigma_t^2}{\sigma_0^2}\left(\mu_0 + b\sigma_0^2 \lambda(t)\right). \tag{3.55}$$

We first examine the case $a = 0, b \neq 0$. Since the variance $\sigma_t^2 = \sigma_0^2$ remains constant and only the mean μ_t changes with time, the flow transport can be achieved

by translations so we expect (3.54) to induce well-defined flow maps. This is indeed the case as $f(t,x) = b\sigma_0^2\lambda'(t)$*, therefore the ODE (3.17) admits a unique solution and flow maps* $T_t(x_0) = x_0 + b\sigma_0^2\lambda(t)$ *satisfy* $\pi_t = T_{t\#}\pi_0$ *for all* $t \in [0,1]$*.*

In contrast, we move to the case $a \neq 0, b = 0$ *and for concreteness we set* $a = -1/2, \mu_0 = 0, \sigma_0^2 = 1, \lambda(t) = t$*. Under this setting, the mean* $\mu_t = 0$ *remains constant while the variance* $\sigma_t^2 = (1+t)^{-1}$ *decreases with time, so we expect any valid flow transport to have a mean-reverting behaviour towards the origin. As*

$$f(t,x) = -\frac{x}{2(1+t)} + \frac{1}{2x(1+t)^2}, \tag{3.56}$$

this is observed for particles that are far enough away from the origin ($|x| > 1/\sqrt{1+t}$*) where the linear term in (3.56) is dominant (see left panel of Figure 3.2). The strong repelling force experienced by particles near the origin, where the second term in (3.56) dominates, is consistent with the incompressibility assumption: particles are pushed towards the origin but the flow cannot be compressed and hence the singularity at the origin. In fact, the flow maps are not well-defined in this case as the corresponding ODE with drift function* f *admits two solutions given by*

$$x(t;x_0) = \frac{\pm\sqrt{2x_0^2 + 2\log(1+t)}}{\sqrt{2}\sqrt{1+t}} \tag{3.57}$$

for $t \in [0,1]$*.*

3.3.5 A solution on $\mathbb{R}$

We now focus on the one-dimensional case $d = 1$ as our main construction for $d > 1$ partially builds upon it. In this case, there is a rather well-known solution to the flow transport problem; for example see Barron and Luo (2007). Moreover, we establish that this in fact coincides with the minimal kinetic energy velocity field described in Section 3.3.1.

Proposition 3.9. *Define the velocity field* $f : [0,1] \times \mathbb{R} \to \mathbb{R}$ *as*

$$f(t,x) := \frac{-\int_{-\infty}^{x} \partial_t \pi_t(y)\,\mathrm{d}y}{\pi_t(x)}. \tag{3.58}$$

If there exists an $\varepsilon > 0$ *such that* $x \mapsto |f(t,x)|\,\pi_t(x) = O(|x|^{-1-\varepsilon})$ *as* $|x| \to \infty$ *with a constant that is independent of* $t \in [0,1]$*, then the velocity field (3.58) lies in* $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$ *and thus solves the flow transport problem on* $\mathbb{R}$*. It is additionally the minimal kinetic energy solution defined in (3.34) for any mass* $M_t \in \mathbb{R}_+$*.*

Proof. Using continuity of π_0, L and positivity of L, an application of the first fundamental theorem of calculus shows that $f \in \mathcal{L}(\mathcal{C}_\pi)$. The assumptions on π_0 and L imply $f \in C^1([0,1] \times \mathbb{R}, \mathbb{R})$; hence for any compact set $A \subset \mathbb{R}$, its derivative is bounded on $[0,1] \times A$ and local Lipschitzness A1 follows. The integrability condition A2 follows from the prescribed tail behaviour of $x \mapsto |f(t,x)|\,\pi_t(x)$ uniformly over $t \in [0,1]$. Hence $f \in \mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$ and appealing to Theorem 3.5 shows that (3.58) solves the flow transport problem.

To see that (3.58) is indeed the minimal kinetic energy solution, we note that the optimality condition in (3.36) requires existence of a function $\phi : [0,1] \times \mathbb{R} \to \mathbb{R}$ such that $f(t,x) = M_t^{-1} \nabla \phi(t,x)$ and $\partial_t \pi_t = -\nabla \cdot (\pi_t M_t^{-1} \nabla \phi)$. The former is trivially satisfied as a consequence of working on $\mathbb{R}$ since we may set $\phi(t,x) = M_t \int_c^x f(t,y)\,dy < \infty$ for any $c < x$ and the latter follows since $f \in \mathcal{L}(\mathcal{C}_\pi)$. □

We note that the velocity field (3.58) satisfies the vanishing property by construction as $x \mapsto \int_{-\infty}^x \partial_t \pi_t(y)\,\mathrm{d}y$ vanishes in the tails. From (3.10), (3.58) may be re-written as

$$f(t,x) = \frac{\lambda'(t) I_t \left(F_t(x) - I_t^x / I_t\right)}{\pi_t(x)}, \tag{3.59}$$

where $I_t^x := \int_{-\infty}^x \log L(y) \pi_t(y)\,\mathrm{d}y$ and $F_t(x) := \int_{-\infty}^x \pi_t(y)\,\mathrm{d}y$ is the CDF of π_t. In the Lagrangian perspective, the velocity field (3.59) may be likened to driving a vehicle. The denominator corresponds to the *accelerator* since, for example, particles in the tails of π_t need to *speed up* to meet the changing schedule of intermediate distributions. Also, it is intuitive that particle speeds are proportional to the rate $\lambda'(t)$ at which we introduce the likelihood. The numerator amounts to the *steering wheel*: a particle's direction of travel is given by the relative difference between its *current location* x, described by the term $F_t(x)$, and *where the particle*

needs to go, prescribed by the term $I_t^x/I_t \in [0,1]$ which contains information from the likelihood. We now investigate the behaviour of this flow in a Gaussian scenario to build intuition.

Example 3.10. *As noted in Proposition 3.9, the velocity field in (3.58) corresponds exactly to (3.41) when $d = 1$. For a more concrete example, we re-visit the setup in the second part of Example 3.8 where the curve $\mathcal{C}_\pi$ of interest was $\pi_t = \mathcal{N}(0, (1+t)^{-1})$ for $t \in [0,1]$. In this case, (3.58) gives a linear mean-reverting drift towards the origin*

$$f(t,x) = -\frac{x}{2(1+t)}. \tag{3.60}$$

The right panel of Figure 3.2 also illustrates this behaviour with the steering property mentioned earlier: since $I_t = -1/(2(1+t)) < 0$ for all $t \in [0,1]$, reversion to the stable stationary point at the origin dictates that $F_t(x) < I_t^x/I_t$ for $x < 0$ and $F_t(x) > I_t^x/I_t$ for $x > 0$.

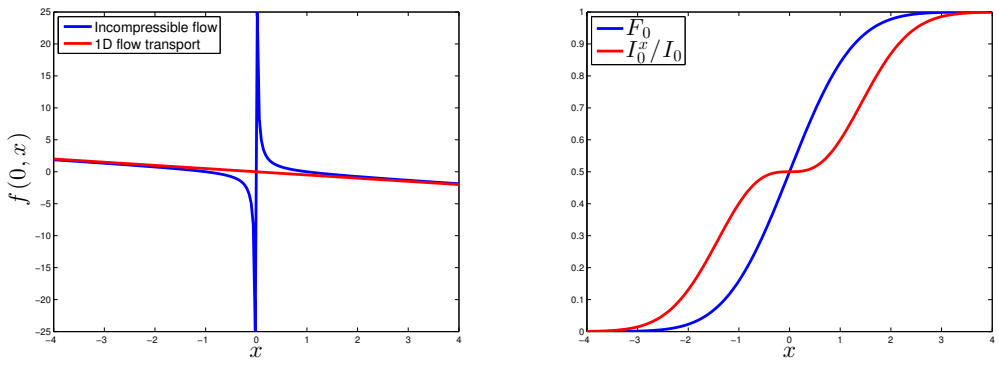

Figure 3.2: (*Left*) Comparing the incompressible flow with the 1D flow transport solution. (*Right*) Illustrating steering property of (3.59) on univariate Gaussian example.

3.3.6 An incorrect solution in $\mathbb{R}^d$, $d \geq 2$

It is tempting to extend the flow transport solution in Proposition 3.9 from $\mathbb{R}$ to $\mathbb{R}^d$, $d \geq 2$, by simply introducing the velocity field $\bar{f} = \left(\bar{f}_1, \ldots, \bar{f}_d\right)^T$ given by

$$\bar{f}_i(t,x) := \frac{-\alpha_i \int_{-\infty}^{x_i} \partial_t \pi_t(y_i, x_{-i})\, \mathrm{d}y_i}{\pi_t(x)} \tag{3.61}$$

for $i = 1, \ldots, d$, where $\alpha_i \in \mathbb{R}$ and the integrand of (3.61) is to be understood as $\partial_t \pi_t(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_d)$. This velocity field has been previously mentioned in Barron and Luo (2007) and using the same arguments in the proof of Proposition 3.9, it follows that $\bar{f} \in \mathcal{L}(\mathcal{C}_\pi)$ whenever $\sum_{i=1}^d \alpha_i = 1$. However it is easy to see that $|\bar{f}_i(t,x)|\, \pi_t(x) \nrightarrow 0$ as $x_i \to \infty$ for any $i = 1, \ldots, d$, so condition A2 does not hold. Hence $\bar{f} \notin \mathcal{E}(\mathcal{C}_\pi)$ so Theorem 3.5 does not apply. On a simple Gaussian example detailed below, we show that an ODE with drift function $\bar{f}$ results in divergent particle trajectories.

Example 3.11. *Consider $\pi_t = \mathcal{N}(\mu_t, \Sigma_t)$ with parameters given by (3.38) where $\mu_0 = 0_2, \Sigma_0 = H = R = I_2$ and $y = 0_2$. This setup corresponds to independent components marginally distributed according to the univariate Gaussian model of Example 3.10. Hence we would expect a particle under a valid flow transport to have a mean-reverting behaviour towards the origin. The velocity field in (3.61) has the form*

$$\begin{pmatrix} \bar{f}_1(t,x) \\ \bar{f}_2(t,x) \end{pmatrix} = \begin{pmatrix} \frac{\alpha_1}{2\pi_t(x_1)} \left(\int_{-\infty}^{x_1} y_1^2 \pi_t(y_1)\, \mathrm{d}y_1 + x_2^2 F_t(x_1) - \frac{F_t(x_1)}{1+t} \right) \\ \frac{\alpha_2}{2\pi_t(x_2)} \left(\int_{-\infty}^{x_2} y_2^2 \pi_t(y_2)\, \mathrm{d}y_2 + x_1^2 F_t(x_2) - \frac{F_t(x_2)}{1+t} \right) \end{pmatrix} \tag{3.62}$$

for $x = (x_1, x_2)^T \in \mathbb{R}^2$ and $t \in [0,1]$, where $\pi_t = \mathcal{N}\left(0_2, (1+t)^{-1} I_2\right)$ and $F_t(x_i)$ denotes the marginal CDFs. We note that the two components of the velocity field are coupled.

Now consider $\alpha_1, \alpha_2 > 0$ with $\alpha_1 + \alpha_2 = 1$. We investigate the behaviour of particles in the upper-right quadrant of the space. For each $t \in [0,1]$, define the sets $\mathcal{S}_t := \left\{ x \in \mathbb{R}^2 : x_1, x_2 > 1/\sqrt{1+t} \right\}$, $\mathcal{P}_t := \left\{ x \in \mathbb{R}^2 : \bar{f}(t,x) > 0_2 \right\}$ and note from (3.62) that $\mathcal{S}_0 \subset \mathcal{S}_t \subset \mathcal{P}_t$ for any $t \in (0,1]$. Since $\pi_0(\mathcal{S}_0) > 0$, we can conclude that there exist particle trajectories which only move farther away from the origin with positive probability. Analytical tractability in this simple example allows us to strengthen the previous statement and show that these trajectories in fact blow up in finite time. We start by seeking a lower bound on $\bar{f}$; by symmetry, it suffices to consider only the first component. On the set $\mathcal{S}_0$ we have $\int_{-\infty}^{x_1} y_1^2 \pi_t(y_1)\, \mathrm{d}y_1 >$

$\frac{1}{2(1+t)} \geq \frac{1}{4}$, *hence*

$$\bar{f}_1(t,x) \geq \frac{c}{4}\exp\left(\frac{1}{2}x_1^2\right) \geq \frac{c}{32}x_1^4 \tag{3.63}$$

with $c := \frac{\alpha_1\sqrt{\pi}}{2} > 0$. *Now consider an uncoupled system of ODE with drift function*

$$\begin{pmatrix}\hat{f}_1(t,x_1)\\ \hat{f}_2(t,x_2)\end{pmatrix} := \begin{pmatrix}\frac{c}{32}x_1^4\\ \frac{c}{32}x_2^4\end{pmatrix} \leq \begin{pmatrix}\bar{f}_1(t,x)\\ \bar{f}_2(t,x)\end{pmatrix}, \tag{3.64}$$

and note that its solution $x_i(t;x_{0,i}) = 1/\sqrt[3]{3\left(\frac{1}{3x_{0,i}^3} - \frac{c}{32}t\right)}$, *corresponding to an initial condition* $x_0 = (x_{0,1}, x_{0,2})^T \in \mathbb{R}^2$, *diverges as* $t \to \frac{32}{3c\,x_{0,i}^3}$. *Define the set* $\mathcal{V} = \left\{x \in \mathbb{R}^2 : x_1, x_2 > \sqrt[3]{\frac{32}{3c}}\right\}$. *Noting that* $\hat{f}$ *is locally Lipschitz and componentwise increasing, the comparison theorem (Walter 1998; Theorem III.10.XII (b)) implies that a particle starting in* $\mathcal{S}_0 \cap \mathcal{V}$ *and evolving under (3.62) has a trajectory that explodes before* $t = 1$. *Since* $\pi_0(\mathcal{S}_0 \cap \mathcal{V}) > 0$, *we conclude the claim that there exist divergent particle trajectories with positive probability.*

3.3.7 A solution in $\mathbb{R}^d$, $d \geq 1$

The main reason why the flow induced by the velocity field $\bar{f}$ in (3.61) fails to solve the flow transport problem for $d \geq 2$ is because $\bar{f}$ does not vanish in the tails. In the following, we show that the introduction of some regularizing functions allows us to resolve this issue.

Proposition 3.12. *For* $i = 1, \ldots, d-1$ *let* $g_i \in C^2([0,1]\times\mathbb{R}, [0,1])$ *be a non-decreasing function with the following tail behaviour:* $g_i(t,x_i) \to 0$ *as* $x_i \to -\infty$ *and* $g_i(t,x_i) \to 1$ *as* $x_i \to \infty$. *Denote partial derivatives* $\partial_{x_i} g_i(t,x_i)$ *by* $g_i'(t,x_i)$ *and define the velocity field* $f : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ *as*

$$\begin{aligned} f_i(t,x) := -\Bigg(&\prod_{j=1}^{i-1} g_j'(t,x_j)\int_{-\infty}^{x_i}\int_{\mathbb{R}^{i-1}} \partial_t\pi_t(y_{1:i-1}, y_i, x_{i+1:d})\,\mathrm{d}y_{1:i-1}\mathrm{d}y_i \\ &- \prod_{j=1}^{i-1} g_j'(t,x_j)g_i(t,x_i)\int_{\mathbb{R}^i}\partial_t\pi_t(y_{1:i}, x_{i+1:d})\,\mathrm{d}y_{1:i}\Bigg)\Big/\pi_t(x) \end{aligned} \tag{3.65}$$

for $i = 1, \ldots, d-1$ *(using the convention* $\prod_1^0 := 1$*) and*

$$f_d(t,x) := -\left(\prod_{j=1}^{d-1} g_j'(t,x_j)\int_{-\infty}^{x_d}\int_{\mathbb{R}^{d-1}}\partial_t\pi_t(y_{1:d-1}, y_d)\,\mathrm{d}y_{1:d-1}\mathrm{d}y_d\right)\Big/\pi_t(x). \tag{3.66}$$

If there exists an $\varepsilon > 0$ such that $x \mapsto |f(t,x)|\,\pi_t(x) = O(|x|^{-1-\varepsilon})$ as $|x| \to \infty$ with a constant that is independent of $t \in [0,1]$, then the velocity field (3.65)-(3.66) lies in $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$ and thus solves the flow transport problem on $\mathbb{R}^d, d \geq 1$.

Proof. The arguments are similar to those in Proposition 3.9. By straightforward verification $f \in \mathcal{L}(\mathcal{C}_\pi)$:

$$
\begin{aligned}
\nabla \cdot (\pi_t f) &= \sum_{i=1}^{d} \partial_{x_i}(\pi_t(x) f_i(t,x)) \qquad (3.67)\\
&= -\sum_{i=1}^{d-1} \partial_{x_i} \Bigg(\prod_{j=1}^{i-1} g_j'(t,x_j) \int_{\mathbb{R}^{i-1}} \int_{-\infty}^{x_i} \partial_t \pi_t(y_{1:i-1}, y_i, x_{i+1:d})\, \mathrm{d}y_{1:i-1} \mathrm{d}y_i \\
&\qquad\qquad - \prod_{j=1}^{i-1} g_j'(t,x_j) g_i(t,x_i) \int_{\mathbb{R}^i} \partial_t \pi_t(y_{1:i}, x_{i+1:d})\, \mathrm{d}y_{1:i} \Bigg) \\
&\qquad\qquad - \partial_{x_d} \Bigg(\prod_{j=1}^{d-1} g_j'(t,x_j) \int_{\mathbb{R}^{d-1}} \int_{-\infty}^{x_d} \partial_t \pi_t(y_{1:d-1}, y_d)\, \mathrm{d}y_{1:d-1} \mathrm{d}y_d \Bigg) \\
&= -\sum_{i=1}^{d-1} \Bigg(\prod_{j=1}^{i-1} g_j'(t,x_j) \int_{\mathbb{R}^{i-1}} \partial_t \pi_t(y_{1:i-1}, x_i, x_{i+1:d})\, \mathrm{d}y_{1:i-1} \\
&\qquad\qquad - \prod_{j=1}^{i} g_j'(t,x_j) \int_{\mathbb{R}^i} \partial_t \pi_t(y_{1:i}, x_{i+1:d})\, \mathrm{d}y_{1:i} \Bigg) \\
&\qquad\qquad - \prod_{j=1}^{d-1} g_j'(t,x_j) \int_{\mathbb{R}^{d-1}} \partial_t \pi_t(y_{1:d-1}, x_d)\, \mathrm{d}y_{1:d-1} \\
&= -\partial_t \pi_t.
\end{aligned}
$$

The penultimate line applies the first fundamental theorem of calculus and the final equality comes from the telescopic sum. The assumptions on π_0, L and $\{g_i\}_{i=1}^{d-1}$ imply $f \in C^1([0,1] \times \mathbb{R}^d, \mathbb{R}^d)$; hence local Lipschitzness A1 follows. The integrability condition A2 follows from the prescribed tail behaviour of $x \mapsto |f(t,x)|\,\pi_t(x)$ uniformly over $t \in [0,1]$. Hence $f \in \mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$ and appealing to Theorem 3.5 completes the proof. □

As before, we note that (3.65)-(3.66) satisfies the vanishing property by construction. To see this, observe that $x \mapsto |f_i(t,x)|\,\pi_t(x)$ vanishes in the tails for $i = 1, \ldots, d-1$ as this is so of $x \mapsto \partial_t \pi_t(x)$ and our assumptions imply that $g_i'(t,x_i) \to 0$ as $|x_i| \to \infty$. This also holds for $x \mapsto |f_d(t,x)|\,\pi_t(x)$ using, additionally, the fact that $\int_{\mathbb{R}^d} \partial_t \pi_t(y)\, \mathrm{d}y = 0$.

Note that Proposition 3.12 recovers Proposition 3.9 in the $d = 1$ case. A careful inspection of (3.65)-(3.66) reveals that the dynamics are constructed to track changes in the underlying conditionals $\{\pi_t(\mathrm{d}x_i|x_{1:i-1})\}_{i=2}^d$ and in this sense may be thought of as the flow transport analogue of the Knothe-Rosenblatt transport (Section 2.3). Our construction is a generalization of a method proposed by Bokanowski and Grébert (1996) to build a compactly supported three-dimensional velocity field solving a flow transport problem in the context of molecular quantum chemistry.

Proposition 3.12 leaves a degree of freedom over the choice of scalar functions $\{g_i\}_{i=1}^{d-1}$. We advocate choosing $\{g_i\}_{i=1}^{d-1}$ so that the velocity field on $\mathbb{R}^d$ reduces to d many independent velocity fields on $\mathbb{R}$ if the posterior distribution π and hence the curve $\mathcal{C}_\pi$ factorizes, i.e. if we can write $\pi_t(x) = \prod_{i=1}^d \pi_t(x_i)$ where $\pi_t(x_i) = \pi_0(x_i)L_i(x_i)^{\lambda(t)}/Z_i(t)$. More precisely, we would like the original Liouville equation defined on $(0,1) \times \mathbb{R}^d$ to simplify to a system of *uncoupled* Liouville PDEs each defined on $(0,1) \times \mathbb{R}$:

$$\partial_t \pi_t(x_i) = -\partial_{x_i}(\pi_t(x_i) f_i(t, x_i)) \tag{3.68}$$

for $i = 1, \ldots, d$ and solved by Proposition 3.9. We shall refer to velocity fields which exhibit this behaviour as having the *factorization under independence property*.

Proposition 3.13. *If $g_i(t, x_i) = F_t(x_i)$ for $i = 1, \ldots, d-1$, then the velocity field defined in (3.65)-(3.66) factorizes if the posterior distribution π factorizes.*

Proof. Note that $g_i'(t, x_i) = \pi_t(x_i)$ and define $I_t^{(i)} := \int_{\mathbb{R}} \log L_i(x_i)\pi_t(x_i)\,\mathrm{d}x_i$. From

(3.65), for $i = 1, \ldots, d-1$

$$
\begin{aligned}
f_i(t,x) = \frac{\lambda'(t)}{\prod_{l=1}^d \pi_t(x_l)} \Bigg(& \prod_{j=1}^{i-1} \pi_t(x_j) \int_{-\infty}^{x_i} \int_{\mathbb{R}^{i-1}} \left(I_t - \sum_{l=1}^{i} \log L_l(y_l) - \sum_{k=i+1}^{d} \log L_k(x_k) \right) \\
& \times \prod_{j=1}^{i} \pi_t(y_j) \prod_{k=i+1}^{d} \pi_t(x_k) \, \mathrm{d}y_{1:i-1} \mathrm{d}y_i \\
& - \prod_{j=1}^{i-1} \pi_t(x_j) \int_{-\infty}^{x_i} \pi_t(y_i) \, \mathrm{d}y_i \int_{\mathbb{R}^i} \left(I_t - \sum_{l=1}^{i} \log L_l(y_l) - \sum_{k=i+1}^{d} \log L_k(x_k) \right) \\
& \times \prod_{j=1}^{i} \pi_t(y_j) \prod_{k=i+1}^{d} \pi_t(x_k) \, \mathrm{d}y_{1:i} \Bigg), \\
= \frac{\lambda'(t)}{\pi_t(x_i)} \Bigg(& \int_{-\infty}^{x_i} \pi_t(y_i) \, \mathrm{d}y_i \left(I_t - \sum_{l=1}^{i-1} I_t^{(l)} - \sum_{k=i+1}^{d} \log L_k(x_k) \right) \\
& - \int_{-\infty}^{x_i} \log L_i(y_i) \pi_t(y_i) \, \mathrm{d}y_i - \int_{-\infty}^{x_i} \pi_t(y_i) \, \mathrm{d}y_i \left(I_t - \sum_{l=1}^{i} I_t^{(l)} - \sum_{k=i+1}^{d} \log L_k(x_k) \right) \Bigg) \\
= \frac{\lambda'(t)}{\pi_t(x_i)} \Bigg(& \int_{-\infty}^{x_i} (I_t^{(i)} - \log L_i(y_i)) \pi_t(y_i) \, \mathrm{d}y_i \Bigg), \qquad (3.69)
\end{aligned}
$$

and from (3.66)

$$
\begin{aligned}
f_d(x,t) &= \frac{\lambda'(t)}{\prod_{l=1}^d \pi_t(x_l)} \left(\prod_{j=1}^{d-1} \pi_t(x_j) \int_{-\infty}^{x_d} \int_{\mathbb{R}^{d-1}} \sum_{l=1}^{d} (I_t^{(l)} - \log L_l(y_l)) \prod_{k=1}^{d} \pi_t(y_k) \, \mathrm{d}y_{1:d-1} \mathrm{d}y_d \right), \\
&= \frac{\lambda'(t)}{\pi_t(x_d)} \left(\int_{-\infty}^{x_d} \pi_t(y_d) \, \mathrm{d}y_d \left(\sum_{l=1}^{d} I_t^{(l)} - \sum_{l=1}^{d-1} I_t^{(l)} \right) - \int_{-\infty}^{x_d} \log L_d(y_d) \pi_t(y_d) \, \mathrm{d}y_d \right) \\
&= \frac{\lambda'(t)}{\pi_t(x_d)} \left(\int_{-\infty}^{x_d} (I_t^{(d)} - \log L_d(y_d)) \pi_t(y_d) \, \mathrm{d}y_d \right). \qquad (3.70)
\end{aligned}
$$

□

The above result is intuitive: although Proposition 3.12 holds for any set of admissible scalar functions $\{g_i\}_{i=1}^{d-1}$, access to marginal information allows us to construct a flow with more structure.

Example 3.14. *Consider the Gaussian curve in Example 3.7 with model parameters* $\mu_0 = 0_2, \Sigma_0 = H = I_2, R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \rho = 0.85, y = (14.25, 14.25)^T$. *We see from (3.38) that as time progresses, the independent prior distribution simultaneously gets deformed and translated. Figure 3.3 illustrates that, on average, particles driven by (3.34) require less kinetic energy than that of (3.65)-(3.66) using scalar function* g_1

specified in Proposition 3.13. However, in the general non-Gaussian case, obtaining the minimal kinetic energy velocity field requires numerical resolution of the elliptic PDE (3.36).

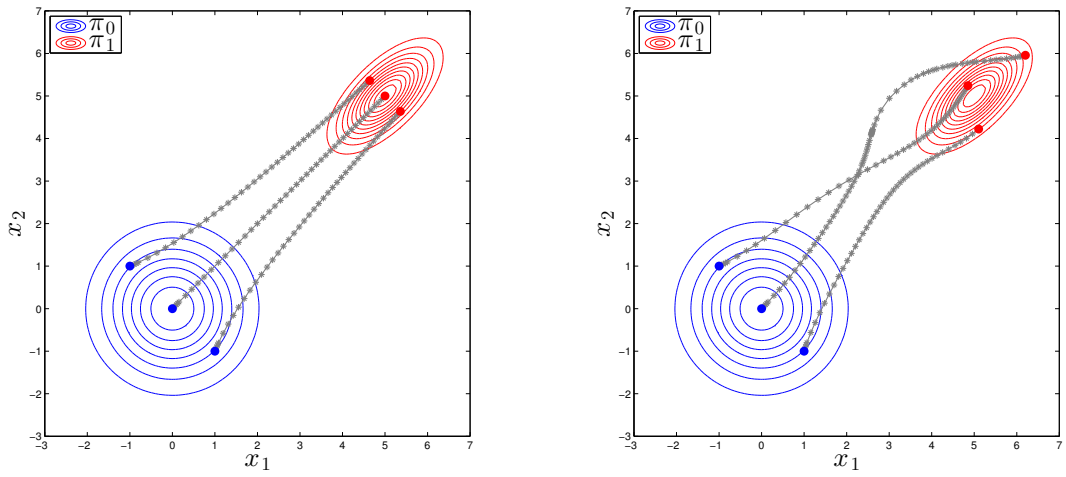

Figure 3.3: Bivariate Gaussian example. Three particle trajectories driven under different velocity fields but with the same initial conditions in both panels: (*left*) minimal kinetic energy velocity field (3.34); (*right*) velocity field (3.65)-(3.66) in Proposition 3.12. The asterisk symbols displayed correspond to time steps taken by an adaptive explicit fourth-order Runge-Kutta numerical integrator.

3.3.8 Gibbs flow approximation

Despite the explicit form of the flow transport solution on $\mathbb{R}^d$ introduced in Proposition 3.12, this flow still lacks tractability as a numerical implementation would require computing integrals of dimension up to d. For computational tractability, we approximate this flow transport by constructing a flow which tracks changes in the underlying full conditional distributions $\{\pi_t(\mathrm{d}x_i|x_{-i})\}_{i=1}^d$ of the distribution π_t instead. This amounts to having a coupled system of d one-dimensional flow transport problems for each of the full conditionals, which can be solved by computing only one-dimensional integrals in view of Proposition 3.9. We shall refer to this approximation as the *Gibbs flow*. Note the price to pay for this tractability: except when the posterior distribution factorizes, this flow does not solve the flow transport problem and only approximately tracks the curve $\mathcal{C}_\pi$. We now make these ideas more precise in the following proposition.

Proposition 3.15. *Consider the Gibbs velocity field* $\tilde{f} = \left(\tilde{f}_1, \ldots, \tilde{f}_d\right)^T$ *defined for* $i = 1, \ldots, d$ *by*

$$\tilde{f}_i(t,x) := \frac{-\int_{-\infty}^{x_i} \partial_t \pi_t(y_i|x_{-i})\,\mathrm{d}y_i}{\pi_t(x_i|x_{-i})} = \frac{\lambda'(t) I_t(x_{-i})\left(F_t(x_i|x_{-i}) - I_t^{x_i}(x_{-i})/I_t(x_{-i})\right)}{\pi_t(x_i|x_{-i})}, \tag{3.71}$$

where

$$\begin{aligned} I_t^{x_i}(x_{-i}) &:= \int_{-\infty}^{x_i} \log L(y_i, x_{-i}) \pi_t(y_i|x_{-i})\,\mathrm{d}y_i, \\ I_t(x_{-i}) &:= \int_{-\infty}^{\infty} \log L(y_i, x_{-i}) \pi_t(y_i|x_{-i})\,\mathrm{d}y_i, \end{aligned} \tag{3.72}$$

and $F_t(x_i|x_{-i}) := \int_{-\infty}^{x_i} \pi_t(y_i|x_{-i})\,\mathrm{d}y_i$ *is the CDF of* $\pi_t(\mathrm{d}x_i|x_{-i})$*. The Gibbs velocity field solves the following system of coupled Liouville equations*

$$\partial_t \pi_t(x_i|x_{-i}) = -\partial_{x_i}(\pi_t(x_i|x_{-i})\tilde{f}_i(t,x)) \tag{3.73}$$

for $i = 1, \ldots, d$*, each of which defined on* $(0,1)\times\mathbb{R}$*. Additionally, if* $\lim_{|x|\to\infty} L(x) = 0$*, then for any initial condition* $x_0 \in \mathbb{R}^d$*, an ODE with drift function (3.71) admits a unique solution on* $[0,1]$*.*

Proof. Using continuity of π_0 and L, routine application of the first fundamental theorem of calculus shows that the coupled system of ODEs corresponding to (3.71) solves (3.73). The assumptions on π_0 and L imply $\tilde{f} \in C^1([0,1]\times\mathbb{R}^d, \mathbb{R}^d)$; hence for any compact set $A \subset \mathbb{R}^d$, its derivative is bounded on $[0,1]\times A$ and local Lipschitzness follows.

Recall that since $\tilde{f}$ is locally Lipschitz, we need to demonstrate that the solution $x(t;x_0)$ of an ODE with drift function $\tilde{f}$ is bounded whenever it exists to complete the proof. Boundedness will be obtained by establishing that $V(x) := |x|^2$ is a Lyapunov function. It can be shown that the time evolution of each full conditional distribution is given by

$$\partial_t \pi_t(x_i|x_{-i}) = \lambda'(t)\left(\log L(x) - I_t(x_{-i})\right)\pi_t(x_i|x_{-i}). \tag{3.74}$$

By assumption $\log L(x) \to -\infty$ as $|x| \to \infty$, so for each $x_{-i} \in \mathbb{R}^{d-1}$, there exists $R_i > 0$ such that $\log L(x) < I_t(x_{-i})$ for $|x_i| > R_i$. This implies that

$x_i \mapsto -x_i \int_{-\infty}^{x_i} \partial_t \pi_t(y_i|x_{-i})\,\mathrm{d}y_i = x_i \int_{x_i}^{\infty} \partial_t \pi_t(y_i|x_{-i})\,\mathrm{d}y_i < 0$. Therefore we may choose a sufficiently large $R > 0$ such that

$$\frac{d}{dt}V(x) = 2\left\langle x, \tilde{f}(t,x)\right\rangle < 0 \tag{3.75}$$

for $x \in \mathbb{R}^d \setminus B(0,R)$, where $B(0,R) := \{x \in \mathbb{R}^d : |x| < R\}$. It follows that $|x(t;x_0)| < \max\{R, |x_0|\}$ for all $t \in [0,1]$. □

Equation (3.71) is the full conditional analogue of (3.59), so the interpretations made in Section 3.3.5 now carry over to each full conditional level. We stress that an evaluation of the Gibbs velocity field only requires computation of one-dimensional integrals as the normalizing constant $Z(t)$ cancels in the expression:

$$\begin{aligned}\tilde{f}_i(t,x) = \lambda'(t)\Bigg\{ & F_t(x_i|x_{-i}) \int_{-\infty}^{\infty} \log L(y_i,x_{-i})\pi_0(y_i,x_{-i})L(y_i,x_{-i})^{\lambda(t)}\,\mathrm{d}y_i \\ & - \int_{-\infty}^{x_i} \log L(y_i,x_{-i})\pi_0(y_i,x_{-i})L(y_i,x_{-i})^{\lambda(t)}\,\mathrm{d}y_i \Bigg\} \Big/ \pi_0(x)L(x)^{\lambda(t)}\end{aligned} \tag{3.76}$$

for $i = 1,\ldots,d$, where

$$F_t(x_i|x_{-i}) = \frac{\int_{-\infty}^{x_i} \pi_t(y_i,x_{-i})\,\mathrm{d}y_i}{\int_{-\infty}^{\infty} \pi_t(z_i,x_{-i})\,\mathrm{d}z_i} = \frac{\int_{-\infty}^{x_i} \pi_0(y_i,x_{-i})L(y_i,x_{-i})^{\lambda(t)}\,\mathrm{d}y_i}{\int_{-\infty}^{\infty} \pi_0(z_i,x_{-i})L(z_i,x_{-i})^{\lambda(t)}\,\mathrm{d}z_i}. \tag{3.77}$$

When initialized at $\tilde{\pi}_0 = \pi_0$, under the conditions of Proposition 3.15, $\tilde{f}$ induces a curve of probability measures $\mathcal{C}_{\tilde{\pi}} = \{\tilde{\pi}_t\}_{t\in[0,1]}$ in the sense that $\tilde{f} \in \mathcal{L}(\mathcal{C}_{\tilde{\pi}})$. In the following proposition, we establish a quantitative bound between $\tilde{\pi}_t$ and π_t as a function of the following time-dependent local error which compares how much the Gibbs flow mimics the desired change in mass (3.10):

$$\varepsilon_t(x) := \left|\partial_t \pi_t(x) + \nabla \cdot (\pi_t(x)\tilde{f}(t,x))\right| \tag{3.78}$$

$$= \left|\partial_t \pi_t(x) - \sum_{i=1}^{d} \partial_t \pi_t(x_i|x_{-i})\pi_t(x_{-i})\right| \tag{3.79}$$

$$= \lambda'(t)\pi_t(x)\left|\log L(x) - I_t - \sum_{i=1}^{d}\left(\log L(x) - I_t(x_{-i})\right)\right| \tag{3.80}$$

for $(t,x) \in (0,1)\times\mathbb{R}^d$. Recall that when deriving Liouville's equation in Section 3.2.1, we summed over all axes in (3.23) to obtain the net rate at which probability mass is accumulating in a given control volume. The sum in (3.79) reveals that

there is no interaction between components of (3.71), i.e. no information about how much probability mass is changing in a particular direction is shared with the other components, in contrast with the telescopic sum in the proof of Proposition 3.12. The latter behaviour is a consequence of breaking down a global problem in d dimensions to d many one-dimensional problems.

Proposition 3.16. *Let $\mathcal{C}_\pi = \{\pi_t\}_{t\in[0,1]}$ be the curve of probability measures defined in (3.1). Assume the conditions of Proposition 3.15 and denote by $\mathcal{C}_{\tilde{\pi}} = \{\tilde{\pi}_t\}_{t\in[0,1]}$ the curve of probability measures induced by the Gibbs velocity field $\tilde{f}$ defined in (3.71) when initialized at $\tilde{\pi}_0 = \pi_0$. Suppose additionally that $|\tilde{f}(t,x)|\,\tilde{\pi}_t(x) \to 0$ as $|x| \to \infty$ for each $t \in [0,1]$. Then the error involved in the Gibbs flow transport approximation is characterized by the following inequality for $t \in (0,1]$*

$$\|\tilde{\pi}_t - \pi_t\|^2_{L^2(\mathrm{d}x)} \leq t \int_0^t \|\varepsilon_s\|^2_{L^2(\mathrm{d}x)}\,\mathrm{d}s \cdot \exp\left(1 + \int_0^t \|\nabla \cdot \tilde{f}(s,\cdot)\|_\infty\,\mathrm{d}s\right). \tag{3.81}$$

Proof. Since $\mathcal{L}(\mathcal{C}_\pi) \cap \mathcal{E}(\mathcal{C}_\pi)$ is non-empty by Proposition 3.12, let f be a velocity field from this set. Note that the time evolution of the distributions $\{\tilde{\pi}_t\}_{t\in[0,1]}$ induced by $\tilde{f}$ is governed by another Liouville equation:

$$\partial_t \tilde{\pi}_t = -\nabla \cdot (\tilde{\pi}_t \tilde{f}). \tag{3.82}$$

Define $\Delta : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ as the difference $\Delta_t := \pi_t - \tilde{\pi}_t$. By taking the difference between

$$\partial_t \pi_t = -\nabla \cdot (\pi_t f) \tag{3.83}$$

and (3.82) and introducing a cross term, we obtain

$$\partial_t \Delta_t = -\nabla \cdot (\pi_t(f - \tilde{f}) + \Delta_t \tilde{f}). \tag{3.84}$$

Multiplying throughout by Δ_t and applying chain rule yields

$$\frac{1}{2}\partial_t \Delta_t^2 = -(\nabla \cdot \tilde{f})\Delta_t^2 - \frac{1}{2}\left\langle \tilde{f}, \nabla \Delta_t^2 \right\rangle - \nabla \cdot (\pi_t(f - \tilde{f}))\Delta_t. \tag{3.85}$$

We then integrate by parts and note that the boundary term vanishes since $x \mapsto |\tilde{f}(t,x)|\,\tilde{\pi}_t(x)$ vanishes in the tails by assumption and $x \mapsto |\tilde{f}(t,x)|\,\pi_t(x)$ also vanishes in the tails under the construction (3.71):

$$\partial_t \|\Delta_t\|^2_{L^2(\mathrm{d}x)} = -\int_{\mathbb{R}^d} (\nabla \cdot \tilde{f})\Delta_t^2 \,\mathrm{d}x - 2\int_{\mathbb{R}^d} \nabla \cdot (\pi_t(f-\tilde{f}))\Delta_t \,\mathrm{d}x. \tag{3.86}$$

Using Young's inequality gives

$$\begin{aligned}\partial_t \|\Delta_t\|^2_{L^2(\mathrm{d}x)} &\leq \left|\int_{\mathbb{R}^d} (\nabla \cdot \tilde{f})\Delta_t^2 \,\mathrm{d}x\right| + 2\left|\int_{\mathbb{R}^d} \nabla \cdot (\pi_t(f-\tilde{f}))\Delta_t \,\mathrm{d}x\right| \\ &\leq \|\nabla \cdot \tilde{f}(t,\cdot)\|_\infty \|\Delta_t\|^2_{L^2(\mathrm{d}x)} + \delta^{-1}\|\Delta_t\|^2_{L^2(\mathrm{d}x)} + \delta\|\varepsilon_t\|^2_{L^2(\mathrm{d}x)}\end{aligned} \tag{3.87}$$

for any $\delta > 0$. Since $\tilde{\pi}_0 = \pi_0$, integrating both sides of (3.87) on $[0,t]$ yields

$$\|\Delta_t\|^2_{L^2(\mathrm{d}x)} \leq \delta \int_0^t \|\varepsilon_s\|^2_{L^2(\mathrm{d}x)} \,\mathrm{d}s + \int_0^t \left(\|\nabla \cdot \tilde{f}(s,\cdot)\|_\infty + \delta^{-1}\right) \|\Delta_s\|^2_{L^2(\mathrm{d}x)} \,\mathrm{d}s. \tag{3.88}$$

Now applying Gronwall's lemma on the time interval $[0,t]$ combined with the fact that $t \mapsto \delta \int_0^t \|\varepsilon_s\|^2_{L^2(\mathrm{d}x)} \,\mathrm{d}s$ is non-decreasing:

$$\|\Delta_t\|^2_{L^2(\mathrm{d}x)} \leq \delta \int_0^t \|\varepsilon_s\|^2_{L^2(\mathrm{d}x)} \,\mathrm{d}s \cdot \exp\left(\frac{t}{\delta} + \int_0^t \|\nabla \cdot \tilde{f}(s,\cdot)\|_\infty \,\mathrm{d}s\right). \tag{3.89}$$

Lastly, minimizing this upper bound with respect to δ gives (3.81). □

The upper bound (3.81) is tight in the sense that it is equal to zero when the posterior distribution factorizes. When this is not the case, we observe that the bound deteriorates with time which is expected as errors accumulate. This bound also suggests that the tempering function $\lambda(t)$ should be chosen such that its derivative $\lambda'(t)$ is small at those time instances when the integrated local error $\|\varepsilon_t\|^2_{L^2(\mathrm{d}x)}$ is large, as this would reduce the magnitude of the resulting L^2-error. Lastly, we note that the result is also applicable to other approximate flow transport as long as the local errors are measured in terms of (3.78).

To illustrate the nature of the Gibbs flow approximation, we return to Example 3.14 and observe the L^2-error at varying degrees of correlation, induced by the parameter ρ, and extremality of the observation y. The left panel of Figure 3.4 shows that while performance degrades with ρ, as expected from our construction, the approximation is able to exploit any local independence structure in the target

distributions, thus keeping the error reasonably small for moderate degrees of correlation. The right panel of Figure 3.4 reveals the inadequacy of the approximation when the overlap between the prior distribution and the likelihood function decreases, which is also to be expected.

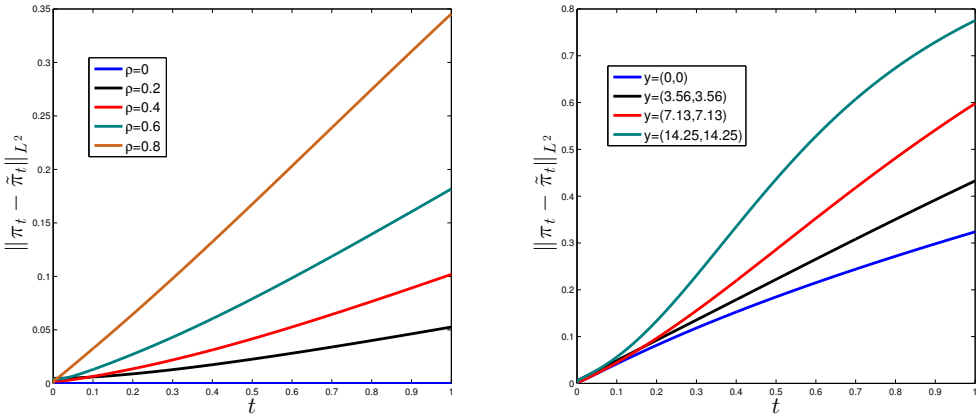

Figure 3.4: Bivariate Gaussian example. Error in L^2-norm at varying degrees of correlation ρ (*left*) and extremality of the observation y (*right*).

3.4 Gibbs flow implementation

3.4.1 Quadrature and numerical integration

Consider an ODE with the Gibbs velocity field (3.71) initialized at $X_0 \sim \pi_0$. A practical implementation of the Gibbs flow involves two source of approximations. Firstly, for most non-trivial problems, the integrals in the expression of the velocity field will not be analytically tractable so numerical approximation is necessary. Secondly, as the resulting ODE is also typically intractable, the use of a numerical integration scheme is also required. We now detail both approximations.

Observe from (3.76) that each evaluation of the Gibbs velocity field requires computing integrals of the form $\int_D \varphi(y_i, x_{-i})\,\mathrm{d}y_i$ for some integrand φ and domain $D \subseteq \mathbb{R}$. Here we consider the class of composite *Newton-Cotes quadrature* rules

$$\int_D \varphi(y_i, x_{-i})\,\mathrm{d}y_i \approx \sum_{r=1}^{R} \omega_r \varphi(z_r, x_{-i}) \tag{3.90}$$

where $\{\omega_r\}_{r=1}^R$ are quadrature weights which depend on the degree of the approximation and $\{z_r\}_{r=1}^R$ are $R \in \mathbb{N}$ many equispaced quadrature points in D (Iserles 2009; p. 34). Non-equispaced quadrature points, corresponding to other quadrature methods such as *Gaussian quadrature* and *Clenshaw-Curtis quadrature*, can also be used and may offer more stability and accuracy. We take (3.90) to be of the closed type, i.e. z_1 and z_R take the endpoints of D.[2] In what follows, the latter choice will be convenient when approximating integrals on domains of the type $D = (-\infty, x_i]$ for $x_i < \infty$. The composite quadrature rule (3.90) is derived by integrating Lagrange interpolation polynomials on subintervals; the degree of which dictates the accuracy of the approximation on each subinterval. We shall henceforth denote a numerical approximation of the Gibbs velocity field $\tilde{f}(t,x)$ by $\hat{f}(t,x)$.

For ease of presentation, here we consider the *forward Euler scheme* to numerically integrate an ODE with the approximate Gibbs velocity field $\hat{f}(t,x)$ at times $0 =: t_0 < \cdots < t_M := 1$ for some $M \in \mathbb{N}$. At time $t = 0$, we initialize a particle by sampling $X_0 \sim \pi_0$. Subsequently, for $m = 1, ..., M$, we move the particle with location X_{m-1} at time t_{m-1} to location X_m at time $t_m = t_{m-1} + \Delta t_m$ using the iteration

$$X_m := \Phi_m(X_{m-1}) := X_{m-1} + \Delta t_m \hat{f}(t_{m-1}, X_{m-1}), \tag{3.91}$$

which can be re-written as

$$X_m := \hat{T}_{t_m}(X_0) := \Phi_m \circ \cdots \circ \Phi_1(X_0). \tag{3.92}$$

More intricate *higher order methods* can also be used to define the mappings $\{\Phi_m\}_{m=1}^M$. However, as the Jacobians of these maps are needed in our context (see (3.93) and (3.95) below), implementation quickly becomes cumbersome. Additional smoothness assumptions would also be needed for these methods to achieve their full potential. For increased stability, *implicit methods* could also be considered but we prefer embedding a potentially less stable explicit scheme in the correction structure afforded by importance sampling to solving the non-linear equations that would otherwise arise.

[2]Unbounded domains are treated with suitable truncation.

3.4.2 Distribution of approximate Gibbs flow samples

We now show that it is possible to compute the distribution $\hat{\pi}_{t_m}$ of X_m generated by the iteration (3.92). This allows us to use the approximate Gibbs flow as a proposal distribution within importance sampling (Section 1.2.2), MCMC (Section 1.2.3) or SMC (Section 1.2.4) methods.

In Proposition 3.15 we showed that, under mild assumptions, an ODE with the Gibbs velocity field admits a unique solution; hence the Gibbs flow maps $\{\tilde{T}_t\}_{t\in[0,1]}$ are well-defined and are by construction C^1-diffeomorphisms. Therefore the maps $\{\hat{T}_{t_m}\}_{m=1}^M$ defined in (3.92), which are consistent approximations of $\{\tilde{T}_t\}_{t\in[0,1]}$, will be injective for sufficiently small step sizes $\{\Delta t_m\}_{m=1}^M$ and an adequate quadrature approximation – see Bunch and Godsill (2016), Liu and Wang (2016) for similar arguments. Under these conditions, it follows from (2.3) that the density of $\hat{\pi}_{t_m} = \hat{T}_{t_m\#}\pi_0$ is given by

$$\hat{\pi}_{t_m}(x) = \pi_0\left(\hat{T}_{t_m}^{-1}(x)\right)\left|\det\left(\nabla\hat{T}_{t_m}\left(\hat{T}_{t_m}^{-1}(x)\right)\right)\right|^{-1}, \tag{3.93}$$

where $\hat{T}_{t_m}^{-1}:\mathbb{R}^d\to\mathbb{R}^d$ and $\nabla\hat{T}_{t_m}:\mathbb{R}^d\to\mathbb{R}^{d\times d}$ denote the inverse and Jacobian matrix of $\hat{T}_{t_m}$ respectively. In numerical implementations, monotonicity may be monitored by checking for any sign changes in the Jacobian determinant.

We see from (3.91)-(3.92) that computing the Jacobian of $\hat{T}_{t_m}$ requires the Jacobians of the mappings Φ_m for $k=1,\ldots,m$, which in turn requires the Jacobian of $\hat{f}$. Analytical tractability of the Gibbs velocity field allows us to obtain an exact expression of the Jacobian of $\tilde{f}$ (see Appendix A.1). When integrals in $\tilde{f}$ are replaced by quadrature approximations (3.90) to yield $\hat{f}$, it turns out that the Jacobian of $\hat{f}$ can be obtained by replacing integrals in the Jacobian of $\tilde{f}$ with approximations based on the same quadrature rule. This result follows straightforwardly for off-diagonal terms of the Jacobian matrix using linearity. For diagonal entries, we have to compute partial derivatives with respect to x_i of approximations of integrals of the form $\int_{-\infty}^{x_i}\varphi(y_i,x_{-i})\,\mathrm{d}y_i$, which can be done using the following argument. Denote by $\hat{\varphi}$ the underlying Lagrange interpolant

giving rise to the quadrature rule (3.90). Using the first fundamental theorem of calculus and the closed property of (3.90) gives

$$\partial_{x_i} \sum_{r=1}^{R} \omega_r \varphi(z_i, x_{-i}) = \partial_{x_i} \int_{-\infty}^{x_i} \hat{\varphi}(y_i, x_{-i}) \, \mathrm{d}y_i = \hat{\varphi}(x_i, x_{-i}) = \varphi(x_i, x_{-i}). \tag{3.94}$$

Hence with $N \in \mathbb{N}$ iid samples $\{X_m^n\}_{n=1}^N$ from $\hat{\pi}_{t_m}$, we can form a particle approximation of π_{t_m} using $\sum_{n=1}^N W_m^n \delta_{X_m^n}$, where the normalized weights are given by

$$W_m^n \propto \frac{\pi_{t_m}(X_m^n)}{\hat{\pi}_{t_m}(X_m^n)} \propto W_{m-1}^n \frac{\pi_{t_m}(X_m^n)}{\pi_{t_{m-1}}(X_{m-1}^n) \left|\det\left(\nabla \Phi_m(X_{m-1}^n)\right)\right|^{-1}} \tag{3.95}$$

for $m = 1, \ldots, M$ with initialization $W_0^n = N^{-1}$. In the SMC sampler framework, we note that (3.95) corresponds to employing the optimal backward kernel (1.49), which is tractable in this case owing to the use of deterministic dynamics.

At each time iteration, the computational cost involved is $O(dR)$ to perform quadrature plus the cost involved in computing the Jacobian determinant of a $d \times d$ matrix. In general, the latter has a computational cost of order $O(d^3)$; that said, there are more efficient implementations such as the Strassen algorithm with slightly lower cost. However, this cost will be significantly lowered in statistical models with conditional independence structures since, by construction, the Gibbs flow will exploit such structures to yield sparse Jacobian matrices – see Equation (3.71). For example, the Jacobian associated to a chain-shaped undirected graphical model is a tridiagonal matrix, so computing its determinant only requires a cost that is linear in d.

3.4.3 Combining the Gibbs flow with annealed importance sampling

The AIS algorithm, described in Section 1.2.4, performs poorly whenever the underlying MCMC kernels $\{K_m\}_{m=1}^M$ mix slowly and/or the intermediate distributions $\{\pi_{t_m}\}_{m=1}^M$ are too distant. In such a scenario, it is natural to combine the Gibbs flow with AIS. Similar ideas were suggested by Vaikuntanathan and Jarzynski (2008; 2011) but the authors did not propose a generic methodology to construct an approximate flow transport.

Practically, for $n = 1, \ldots, N$, one initializes by sampling $X_0^n \sim \pi_0$ and setting $\tilde{X}_0^n = X_0^n$, and for $m = 1, \ldots, M$, iterate by setting $X_m^n = \Phi_m(\tilde{X}_{m-1}^n)$ and sampling $\tilde{X}_m^n \sim K_m(X_m^n, \cdot)$. Such a procedure falls under the SMC framework: by choosing the optimal backward kernel (1.49) for the deterministic maps $\{\Phi_m\}_{m=1}^M$ and the time reversed backward kernel (1.54) for the MCMC kernels $\{K_m\}_{m=1}^M$, a particle approximation of π_{t_m} is given by $\sum_{n=1}^N W_m^n \delta_{\tilde{X}_m^n}$, where the normalized weights are defined by the following recursion

$$W_m^n \propto W_{m-1}^n \frac{\pi_{t_m}(X_m^n)}{\pi_{t_{m-1}}(\tilde{X}_{m-1}^n) \left| \det \left(\nabla \Phi_m(\tilde{X}_{m-1}^n) \right) \right|^{-1}} \tag{3.96}$$

for $m = 1, \ldots, M$ with initialization $W_0^n = N^{-1}$. We summarize the resulting sampler in Algorithm 2.

Re-visiting Example 3.14, Figure 3.5 illustrates the difference in terminal particle locations when running solely Gibbs flow, AIS and combining Gibbs flow with AIS. We observe that the combination of the diffusive behaviour of RWMH kernels used within AIS moves and the deterministic mappings obtained by approximating the Gibbs flow provides particles whose terminal positions overlap much better with the support of the posterior distribution than using solely Gibbs flow or AIS.

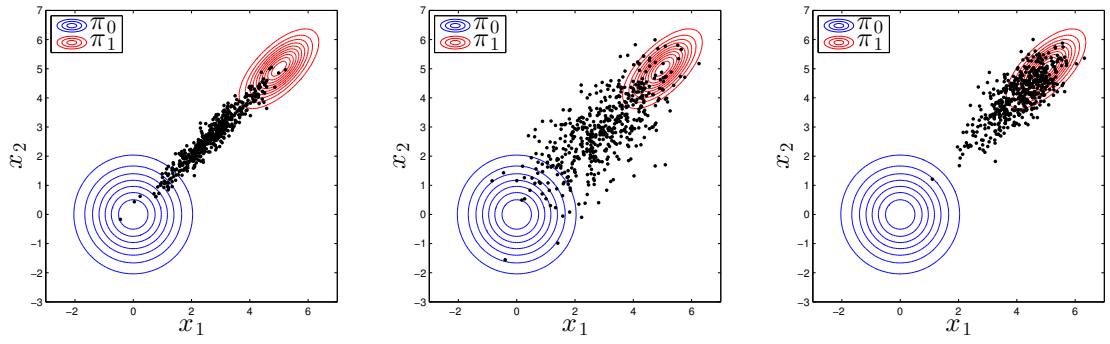

Figure 3.5: Bivariate Gaussian example. Terminal particle positions of $N = 500$ prior samples whose time evolution were prescribed by: (*left*) Gibbs flow iteration in (3.91); (*middle*) AIS with random walk Metropolis-Hastings (RWMH) kernels; (*right*) combining (3.91) with the corresponding RWMH kernel used in AIS.

Algorithm 2 Gibbs flow AIS sampler

Input: particles N, time steps M, step sizes $\{\Delta t_m\}_{m=1}^M$, quadrature points R, Markov kernels $\{K_m\}_{m=1}^M$, resampling threshold $\theta \in (0,1)$.

1. Initialization: sample $X_0^n \sim \pi_0$ and set $\tilde{X}_0^n = X_0^n, W_0^n = N^{-1}$ for $n = 1, \ldots, N$.
2. For $m = 1, \ldots, M$,
   (a) set $X_m^n = \Phi_m(\tilde{X}_{m-1}^n)$ using (3.91);
   (b) update normalized weights $\{W_m^n\}_{n=1}^N$ using (3.96);
   (c) if $\mathrm{ESS}_m < \theta N$, resample particles and set $W_m^n = N^{-1}$ for $n = 1, \ldots, N$;
   (d) sample $\tilde{X}_m^n \sim K_m(X_m^n, \cdot)$ for $n = 1, \ldots, N$;
   (e) compute ratio of normalizing constants estimate $\hat{R}_m^N$ using (1.47).

Output: particles $\{\tilde{X}_M^n\}_{n=1}^N$, normalized weights $\{W_M^n\}_{n=1}^N$ and normalizing constant estimate $\hat{Z}_M^N = \prod_{m=1}^M \hat{R}_m^N$.

3.4.4 Selecting the tempering schedule

The tempering function λ in (3.1), that controls the rate at which we want to introduce the likelihood, has significant impact on the performance of both AIS and the aforementioned methodology. Various methods have been proposed to select this function in different settings: see Gelman and Meng (1998) in the context of path sampling, Section 1.2.4.4 for SMC samplers when the time reversed backward kernel (1.54) is employed and Betancourt (2014) for a novel Hamiltonian flow, based on adiabatic processes in thermodynamics, where the schedule is determined dynamically.

For our purposes, recall from the discussion after Proposition 3.16 that λ should be chosen so that its derivative is small whenever the time-dependent integrated local error (3.78) is large. The latter is typically substantial whenever there are large changes between intermediate distributions. Noting that large changes along $\mathcal{C}_\pi$ necessarily imply large changes in the corresponding full conditionals, the time steps $\{\Lambda_m\}$ taken by an adaptive scheme to numerically integrate the Gibbs ODE may be used to guide the choice of a suitable tempering function λ. This is because large variations in (3.71) will require smaller step sizes to keep estimates of numerical integration error below a pre-specified tolerance.

We demonstrate this on the curve of probability measures (3.97) arising from a Bayesian mixture modelling application detailed later in Section 3.5. Observe from Figure 3.6 that with a linear tempering function, large changes along the curve $\mathcal{C}_\pi$ occur at very early times. We advocate selecting λ so that the time steps $\{\Lambda_m\}$ taken by an adaptive numerical integrator is as close as possible to being equispaced on $[0, 1]$ – up to some variability between different initial conditions. Figure 3.7 shows that this can be achieved for this example by setting $\lambda(t) = t^6$.

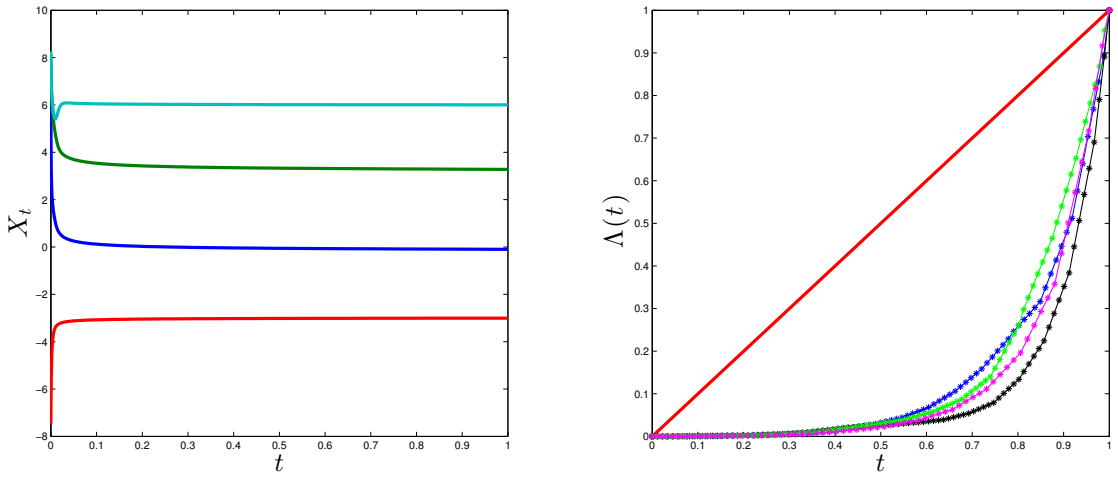

Figure 3.6: Mixture modelling example with $\lambda(t) = t$. (*Left*) Trajectory of a particle under the Gibbs flow with different colors representing each dimension. (Right) Colored lines with asterisk symbols correspond to the time steps taken by an adaptive numerical integrator for four different prior samples evolving under the Gibbs flow to be compared against the red identity line.

3.5 Applications

3.5.1 Bayesian mixture modelling

We now demonstrate the performance of Gibbs flow based algorithms on a Bayesian mixture model where the posterior distribution of mixture means is inferred. This is a canonical example of distributions with multiple well-separated modes.

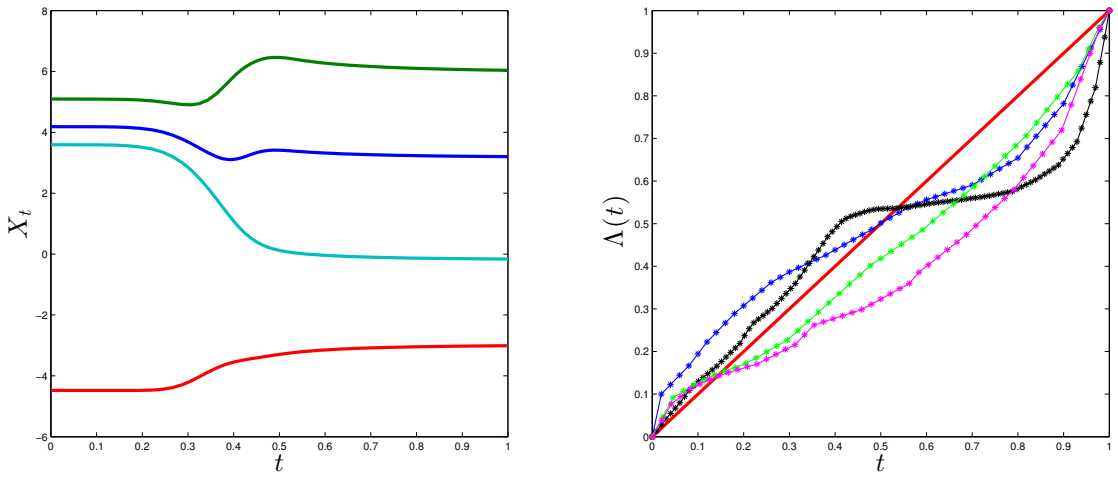

Figure 3.7: Mixture modelling example with $\lambda(t) = t^6$. (*Left*) Trajectory of a particle under the Gibbs flow with different colors representing each dimension. (Right) Colored lines with asterisk symbols correspond to the time steps taken by an adaptive numerical integrator for four different prior samples evolving under the Gibbs flow to be compared against the red identity line.

3.5.1.1 Model description

Consider $n \in \mathbb{N}$ independent observations $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ from a univariate Gaussian mixture model with $d \in \mathbb{N}$ components, i.e. y_i is a realization from $\frac{1}{d}\sum_{j=1}^{d} \mathcal{N}(x_j, \sigma_j^2) \in \mathcal{P}(\mathbb{R})$. Following Lee et al. (2010), we set $d = 4$, $\sigma_j = \sigma = 0.55$ for $j = 1, \ldots, d$ and perform inference only on the mean parameters $x \in \mathbb{R}^4$. We generate the data y using $n = 100$ simulations from the model with parameter value $x^* = (-3, 0, 3, 6)^T$ and stratification between components. We prescribe a uniform prior distribution on the d-dimensional hypercube $[-10, 10]^d$. The curve of probability measures $\mathcal{C}_\pi$ is simply

$$\pi_t(\mathrm{d}x) = \frac{\mathbb{1}_{[-10,10]^d}(x) L(x; y)^{\lambda(t)} \mathrm{d}x}{20^d Z(t)} \tag{3.97}$$

for $t \in [0, 1]$, where

$$L(x; y) = \frac{1}{d^n} \prod_{i=1}^{n} \sum_{j=1}^{d} \mathcal{N}(y_i; x_j, \sigma^2) \tag{3.98}$$

for $x \in \mathbb{R}^d$. It follows from exchangeability of the prior and non-identifiability of mixture components that the posterior distribution is invariant under "label

permutation". Therefore $\pi = \pi_1$ admits $d! = 24$ well-separated modes centered approximately around all permutations of x^*.

3.5.1.2 The Gibbs flow approximation

Firstly, we investigate the quality of the Gibbs flow approximation, before employing any importance sampling correction. We do so by comparing the time evolution of $N = 1000$ prior samples under the Gibbs flow with the output of a SMC sampler, using the configuration described in Lee et al. (2010), as the reference truth in Figure 3.8.

The performance of the approximation for this challenging problem is striking; particles were able to reach all 24 modes in $\mathbb{R}^4$. This is corroborated in Figure 3.9 which plots all pairs of marginal posterior distributions on $\mathbb{R}^2$ (note that each of these admits 12 well-separated modes) and in Figure 3.10 which displays the proportion of particles in each of the 24 modes when particles were initialized as a latin hypercube sample of size $N = 1000$ (to reduce the variance from prior sampling). We note that the similarity in the proportions observed at each mode demonstrates a "global" nature of the Gibbs flow approximation.

3.5.1.3 Comparison of algorithmic performance

We now compare a SMC sampler based on Gibbs flow (Section 3.4.2), AIS with MALA moves (Section 1.2.4 & 1.2.3.3) and a SMC sampler which combines Gibbs flow with RWMH moves (Section 3.4.3). Following the discussion in Section 3.4.4, we select the tempering function as $\lambda(t) = t^6$. The choice of numerical integration scheme is the forward Euler method with step sizes selected to ensure monotonicity of the mappings defined in (3.92). Using the left panel of Figure 3.7, we prescribe a piecewise linear time discretization to focus our computational effort at times with more particle motion. At each time iteration, we allow the AIS sampler and the Gibbs-AIS sampler to take 10 MCMC moves, which are tuned to achieve suitable acceptance probabilities. All one-dimensional integrals involved in evaluations of the Gibbs velocity field and its Jacobian were computed using a composite Simpsons rule with 50 quadrature points.

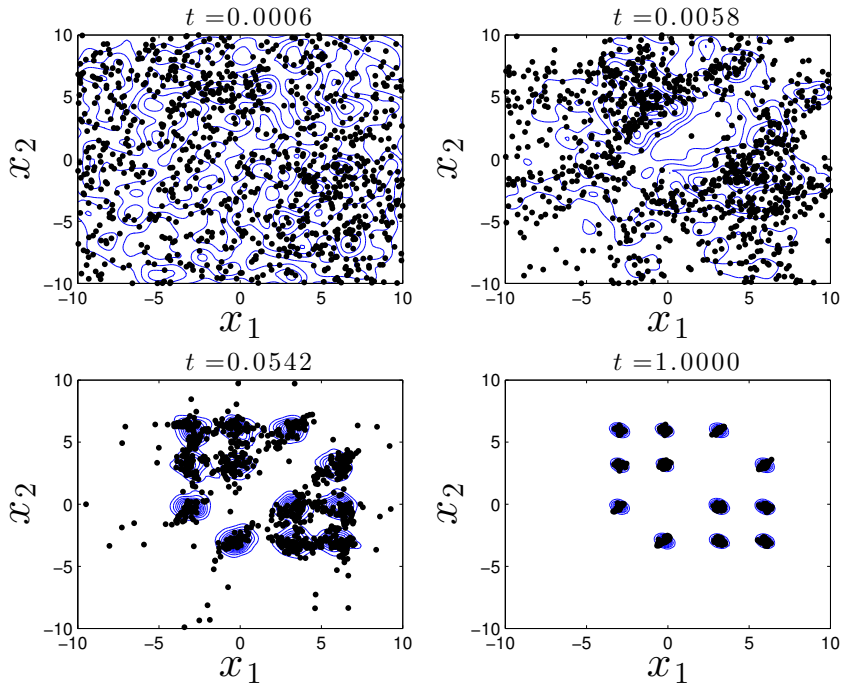

Figure 3.8: Time evolution of $N = 1000$ prior samples under the Gibbs flow (*black dots*) up to time $t = 1$. For each time instance, the superimposed blue contours represent the target distribution obtained as a kernel density estimate from the output of a SMC sampler.

We measure algorithmic performance in terms of ESS (Section 1.2.2.4); for the comparison to be meaningful, we do not perform resampling. To yield a fair comparison, we set the number of time steps taken by each algorithm so as to match computational cost, measured in terms of run time. The results displayed in Figure 3.11 show that the Gibbs-AIS sampler outperforms the other algorithms. The reason for poor performance of the sampler based solely on Gibbs flow can be seen in Figure 3.8 and 3.9; the distribution of samples under the Gibbs flow is a poor importance distribution as it has thinner tails than the target distribution. The latter is not a difficulty when one combines Gibbs flow with AIS owing to the diffusivity introduced in the MCMC moves.

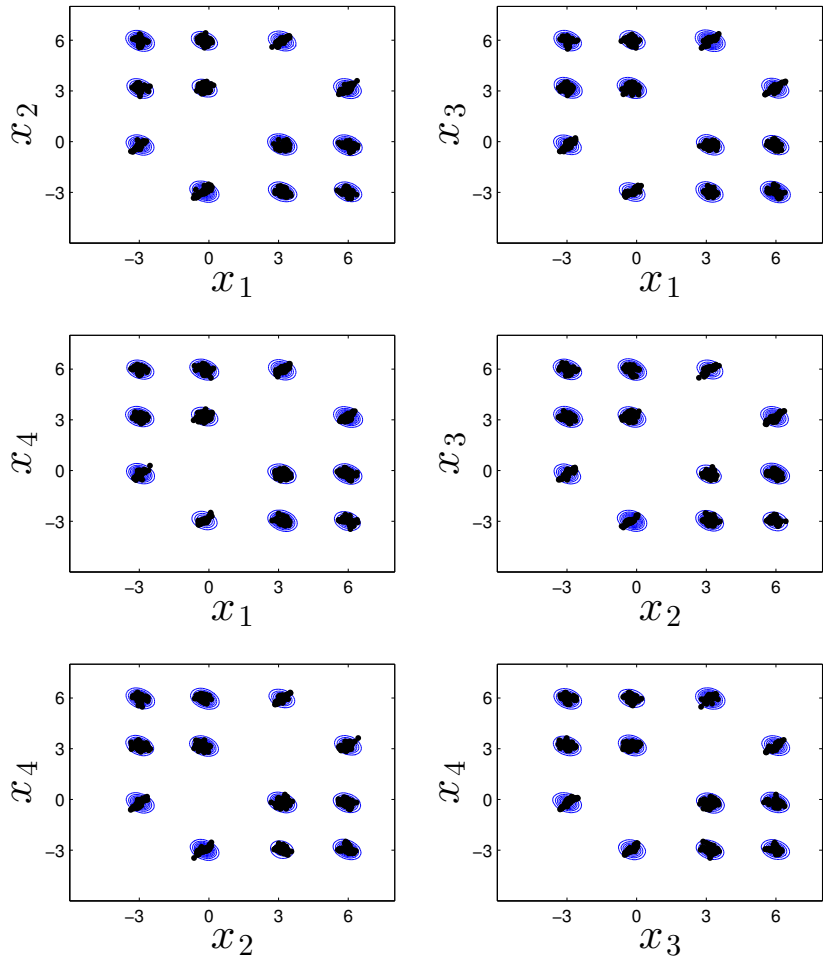

Figure 3.9: All pairs of marginal posterior distributions on $\mathbb{R}^2$.

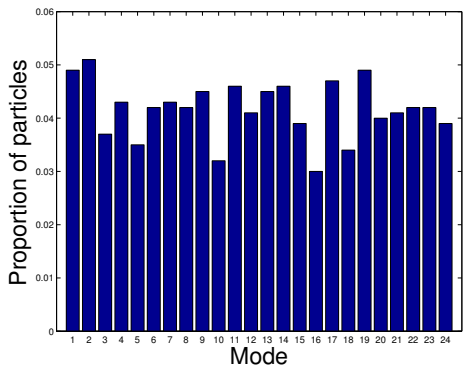

Figure 3.10: Proportion of particles in each of the 24 modes in $\mathbb{R}^4$.

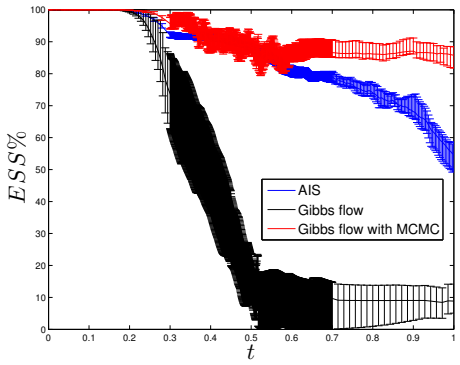

Figure 3.11: Time evolution of ESS%. Lines and error bars indicate median and interquartile range of 20 repetitions respectively.

3.5.2 Sampling truncated multivariate Gaussians with applications to probit models

So far we have restricted our attention to problems where we build the curve of probability measures $\mathcal{C}_\pi$ by tempering the likelihood function L in (3.1). We now show that these assumptions can be relaxed by adapting the Gibbs flow approximation to sample from truncated multivariate Gaussian distributions and illustrate in Section 3.5.2.3 how this procedure can be included in a MCMC algorithm to perform Bayesian inference for a multivariate probit model.

3.5.2.1 Model and Gibbs flow construction

Let $\pi_0 = \mathcal{N}(\mu, \Sigma)$ be a multivariate Gaussian on $\mathbb{R}^d$ and denote the truncated distribution as π. Assume that the truncation happens component-wise, i.e. the support of π is $\text{supp}(\pi) := \prod_{i=1}^d (a_i, b_i)$ where $a_i, b_i \in \bar{\mathbb{R}}$ and $a_i < b_i$ for all $i = 1, \ldots, d$. If the truncation is extreme, it is natural to introduce a sequence of bridging distributions by performing the truncation gradually. More precisely, we build a curve of probability measures $\mathcal{C}_\pi$ via

$$\pi_t(\mathrm{d}x) := \frac{\pi_0(\mathrm{d}x) \prod_{i=1}^d \mathbb{1}_{(\alpha_i(t), \beta_i(t))}(x_i)}{Z(t)}, \tag{3.99}$$

where for all $i = 1, \ldots, d$, $\alpha_i : [0,1] \to \bar{\mathbb{R}}$ is non-decreasing with boundary conditions $\alpha_i(0) = -\infty$, $\alpha_i(1) = a_i$, $\beta_i : [0,1] \to \bar{\mathbb{R}}$ is non-increasing with boundary conditions $\beta_i(0) = \infty$, $\beta_i(1) = b_i$, $\alpha_i(t) < \beta_i(t)$ for all $t \in [0,1]$

and $Z(t) := \pi_0\left(\prod_{i=1}^{d}(\alpha_i(t), \beta_i(t))\right)$. In some applications, the tail probability $Z := Z(1) = \pi_0\left(\prod_{i=1}^{d}(a_i, b_i)\right)$ is the quantity of interest. From these assumptions, it is clear that the curve $\mathcal{C}_\pi$ connects π_0 to $\pi_1 = \pi$. In contrast to having a tempering function, $\{\alpha_i\}_{i=1}^{d}$ and $\{\beta_i\}_{i=1}^{d}$ now control the rate of truncation.

It can be shown that

$$\partial_t \pi_t(x_i|x_{-i}) = \left(\alpha_i'(t)\pi_t(\alpha_i(t)|x_{-i}) - \beta_i'(t)\pi_t(\beta_i(t)|x_{-i})\right)\pi_t(x_i|x_{-i}) \tag{3.100}$$

for $x \in \text{supp}(\pi_t) = \prod_{i=1}^{d}(\alpha_i(t), \beta_i(t))$, where α_i' and β_i' denote the time derivatives of α_i and β_i respectively. In the same manner as in Proposition 3.15, we can solve the system of Liouville equations (3.73) with

$$\begin{aligned}\tilde{f}_i(t,x) := \Bigg(&\alpha_i'(t)\pi_0(\alpha_i(t), x_{-i})\int_{x_i}^{\beta_i(t)}\pi_0(y_i, x_{-i})\,\mathrm{d}y_i \\ &+ \beta_i'(t)\pi_0(\beta_i(t), x_{-i})\int_{\alpha_i(t)}^{x_i}\pi_0(y_i, x_{-i})\,\mathrm{d}y_i\Bigg) \Bigg/ \pi_0(x)\int_{\alpha_i(t)}^{\beta_i(t)}\pi_0(y_i, x_{-i})\,\mathrm{d}y_i.\end{aligned} \tag{3.101}$$

The expression of the Jacobian of $\tilde{f}$ is given in Appendix A.2. As in Section 3.4.2, the Jacobian of $\tilde{f}$ with quadrature approximation can simply be computed by replacing the integrals in the Jacobian of $\tilde{f}$ by their quadrature approximations under the same rule.

3.5.2.2 Comparison of algorithmic performance

To address a similar problem, Moffa and Kuipers (2014) proposed a method based on SMC sampler (Section 1.2.4) and reported computational gains in comparison to the one-at-a-time Gibbs sampler when the degree of correlation in the multivariate Gaussian is significant. We adopt the simulation study in Moffa and Kuipers (2014; Section 4) and compare a SMC sampler based solely on Gibbs flow, AIS with RWMH moves and a SMC sampler which combines Gibbs flow with RWMH moves. Although it would also be interesting to compare these algorithms against the SMC sampler developed in Ridgway (2016), we will leave this to future work. Before proceeding, we note that MALA moves were not employed within AIS; the use of gradient information of π is not appropriate in this context as the gradient might

point to directions of zero probability mass. In contrast, flow transport provides a principled way to drift particles towards the right regions of the state space.

The effect of correlation. Consider $d = 4$, a mean vector of $\mu = (-\xi, -\xi, \xi, \xi)^T$ for $\xi > 0$, which keeps two components in the truncation region of $\text{supp}(\pi) = [0, \infty)^d$. For starters, we set $\xi = 1$. The off-diagonal elements of Σ are set to a value such that all pairwise correlations are equal to $\rho \in [0, 1]$. As before, the numerical integration scheme used is the forward Euler method with step sizes selected to ensure monotonicity of the mappings in (3.92).

We perform the truncation with $\alpha_i(t) = -t^{-1} + 1$ for all $i = 1, \ldots, d$. Using insight from preliminary simulations of the Gibbs flow, we select a piecewise linear time discretization to focus our computational effort at times with more particle motion. At each time iteration, we allow AIS and the Gibbs flow-AIS sampler to take 50 RWMH moves. The covariance of the Gaussian random walk is set as $\sigma\Sigma$, with $\sigma > 0$ tuned to achieve suitable acceptance probabilities. All one-dimensional integrals involved in evaluations of the Gibbs velocity field and its Jacobian are computed using a composite Simpsons rule with 40 quadrature points.

We perform the same ESS comparison as before by not resampling and setting the number of time steps taken by each algorithm to match computational cost, measured in terms of run time. The left panel of Figure 3.12 shows how the ESS of each sampler varies with the correlation parameter ρ. The results are striking and interesting: the performance of samplers based on Gibbs flow degrade with ρ whilst that of AIS which uses only RWMH moves improves with ρ (for this particular example, the overlap between π_0 and $\text{supp}(\pi) = [0, \infty)^d$ increases with ρ). This behaviour clearly illustrates the Gibbs flow's ability to exploit any local independence structure in π_0.

The effect of truncation extremality. Again for dimension $d = 4$, we now fix the correlation parameter at $\rho = 0.5$ and vary the location parameter ξ in the middle panel of Figure 3.12. All other algorithmic settings are the same as before. The

results show that as the truncation becomes extreme, the Gibbs flow can mitigate particle degeneracy by moving particles towards the right regions of the state space.

The effect of dimension. We now set correlation at $\rho = 0.5$, truncation at $\xi = 1$ and vary dimension d. Algorithmic settings are the same as before except that we now allow the number of RWMH moves taken at each time iteration to increase linearly with dimension. The results, summarized in the right panel of Figure 3.12, show that while the performance of all algorithms degrade with dimension, which is to be expected, combining flow transport with MCMC has the potential to allow SMC samplers to remain competitive in high dimensions.

Normalizing constant estimation. Lastly, we compare the performance of these algorithms to estimate the normalizing constant $Z = \pi_0\left([0,\infty)^d\right)$ as correlation parameter ρ, location parameter ξ and dimension d vary one at a time. Algorithmic settings are the same as above with the exception of applying systematic resampling whenever ESS falls below half of the number of particles used. As performance measure, in Figure 3.13 we plot the estimated standard deviation of the normalizing constant estimator (1.48) of AIS (with resampling) relative to the other two algorithms based on Gibbs flow. The results are similar to those obtained in the ESS comparisons and show that the sampler combining Gibbs flow with AIS provides an estimator with significantly lower variance.

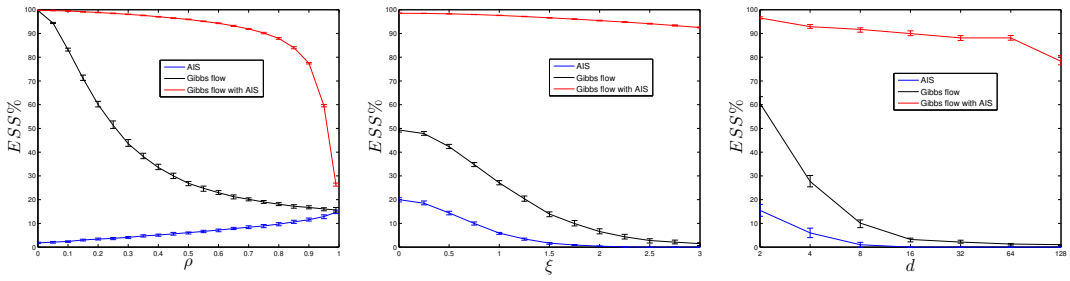

Figure 3.12: Comparison of ESS% between algorithms as the correlation parameter ρ (*left*), the location parameter ξ (*middle*) and dimension d (*right*) vary one at a time. Lines and error bars indicate median and interquartile range of 100 repetitions respectively.

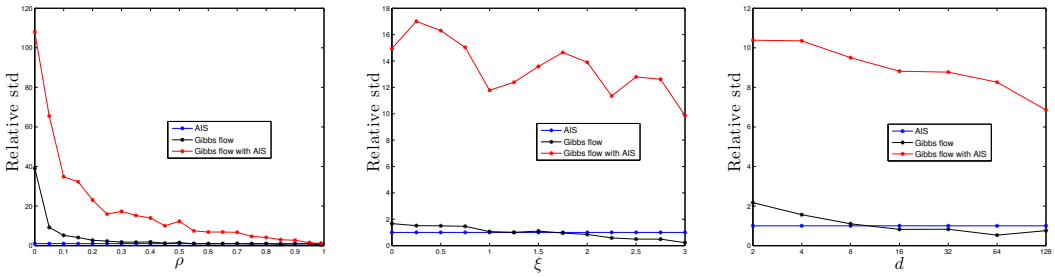

Figure 3.13: Estimated relative standard deviation (with AIS as benchmark) of normalizing constant estimators based on 100 repetitions as the correlation parameter ρ (*left*), the location parameter ξ (*middle*) and dimension d (*right*) vary one at a time.

3.5.2.3 Bayesian multivariate probit model

We now apply the above procedure to the Bayesian multivariate probit model discussed in Talhouk et al. (2012). Denote by $Y \in \{0,1\}^{n\times d}$ the $d \in \mathbb{N}$ dimensional binary responses on $n \in \mathbb{N}$ subjects, $X \in \mathbb{R}^{n\times p}$ the design matrix, $\beta \in \mathbb{R}^{p\times d}$ the regression coefficients and $R \in \mathbb{R}^{d\times d}$ a correlation matrix. For each subject $i = 1, \ldots, n$, the multivariate probit model specifies that the probability distribution of $Y_i \in \{0,1\}^d$ is given by

$$\mathbb{P}(Y_i = y_i | X, \beta, R) := \int_{\mathcal{I}_i} \mathcal{N}\left(z_i; (X\beta)_i, R\right) \mathrm{d}z_i, \tag{3.102}$$

where $(X\beta)_i$ is the i^{th} row of $X\beta$ and $\mathcal{I}_i := \mathcal{I}_{i1} \times \cdots \times \mathcal{I}_{id}$ with $\mathcal{I}_{ij} = [0,\infty)$ if $y_{ij} = 1$ and $(-\infty, 0)$ otherwise. We note that the restriction of R to correlation matrices in (3.102) ensures likelihood identifiability (Chib and Greenberg 1998). Equation (3.102) also prompts characterization of the model using Gaussian latent variables $Z \in \mathbb{R}^{n\times d}$ with the relation $Y_{ij} = \mathbb{1}_{\{Z_{ij}\geq 0\}}$.

We assign a prior distribution to β, R and the graph structure $G \in \mathcal{G}$ of the inverse correlation matrix R^{-1}. Interest here is to sample from the resulting posterior distribution

$$\pi(G, R, \beta, z|y) \propto \pi(G)\pi(R|G)\pi(\beta|R)\pi(z|\beta, R) \prod_{i=1}^{n} \mathbb{1}_{\{z_i \in \mathcal{I}_i\}}. \tag{3.103}$$

Our choice of prior is similar to Talhouk et al. (2012) which showed that it is possible to sample from the posterior using a Gibbs update for Z and β, a simple MH random

walk for G on the space of graphs $\mathcal{G}$ and a parameter expansion data augmentation step for R. For the latent Gaussian variables, the full conditional density factorizes as

$$\pi(z|\beta, R, y, G) \propto \prod_{i=1}^{n} \mathcal{N}\left(z_i; (X\beta)_i, R\right) \mathbb{1}_{\{z_i \in \mathcal{I}_i\}}. \tag{3.104}$$

The sampling scheme used in Talhouk et al. (2012) samples each $Z_i \in \mathbb{R}^d$ by updating its components one-at-a-time using a Gibbs sampler which leads to slow convergence of the resulting algorithm. To speed up convergence, we employ the above Gibbs-AIS sampler for truncated Gaussians, implemented as a conditional SMC update (Andrieu et al. 2010) to ensure validity of the resulting Gibbs sampler.

3.5.2.4 Six cities data set

We now apply the above methodology to analyze a well-known data set from the Six Cities longitudinal study on the health effects of air pollution.

Description. The data set concerned contains repeated binary measurements of $n = 537$ children's wheezing status from Steubenville, Ohio. Interest here is on modelling the probabilistic relation over time of the wheezing status of a child as a function of their age and their mother's smoking habit during the first year of the study. Notationally, the binary response y_{ij} indicates if child $i = 1, \ldots, n$ was wheezing in the $j = 1, 2, 3, 4$ year of the study (corresponding to when the subject was of age $7, 8, 9, 10$ respectively).

The nature of the data suggest that using a multivariate probit model to account for the structure of association between components of the multivariate binary response is appropriate. Table 3.1 also supports having mothers' smoking habits as a covariate.

We note that similar analyses have been conducted on this particular data set with differing inference procedures; see Chib and Greenberg (1998), Talhouk et al. (2012), Moffa and Kuipers (2014).

Mother's smoking status	7	8	9	10
Smoker	32 (17.0%)	40 (21.3%)	36 (19.1%)	27 (14.4%)
Non-smoker	55 (15.8%)	51 (14.6%)	49 (14.0%)	36 (10.0%)
Total	87 (16.2%)	91 (16.9%)	85 (15.8%)	63 (11.7%)

Table 3.1: Breakdown of wheezing cases by age group and initial smoking status of mothers. Percentages are with respect to each age group.

Algorithmic settings. Settings within the conditional Gibbs-AIS sampler used to update Z involved 10 particles with multinomial resampling triggered whenever the ESS falls below 5; a linear time discretization with 50 steps and 20 RWMH moves. We run $22,000$ iterations of the Gibbs sampler described earlier with estimation of the graph structure and use a burn-in of 2000 samples.

Results. Table 3.2 gives the posterior mean and standard deviation of parameters in the model. The results obtained are similar over independent runs of the algorithm with different initial values and to those reported in Talhouk et al. (2012). Figure 3.14 displays the most probable graphs under the posterior that were identified by the procedure. While the second most probable structure is the saturated model, the maximum a posteriori graph has a conditional independence structure between wheezing at age 7 and 9.

	Posterior mean	**Posterior standard deviation**
β_{11}	-0.9207	0.091
β_{12}	0.030	0.154
β_{21}	-0.9845	0.092
β_{21}	0.222	0.151
β_{31}	-1.012	0.096
β_{32}	0.180	0.156
β_{41}	-1.179	0.100
β_{42}	0.170	0.165
R_{12}	0.496	0.069
R_{13}	0.423	0.073
R_{14}	0.483	0.073
R_{23}	0.588	0.058
R_{24}	0.470	0.075
R_{34}	0.549	0.066

Table 3.2: Posterior mean and standard deviation of parameters in the multivariate probit model.

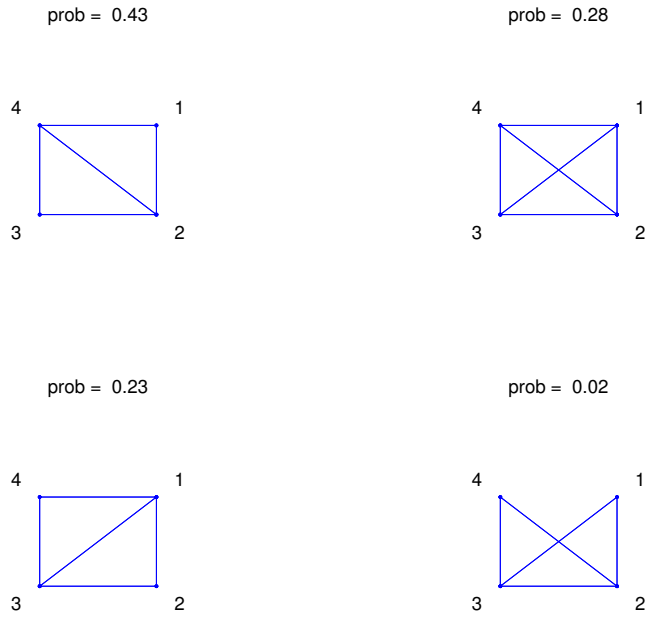

Figure 3.14: The four most probable graph structures and their corresponding posterior probabilities.

4

Controlled sequential Monte Carlo samplers

Contents

Although AIS and SMC samplers discussed in Section 1.2.4 have been employed in a wide range of applications, they can still perform poorly if the forward transition kernels $\{K_t\}_{t=1}^T$ are such that the induced proposal distribution $Q(\mathrm{d}x_{0:T})$ in (1.39) differs significantly from the extended target distribution $P(\mathrm{d}x_{0:T})$ defined in (1.38). To improve the performance of SMC samplers, in Section 4.1.1 we consider 'twisting' these transition kernels $\{K_t\}_{t=1}^T$ to better approximate $P(\mathrm{d}x_{0:T})$. After introducing this class of twisted SMC samplers, in Section 4.1.2 we identify a sequence of functions that induces an optimal sampler in the sense of having distribution $P(\mathrm{d}x_{0:T})$. In Section 4.1.3, we show that this optimal sequence of twisting functions can also be viewed as the solution of a Kullback-Leibler optimal control problem and the optimal value functions of this problem are given by a logarithmic transformation of the optimal twisting functions. Drawing on ideas from the control literature, in Section 4.1.4 we describe general algorithms to approximate these optimal twisting and value functions. Using these methods, we then develop an iterative scheme to build better approximations of $P(\mathrm{d}x_{0:T})$ in Section 4.1.5. Algorithmic settings and connections to related work are discussed in Section 4.2. The resulting algorithm, which we will refer to as the controlled SMC sampler, can be thought of as a type of adaptive importance sampler that is trained using reinforcement learning. We provide some theoretical analysis of these methods in Section 4.3 and conclude with some examples in Section 4.4.

4.1 Optimal importance sampling as optimal control

We begin with some necessary notation and concepts.

4.1.1 Twisted probability measures

Let $\mathcal{M}(\Omega)$ denote the set of Markov transition kernels on a measurable space $(\Omega, \mathcal{F})$. For any measurable function $\varphi : \Omega \times \Omega \to \mathbb{R}$ and $K \in \mathcal{M}(\Omega)$, we will write $K(\varphi)(x) := \int_\Omega \varphi(x, y)K(x, \mathrm{d}y)$. Our methodology is based on the notion of '*twisted*' probability measures on path space which we now define.

Definition 4.1. *Given $R \in \mathcal{P}(\mathcal{X}^{T+1})$ of the form*

$$R(\mathrm{d}x_{0:T}) = \mu_0(\mathrm{d}x_0) \prod_{t=1}^{T} M_t(x_{t-1}, \mathrm{d}x_t) \tag{4.1}$$

for some $\mu_0 \in \mathcal{P}(\mathcal{X})$, $M_t \in \mathcal{M}(\mathcal{X}), t = 1, \ldots, T$ and a sequence of positive 'twisting' functions $\psi = \{\psi_t\}_{t=0}^T \in \Psi(R)$, we define $R^\psi \in \mathcal{P}(\mathcal{X}^{T+1})$ the ψ-twisted version of R as

$$R^\psi(\mathrm{d}x_{0:T}) := \mu_0^\psi(\mathrm{d}x_0) \prod_{t=1}^{T} M_t^\psi(x_{t-1}, \mathrm{d}x_t) \tag{4.2}$$

where

$$\mu_0^\psi(\mathrm{d}x_0) := \frac{\mu_0(\mathrm{d}x_0)\psi_0(x_0)}{\mu_0(\psi_0)}, \quad M_t^\psi(x_{t-1}, \mathrm{d}x_t) := \frac{M_t(x_{t-1}, \mathrm{d}x_t)\psi_t(x_{t-1}, x_t)}{M_t(\psi_t)(x_{t-1})}. \tag{4.3}$$

The set of admissible twisting functions is

$$\Psi(R) := \Big\{ \psi_0 : \mathcal{X} \to \mathbb{R}_+, \psi_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \, t = 1, \ldots, T : \psi_0 \in L^1(R_0), \\ \psi_t \in L^1(R_{t-1,t}) \Big\}, \tag{4.4}$$

where $R_t(\mathrm{d}x_t) := R(\mathrm{d}x_t)$ and $R_{t-1,t}(\mathrm{d}x_{t-1}, \mathrm{d}x_t) := R(\mathrm{d}x_{t-1}, \mathrm{d}x_t)$ denote the one-time and two-time marginal distributions of R respectively.

4.1.2 Twisted sequential Monte Carlo samplers

For ease of presentation, we will henceforth assume that the backward kernels $\{L_t\}_{t=0}^{T-1}$ are selected so that $P(A) > 0$ for any set $A \in \mathcal{B}(\mathcal{X}^{\otimes T+1})$ of positive Lebesgue measure and the forward kernels $\{K_t\}_{t=1}^{T}$ are chosen such that the weight

function (1.41) satisfies $0 < W(x_{0:T}) < \infty$ for any $x_{0:T} \in \mathcal{X}^{T+1}$. For any $\psi \in \Psi(Q)$, since $P \ll Q \ll Q^\psi$ by positivity of ψ, we have

$$W^\psi(x_{0:T}) := Z\frac{\mathrm{d}P}{\mathrm{d}Q^\psi}(x_{0:T}) = Z\frac{\mathrm{d}P}{\mathrm{d}Q}(x_{0:T})\frac{\mathrm{d}Q}{\mathrm{d}Q^\psi}(x_{0:T}) \tag{4.5}$$
$$= w_0^\psi(x_0)\prod_{t=1}^{T} w_t^\psi(x_{t-1}, x_t),$$

with

$$w_0^\psi(x_0) := \frac{\pi_0(\psi_0)}{\psi_0(x_0)}, \quad w_t^\psi(x_{t-1}, x_t) := \frac{w_t(x_{t-1}, x_t)K_t(\psi_t)(x_{t-1})}{\psi_t(x_{t-1}, x_t)}. \tag{4.6}$$

If the weights (4.5)-(4.6) can be computed, then like in Section 1.2.4.2, we can use Q^ψ as an importance distribution to approximate P and obtain the following unbiased estimator of $Z = Q^\psi(W^\psi)$

$$\hat{Z}^\psi = \frac{1}{N}\sum_{n=1}^{N} W^\psi(X_{0:T}^n), \tag{4.7}$$

where $\{X_{0:T}^n\}_{n=1}^N$ are $N \in \mathbb{N}$ iid samples from Q^ψ. To simplify notation, we will write $\mathbb{E}_{Q^\psi}^{t,x}$ to denote expectation with respect to the law of Q^ψ initialized at $X_t = x$.

It is natural to consider an iterative scheme to build better approximations of P: i.e. given a current approximation Q^ψ of P for some $\psi \in \Psi(Q)$, we could *further twist* Q^ψ by functions $\phi = \{\phi_t\}_{t=0}^T \in \Psi(Q^\psi)$ to obtain another approximation $(Q^\psi)^\phi$ of P. Note that $(Q^\psi)^\phi = Q^{\psi\cdot\phi}$ where $\psi \cdot \phi := \{\psi_t \cdot \phi_t\}_{t=0}^T \in \Psi(Q)$ denotes the pointwise product of functions. The choice of ϕ is guided by the following key result.

Proposition 4.2. *For any $\psi \in \Psi(Q)$, we have $P = (Q^\psi)^{\phi^*}$ where the optimal*[1] *sequence of twisting functions $\phi^* = \{\phi_t^*\}_{t=0}^T \in \Psi(Q^\psi)$ with respect to Q^ψ are given by*

$$\phi_0^*(x_0) := w_0^\psi(x_0)\mathbb{E}_{Q^\psi}^{0,x_0}\left[\prod_{k=1}^{T} w_k^\psi(X_{k-1}, X_k)\right], \tag{4.8}$$
$$\phi_t^*(x_{t-1}, x_t) := w_t^\psi(x_{t-1}, x_t)\mathbb{E}_{Q^\psi}^{t,x_t}\left[\prod_{k=t+1}^{T} w_k^\psi(X_{k-1}, X_k)\right], \quad t = 1, \ldots, T-1,$$
$$\phi_T^*(x_{T-1}, x_T) := w_T^\psi(x_{T-1}, x_T).$$

[1] Equation (4.8) should be understood as a definition of optimality as there is more than one sequence of twisting functions ϕ satisfying $P = (Q^\psi)^\phi$ (for example scaling ϕ^* by a constant).

Proof. From (4.5), we have

$$P(\mathrm{d}x_{0:T}) = Z^{-1} Q^{\psi}(\mathrm{d}x_{0:T}) W^{\psi}(x_{0:T}), \tag{4.9}$$

where

$$Q^{\psi}(\mathrm{d}x_{0:T}) = \pi_0^{\psi}(\mathrm{d}x_0) \prod_{t=1}^{T} K_t^{\psi}(x_{t-1}, \mathrm{d}x_t). \tag{4.10}$$

By Fubini's theorem, ϕ^* is well-defined and is an element of $\Psi(Q^{\psi})$ as the integrals in (4.8) exist since $Z = Q^{\psi}(W^{\psi}) < \infty$. From (4.5) and (4.10), we have for $t = 0, \ldots, T$ that

$$\begin{aligned} P(\mathrm{d}x_{0:t}) = Z^{-1} \pi_0^{\psi}(\mathrm{d}x_0) w_0^{\psi}(x_0) \prod_{k=1}^{t} K_k^{\psi}(x_{k-1}, \mathrm{d}x_k) w_k^{\psi}(x_{k-1}, x_k) \\ \times \int_{\mathcal{X}^{T-t}} \prod_{k=t+1}^{T} K_k^{\psi}(x_{k-1}, \mathrm{d}x_k) w_k^{\psi}(x_{k-1}, x_k) \end{aligned} \tag{4.11}$$

with the convention $\prod_i^j := 1$ for $i > j$. Setting $t = 0$ gives the initial distribution $(\pi_0^{\psi})^{\phi^*}$ since $Z = \pi_0^{\psi}(\phi_0^*)$ and for the transition kernels $(K_t^{\psi})^{\phi^*}$ we note that for $t = 1, \ldots, T$

$$\frac{P(\mathrm{d}x_{0:t})}{P(\mathrm{d}x_{0:t-1})} = \frac{K_t^{\psi}(x_{t-1}, \mathrm{d}x_t) w_t^{\psi}(x_{t-1}, x_t) \int_{\mathcal{X}^{T-t}} \prod_{k=t+1}^{T} K_k^{\psi}(x_{k-1}, \mathrm{d}x_k) w_k^{\psi}(x_{k-1}, x_k)}{\int_{\mathcal{X}^{T-t+1}} \prod_{k=t}^{T} K_k^{\psi}(x_{k-1}, \mathrm{d}x_k) w_k^{\psi}(x_{k-1}, x_k)}. \tag{4.12}$$

□

For any approximation Q^{ψ} of P with $\psi \in \Psi(Q)$, note that an application of Proposition 4.2 gives us the optimal sequence of twisting functions $\psi^* := \psi \cdot \phi^*$ with respect to Q. In the following, it will be useful to consider a logarithmic transformation of ϕ^*, i.e. define $V^* = \{V_t^*\}_{t=0}^T$ as $V_0^* := -\log \phi_0^*(x_0)$ and $V_t^*(x_{t-1}, x_t) := -\log \phi_t^*(x_{t-1}, x_t)$ for $t = 1, \ldots, T$. This is because V^* can be thought of as the optimal sequence of *value functions* with respect to Q^{ψ} of an associated optimal control problem to be introduced in Section 4.1.3. We now define some operators, commonly known as *Bellman operators* in the control literature, which will play an important role in our analysis and greatly simplify notation.

Definition 4.3. *Given $\psi \in \Psi(Q)$ and a measurable function $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we define the operators $\left\{\mathsf{B}_t^\psi\right\}_{t=0}^{T-1}$ by*

$$\begin{aligned}(\mathsf{B}_0^\psi \varphi)(x) &:= w_0^\psi(x) K_1^\psi(\varphi)(x), \\ (\mathsf{B}_t^\psi \varphi)(x,y) &:= w_t^\psi(x,y) K_{t+1}^\psi(\varphi)(y), \quad t = 1, \ldots, T-1,\end{aligned} \tag{4.13}$$

and $\left\{\mathsf{T}_t^\psi\right\}_{t=0}^{T-1}$ by

$$\begin{aligned}(\mathsf{T}_0^\psi \varphi)(x) &:= -\log w_0^\psi(x) - \log K_1^\psi(e^{-\varphi})(x), \\ (\mathsf{T}_t^\psi \varphi)(x,y) &:= -\log w_t^\psi(x,y) - \log K_{t+1}^\psi(e^{-\varphi})(y), \quad t = 1, \ldots, T-1.\end{aligned} \tag{4.14}$$

From Proposition 4.2, the following result is straightforward.

Proposition 4.4. *For any $\psi \in \Psi(Q)$, the optimal sequence of twisting functions $\phi^* = \{\phi_t^*\}_{t=0}^T$ with respect to Q^ψ satisfies the following backward recursion*

$$\begin{aligned}\phi_T^* &= w_T^\psi, \\ \phi_t^* &= \mathsf{B}_t^\psi \phi_{t+1}^*, \quad t = T-1, \ldots, 0,\end{aligned} \tag{4.15}$$

and $\pi_0^\psi(\phi_0^*) = Z$. Equivalently, the optimal sequence of value functions $V^* = \{V_t^*\}_{t=0}^T$ with respect to Q^ψ satisfies

$$\begin{aligned}V_T^* &= -\log w_T^\psi, \\ V_t^* &= \mathsf{T}_t^\psi V_{t+1}^*, \quad t = T-1, \ldots, 0.\end{aligned} \tag{4.16}$$

Consider recursion (4.15) applied to the AIS setting with $\psi_t = 1, t = 0, \ldots, T$, in which case, the incremental weights (1.55) have a simpler form $w_t(x_{t-1}, x_t) = w_t(x_{t-1})$. Substituting into (4.8) and noting that the factor $w_t(x_{t-1})$ in ϕ_t^* can be ignored owing to the form of $K_t^{\phi^*}$, it follows that $P = Q^{\phi^*}$ also holds if we have $\phi_T^*(x_T) = 1$ and the recursion

$$\phi_t^*(x_t) = w_{t+1}(x_t) K_{t+1}(\phi_{t+1}^*)(x_t), \quad t = 0, \ldots, T-1. \tag{4.17}$$

Equation (4.17) is structurally similar to the *backward information filter* recursion that arises in smoothing problems for state space models (Briers et al. 2010, Pieralberto et al. 2016). In particular, we note that it admits a simpler structure than (4.15) as ϕ_t^* is a function of only one argument x_t instead of both x_{t-1} and x_t.

4.1.3 Optimal importance sampling as Kullback-Leibler control

We now show that the optimal sequence of twisting functions in Proposition 4.2 can also be viewed as the optimal policy of a *Kullback-Leibler optimal control problem* (Kappen et al. 2012). While the re-formulation might appear superfluous at first glance, making this connection allows us to exploit numerical methods developed in the *approximate dynamic programming* literature (Bertsekas and Tsitsiklis 1996) in Section 4.1.4 and draw ideas from existing results (Tsitsiklis and Van Roy 2001) to analyze these algorithms in Section 4.3.

Suppose $\psi \in \Psi(Q)$ and consider the following optimal control problem

$$\inf_{\phi \in \Phi(Q^\psi)} \mathrm{KL}\left((Q^\psi)^\phi | P\right) = \inf_{\phi \in \Phi(Q^\psi)} \mathbb{E}_{(Q^\psi)^\phi}\left[C(X_{0:T})\right] \tag{4.18}$$

where the cost functional $C : \mathcal{X}^{T+1} \to \mathbb{R}$ can be written as

$$\begin{aligned} C(x_{0:T}) &:= \log \frac{\mathrm{d}(Q^\psi)^\phi}{\mathrm{d}Q^\psi}(x_{0:T}) - \log \frac{\mathrm{d}P}{\mathrm{d}Q^\psi}(x_{0:T}) \\ &= \log \frac{\mathrm{d}(\pi_0^\psi)^\phi}{\mathrm{d}\pi_0^\psi}(x_0) + \sum_{t=1}^{T} \log \frac{\mathrm{d}(K_t^\psi)^\phi(x_{t-1},\cdot)}{\mathrm{d}K_t^\psi(x_{t-1},\cdot)}(x_t) \\ &\quad - \log w_0^\psi(x_0) - \sum_{t=1}^{T} \log w_t^\psi(x_{t-1}, x_t) + \log Z \end{aligned} \tag{4.19}$$

and

$$\begin{aligned} \Phi(Q^\psi) := \Bigg\{ \phi \in \Psi(Q^\psi) : \mathrm{KL}\left((\pi_0^\psi)^\phi | \pi_0^\psi\right) < \infty, \mathrm{KL}\left((K_t^\psi)^\phi | K_t^\psi\right)(x_{t-1}) < \infty \\ Q^\psi(\mathrm{d}x_{t-1}) - \text{a.e. for } t = 1, \ldots, T \Bigg\} \subseteq \Psi(Q^\psi) \end{aligned} \tag{4.20}$$

denotes the set of admissible sequences of twisting functions or policies for the control problem. It follows that

$$\begin{aligned} \mathbb{E}_{(Q^\psi)^\phi}\left[C(X_{0:T})\right] &= \mathrm{KL}\left((\pi_0^\psi)^\phi | \pi_0^\psi\right) + \sum_{t=1}^{T} \mathbb{E}_{(Q^\psi)^\phi}\left[\mathrm{KL}\left((K_t^\psi)^\phi | K_t^\psi\right)(X_{t-1})\right] \\ &\quad - \mathbb{E}_{(\pi_0^\psi)^\phi}\left[\log w_0^\psi(X_0)\right] - \sum_{t=1}^{T} \mathbb{E}_{(Q^\psi)^\phi}\left[\log w_t^\psi(X_{t-1}, X_t)\right] + \log Z. \end{aligned} \tag{4.21}$$

We know from Proposition 4.2 that ϕ^* defined in (4.8) solves the optimal control problem (4.18) since we have $\mathrm{KL}(\mu|\nu) \geq 0$ for any probability measures μ, ν and

$\mathrm{KL}(\mu|\nu) = 0$ if and only if $\mu = \nu$. We shall henceforth re-define the cost functional (4.19) to remove the intractable term $\log Z$.

Given a sequence of twisting functions $\phi = \{\phi_t\}_{t=0}^T$, also known as a policy in optimal control terminology, the corresponding value functions $\left\{V_t^\phi\right\}_{t=0}^T$ of this control problem are given by the expected cost-to-go from a fixed time and state (Bertsekas and Tsitsiklis 1996; Section 2.1)

$$\begin{aligned} V_0^\phi(x_0) &:= \mathrm{KL}\left((K_1^\psi)^\phi | K_1^\psi\right)(x_0) + \sum_{k=1}^{T-1} \mathbb{E}_{(Q^\psi)^\phi}^{0,x_0}\left[\mathrm{KL}\left((K_{k+1}^\psi)^\phi | K_{k+1}^\psi\right)(X_k)\right] \\ &\quad - \log w_0^\psi(x_0) - \sum_{k=1}^{T} \mathbb{E}_{(Q^\psi)^\phi}^{0,x_0}\left[\log w_k^\psi(X_{k-1}, X_k)\right], \\ V_t^\phi(x_{t-1}, x_t) &:= \mathrm{KL}\left((K_{t+1}^\psi)^\phi | K_{t+1}^\psi\right)(x_t) + \sum_{k=t+1}^{T-1} \mathbb{E}_{(Q^\psi)^\phi}^{t,x_t}\left[\mathrm{KL}\left((K_{k+1}^\psi)^\phi | K_{k+1}^\psi\right)(X_k)\right] \\ &\quad - \log w_t^\psi(x_{t-1}, x_t) - \sum_{k=t+1}^{T} \mathbb{E}_{(Q^\psi)^\phi}^{t,x_t}\left[\log w_k^\psi(X_{k-1}, X_k)\right], \quad t = 1, \ldots, T-1, \\ V_T^\phi(x_{T-1}, x_T) &:= -\log w_T^\psi(x_{T-1}, x_T). \end{aligned} \tag{4.22}$$

In this notation, the total cost of using ϕ is given by $v(\phi) := (\pi_0^\psi)^\phi(V_0^\phi) + \mathrm{KL}\left((\pi_0^\psi)^\phi | \pi_0^\psi\right) = \mathrm{KL}\left((Q^\psi)^\phi | P\right) - \log Z$. We now define the optimal sequence of value functions $V^* = \{V_t^*\}_{t=0}^T$ with respect to Q^ψ by taking the infimum over the set $\Phi(Q^\psi)$:

$$\begin{aligned} v^* &:= \inf_\phi v(\phi), \\ V_0^*(x_0) &:= \inf_{\{\phi_k, 1 \le k \le T\}} V_0^\phi(x_0), \\ V_t^*(x_{t-1}, x_t) &:= \inf_{\{\phi_k, t+1 \le k \le T\}} V_t^\phi(x_{t-1}, x_t), \quad t = 1, \ldots, T-1, \\ V_T^*(x_{T-1}, x_T) &:= -\log w_T^\psi(x_{T-1}, x_T), \end{aligned} \tag{4.23}$$

and denote the minimizer (if it exists) as $\phi^* = \{\phi_t^*\}_{t=0}^T$. We stress the dependence of both V^* and ϕ^* on the given sequence of twisting functions $\psi \in \Psi(Q)$ as this is not done notationally. These minimization problems can be solved using a backward *dynamic programming* approach. From Definition (4.22) and (4.23), we

have the dynamic programming recursion

$$
\begin{aligned}
V_T^*(x_{T-1}, x_T) &= -\log w_T^\psi(x_{T-1}, x_T), \qquad (4.24)\\
V_t^*(x_{t-1}, x_t) &= -\log w_t^\psi(x_{t-1}, x_t) + \inf_{\phi_{t+1}} \Big\{ (K_{t+1}^\psi)^\phi (V_{t+1}^*)(x_t) \\
&\qquad + \mathrm{KL}\left((K_{t+1}^\psi)^\phi | K_{t+1}^\psi \right)(x_t) \Big\}, \quad t = T-1, \ldots, 1,\\
V_0^*(x_0) &= -\log w_0^\psi(x_0) + \inf_{\phi_1} \left\{ (K_1^\psi)^\phi (V_1^*)(x_0) + \mathrm{KL}\left((K_1^\psi)^\phi | K_1^\psi \right)(x_0) \right\},\\
v^* &= \inf_{\phi_0} \left\{ (\pi_0^\psi)^\phi (V_0^*) + \mathrm{KL}\left((\pi_0^\psi)^\phi | \pi_0^\psi \right) \right\}.
\end{aligned}
$$

The above is commonly referred to as the discrete time *Hamilton-Jacobi-Bellman* equation. We now state a well-known lemma (Dupuis and Ellis 2011, Dai Pra et al. 1996, Bierkens and Kappen 2014) which allows us to solve the weighted Kullback-Leibler minimization problems in (4.24).

Lemma 4.5. *Let μ be a probability measure on a measurable space $(\Omega, \mathcal{F})$ and $\varphi : \Omega \to \mathbb{R}$ be a measurable function such that $\mu(|\varphi| e^{-\varphi}) < \infty$. The following equality holds*

$$
-\log \mu(e^{-\varphi}) = \inf_{\nu \in \mathcal{P}_{\mu,\varphi}(\Omega)} \{ \nu(\varphi) + \mathrm{KL}(\nu|\mu) \} \qquad (4.25)
$$

where $\mathcal{P}_{\mu,\varphi}(\Omega) := \{\nu \in \mathcal{P}(\Omega) : \mathrm{KL}(\nu|\mu) < \infty$ and $\nu(|\varphi|) < \infty\}$. Moreover, the infimum in (4.25) is attained at $\nu^* \ll \mu$ with $\mathrm{d}\nu^*/\mathrm{d}\mu = e^{-\varphi}/\mu(e^{-\varphi})$.

Proof. We first note that $\mu(|\varphi| e^{-\varphi}) < \infty$ implies that the left hand side of (4.25) is finite. Since $0 < \mathrm{d}\nu^*/\mathrm{d}\mu < \infty$, for each $\nu \in \mathcal{P}_{\mu,\varphi}(\Omega)$, we have $\nu \ll \mu \ll \nu^*$ and hence

$$
\begin{aligned}
\nu(\varphi) + \mathrm{KL}(\nu|\mu) &= \nu(\varphi) + \int_\Omega \log\left(\frac{\mathrm{d}\nu^*}{\mathrm{d}\mu} \right) \mathrm{d}\nu + \int_\Omega \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu^*} \right) \mathrm{d}\nu \qquad (4.26)\\
&= -\log \mu(e^{-\varphi}) + \mathrm{KL}(\nu|\nu^*).
\end{aligned}
$$

The proof is complete by noting that $\mathrm{KL}(\nu|\nu^*) \geq 0$ and $\mathrm{KL}(\nu|\nu^*) = 0$ if and only if $\nu = \nu^*$. □

We note that although Equation (4.25) holds without the assumption $\mu(|\varphi| e^{-\varphi}) < \infty$, there does not exist a unique minimizer in this case (Bierkens and Kappen

2014; Proposition 2.6, Corollary 2.7). The integrability assumption, which holds if and only if $\mathrm{KL}(\nu^*|\mu) < \infty$, is sufficient to guarantee existence of a minimizer. The following is a straightforward consequence of Lemma 4.5.

Corollary 4.6. *If $\psi \in \Psi(Q)$ is such that $\mathrm{KL}(P|Q^\psi) < \infty$, then the infimums in (4.24) are attained at $\phi_t^* = e^{-V_t^*}, t = 0, \ldots, T$ and at the minimum, the dynamic programming recursion (4.24) recovers the backward recursion (4.16).*

From the conclusions of Corollary 4.6, we can infer the backward recursion (4.15) and the form of ϕ^* in (4.8). We note that the optimal cost is $v^* = -\log Z$ as we have adjusted the cost functional (4.19) and that the finite Kullback-Leibler assumption ensures that the minimizer ϕ^* lies in $\Phi(Q^\psi) \subseteq \Psi(Q^\psi)$. Although it is unsurprising for such a condition to be required when we formulate ϕ^* as the solution of a Kullback-Leibler control problem, it should be clear from Section 4.1.2 that this is not necessary for our purposes.

4.1.4 Approximation dynamic programming

For most problems of practical interest, the recursions (4.15)-(4.16) are intractable so we need to rely on approximations.

Definition 4.7. *Let μ be a probability measure on a measurable space $(\Omega, \mathcal{F})$, $\xi : \Omega \to \mathbb{R}$ be a measurable function in $L^2(\mu)$ and F be a closed linear subspace of $L^2(\mu)$. We define the (F, μ)-projection operator as*

$$\mathsf{P}^\mu \xi := \arg\min_{\varphi \in \mathsf{F}} \|\varphi - \xi\|^2_{L^2(\mu)}. \tag{4.27}$$

The projection theorem gives existence of a unique $\mathsf{P}^\mu \xi \in \mathsf{F}$ and that $\mathsf{P}^\mu \xi - \xi$ is orthogonal to F. Hence $\mathsf{P}^\mu \xi$ is typically referred to as the *orthogonal projection* of ξ unto F.

As before, $\left\{Q^\psi_{t-1,t}\right\}_{t=1}^T$ will denote the marginal distributions of Q^ψ and for notational convenience, we will occasionally identify $Q^\psi_{-1,0} := Q^\psi_0$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For each time $t = 0, \ldots, T$, we consider the probability space $\left(\mathcal{X}^2, \mathcal{B}(\mathcal{X})^{\otimes 2}, Q^\psi_{t-1,t}\right)$ and given a pre-specified closed and linear function class $\mathsf{F}_t \subset L^2(Q^\psi_{t-1,t})$, the

$\left(\mathsf{F}_t, Q^{\psi}_{t-1,t}\right)$-projection operator denoted by P^{ψ}_t will be of interest. It is natural to approximate the recursion (4.15) by

$$\begin{aligned} \tilde{\phi}_T &= \mathsf{P}^{\psi}_T w^{\psi}_T, \\ \tilde{\phi}_t &= \mathsf{P}^{\psi}_t \mathsf{B}^{\psi}_t \tilde{\phi}_{t+1}, \quad t = T-1, \ldots, 0. \end{aligned} \tag{4.28}$$

The quality of the sequence of twisting functions $\tilde{\phi} = \left\{\tilde{\phi}_t\right\}_{t=0}^{T}$ as an approximation of ϕ^* will depend on how 'rich' the chosen function classes $\{\mathsf{F}_t\}_{t=0}^{T}$ are. Of course, (4.28) is only an idealized algorithm as projections are typically intractable. In practice, we would have to approximate such a procedure using Monte Carlo sampling.

Definition 4.8. *Following the notation in Definition 4.7, for each $N \in \mathbb{N}$, we define the approximate (F, μ)-projection operator $\mathsf{P}^{\mu,N}$ as the (F, μ^N)-projection operator, where $\mu^N = N^{-1}\sum_{n=1}^{N} \delta_{X^n}$ denotes the random probability measure based on N iid samples $\{X^n\}_{n=1}^{N}$ from μ. If Ω is a topological space with σ-algebra $\mathcal{F} = \mathcal{B}(\Omega)$, we shall additionally assume the function class F is such that $\left(x, \{X^n\}_{n=1}^{N}\right) \mapsto (\mathsf{P}^{\mu,N}\xi)(x)$ is Borel measurable for all $\xi \in L^2(\mu)$.*

Given a fixed number of samples $N \in \mathbb{N}$, we denote $\mathsf{P}^{\psi,N}_t$ as the approximate $\left(\mathsf{F}_t, Q^{\psi}_{t-1,t}\right)$-projection operator for each $t = 0, \ldots, T$ and define an approximate dynamic programming (ADP) algorithm for learning the optimal sequence of twisting functions ϕ^* as

$$\begin{aligned} \hat{\phi}_T &= \mathsf{P}^{\psi,N}_T w^{\psi}_T, \\ \hat{\phi}_t &= \mathsf{P}^{\psi,N}_t \mathsf{B}^{\psi}_t \hat{\phi}_{t+1}, \quad t = T-1, \ldots, 0. \end{aligned} \tag{4.29}$$

In Section 4.3.4, for a common choice of function class, we will show that $\hat{\phi}$ converges (in a suitable sense) to $\tilde{\phi}$ as $N \to \infty$. Similarly, by approximating (4.16), we could also have an ADP algorithm to learn the optimal sequence of value functions V^*:

$$\begin{aligned} \hat{V}_T &= \mathsf{P}^{\psi,N}_T(-\log w^{\psi}_T), \\ \hat{V}_t &= \mathsf{P}^{\psi,N}_t \mathsf{T}^{\psi}_t \hat{V}_{t+1}, \quad t = T-1, \ldots, 0, \end{aligned} \tag{4.30}$$

and define the corresponding approximation $\hat{\phi} = \left\{\hat{\phi}_t\right\}_{t=0}^{T}$ of ϕ^* as $\hat{\phi}_t := e^{-\hat{V}_t}, t = 0, \ldots, T$. A more detailed outline of (4.29) and (4.30) are given in Algorithm 3 and 4 respectively. While an analysis of both algorithms will be considered in Section 4.3, we only performed value function learning in our implementation as it is numerically more stable to compute in logarithmic scale. We defer a detailed discussion on the choice of function classes $\{\mathsf{F}_t\}_{t=0}^{T}$ to Section 4.2.2 and shall assume for now this is such that (i) $\hat{\phi} \in \Psi(Q^\psi)$; (ii) sampling from $Q^{\psi\cdot\hat{\phi}}$ is possible if sampling from Q^ψ is; (iii) the weight $W^{\psi\cdot\hat{\phi}}$ can be computed if W^ψ is computable.

Algorithm 3 Approximate dynamic programming for learning optimal sequence of twisting functions

Input: twisting functions $\psi \in \Psi(Q)$, $N \in \mathbb{N}$ iid samples $\{X_{0:T}^n\}_{n=1}^N$ from Q^ψ.

1. Initialization: set $K_{T+1}^\psi(\hat{\phi}_{T+1})(X_T^n) = 1$ for $n = 1, \ldots, N$.
2. For $t = T, T-1, \ldots, 1$,
   (a) set $\bar{\phi}_t(X_{t-1}^n, X_t^n) = w_t^\psi(X_{t-1}^n, X_t^n) K_{t+1}^\psi(\hat{\phi}_{t+1})(X_t^n)$;
   (b) set $\hat{\phi}_t = \arg\min_{\xi \in \mathsf{F}_t} \sum_{n=1}^N \left(\xi(X_{t-1}^n, X_t^n) - \bar{\phi}_t(X_{t-1}^n, X_t^n)\right)^2$.
3. For $t = 0$,
   (a) set $\bar{\phi}_0(X_0^n) = w_0^\psi(X_0^n) K_1^\psi(\hat{\phi}_1)(X_0^n)$;
   (b) set $\hat{\phi}_0 = \arg\min_{\xi \in \mathsf{F}_0} \sum_{n=1}^N \left(\xi(X_0^n) - \bar{\phi}_0(X_0^n)\right)^2$.

Output: twisting functions $\hat{\phi} = \left\{\hat{\phi}_t\right\}_{t=0}^{T}$.

4.1.5 Controlled sequential Monte Carlo samplers

If the recursions in Proposition 4.4 can be performed exactly, then having an iterative procedure to approximate P is unnecessary. We would simply initialize with $Q^\psi = Q$, i.e. $\psi_t = 1$ for $t = 0, \ldots, T$, and obtain the optimal sequence of twisting functions ψ^* with respect to Q satisfying $Q^{\psi^*} = P$. As alluded earlier, when approximations (4.29)-(4.30) are employed, it is worth considering an iterative scheme to build better approximations of P which we now describe.

Suppose that at iteration $i \in \mathbb{N}_0$, we have a sequence of twisting functions $\hat{\psi}(i) \in \Psi(Q)$ (with initialization $\hat{\psi}(0) = 1$ to mean the constant one function). We

Algorithm 4 Approximate dynamic programming for learning optimal sequence of value functions

Input: twisting functions $\psi \in \Psi(Q)$, $N \in \mathbb{N}$ iid samples $\{X_{0:T}^n\}_{n=1}^N$ from Q^ψ.

1. Initialization: set $K_{T+1}^\psi(e^{-\hat{V}_{T+1}})(X_T^n) = 1$ for $n = 1, \ldots, N$.
2. For $t = T, T-1, \ldots, 1$,
   (a) set $\bar{V}_t(X_{t-1}^n, X_t^n) = -\log w_t^\psi(X_{t-1}^n, X_t^n) - \log\left(K_{t+1}^\psi(e^{-\hat{V}_{t+1}})(X_t^n)\right)$;
   (b) set $\hat{V}_t = \arg\min_{\xi \in \mathsf{F}_t} \sum_{n=1}^N \left(\xi(X_{t-1}^n, X_t^n) - \bar{V}_t(X_{t-1}^n, X_t^n)\right)^2$.
3. For $t = 0$,
   (a) set $\bar{V}_0(X_0^n) = -\log w_0^\psi(X_0^n) - \log\left(K_1^\psi(e^{-\hat{V}_1})(X_0^n)\right)$;
   (b) set $\hat{V}_0 = \arg\min_{\xi \in \mathsf{F}_0} \sum_{n=1}^N \left(\xi(X_0^n) - \bar{V}_0(X_0^n)\right)^2$.

Output: value functions $\hat{V} = \left\{\hat{V}_t\right\}_{t=0}^T$ and twisting functions $\hat{\phi} = \left\{\hat{\phi}_t = e^{-\hat{V}_t}\right\}_{t=0}^T$.

then run a twisted SMC sampler with the current approximation $Q^{\hat{\psi}(i)}$ of P and use its output to perform the ADP procedure described in either Algorithm 3 or 4. This returns a sequence of twisting functions $\hat{\phi}(i+1) \in \Psi(Q^{\hat{\psi}(i)})$, which we then use to obtain our new approximation $Q^{\hat{\psi}(i+1)}$ of P with the update $\hat{\psi}(i+1) := \hat{\psi}(i) \cdot \hat{\phi}(i+1) \in \Psi(Q)$. The resulting algorithm, summarized in Algorithm 5, will be referred to as the *controlled* SMC sampler. To maintain a coherent terminology, we will call a twisted SMC sampler with $\hat{\psi}(0) = 1$ and ψ^* as the *uncontrolled* and *optimally controlled* SMC sampler respectively.

We repeat this procedure for a number of iterations $I \in \mathbb{N}$, when successive twisting yield no improvement in performance. Possible measures of performance here include ESS defined in Section 1.2.2.4 or the variance of particle weights. In Section 4.3.5, under appropriate regularity assumptions, we will show that this iterative scheme generates a geometrically ergodic Markov chain on the closure of $\Psi(Q)$ that converges to its unique invariant distribution. From our numerical implementations, we observe that convergence happens very rapidly so only a small number of iterations is necessary.

Contrary to common practice (Section 1.2.4.3), the absence of the resampling step

in Algorithm 5 is intentional. The rationale here is that for time $t < T$, the artificially constructed target distribution $P_t(\mathrm{d}x_{0:t}) = \pi_t(\mathrm{d}x_t) \prod_{k=1}^{t} L_{k-1}(x_k, \mathrm{d}x_{k-1})$ might differ significantly from $Q^{\psi^*}(\mathrm{d}x_{0:t}) = P(\mathrm{d}x_{0:t}) = \pi_0^{\psi^*}(\mathrm{d}x_0) \prod_{k=1}^{t} K_k^{\psi^*}(x_{k-1}, \mathrm{d}x_k)$ – see Figure 4.2 for a concrete illustration. If the latter is the case, under an adaptive resampling strategy, based for example on the ESS criterion, frequent resampling will result in higher variance of Monte Carlo estimates. Lastly, to avoid repeating computations, we note that the incremental weights required in step 2(b) can be pre-computed when evaluating the weights in step 2(a)ii.

Algorithm 5 Controlled sequential Monte Carlo sampler

Input: time steps T, sequence of distributions $\{\pi_t\}_{t=0}^{T}$, particles N, iterations I.

1. Initialization: set twisting functions $\hat{\psi}(0) = 1$.
2. For $i = 0, \ldots, I-1$,
   (a) run a twisted SMC sampler with twisting functions $\hat{\psi}(i)$:
      i. sample N paths independently, i.e. $X_{0:T}^n(i) \sim Q^{\hat{\psi}(i)}$ for $n = 1, \ldots, N$;
      ii. compute weights $\left\{W^{\hat{\psi}(i)}(X_{0:T}^n(i))\right\}_{n=1}^{N}$ using (4.5)-(4.6);
      iii. estimate normalizing constant $\hat{Z}^{\hat{\psi}(i)} = \frac{1}{N}\sum_{n=1}^{N} W^{\hat{\psi}(i)}(X_{0:T}^n(i))$;
   (b) perform approximate dynamic programming to obtain updated twisting functions $\hat{\psi}(i+1)$:
      i. compute twisting functions $\hat{\phi}(i+1)$ with respect to $Q^{\hat{\psi}(i)}$ using either Algorithm 3 or 4 with $\hat{\psi}(i)$ and $\{X_{0:T}^n(i)\}_{n=1}^{N}$ as input;
      ii. update twisting functions by setting $\hat{\psi}(i+1) = \hat{\psi}(i) \cdot \hat{\phi}(i+1)$.
3. For $i = I$,
   (a) run a twisted SMC sampler with twisting functions $\hat{\psi}(I)$.

Output: trajectories $\{X_{0:T}^n(I)\}_{n=1}^{N}$, weights $\left\{W^{\hat{\psi}(I)}(X_{0:T}^n(I))\right\}_{n=1}^{N}$ and normalizing constant estimate $\hat{Z}^{\hat{\psi}(I)}$.

4.2 Implementation

We now discuss algorithmic settings behind an implementation of the controlled SMC sampler.

4.2.1 Uncontrolled sequential Monte Carlo sampler settings

To implement SMC samplers, we need to specify the transition kernels $\{K_t, L_{t-1}\}_{t=1}^T$. For the forward kernel, we select

$$K_t(x_{t-1}, \mathrm{d}x_t) := \mathcal{N}\left(x_t; x_{t-1} + \frac{\Delta_t}{2}\nabla \log \pi_t(x_{t-1}), \Delta_t I_d\right) \mathrm{d}x_t, \tag{4.31}$$

where $\Delta_t > 0$ denotes the step size at time $t = 1, \ldots, T$. The rationale for using this transition kernel is that it corresponds to an Euler-Maruyama discretization of the Langevin diffusion (1.33) with invariant distribution π_t. Under appropriate regularity conditions, for sufficiently small Δ_t, K_t admits an invariant distribution that is close to π_t (Mattingly et al. 2002). Moreover, as (1.33) is also reversible with respect to π_t, this suggests that K_t will also be *approximately* reversible with respect to π_t for small Δ_t. This prompts the following choice of backward kernel

$$L_{t-1}(x_t, \mathrm{d}x_{t-1}) := K_t(x_t, \mathrm{d}x_{t-1}) = \mathcal{N}\left(x_{t-1}; x_t + \frac{\Delta_t}{2}\nabla \log \pi_t(x_t), \Delta_t I_d\right) \mathrm{d}x_{t-1}, \tag{4.32}$$

in which case we expect the incremental weights (1.42) to be close to that of AIS (1.55) when the step size is small.

Contrary to AIS, we stress that K_t is not π_t-invariant and L_{t-1} is not the time reversed backward kernel (1.54) associated to K_t. Although one could augment an application of K_t with a MH accept-reject step to enforce reversibility, we do not adopt this approach here as it is difficult to sample from the resulting twisted kernel K_t^ψ and compute $x \mapsto K_t(\psi_t)(x)$ pointwise. Under the above choice of kernels, it is straightforward to show that the incremental weights (1.42) have the form

$$\begin{aligned} -\log w_t(x_{t-1}, x_t) = \log \gamma_{t-1}(x_{t-1}) &- \frac{1}{2}\nabla \log \pi_t(x_{t-1})^T x_{t-1} - \frac{\Delta_t}{8}\left|\nabla \log \pi_t(x_{t-1})\right|^2 \\ &- \log \gamma_t(x_t) + \frac{1}{2}\nabla \log \pi_t(x_t)^T x_t + \frac{\Delta_t}{8}\left|\nabla \log \pi_t(x_t)\right|^2 \\ &+ \frac{1}{2}\nabla \log \pi_t(x_{t-1})^T x_t - \frac{1}{2}\nabla \log \pi_t(x_t)^T x_{t-1}, \end{aligned} \tag{4.33}$$

for $t = 1, \ldots, T$. This decomposition will provide some guidelines on how to parameterize the approximate value functions in the next section. For the sake of brevity, we have limited our discussion here to discretized Langevin dynamics;

extension to the case of generalized Langevin dynamics (Leimkuhler and Matthews 2015) and other non-reversible dynamics are also possible.

4.2.2 Approximate dynamic programming settings

In our setup, in addition to obtaining an approximation $\hat{\psi} \in \Psi(Q)$ of the optimal sequence of twisting functions with respect to Q, it is also necessary to be able to sample from the resulting approximation $Q^{\hat{\psi}}$ of P and perform importance sampling between the two distributions. Hence the choice of function classes $\{\mathsf{F}_t\}_{t=0}^T$ used in the ADP procedure should be such that sampling from the initial distribution $\pi_0^{\hat{\psi}}$, transition kernels $\left\{K_t^{\hat{\psi}}\right\}_{t=1}^T$ and computing the importance weights (4.5)-(4.6) are tractable.

To deal with these constraints, we shall restrict our attention to the following function classes when employing Algorithm 4:

$$\mathsf{F}_0 := \Big\{\varphi : \mathcal{X} \to \mathbb{R} : \varphi(x_0) = x_0^T A_0 x_0 + x_0^T b_0 + c_0, \text{ for some symmetric } A_0 \in \mathbb{R}^{d\times d}, \\ b_0 \in \mathbb{R}^d, c_0 \in \mathbb{R}\Big\}, \tag{4.34}$$

$$\mathsf{F}_t := \Big\{\varphi : \mathcal{X}\times\mathcal{X} \to \mathbb{R} : \varphi(x_{t-1}, x_t) = x_t^T A_t(x_{t-1}) x_t + x_t^T b_t(x_{t-1}) + c_t + d_t(x_{t-1}), \\ \text{for some symmetric } A_t : \mathcal{X} \to \mathbb{R}^{d\times d}, b_t : \mathcal{X} \to \mathbb{R}^d, c_t \in \mathbb{R}, d_t : \mathcal{X} \to \mathbb{R}\Big\}, \tag{4.35}$$

for $t = 1, \ldots, T$. With the choice (4.34)-(4.35) in Algorithm 4, the sequence of twisting functions $\hat{\psi}(i) = \left\{\hat{\psi}_t^{(i)}\right\}_{t=0}^T \in \Psi(Q)$ at iteration $i \in \mathbb{N}$ has the form

$$\hat{\psi}_0^{(i)}(x_0) = \exp\left(-x_0^T A_0^{(i)} x_0 - x_0^T b_0^{(i)} - c_0^{(i)}\right), \tag{4.36}$$

$$\hat{\psi}_t^{(i)}(x_{t-1}, x_t) = \exp\left(-x_t^T A_t^{(i)}(x_{t-1}) x_t - x_t^T b_t^{(i)}(x_{t-1}) - c_t^{(i)} - d_t^{(i)}(x_{t-1})\right), \tag{4.37}$$

for $t = 1, \ldots, T$, with $A_t^{(i)} := \sum_{k=1}^i A_t^k, b_t^{(i)} := \sum_{k=1}^i b_t^k, c_t^{(i)} := \sum_{k=1}^i c_t^k, d_t^{(i)} := \sum_{k=1}^i d_t^k$ where $\left\{A_t^k, b_t^k, c_t^k, d_t^k\right\}$ denote the coefficients and functions estimated at the k^{th} iteration of Algorithm 5 (step 2(b)i). For simplicity, we assume that the initial distribution is a Gaussian mixture, i.e. $\pi_0 = \sum_{j=1}^J \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$ with $J \in \mathbb{N}$ components, weights $\alpha_j \in \mathbb{R}_+$ satisfying $\sum_{j=1}^J \alpha_j = 1$ and $\mu_j \in \mathbb{R}^d, \Sigma_j \in \mathbb{R}^{d\times d}$ denote the mean and covariance matrix of the j^{th} component respectively. For

sampling and importance weights to be tractable, we need to additionally impose the following inequalities (in the positive definite sense)

$$A_0^{(i)} > -\frac{1}{2}\Sigma_j^{-1} \text{ for all } j = 1, \ldots, J, \tag{4.38}$$

$$A_t^{(i)}(x_{t-1}) > -I_d/(2\Delta_t) \text{ for all } x_{t-1} \in \mathcal{X}, t = 1, \ldots, T. \tag{4.39}$$

With the form (4.36) and constraint (4.38), it follows that

$$\pi_0^{\hat{\psi}(i)} = \sum_{j=1}^{J} \tilde{\alpha}_j^{(i)} \mathcal{N}\left(\tilde{\Sigma}_j^{(i)}(\Sigma_j^{-1}\mu_j - b_0^{(i)}), \tilde{\Sigma}_j^{(i)}\right) \tag{4.40}$$

and $\pi_0(\hat{\psi}_0^{(i)}) = \sum_{j=1}^{J} \alpha_j \zeta_j^{(i)}$ where $\tilde{\Sigma}_j^{(i)} := \left(\Sigma_j^{-1} + 2A_0^{(i)}\right)^{-1} > 0$,

$$\zeta_j^{(i)} := \det(\Sigma_j)^{-1/2} \det\left(\tilde{\Sigma}_j^{(i)}\right)^{1/2} \exp\left(\frac{1}{2}\left(\Sigma_j^{-1}\mu_j - b_0^{(i)}\right)^T \tilde{\Sigma}_j^{(i)}\left(\Sigma_j^{-1}\mu_j - b_0^{(i)}\right)\right)$$
$$\times \exp\left(-\frac{1}{2}\mu_j^T \Sigma_j^{-1}\mu_j - c_0^{(i)}\right) \tag{4.41}$$

and $\tilde{\alpha}_j^{(i)} := \alpha_j \zeta_j^{(i)} / \sum_{k=1}^{J} \alpha_k \zeta_k^{(i)}$. For π_0 outside the Gaussian mixture family, we note that our methodology requires sampling from $\pi_0^{\hat{\psi}(i)}$ and computing $\pi_0(\hat{\psi}_0^{(i)})$ to be tractable.

Let $f_t(x_{t-1}) := x_{t-1} + \frac{\Delta_t}{2}\nabla \log \pi_t(x_{t-1})$ denote the forward Euler discretization and note that $\Theta_t^{(i)}(x_{t-1}) := \left(I_d + 2\Delta_t A_t^{(i)}(x_{t-1})\right)^{-1}$ is positive definite under the constraint (4.39). With the forward transition kernel (4.31) and the form of (4.37), we have

$$K_t^{\hat{\psi}(i)}(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}\left(x_t; \Theta_t^{(i)}(x_{t-1})\left(f_t - \Delta_t b_t^{(i)}\right)(x_{t-1}), \Delta_t \Theta_t^{(i)}(x_{t-1})\right) \mathrm{d}x_t, \tag{4.42}$$

and

$$K_t(\hat{\psi}_t^{(i)})(x_{t-1}) = \det\left(\Theta_t^{(i)}(x_{t-1})\right)^{1/2} \exp\left(\frac{1}{2\Delta_t}\left(f_t - \Delta_t b_t^{(i)}\right)^T \Theta_t^{(i)}\left(f_t - \Delta_t b_t^{(i)}\right)(x_{t-1})\right)$$
$$\times \exp\left(-\frac{1}{2\Delta_t} f_t^T f_t(x_{t-1}) - c_t^{(i)} - d_t^{(i)}(x_{t-1})\right) \tag{4.43}$$

for $t = 1, \ldots, T$. The parameterization (4.35) still leaves a great deal of flexibility over the choice of functions $\{A_t^i, b_t^i, d_t^i\}$. The use of constant functions, considered in Pieralberto et al. (2016), can be effective when V^* is adequately approximated

by quadratic functions (see Section 4.4.2 for an application). Setting A_t^i to zero is closely related to the continuous time approach in Kappen and Ruiz (2016) and Ruiz and Kappen (2016), where (4.42) corresponds to an Euler-Maruyama discretization of a controlled diffusion with additive control $-b_t^{(i)}$ learned through an iterative procedure. By comparing (4.35) with step 2(a) of Algorithm 4, we see that at iteration $i \in \mathbb{N}$, we should select $d_t^i(x_{t-1})$ so as to match all additive x_{t-1} terms in $-\log w_t^{\hat{\psi}(i-1)}(x_{t-1}, x_t)$. Hence from (4.33) and (4.43), we set

$$\begin{aligned} d_t^i(x_{t-1}) = \log \gamma_{t-1}(x_{t-1}) - \frac{1}{2}(\nabla \log \pi_t(x_{t-1}) + \kappa_t)^T x_{t-1} - \frac{\Delta_t}{8} \left|\nabla \log \pi_t(x_{t-1})\right|^2 \\ - \frac{1}{2\Delta_t} \left(f_t - \Delta_t b_t^{(i-1)}\right)^T \Theta_t^{(i-1)} \left(f_t - \Delta_t b_t^{(i-1)}\right)(x_{t-1}) + \frac{1}{2\Delta_t} f_t^T f_t(x_{t-1}) \\ - \frac{1}{2} \log \det \left(\Theta_t^{(i-1)}(x_{t-1})\right) \end{aligned} \tag{4.44}$$

where κ_t denotes all additive constants in the function $x \mapsto \nabla \log \pi_t(x)$. Lastly, we note that although the terms $\left\{c_t^{(i)}, d_t^{(i)}\right\}$ do not affect importance weights (4.5)-(4.6), they indirectly influence algorithmic performance through the estimation of twisting functions in Algorithm 4.

4.2.3 Connections to other work

Similar ideas have been proposed recently to develop efficient *smoothing* algorithms for discrete time state space models in Pieralberto et al. (2016) and continuous time models in Kappen and Ruiz (2016), Ruiz and Kappen (2016). In the last two references, the connection to optimal control was drawn explicitly.

We now note how the above methodology differs from these earlier contributions. Firstly, although the methods developed in these references can be extended to the AIS setting, these extensions are only formal as they result in algorithms that are not implementable. The choices made in Section 4.2.1 ensure that the resulting algorithm is practical.

Secondly, the iterative nature of Algorithm 5 differs from previously proposed methods as it approximates a different optimal control problem at each iteration. In particular, although the algorithm in Pieralberto et al. (2016) samples from different twisted versions of Q between iterations, it always approximates (4.18)

with $\psi = 1$. Therefore Algorithm 5 would offer a new methodology when applied to state space models.

Lastly, we provide a theoretical analysis of our proposed methodology in the following section. Some of these results can be applied or easily adapted to study the algorithm proposed by Pieralberto et al. (2016).

4.3 Analysis

This section is organized as follows. We begin in Section 4.3.1 by characterizing the optimal twisting and value functions of Proposition 4.4 in a particular setting of practical interest. In Section 4.3.2 and 4.3.3, we study ADP algorithm (4.29) and (4.30), respectively, in terms of approximate projection errors. For a common choice of function class, in Section 4.3.4, we give a law of large numbers and a central limit theorem to describe how these algorithms converge to their counterpart with orthogonal projections (i.e. Equation (4.28) for twisting functions) as the number of samples N converges to infinity. In Section 4.3.5, we then analyze the asymptotic behaviour of twisting functions generated by Algorithm 5 as the number of iterations I converges to infinity. Lastly in Section 4.3.6, we characterize the distance between the terminal distribution of a twisted SMC sampler and the target distribution π in terms of value function approximation errors. Before proceeding, we remark that although some results will be established under strong assumptions, which might be difficult to verify in practice, they still offer some insight into the properties of the proposed methodology. Weakening of such assumptions will be left as future work.

4.3.1 Log-concavity of optimal twisting functions and convexity of optimal value functions

Proposition 4.9. *Let $\psi \in \Psi(Q)$ and assume that the weight functions $\left\{w_t^\psi\right\}_{t=0}^T$ in (4.6) and densities of the transition kernels $\left\{K_t^\psi\right\}_{t=0}^T$ are log-concave on their domain of definition. Then the optimal twisting functions $\phi^* = \{\phi_t^*\}_{t=0}^T$ with respect to Q^ψ are log-concave and the optimal value functions $V^* = \{V_t^*\}_{t=0}^T$ with respect to Q^ψ are convex.*

Proof. We establish that the optimal twisting functions are log-concave; convexity of the optimal value functions follow from the relationship $\phi_t^* = e^{-V_t^*}, t = 0, \ldots, T$. For $t = T$, log-concavity of $\phi_T^* = w_T^\psi$ follows by assumption. Assuming that $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is log-concave, note that $y \mapsto K_{t+1}^\psi(\varphi)(y)$ is log-concave since the product $(x, y) \mapsto \varphi(x, y) K_{t+1}^\psi(x, y)$ is and log-concavity is preserved by marginalization. It follows that the Bellman operator B_t^ψ preserves log-concavity as the product $(x, y) \mapsto w_t^\psi(x, y) K_{t+1}^\psi(\varphi)(y)$ is also log-concave. By induction, the proof is complete using the backward recursion (4.15). □

4.3.2 Approximate dynamic programming for learning optimal sequence of twisting functions

We first note that the Bellman operators (4.13) are Lipschitz continuous maps in L^2-norm.

Lemma 4.10. *Let $\psi \in \Psi(Q)$ be such that $\|w_t^\psi\|_\infty < \infty$ for all $t = 0, \ldots, T-1$. For each $t = 0, \ldots, T-1$, the Bellman operator $\mathsf{B}_t^\psi : L^2(Q_{t,t+1}^\psi) \to L^2(Q_{t-1,t}^\psi)$ defined in (4.13) satisfies*

$$\|\mathsf{B}_t^\psi \varphi - \mathsf{B}_t^\psi \xi\|_{L^2(Q_{t-1,t}^\psi)} \leq \|w_t^\psi\|_\infty \|\varphi - \xi\|_{L^2(Q_{t,t+1}^\psi)} \tag{4.45}$$

for all $\varphi, \xi \in L^2(Q_{t,t+1}^\psi)$.

Proof. By Jensen's inequality, we have

$$\begin{aligned}
\|\mathsf{B}_t^\psi \varphi - \mathsf{B}_t^\psi \xi\|_{L^2(Q_{t-1,t}^\psi)}^2 &= \int_{\mathcal{X}^2} w_t^\psi(x_{t-1}, x_t)^2 K_{t+1}^\psi(\varphi - \xi)(x_t)^2 Q_{t-1,t}^\psi(\mathrm{d}x_{t-1}, \mathrm{d}x_t) \\
&\leq \|w_t^\psi\|_\infty^2 \int_{\mathcal{X}^2} (\varphi - \xi)^2(x_t, x_{t+1}) K_{t+1}^\psi(x_t, \mathrm{d}x_{t+1}) Q_t^\psi(\mathrm{d}x_t) \\
&= \|w_t^\psi\|_\infty^2 \int_{\mathcal{X}^2} (\varphi - \xi)^2(x_t, x_{t+1}) Q_{t,t+1}^\psi(\mathrm{d}x_t, \mathrm{d}x_{t+1}) \\
&= \|w_t^\psi\|_\infty^2 \|\varphi - \xi\|_{L^2(Q_{t,t+1}^\psi)}^2
\end{aligned} \tag{4.46}$$

for $t = 1, \ldots, T-1$. The same arguments show that (4.45) also holds for $t = 0$ and $\mathsf{B}_t^\psi \varphi \in L^2(Q_{t-1,t}^\psi)$ if $\varphi \in L^2(Q_{t,t+1}^\psi)$. □

The next result characterizes the error of ADP algorithm (4.29) for learning the optimal sequence of twisting functions $\phi^* = \{\phi_t^*\}_{t=0}^T$ with respect to Q^ψ in terms of approximate projection errors. In the following, we will write $\mathbb{E}$ to denote expectation with respect to $N \in \mathbb{N}$ iid random variables $\{X_{0:T}^n\}_{n=1}^N$ with distribution Q^ψ produced in the algorithm of interest.

Proposition 4.11. *Let $\psi \in \Psi(Q)$ be such that $\|w_t^\psi\|_\infty < \infty$ for all $t = 0, \ldots, T$. For each $t = 0, \ldots, T$, given $N \in \mathbb{N}$ and a pre-specified closed and linear function class $\mathsf{F}_t \subset L^2(Q_{t-1,t}^\psi)$, we assume that the approximate $(\mathsf{F}_t, Q_{t-1,t}^\psi)$-projection operator satisfies*

$$\sup_{\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}} \mathbb{E}\|\mathsf{P}_t^{\psi,N}\xi - \xi\|_{L^2(Q_{t-1,t}^\psi)} \leq \varepsilon_t(N) \tag{4.47}$$

where $\mathsf{B}_t^\psi \mathsf{F}_{t+1} := \{\mathsf{B}_t^\psi \varphi : \varphi \in \mathsf{F}_{t+1}\}$ for $t = 0, \ldots, T-1$ and $\mathsf{B}_T^\psi \mathsf{F}_{T+1} := \{w_T^\psi\}$. Let $\hat{\phi} = \left\{\hat{\phi}_t\right\}_{t=0}^T$ denote the twisting functions generated by ADP algorithm (4.29) and suppose that $\left(x, \{X_{t-1:T}^n\}_{n=1}^N\right) \mapsto \hat{\phi}_t(x)$ is Borel measurable for all t. Then

$$\mathbb{E}\|\hat{\phi}_t - \phi_t^*\|_{L^2(Q_{t-1,t}^\psi)} \leq \sum_{k=t}^T \varepsilon_k(N) \prod_{i=t}^{k-1} \|w_i^\psi\|_\infty \tag{4.48}$$

for $t = 0, \ldots, T$ (with the convention that $\prod_{i=j}^l := 1$ for $l < j$).

Proof. Using triangle inequality, the assumption on approximate projection errors and Lemma 4.10 (since $\mathsf{F}_{t+1} \subset L^2(Q_{t,t+1}^\psi)$ and $\phi_{t+1}^* \in L^2(Q_{t,t+1}^\psi)$), we have

$$\begin{aligned} e_t &:= \mathbb{E}\|\hat{\phi}_t - \phi_t^*\|_{L^2(Q_{t-1,t}^\psi)} \\ &= \mathbb{E}\|\hat{\phi}_t - \mathsf{B}_t^\psi \phi_{t+1}^*\|_{L^2(Q_{t-1,t}^\psi)} \\ &\leq \mathbb{E}\|\mathsf{P}_t^{\psi,N}\mathsf{B}_t^\psi \hat{\phi}_{t+1} - \mathsf{B}_t^\psi \hat{\phi}_{t+1}\|_{L^2(Q_{t-1,t}^\psi)} + \mathbb{E}\|\mathsf{B}_t^\psi \hat{\phi}_{t+1} - \mathsf{B}_t^\psi \phi_{t+1}^*\|_{L^2(Q_{t-1,t}^\psi)} \\ &\leq \varepsilon_t(N) + \|w_t^\psi\|_\infty e_{t+1} \end{aligned} \tag{4.49}$$

for $t = 0, \ldots, T-1$. Equation (4.48) holds for $t = T$ by assumption since $\mathsf{B}_T^\psi \mathsf{F}_{T+1} = \{w_T^\psi\}$ and for all $t = 0, \ldots, T-1$ by induction. □

Although the assumption that the weights $\left\{w_t^\psi\right\}_{t=0}^T$ are bounded is typically used in similar ADP error analyses (Gobet 2016), in our context, such a condition might be quite difficult to verify in practice. Equation (4.48) reveals how the approximate projection errors propagate backwards in time; in particular, it shows that if these errors can be kept small then ADP algorithm (4.29) provides a good approximation of ϕ^*. As these errors are measured in L^2-norm, it is possible to provide a more precise description of $\varepsilon_t(N)$ using results on least squares approximations which we collect in Appendix B. The following is a corollary of Proposition B.1 which deals with the case of linear least squares; see Appendix B.1.2 for the non-linear case. For any real square matrix A, we denote its minimum eigenvalue by $\lambda_{\min}(A)$.

Corollary 4.12. *(Linear least squares) For each $t = 0, \dots, T$, suppose that the function class is of the form $\mathsf{F}_t = \left\{\Phi_t(x)^T\beta_t : \beta_t \in \mathbb{R}^{M_t}\right\}$ where $\Phi_t := \left(\varphi_t^1, \dots, \varphi_t^{M_t}\right)^T$ is a vector of $M_t \in \mathbb{N}$ pre-specified basis functions satisfying $\|\varphi_t^m\|_\infty \leq U_t$ for all $m = 1, \dots, M_t$. Define the matrices*

$$A_0^N := \frac{1}{N}\sum_{n=1}^N \left(\Phi_0\Phi_0^T\right)(X_0^n), \quad A_0 := \mathbb{E}_{Q^\psi}\left[\left(\Phi_0\Phi_0^T\right)(X_0)\right], \tag{4.50}$$
$$A_t^N := \frac{1}{N}\sum_{n=1}^N \left(\Phi_t\Phi_t^T\right)(X_{t-1}^n, X_t^n), \quad A_t := \mathbb{E}_{Q^\psi}\left[\left(\Phi_t\Phi_t^T\right)(X_{t-1}, X_t)\right],$$

for $t = 1, \dots, T$, assume that $\lambda_{\min}(A_t) > 0$ and $N \in \mathbb{N}$ is such that $\mathbb{E}\left[\lambda_{\min}^{-4}(A_t^N)\right] < \infty$ for all t. For $t = 0, \dots, T$ and $p \geq 1$, let $\|\mathsf{F}_t\|_{L^p(Q_{t-1,t}^\psi)} := \sup_{\xi\in\mathsf{F}_t}\|\xi\|_{L^p(Q_{t-1,t}^\psi)}$ (set $\|\mathsf{F}_{T+1}\|_{L^p(Q_{T,T+1}^\psi)} := 1$ for notational convenience) and suppose this is finite for $p = 4$. Then

$$\varepsilon_t^2(N) = \sup_{\xi\in\mathsf{B}_t^\psi\mathsf{F}_{t+1}} \|\mathsf{P}_t^\psi\xi - \xi\|^2_{L^2(Q_{t-1,t}^\psi)} + C_t(N)U_t^4M_t^2\mathbb{E}\left[\lambda_{\min}^{-4}(A_t^N)\right]^{1/2}\|w_t^\psi\|_\infty^2 \tag{4.51}$$

where $C_t(N) := \left(\varrho_t(N) + \lambda_{\min}^{-1}(A_t)\|\mathsf{F}_{t+1}\|_{L^2(Q_{t,t+1}^\psi)}\theta_t(N)\right)^2$ with

$$\varrho_t^2(N) := \left(3\left(N^{-2}+N^{-3}\right)\|\mathsf{F}_{t+1}\|^4_{L^2(Q_{t,t+1}^\psi)} + 5N^{-3}\|\mathsf{F}_{t+1}\|^4_{L^4(Q_{t,t+1}^\psi)}\right)^{1/2}, \tag{4.52}$$
$$\theta_t^2(N) := \sum_{n,m=1}^M \left(3\left(N^{-2}-N^{-3}\right)Q_{t-1,t}^\psi\left(\left[\varphi_t^n\varphi_t^m - Q_{t-1,t}^\psi(\varphi_t^n\varphi_t^m)\right]^2\right)^2 + N^{-3}Q_{t-1,t}^\psi\left(\left[\varphi_t^n\varphi_t^m - Q_{t-1,t}^\psi(\varphi_t^n\varphi_t^m)\right]^4\right)\right)^{1/2}.$$

The upper bound (4.51) gives a decomposition of the error involved in propagating approximate projections. The first term is the largest norm of residuals $\mathsf{P}_t^\psi \xi - \xi \in \mathsf{F}_t^\perp$ for $\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}$, where $\mathsf{F}_t^\perp$ denotes the orthogonal complement to F_t in $L^2(Q_{t-1,t}^\psi)$. This should be thought of as the bias term that depends on the choice of function class. The second term describes the variance that comes from learning with only a sample of size N (note that the error is of order N^{-1}) with constants that depend on the number of basis functions M_t and their magnitude U_t, the moments in (4.52) and how well-conditioned the projection and its approximation are, described by the size of $\lambda_{\min}^{-1}(A_t)$ and the inverse moments of $\lambda_{\min}(A_t^N)$ respectively.

4.3.3 Approximate dynamic programming for learning optimal sequence of value functions

Like before, we first note that the Bellman operators (4.14) are non-expansive maps in the supremum norm.

Lemma 4.13. *Let $\psi \in \Psi(Q)$ be such that $\|w_t^\psi\|_\infty < \infty$ for all $t = 0, \ldots, T-1$ and define $D := \left\{\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : \inf_{(x,y)\in\mathcal{X}^2} \varphi(x,y) > -\infty\right\}$. For each $t = 0, \ldots, T-1$, the Bellman operator $\mathsf{T}_t^\psi : D \to D$ defined in (4.14) satisfies*

$$\|\mathsf{T}_t^\psi \varphi - \mathsf{T}_t^\psi \xi\|_\infty \leq \|\varphi - \xi\|_\infty \tag{4.53}$$

for all $\varphi, \xi \in D$.

Proof. For any $\varphi \in D$, we have $(\mathsf{T}_t^\psi \varphi)(x,y) \geq -\log \|w_t^\psi\|_\infty + \inf_{(u,v)\in\mathcal{X}^2} \varphi(u,v)$ hence $\mathsf{T}_t^\psi \varphi \in D$. Note that

$$\begin{aligned}(\mathsf{T}_t^\psi \varphi - \mathsf{T}_t^\psi \xi)(x,y) &= \log K_{t+1}^\psi(e^{-\xi})(y) - \log K_{t+1}^\psi(e^{-\varphi})(y) \\ &\leq \log K_{t+1}^\psi(e^{-\varphi+\|\varphi-\xi\|_\infty})(y) - \log K_{t+1}^\psi(e^{-\varphi})(y) \\ &= \|\varphi - \xi\|_\infty.\end{aligned} \tag{4.54}$$

By symmetry, we have $|\mathsf{T}_t^\psi \varphi - \mathsf{T}_t^\psi \xi|(x,y) \leq \|\varphi - \xi\|_\infty$ hence taking supremum gives the desired inequality. □

To measure errors in supremum norm, we will rely on the following lemma to deal with measurability issues.

Lemma 4.14. *Let Ω_1 be a topological space, $\mathcal{B}(\Omega_1)$ be its corresponding Borel σ-algebra and Ω_2 be an arbitrary set. If $\varphi : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is a function such that $x \mapsto \varphi(x, y)$ is continuous for all $y \in \Omega_2$, then $x \mapsto \sup_{y\in\Omega_2} \varphi(x, y)$ is Borel measurable.*

Proof. It suffices to check that $A_\alpha := \left\{x \in \Omega_1 : \sup_{y\in\Omega_2} \varphi(x, y) \leq \alpha\right\} \in \mathcal{B}(\Omega_1)$ for all $\alpha \in \mathbb{R}$. Write $\xi_y(x) := \varphi(x, y)$ and note that since $(-\infty, \alpha]$ is closed in $\mathbb{R}$, by continuity $\xi_y^{-1}((-\infty, \alpha])$ is closed in Ω_1 for all $y \in \Omega_2$. Hence $A_\alpha = \{x \in \Omega_1 : \varphi(x, y) \leq \alpha \, \forall y \in \Omega_2\} = \cap_{y\in\Omega_2} \xi_y^{-1}((-\infty, \alpha])$ is a closed set in $\mathcal{B}(\Omega_1)$. □

Analogous to Proposition 4.11 for twisting functions, the next result characterizes the error of ADP algorithm (4.30) for learning the optimal sequence of value functions $V^* = \{V_t^*\}_{t=0}^T$ with respect to Q^ψ in terms of approximate projection errors.

Proposition 4.15. *Let $\psi \in \Psi(Q)$ be such that $\|w_t^\psi\|_\infty < \infty$ for all $t = 0, \ldots, T$. Given pre-specified closed and linear function classes $\mathsf{F}_t \subset D, t = 0, \ldots, T$, we define the sets $\mathsf{T}_t^\psi \mathsf{F}_{t+1} := \{\mathsf{T}_t^\psi \varphi : \varphi \in \mathsf{F}_{t+1}\}$ for $t = 0, \ldots, T-1$ and $\mathsf{T}_T^\psi \mathsf{F}_{T+1} := \{-\log w_T^\psi\}$. Given $N \in \mathbb{N}$, for each $t = 0, \ldots, T$, we assume that the approximate $(\mathsf{F}_t, Q_{t-1,t}^\psi)$-projection operator is such that $\left\{X_{t-1:t}^n\right\}_{n=1}^N \mapsto (\mathsf{P}_t^{\psi,N}\xi)(x)$ is continuous for all $x \in \mathcal{X}^2, \xi \in \mathsf{T}_t^\psi \mathsf{F}_{t+1}$ and*

$$\sup_{\xi\in\mathsf{T}_t^\psi \mathsf{F}_{t+1}} \mathbb{E}\|\mathsf{P}_t^{\psi,N}\xi - \xi\|_\infty \leq \varepsilon_t(N). \tag{4.55}$$

Let $\hat{V} = \left\{\hat{V}_t\right\}_{t=0}^T$ denote the value functions generated by ADP algorithm (4.30) and suppose that $\left\{X_{t-1:T}^n\right\}_{n=1}^N \mapsto \hat{V}_t(x)$ is continuous for all $x \in \mathcal{X}^2$ and t. Then

$$\mathbb{E}\|\hat{V}_t - V_t^*\|_\infty \leq \sum_{k=t}^T \varepsilon_k(N) \tag{4.56}$$

for $t = 0, \ldots, T$.

Proof. Continuity assumptions and Lemma 4.14 ensure that the left hand side of (4.55) and (4.56) are well-defined. Noting that we have $\mathsf{F}_t \subset D$ and $V_t^* \in D$ since $V_t^* \geq -\sum_{k=t}^T \log \|w_k^\psi\|_\infty$, the rest of the proof follows from Lemma 4.13 and the same arguments in Proposition 4.11. □

We note that it is also possible to apply Proposition 4.11 to quantify the error of the corresponding approximation $\hat{\phi}_t := e^{-\hat{V}_t}, t = 0, \ldots, T$ of the optimal sequence of twisting functions $\phi^* = \{\phi_t^*\}_{t=0}^T$ with respect to Q^ψ.

4.3.4 Limit theorems

We now examine the asymptotic behaviour of ADP algorithm (4.29) and (4.30) with linear least squares approximations as the number of samples N converges to infinity. For notational simplicity, given a vector of $M \in \mathbb{N}$ basis functions $\Phi := (\varphi_1, \ldots, \varphi_M)^T, G \in \mathcal{M}(\mathcal{X})$ and $c \in \mathbb{R}^M$, we denote $G(\Phi)(x) := (G(\varphi_1)(x), \ldots, G(\varphi_M)(x))^T$ as the vector-valued function and $\langle G(\Phi), c\rangle(x) := \sum_{m=1}^M c_m G(\varphi_m)(x)$ as the linear combination. Also for any $p \in \mathbb{N}$, we write 0_p as the vector of p zeros and $0_{p\times p}$ as the $p \times p$ matrix of zeros.

We first look at algorithm (4.29), in which case for each $t = 0, \ldots, T$, given $M_t \in \mathbb{N}$ basis functions $\Phi_t := \left(\varphi_t^1, \ldots, \varphi_t^{M_t}\right)^T$, the approximate twisting function has the form $\hat{\phi}_t = \Phi_t^T \beta_t^N$, where the least squares estimator $\beta_t^N = \left(A_t^N\right)^{-1} b_t^N$ is given by the matrix A_t^N in (4.50) and vector b_t^N defined by the backward recursion

$$
\begin{aligned}
b_T^N &:= \frac{1}{N}\sum_{n=1}^N w_T^\psi(X_{T-1}^n, X_T^n)\Phi_T(X_{T-1}^n, X_T^n), \qquad (4.57)\\
b_t^N &:= \frac{1}{N}\sum_{n=1}^N w_t^\psi(X_{t-1}^n, X_t^n)\left\langle K_{t+1}^\psi(\Phi_{t+1}), \left(A_{t+1}^N\right)^{-1} b_{t+1}^N\right\rangle (X_t^n)\Phi_t(X_{t-1}^n, X_t^n),\\
&\qquad t = T-1, \ldots, 1,\\
b_0^N &:= \frac{1}{N}\sum_{n=1}^N w_0^\psi(X_0^n)\left\langle K_1^\psi(\Phi_1), \left(A_1^N\right)^{-1} b_1^N\right\rangle (X_0^n)\Phi_0(X_0^n).
\end{aligned}
$$

For notational ease, here we consider the setting where the number of basis functions used remains constant over time, i.e. $M_t = M$; extension to the time varying case is straightforward.

Theorem 4.16. *Let $\psi \in \Psi(Q)$ and $\beta^N := \left(\beta_0^N, \ldots, \beta_T^N\right)^T \in \mathbb{R}^{(T+1)M}$ be the vector of least squares estimators generated by ADP algorithm (4.29) under linear least squares approximations with basis functions $\{\Phi_t\}_{t=0}^T$. Let $\{A_t\}_{t=0}^T$ be the matrices defined in (4.50) and assume that $\lambda_{\min}(A_t) > 0$ for all $t = 0, \ldots, T$. Suppose $x \mapsto K_t^\psi(\Phi_t)(x)$ is Borel measurable for $t = 1, \ldots, T$ and*

$$\begin{aligned}
&\mathbb{E}_{Q^\psi}\left[w_0^\psi(X_0)\left|K_1^\psi(\Phi_1)(X_0)\right|\left|\Phi_0(X_0)\right|\right] < \infty, \\
&\mathbb{E}_{Q^\psi}\left[w_t^\psi(X_{t-1}, X_t)\left|K_{t+1}^\psi(\Phi_{t+1})(X_t)\right|\left|\Phi_t(X_{t-1}, X_t)\right|\right] < \infty,\ t = 1, \ldots, T-1, \\
&\mathbb{E}_{Q^\psi}\left[w_T^\psi(X_{T-1}, X_T)\left|\Phi_T(X_{T-1}, X_T)\right|\right] < \infty.
\end{aligned} \tag{4.58}$$

Define the vectors $\{b_t\}_{t=0}^T$ by the backward recursion

$$\begin{aligned}
b_T &:= \mathbb{E}_{Q^\psi}\left[w_T^\psi(X_{T-1}, X_T)\Phi_T(X_{T-1}, X_T)\right], \\
b_t &:= \mathbb{E}_{Q^\psi}\left[w_t^\psi(X_{t-1}, X_t)\left\langle K_{t+1}^\psi(\Phi_{t+1}), A_{t+1}^{-1}b_{t+1}\right\rangle(X_t)\Phi_t(X_{t-1}, X_t)\right], \\
&\qquad t = T-1, \ldots, 1, \\
b_0 &:= \mathbb{E}_{Q^\psi}\left[w_0^\psi(X_0)\left\langle K_1^\psi(\Phi_1), A_1^{-1}b_1\right\rangle(X_0)\Phi_0(X_0)\right].
\end{aligned} \tag{4.59}$$

As $N \to \infty$, β^N converges in probability to $\beta^* := (\beta_0^*, \ldots, \beta_T^*)^T$ where $\beta_t^* := A_t^{-1}b_t$ for $t = 0, \ldots, T$.

Define the vector-valued functions

$$\begin{aligned}
f_0(x_0) &:= A_0^{-1}\left(w_0^\psi(x_0)\left\langle K_1^\psi(\Phi_1), \beta_1^*\right\rangle(x_0)\Phi_0(x_0) - \left(\Phi_0\Phi_0^T\right)(x_0)\beta_0^*\right), \\
f_t(x_{t-1}, x_t) &:= A_t^{-1}\left(w_t^\psi(x_{t-1}, x_t)\left\langle K_{t+1}^\psi(\Phi_{t+1}), \beta_{t+1}^*\right\rangle(x_t)\Phi_t(x_{t-1}, x_t)\right) \\
&\qquad - A_t^{-1}\left(\Phi_t\Phi_t^T\right)(x_{t-1}, x_t)\beta_t^*, \quad t = 1, \ldots, T-1, \\
f_T(x_{T-1}, x_T) &:= A_T^{-1}\left(w_T^\psi(x_{T-1}, x_T)\Phi_T(x_{T-1}, x_T) - \left(\Phi_T\Phi_T^T\right)(x_{T-1}, x_T)\beta_T^*\right),
\end{aligned} \tag{4.60}$$

and write $f := (f_0, \ldots, f_T)^T$ as the vector of functions. If we additionally assume that $\mathbb{E}_{Q^\psi}|f_0(X_0)|^2 < \infty$, $\mathbb{E}_{Q^\psi}|f_t(X_{t-1}, X_t)|^2 < \infty$ for all $t = 1, \ldots, T$, then

$$\sqrt{N}\left(\beta^N - \beta^*\right) \xrightarrow{d.} \mathcal{N}\left(0_{(T+1)M}, \Sigma\right) \tag{4.61}$$

where $\Sigma = U\Sigma_f U^T$ *is given in terms of* $\Sigma_f := \mathbb{E}_{Q^\psi}\left[\left(ff^T\right)(X_{0:T})\right]$ *and a block upper triangular matrix* $U \in \mathbb{R}^{(T+1)M\times(T+1)M}$ *described by its blocks of size* $M \times M$

$$U_{ij} := \begin{cases} 0_{M\times M}, & i > j, \\ I_M, & i = j, \\ \prod_{k=i-1}^{j-2} G_k, & i < j, \end{cases} \tag{4.62}$$

for $i, j = 0, \ldots, T$, *with*

$$G_0 := A_0^{-1}\mathbb{E}_{Q^\psi}\left[w_0^\psi(X_0)\Phi_0(X_0)K_1^\psi(\Phi_1)(X_0)^T\right], \tag{4.63}$$
$$G_t := A_t^{-1}\mathbb{E}_{Q^\psi}\left[w_t^\psi(X_{t-1}, X_t)\Phi_t(X_{t-1}, X_t)K_{t+1}^\psi(\Phi_{t+1})(X_t)^T\right], \quad t = 1, \ldots, T-1.$$

Proof. Note that for each $t = 0, \ldots, T$, by strong LLN $A_t^N \to A_t$ almost surely as $N \to \infty$, hence using continuity of matrix inversion and the continuous mapping theorem, we have $\left(A_t^N\right)^{-1} \to A_t^{-1}$ almost surely. Using continuity of the spectral matrix norm and another application of the continuous mapping theorem, we see that $\lambda_{\min}\left(A_t^N\right) \to \lambda_{\min}\left(A_t\right) > 0$ almost surely, so for sufficiently large N we have invertibility of A_t^N with probability one.

Starting with time $t = T$, by LLN $b_T^N \to b_T$ in probability so by Slutsky's lemma, it follows that $\beta_T^N = \left(A_T^N\right)^{-1} b_T^N \to \beta_T^*$ in probability. Consider

$$\beta_T^N - \beta_T^* = \left(A_T^N\right)^{-1}\left(b_T^N - A_T^N\beta_T^*\right) = \left(A_T^{-1} + o_P(1)\right)\left(b_T^N - A_T^N\beta_T^*\right). \tag{4.64}$$

Since $A_T^{-1}\left(b_T^N - A_T^N\beta_T^*\right) = N^{-1}\sum_{n=1}^N f_T(X_{T-1}^n, X_T^n)$, $\mathbb{E}_{Q^\psi}\left[f_T(X_{T-1}, X_T)\right] = 0_M$ and $\mathbb{E}_{Q^\psi}|f_T(X_{T-1}, X_T)|^2 < \infty$, we have $b_T^N - A_T^N\beta_T^* = O_P(N^{-1/2})$. Therefore

$$\beta_T^N - \beta_T^* = \frac{1}{N}\sum_{n=1}^N f_T(X_{T-1}^n, X_T^n) + o_P(N^{-1/2}), \tag{4.65}$$

hence applying CLT gives $\sqrt{N}\left(\beta_T^N - \beta_T^*\right) \xrightarrow{d.} \mathcal{N}\left(0_M, \Sigma_{f_T}\right)$ with $\Sigma_{f_T} := \mathbb{E}_{Q^\psi}\left[\left(f_T f_T^T\right)(X_{T-1}, X_T)\right]$.

We now argue inductively: for time $t = 1, \ldots, T-1$ (same argument for $t = 0$), we decompose $b_t^N = c_t^N + d_t^N$ with

$$c_t^N := \frac{1}{N}\sum_{n=1}^N w_t^\psi(X_{t-1}^n, X_t^n)\left\langle K_{t+1}^\psi(\Phi_{t+1}), \beta_{t+1}^*\right\rangle(X_t^n)\Phi_t(X_{t-1}^n, X_t^n), \tag{4.66}$$
$$d_t^N := \frac{1}{N}\sum_{n=1}^N w_t^\psi(X_{t-1}^n, X_t^n)\left\langle K_{t+1}^\psi(\Phi_{t+1}), \beta_{t+1}^N - \beta_{t+1}^*\right\rangle(X_t^n)\Phi_t(X_{t-1}^n, X_t^n).$$

Assuming that $\beta_{t+1}^N \to \beta_{t+1}^*$ in probability, by Cauchy-Schwarz inequality we see that $|d_t^N| = o_P(1)$ so $b_t^N \to b_t$ in probability. Therefore by Slutsky's lemma, it follows that $\beta_t^N = \left(A_t^N\right)^{-1} b_t^N \to \beta_t^*$ in probability.

We now examine

$$\beta_t^N - \beta_t^* = \left(A_t^N\right)^{-1}\left(b_t^N - A_t^N\beta_t^*\right) = \left(A_t^{-1} + o_P(1)\right)\left(c_t^N + d_t^N - A_t^N\beta_t^*\right). \quad (4.67)$$

Firstly, since $A_t^{-1}\left(c_t^N - A_t^N\beta_t^*\right) = N^{-1}\sum_{n=1}^N f_t(X_{t-1}^n, X_t^n)$, $\mathbb{E}_{Q^\psi}\left[f_t(X_{t-1}, X_t)\right] = 0_M$ and $\mathbb{E}_{Q^\psi}|f_t(X_{t-1}, X_t)|^2 < \infty$, we have $c_t^N - A_t^N\beta_t^* = O_P(N^{-1/2})$. Secondly, assuming that $\sqrt{N}\left(\beta_{t+1}^N - \beta_{t+1}^*\right) \xrightarrow{d.} \mathcal{N}\left(0_M, \Sigma_{f_{t+1}}\right)$ for some covariance matrix $\Sigma_{f_{t+1}}$, we can write $A_t^{-1}d_t^N = G_t(\beta_{t+1}^N - \beta_{t+1}^*) + o_P(N^{-1/2})$ so $d_t^N = O_P(N^{-1/2})$. Combining these observations and noting that the least squares estimators β_t^N and β_{t+1}^N depend on common random variables, we consider the difference

$$\beta_t^N - \beta_t^* - G_t(\beta_{t+1}^N - \beta_{t+1}^*) = \frac{1}{N}\sum_{n=1}^N f_t(X_{t-1}^n, X_t^n) + o_P(N^{-1/2}), \quad (4.68)$$

which also holds for the case $t = 0$ using the same arguments. Stacking (4.68) for $t = 0, \ldots, T-1$ and (4.65) for $t = T$ as a $(T+1)M$-dimensional vector gives

$$Y_N := \begin{pmatrix} \beta_0^N - \beta_0^* - G_0\left(\beta_1^N - \beta_1^*\right) \\ \beta_1^N - \beta_1^* - G_1\left(\beta_2^N - \beta_2^*\right) \\ \vdots \\ \beta_{T-1}^N - \beta_{T-1}^* - G_{T-1}\left(\beta_T^N - \beta_T^*\right) \\ \beta_T^N - \beta_T^* \end{pmatrix} = \frac{1}{N}\sum_{n=1}^N f(X_{0:T}^n) + o_P(N^{-1/2}). \quad (4.69)$$

Noting that $UY_N = \beta^N - \beta^*$ for any $N \in \mathbb{N}$ and by CLT $\sqrt{N}Y_N \xrightarrow{d.} \mathcal{N}\left(0_{(T+1)M}, \Sigma_f\right)$ as $N \to \infty$, (4.61) follows from an application of the continuous mapping theorem.

□

The above result reveals that in the large N regime, the distribution of twisting functions generated by ADP algorithm (4.29) concentrates around (4.28) with Gaussian fluctuations. We note that the block upper triangular structure of matrix U is inherited from the backward nature of ADP recursions. The form of the asymptotic variance $\Sigma = U\Sigma_f U^T$ also suggests that in general, we expect the variance of the least squares estimator to be larger at earlier times.

We now analyze algorithm (4.30). Like before, for each $t = 0, \ldots, T$, the approximate value function has the form $\hat{V}_t = \Phi_t^T \beta_t^N$, where the least squares estimator $\beta_t^N = \left(A_t^N\right)^{-1} b_t^N$ is given by the matrix A_t^N in (4.50) and vector b_t^N defined by the backward recursion

$$b_T^N := -\frac{1}{N} \sum_{n=1}^{N} \log w_T^\psi(X_{T-1}^n, X_T^n) \Phi_T(X_{T-1}^n, X_T^n), \tag{4.70}$$

$$b_t^N := -\frac{1}{N} \sum_{n=1}^{N} \left(\log w_t^\psi(X_{t-1}^n, X_t^n) + \log K_{t+1}^\psi(e^{-\Phi_{t+1}^T \left(A_{t+1}^N\right)^{-1} b_{t+1}^N})(X_t^n) \right) \Phi_t(X_{t-1}^n, X_t^n),$$

$$t = T-1, \ldots, 1,$$

$$b_0^N := -\frac{1}{N} \sum_{n=1}^{N} \left(\log w_0^\psi(X_0^n) + \log K_1^\psi(e^{-\Phi_1^T \left(A_1^N\right)^{-1} b_1^N})(X_0^n) \right) \Phi_0(X_0^n).$$

Theorem 4.17. *Let $\psi \in \Psi(Q)$ and $\beta^N := \left(\beta_0^N, \ldots, \beta_T^N\right)^T \in \mathbb{R}^{(T+1)M}$ be the vector of least squares estimators generated by ADP algorithm (4.30) under linear least squares approximations with basis functions $\{\Phi_t\}_{t=0}^T$. Let $\{A_t\}_{t=0}^T$ be the matrices defined in (4.50) and assume that $\lambda_{\min}(A_t) > 0$ for all $t = 0, \ldots, T$. Suppose $x \mapsto K_t^\psi(e^{-\Phi_t^T \beta})(x)$ is Borel measurable for all $\beta \in \mathbb{R}^M$ and $t = 1, \ldots, T$, define the vectors $\{b_t\}_{t=0}^T$ by the backward recursion*

$$b_T := -\mathbb{E}_{Q^\psi} \left[\log w_T^\psi(X_{T-1}, X_T) \Phi_T(X_{T-1}, X_T) \right], \tag{4.71}$$

$$b_t := -\mathbb{E}_{Q^\psi} \left[\left(\log w_t^\psi(X_{t-1}, X_t) + \log K_{t+1}^\psi(e^{-\Phi_{t+1}^T A_{t+1}^{-1} b_{t+1}})(X_t) \right) \Phi_t(X_{t-1}, X_t) \right],$$

$$t = T-1, \ldots, 1,$$

$$b_0 := -\mathbb{E}_{Q^\psi} \left[\left(\log w_0^\psi(X_0) + \log K_1^\psi(e^{-\Phi_1^T A_1^{-1} b_1})(X_0) \right) \Phi_0(X_0) \right],$$

and assume that these expectations are finite. Assume also that for each $t = 0, \ldots, T-1$, there exist $C_t : \mathcal{X} \to \mathbb{R}_+$ and $\rho_t \in C(\mathbb{R}_+, \mathbb{R}_+)$ satisfying

$$\mathbb{E}_{Q^\psi} \left[C_t(X_t) \left| \Phi_t(X_{t-1}, X_t) \right| \right] < \infty \quad \text{and} \quad \lim_{x \to 0} \rho_t(x) = 0 \tag{4.72}$$

such that

$$\left| \log K_{t+1}^\psi(e^{-\Phi_{t+1}^T \beta})(x) - \log K_{t+1}^\psi(e^{-\Phi_{t+1}^T \beta'})(x) \right| \leq C_t(x) \rho_t\left(|\beta - \beta'|\right) \tag{4.73}$$

holds $Q_t^\psi(\mathrm{d}x)$ − a.e. for all $\beta, \beta' \in \mathbb{R}^M$. As $N \to \infty$, β^N converges in probability to $\beta^* := (\beta_0^*, \ldots, \beta_T^*)^T$ where $\beta_t^* := A_t^{-1} b_t$ for $t = 0, \ldots, T$.

Suppose that for each $t = 0, \ldots, T-1$, $\beta \mapsto \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T \beta})(x)$ *is twice continuously differentiable for all* $x \in \mathcal{X}$, *the* $i, j = 1, \ldots, M$ *element of its Hessian matrix* $x \mapsto H_{t+1}^{i,j}(\beta, x) := \partial_{\beta_i} \partial_{\beta_j} \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T \beta})(x)$ *is Borel measurable for all* $\beta \in \mathbb{R}^M$, *satisfies* $\mathbb{E}_{Q^\psi} |H_{t+1}^{i,j}(\beta_{t+1}^*, X_t) \Phi_t(X_{t-1}, X_t)| < \infty$ *and*

$$\frac{1}{N} \sum_{n=1}^{N} H_{t+1}^{i,j}(\tilde{\beta}_{t+1}^N, X_t^n) \Phi_t(X_{t-1}^n, X_t^n) \to \mathbb{E}_{Q^\psi} \left[H_{t+1}^{i,j}(\beta_{t+1}^*, X_t) \Phi_t(X_{t-1}, X_t) \right] \tag{4.74}$$

in probability whenever $\tilde{\beta}_{t+1}^N \to \beta_{t+1}^*$ *in probability as* $N \to \infty$. *Define the vector-valued functions*

$$\begin{aligned}
f_0(x_0) &:= -A_0^{-1} \left(\left[\log w_0^{\psi}(x_0) + \log K_1^{\psi}(e^{-\Phi_1^T \beta_1^*})(x_0) \right] \Phi_0(x_0) + \left(\Phi_0 \Phi_0^T \right)(x_0) \beta_0^* \right), \\
f_t(x_{t-1}, x_t) &:= -A_t^{-1} \left[\log w_t^{\psi}(x_{t-1}, x_t) + \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T \beta_{t+1}^*})(x_t) \right] \Phi_t(x_{t-1}, x_t) \\
&\quad - A_t^{-1} \left(\Phi_t \Phi_t^T \right)(x_{t-1}, x_t) \beta_t^*, \quad t = 1, \ldots, T-1, \\
f_T(x_{T-1}, x_T) &:= -A_T^{-1} \left(\log w_T^{\psi}(x_{T-1}, x_T) \Phi_T(x_{T-1}, x_T) + \left(\Phi_T \Phi_T^T \right)(x_{T-1}, x_T) \beta_T^* \right),
\end{aligned} \tag{4.75}$$

and write $f := (f_0, \ldots, f_T)^T$ *as the vector of functions. If we additionally assume that* $\mathbb{E}_{Q^\psi} |f_0(X_0)|^2 < \infty$, $\mathbb{E}_{Q^\psi} |f_t(X_{t-1}, X_t)|^2 < \infty$ *for all* $t = 1, \ldots, T$, *then*

$$\sqrt{N} \left(\beta^N - \beta^* \right) \xrightarrow{d.} \mathcal{N} \left(0_{(T+1)M}, \Sigma \right) \tag{4.76}$$

where $\Sigma = U \Sigma_f U^T$ *is given in terms of* $\Sigma_f := \mathbb{E}_{Q^\psi} \left[\left(f f^T \right)(X_{0:T}) \right]$ *and a block upper triangular matrix* $U \in \mathbb{R}^{(T+1)M \times (T+1)M}$ *described by its blocks of size* $M \times M$

$$U_{ij} := \begin{cases} 0_{M \times M}, & i > j, \\ I_M, & i = j, \\ \prod_{k=i-1}^{j-2} G_k, & i < j, \end{cases} \tag{4.77}$$

for $i, j = 0, \ldots, T$, *with*

$$\begin{aligned}
G_0 &:= -A_0^{-1} \mathbb{E}_{Q^\psi} \left[\Phi_0(X_0) \nabla_\beta \log K_1^{\psi}(e^{-\Phi_1^T \beta_1^*})(X_0)^T \right], \\
G_t &:= -A_t^{-1} \mathbb{E}_{Q^\psi} \left[\Phi_t(X_{t-1}, X_t) \nabla_\beta \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T \beta_{t+1}^*})(X_t)^T \right], \quad t = 1, \ldots, T-1.
\end{aligned} \tag{4.78}$$

Proof. Most arguments in the proof of Theorem 4.16 apply here with necessary changes to deal with intractability of the function $(\beta, x) \mapsto \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T \beta})(x)$

which we outline. For time $t = T$, the same arguments yield (4.65), consistency and a CLT for the estimator β_T^N as $N \to \infty$.

For the inductive step at time $t = 1, \ldots, T-1$ (and similarly for $t = 0$), like before we decompose $b_t^N = c_t^N + d_t^N$ with

$$c_t^N := -\frac{1}{N}\sum_{n=1}^{N}\left(\log w_t^{\psi}(X_{t-1}^n, X_t^n) + \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T\beta_{t+1}^*})(X_t^n)\right)\Phi_t(X_{t-1}^n, X_t^n), \tag{4.79}$$

$$d_t^N := \frac{1}{N}\sum_{n=1}^{N}\left(\log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T\beta_{t+1}^*})(X_t^n) - \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T\beta_{t+1}^N})(X_t^n)\right)\Phi_t(X_{t-1}^n, X_t^n).$$

If $\beta_{t+1}^N \to \beta_{t+1}^*$ in probability, by assumption (4.73) we have

$$|d_t^N| \leq \frac{1}{N}\sum_{n=1}^{N} C_t(X_t^n)|\Phi_t(X_{t-1}^n, X_t^n)|\rho_t(|\beta_{t+1}^N - \beta_{t+1}^*|) = o_P(1) \tag{4.80}$$

hence $\beta_t^N \to \beta_t^*$ in probability. By Taylor's theorem, we have

$$d_t^N = -\frac{1}{N}\sum_{n=1}^{N}\left\langle \nabla_\beta \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T\beta_{t+1}^*})(X_t^n), \beta_{t+1}^N - \beta_{t+1}^*\right\rangle \Phi_t(X_{t-1}^n, X_t^n) + R_t^N \tag{4.81}$$

with remainder

$$R_t^N := -\frac{1}{N}\sum_{n=1}^{N}\left(\beta_{t+1}^N - \beta_{t+1}^*\right)^T H_{t+1}(\tilde{\beta}_{t+1}^N, X_t^n)\left(\beta_{t+1}^N - \beta_{t+1}^*\right)\Phi_t(X_{t-1}^n, X_t^n) \tag{4.82}$$

for some $\tilde{\beta}_{t+1}^N$ between β_{t+1}^N and β_{t+1}^*. By squeeze theorem $\tilde{\beta}_{t+1}^N \to \beta_{t+1}^*$ in probability, so using assumption (4.74) and $\sqrt{N}\left(\beta_{t+1}^N - \beta_{t+1}^*\right) \xrightarrow{d.} \mathcal{N}\left(0_M, \Sigma_{f_{t+1}}\right)$ for some covariance matrix $\Sigma_{f_{t+1}}$ establishes that $R_t^N = O_P(N^{-1})$ and hence $d_t^N = O_P(N^{-1/2})$. The rest of the proof follow the same arguments in Theorem 4.16. □

The conclusions of this result is similar to Theorem 4.16. The assumptions (4.73) and (4.74), needed for consistency and central limit theorem respectively, can be verified when the form of $(\beta, x) \mapsto \log K_{t+1}^{\psi}(e^{-\Phi_{t+1}^T\beta})(x)$ is available. Using Newey and McFadden (1994; Lemma 2.4), a sufficient condition for (4.74) is $\mathbb{E}_{Q^\psi}\left[\sup_{\beta\in\Theta_{t+1}}\left|H_{t+1}^{i,j}(\beta, X_t)\Phi_t(X_{t-1}, X_t)\right|\right] < \infty$ for some compact set Θ_{t+1} containing β_{t+1}^*.

4.3.5 Iterated approximate dynamic programming

We now study the asymptotic behaviour of twisting functions $\hat{\psi}(i) \in \Psi(Q)$ generated by Algorithm 5, when either ADP (4.29) or (4.30) is employed in step 2(b)i, as the number of iterations $i \to \infty$. We will work on the closure of $\Psi(Q)$

$$\overline{\Psi}(Q) := \Big\{ \psi_0 : \mathcal{X} \to [0,\infty), \psi_t : \mathcal{X} \times \mathcal{X} \to [0,\infty)\, t = 1,\dots,T : \psi_0 \in L^1(Q_0), \\ \psi_t \in L^1(Q_{t-1,t}) \Big\}, \tag{4.83}$$

which can be seen as a closed subset of the complete separable metric space $L^1(Q_0)\prod_{t=1}^T L^1(Q_{t-1,t})$ (Cohn 2013; Theorem 3.4.1, Proposition 3.4.5) with the metric

$$\rho(\varphi,\xi) := \|\varphi_0 - \xi_0\|_{L^1(Q_0)} + \sum_{t=1}^T \|\varphi_t - \xi_t\|_{L^1(Q_{t-1,t})} \tag{4.84}$$

for $\varphi = \{\varphi_t\}_{t=0}^T, \xi = \{\xi_t\}_{t=0}^T$, and hence also a complete separable metric space with the inherited metric ρ.

We assume that the choice of function classes $\{\mathsf{F}_t\}_{t=0}^T$ are such that if $\psi \in \overline{\Psi}(Q)$, then $\hat{\phi}$ estimated from ADP algorithm (4.29) or (4.30) satisfies $\psi \cdot \hat{\phi} \in \overline{\Psi}(Q)$. For example, this is satisfied when we select $\mathsf{F}_t \subset L^\infty(\mathcal{X}^2)$ in (4.29) and $\mathsf{F}_t \subset D$ in (4.30). We begin by writing the iterative scheme as a mapping. Let $F : \mathcal{U} \times \overline{\Psi}(Q) \to \overline{\Psi}(Q)$ be defined as $F_U(\psi) := \psi \cdot \hat{\phi}$ where $\hat{\phi}$ is the output of (4.29) or (4.30) and $U \in \mathcal{U}$ encodes all uniform random variables needed for the simulation of $N \in \mathbb{N}$ iid samples from Q^ψ to perform approximate projections. As the uniform variables used at every iteration are iid, F is an *iterated random function* which defines a Markov chain $\left\{\hat{\psi}(i)\right\}_{i=0}^\infty$ on $\overline{\Psi}(Q)$. Under appropriate regularity conditions on F, we can establish existence of a unique invariant distribution and convergence to the latter at a geometric rate.

Theorem 4.18. *(Diaconis and Freedman 1999; Theorem 5.1) Let $\mu \in \mathcal{P}(\mathcal{U})$ denote the uniform distribution on $\mathcal{U}$ and suppose that there exists a function $C : \mathcal{U} \to [0,\infty)$ such that*

$$\rho(F_U(\varphi), F_U(\xi)) \le C(U)\rho(\varphi,\xi) \tag{4.85}$$

for all $\varphi, \xi \in \overline{\Psi}(Q)$, $\int_{\mathcal{U}} C(u)\mu(\mathrm{d}u) < \infty$, $\int_{\mathcal{U}} \rho(F_u(\varphi_0), \varphi_0)\mu(\mathrm{d}u) < \infty$ *for some* $\varphi_0 \in \overline{\Psi}(Q)$ *and* $\int_{\mathcal{U}} \log C(u)\mu(\mathrm{d}u) < 0$. *Then the Markov chain* $\left\{\hat{\psi}(i)\right\}_{i=0}^{\infty}$ *generated by Algorithm 5, when either ADP (4.29) or (4.30) is employed, admits a unique invariant distribution* $\eta \in \mathcal{P}\left(\overline{\Psi}(Q)\right)$. *Moreover, if* $G^i(\varphi, \cdot)$ *denotes the law of the Markov chain after* $i \in \mathbb{N}$ *iterations from* $\varphi \in \overline{\Psi}(Q)$, *then* $\varrho\left(G^i(\varphi, \cdot), \eta\right) \leq D(\varphi) r^i$ *for some* $D : \overline{\Psi}(Q) \to \mathbb{R}_+$ *and* $r \in (0, 1)$, *where* ϱ *is the Prokhorov metric on* $\mathcal{P}\left(\overline{\Psi}(Q)\right)$ *induced by the metric* ρ.

Equation (4.85) is an assumption on the regularity of the ADP procedure: i.e. for two sequences of twisting functions $\varphi \in \overline{\Psi}(Q)$ and $\xi \in \overline{\Psi}(Q)$ that are close, given the same uniform random variables U, the approximated twisting functions $\hat{\varphi} \in \Psi(Q^{\varphi})$ and $\hat{\xi} \in \Psi(Q^{\xi})$ should also be sufficiently close to keep the Lipschitz constant $C(U)$ small.

4.3.6 Distance from target distribution

We are now interested in characterizing how far the terminal distribution of a twisted SMC sampler is from the target distribution π. For a real-valued function $\varphi : \Omega \to \mathbb{R}$, we define its oscillation as $\mathrm{osc}(\varphi) := \sup_{(x,y)\in\Omega\times\Omega} |\varphi(x) - \varphi(y)|$. We will use the following sensitivity result.

Lemma 4.19. *(Del Moral 2004; Lemma 6.2.1) Let* μ *be a probability measure on a measurable space* $(\Omega, \mathcal{F})$ *and* $\varphi, \xi : \Omega \to \mathbb{R}_+$ *be two measurable functions in* $L^1(\mu)$. *The probability measures* $\mu_{\varphi}(\mathrm{d}x) := \mu(\mathrm{d}x)\varphi(x)/\mu(\varphi)$ *and* $\mu_{\xi}(\mathrm{d}x) := \mu(\mathrm{d}x)\xi(x)/\mu(\xi)$ *satisfy*

$$\|\mu_{\varphi} - \mu_{\xi}\|_{\mathrm{TV}} \leq \frac{1}{2}\mathrm{osc}\left(\log\varphi - \log\xi\right). \tag{4.86}$$

Proposition 4.20. *Let* $\hat{\psi} = \left\{\hat{\psi}_t = e^{-\hat{V}_t}\right\}_{t=0}^{T} \in \Psi(Q)$ *denote a sequence of twisting functions and* $\hat{\pi}_T := \pi_0^{\hat{\psi}} K_1^{\hat{\psi}} \cdots K_T^{\hat{\psi}}$ *the terminal distribution of the corresponding twisted SMC sampler. We have*

$$\|\hat{\pi}_T - \pi\|_{\mathrm{TV}} \leq \sum_{t=0}^{T} \|\hat{V}_t - V_t^*\|_{\infty} \tag{4.87}$$

where $\psi^* = \left\{\psi_t^* = e^{-V_t^*}\right\}_{t=0}^T$ *denotes the optimal sequence of twisting functions with respect to* Q*.*

Proof. We write the marginal distribution at time $t = 1, \ldots, T$ of the twisted and optimally controlled SMC samplers as $\hat{\pi}_t := \pi_0^{\hat{\psi}} K_1^{\hat{\psi}} \cdots K_t^{\hat{\psi}}$ and $\pi_t^* := \pi_0^{\psi^*} K_1^{\psi^*} \cdots K_t^{\psi^*}$ respectively. Using triangle inequality and properties of Markov operators, we have

$$\begin{aligned} e_t &:= \|\hat{\pi}_t - \pi_t^*\|_{\mathrm{TV}} \\ &= \|\hat{\pi}_{t-1} K_t^{\hat{\psi}} - \pi_{t-1}^* K_t^{\psi^*}\|_{\mathrm{TV}} \\ &\leq \|\hat{\pi}_{t-1} K_t^{\hat{\psi}} - \pi_{t-1}^* K_t^{\hat{\psi}}\|_{\mathrm{TV}} + \|\pi_{t-1}^* K_t^{\hat{\psi}} - \pi_{t-1}^* K_t^{\psi^*}\|_{\mathrm{TV}} \\ &\leq e_{t-1} + \int_{\mathcal{X}} \|K_t^{\hat{\psi}}(x_{t-1}, \cdot) - K_t^{\psi^*}(x_{t-1}, \cdot)\|_{\mathrm{TV}} \, \pi_{t-1}^*(\mathrm{d}x_{t-1}) \end{aligned} \tag{4.88}$$

for $t = 1, \ldots, T$. Applying Lemma 4.19 gives

$$\|K_t^{\hat{\psi}}(x_{t-1}, \cdot) - K_t^{\psi^*}(x_{t-1}, \cdot)\|_{\mathrm{TV}} \leq \frac{1}{2} \mathrm{osc}\left(\hat{V}_t(x_{t-1}, \cdot) - V_t^*(x_{t-1}, \cdot)\right) \leq \|\hat{V}_t - V_t^*\|_\infty \tag{4.89}$$

for all $x_{t-1} \in \mathcal{X}, t = 1, \ldots, T$ and similarly $e_0 := \|\pi_0^{\hat{\psi}} - \pi_0^{\psi^*}\|_{\mathrm{TV}} \leq \|\hat{V}_0 - V_0^*\|_\infty$. The proof is complete by induction and noting that $\pi_T^* = \pi$ by construction. □

4.4 Examples

4.4.1 Linear quadratic Gaussian control

We consider a particular setup of Example 3.7 where $\pi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $L(x; y) = \exp\left(-\frac{1}{2}(y - x)^T R^{-1}(y - x)\right)$ for some symmetric positive definite $R \in \mathbb{R}^{d \times d}$ and observation $y \in \mathbb{R}^d$. For this model, given a tempering schedule $\{\lambda_t\}_{t=0}^T$, we have $\pi_t = \mathcal{N}(\mu_t, \Sigma_t)$ for $t = 0, \ldots, T$ with $\Sigma_t^{-1} = \Sigma_0^{-1} + \lambda_t R^{-1}$ and $\mu_t = \Sigma_t\left(\Sigma_0^{-1}\mu_0 + \lambda_t R^{-1} y\right)$. Although conjugacy renders this model trivial, the analytical tractability available here is valuable as it provides insight and allows us to draw connections to concepts from the *linear quadratic Gaussian* (LQG) control literature.

4.4.1.1 Riccati equation

We now show that the backward recursion (4.16) can be performed exactly thus giving analytic expressions of the optimal sequence of value and twisting functions with respect to Q. We first note that under the choice of transition kernels in Section 4.2.1, the incremental weights (1.42) have the form

$$-\log w_t(x_{t-1}, x_t) = x_t^T \tilde{A}_t x_t + x_t^T \tilde{b}_t + \tilde{c}_t + x_{t-1}^T \tilde{D}_t x_{t-1} + x_{t-1}^T \tilde{e}_t \tag{4.90}$$

for $t = 1, \ldots, T$, where

$$\begin{aligned}
\tilde{A}_t &:= \frac{\Delta_t}{8}\Sigma_t^{-2}, \\
\tilde{b}_t &:= -\frac{\Delta_t}{4}\Sigma_t^{-2}\mu_t, \\
\tilde{c}_t &:= \frac{(\lambda_t - \lambda_{t-1})}{2} y^T R^{-1} y, \\
\tilde{D}_t &:= -\frac{\Delta_t}{8}\Sigma_t^{-2} + \frac{(\lambda_t - \lambda_{t-1})}{2} R^{-1}, \\
\tilde{e}_t &:= -(\lambda_t - \lambda_{t-1}) R^{-1} y + \frac{\Delta_t}{4}\Sigma_t^{-2}\mu_t.
\end{aligned} \tag{4.91}$$

For sufficiently small step size, observe that dropping $O(\Delta_t)$ terms in (4.90) gives $\log w_t(x_{t-1}, x_t) \approx (\lambda_t - \lambda_{t-1}) \log L(x_{t-1})$ which, as expected, recovers the AIS incremental weight (1.55). For notational convenience, we set $\left\{\tilde{A}_0, \tilde{b}_0, \tilde{c}_0, \tilde{D}_0, \tilde{e}_0\right\}$ as the zero matrix or vector of the appropriate size and write the forward Euler move as $f_t(x_{t-1}) = x_{t-1} + \frac{\Delta_t}{2}\nabla \log \pi_t(x_{t-1}) = G_t x_{t-1} + h_t$, with $G_t := I_d - \frac{\Delta_t}{2}\Sigma_t^{-1}$ and $h_t := \frac{\Delta_t}{2}\Sigma_t^{-1}\mu_t$.

Proposition 4.21. *The optimal sequence of value functions $V^* = \{V_t^*\}_{t=0}^T$ with respect to Q is given by*

$$\begin{aligned}
V_0^*(x_0) &= x_0^T A_0^* x_0 + x_0^T b_0^* + c_0^*, \\
V_t^*(x_{t-1}, x_t) &= x_t^T A_t^* x_t + x_t^T b_t^* + c_t^* + x_{t-1}^T D_t^* x_{t-1} + x_{t-1}^T e_t^*, \quad t = 1, \ldots, T,
\end{aligned} \tag{4.92}$$

where the coefficients $\{A_t^*, b_t^*, c_t^*, D_t^*, e_t^*\}_{t=0}^T$ *are determined by the backward recursion*

$$\begin{aligned}
A_t^* &= \tilde{A}_t + \frac{1}{2\Delta_{t+1}} G_{t+1}\left(I_d - \Theta_{t+1}^*\right) G_{t+1} + D_{t+1}^*, \\
b_t^* &= \tilde{b}_t + G_{t+1}\Theta_{t+1}^* b_{t+1}^* + e_{t+1}^* + \frac{1}{2} G_{t+1}\left(I_d - \Theta_{t+1}^*\right)\Sigma_{t+1}^{-1}\mu_{t+1}, \\
c_t^* &= \tilde{c}_t + c_{t+1}^* - \frac{1}{2}\log\det\left(\Theta_{t+1}^*\right) + \frac{1}{2\Delta_{t+1}} h_{t+1}^T h_{t+1} \\
&\quad - \frac{1}{2\Delta_{t+1}}\left(h_{t+1} - \Delta_{t+1} b_{t+1}^*\right)^T \Theta_{t+1}^*\left(h_{t+1} - \Delta_{t+1} b_{t+1}^*\right), \\
D_t^* &= \tilde{D}_t, \\
e_t^* &= \tilde{e}_t,
\end{aligned} \tag{4.93}$$

for $t = 0, \ldots, T-1$, *with* $\Theta_t^* := (I_d + 2\Delta_t A_t^*)^{-1}$ *and initialization at* $\{A_T^*, b_T^*, c_T^*, D_T^*, e_T^*\}$ $= \left\{\tilde{A}_T, \tilde{b}_T, \tilde{c}_T, \tilde{D}_T, \tilde{e}_T\right\}$. *The corresponding optimal sequence of twisting functions with respect to* Q *is given by* $\psi_t^* = e^{-V_t^*}, t = 0, \ldots, T$.

Proof. We proceed by induction. Clearly, (4.92) holds for $t = T$ since $V_T^* = -\log w_T$. Suppose that V_{t+1}^* has the form (4.92) with coefficients $\left\{A_{t+1}^*, b_{t+1}^*, c_{t+1}^*, D_{t+1}^*, e_{t+1}^*\right\}$. The recursion (4.16) has the form

$$V_t^*(x_{t-1}, x_t) = -\log w_t(x_{t-1}, x_t) - \log K_{t+1}(\psi_{t+1}^*)(x_t), \tag{4.94}$$

with $\psi_{t+1}^* = e^{-V_{t+1}^*}$. Using (4.43) and some manipulations, we obtain

$$\begin{aligned}
-\log K_{t+1}(\psi_{t+1}^*)(x_t) &= x_t^T\left(\frac{1}{2\Delta_{t+1}} G_{t+1}(I_d - \Theta_{t+1}^*) G_{t+1} + D_{t+1}^*\right) x_t \\
&+ x_t^T\left(G_{t+1}\Theta_{t+1}^* b_{t+1}^* + e_{t+1}^* + \frac{1}{2} G_{t+1}\left(I_d - \Theta_{t+1}^*\right)\Sigma_{t+1}^{-1}\mu_{t+1}\right) + c_{t+1}^* - \frac{1}{2}\log\det\left(\Theta_{t+1}^*\right) \\
&+ \frac{1}{2\Delta_{t+1}} h_{t+1}^T h_{t+1} - \frac{1}{2\Delta_{t+1}}\left(h_{t+1} - \Delta_{t+1} b_{t+1}^*\right)^T \Theta_{t+1}^*\left(h_{t+1} - \Delta_{t+1} b_{t+1}^*\right).
\end{aligned} \tag{4.95}$$

Adding this to (4.90) shows that V_t^* has the desired form (4.92) and equating coefficients of the polynomial gives (4.93). □

The backward recursion (4.93) is analogous to the *Riccati equation* in the context of LQG control. To illustrate the behaviour of these coefficients, we shall henceforth set $\mu_0 = 0_d, \Sigma_0 = I_d, y = (\xi, \ldots, \xi)^T$ for some $\xi \in \mathbb{R}$ and $R_{ij} = \delta_{ij} + (1 - \delta_{ij})\rho$ for $i, j = 1, \ldots, d$ and some $\rho \in [-1, 1]$ (here δ_{ij} denotes the Kronecker delta).

In the top row of Figure 4.1, we plot the time evolution of these coefficients for $d = 2, \xi = 4, \rho = 0.8$.

Noting that the optimal cost of the Kullback-Leibler control problem (4.23) is $v^* = -\log Z = -\log \pi_0(\psi_0^*)$, as the top right plot shows that the constant c_0^* has a dominant contribution to v^*, this suggests that it is important to estimate the constants in (4.34)-(4.35) to learn good twisting functions. Moving from the bottom left to top left plot, we observe that increasing the location parameter ξ increases the magnitude of $\{b_t^*, e_t^*\}_{t=0}^T$ but keeps $\{A_t^*, D_t^*\}_{t=0}^T$ unchanged. This is evident from the expressions of $\{D_t^*, e_t^*\}_{t=0}^T$. For the coefficients $\{b_t\}_{t=0}^T$, it is intuitive that its magnitude should also increase since, for the optimally controlled sampler to have the desired terminal distribution, one would have to initialize (4.40) closer to the target distribution and take larger drifts (4.42) in the appropriate direction. As the parameter ξ does not alter the 'structure' of the problem, it is also unsurprising that changing ξ has no effect on $\{A_t^*\}_{t=0}^T$. Comparing the plots on the bottom row, we see that the off-diagonal elements of $\{A_t^*, D_t^*\}_{t=0}^T$ vanish under independence. This indicates poor performance for very correlated target distributions if we do not take these terms into account.

Having obtained the optimal sequence of twisting functions with respect to Q in a backward sweep, we may then simulate the optimally controlled SMC sampler in a forward pass. In Figure 4.2, we contrast the output of the uncontrolled SMC sampler with that of the optimally controlled.

4.4.1.2 Implementation details

We now discuss some implementation issues. Here we choose constant functions in the parameterization (4.35), i.e. $A_t(x_{t-1}) = A_t \in \mathbb{R}^{d \times d}$, $b_t(x_{t-1}) = b_t \in \mathbb{R}^d$, and select $d_t(x_{t-1})$ as discussed in (4.44) (note that for this example, the additive constant needed is $\kappa_t = \Sigma_0^{-1}\mu_0 + \lambda_t R^{-1} y$). For each $t = 0, \ldots, T$, we denote the vector of unknown coefficients as $\beta_t := (A_t, b_t, c_t) \in \mathbb{R}^p$ and note that this has dimension $p = \frac{1}{2}d^2 + \frac{3}{2}d + 1$. As we can compute the optimal sequence of value functions with respect to Q in this setting, this allows us to evaluate the

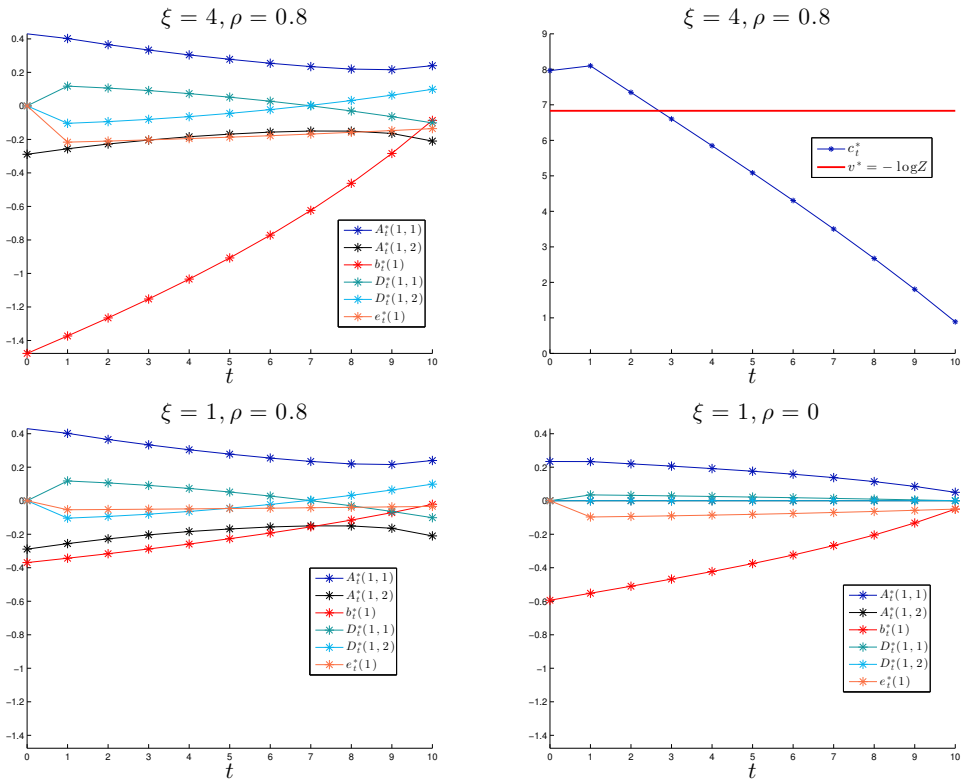

Figure 4.1: Coefficients of the optimal sequence of value functions with respect Q in LQG control under various problem settings. The configuration of the uncontrolled SMC sampler is $T = 10, \Delta_t = 0.1, \lambda_t = t/T$. Note that all except the top right plot share the same axes.

effectiveness of ADP routine (4.30) and the iterative behaviour of Algorithm 5 under *correct parameterization*.

Instead of employing a numerical optimization routine to perform the minimization steps in Algorithm 4, we observe that each of these linear least squares problems is solved by

$$\beta_t^N = \left(\mathbf{X}_t^T \mathbf{X}_t\right)^{-1} \mathbf{X}_t^T \left(\bar{\mathbf{V}}_t - \mathbf{d}_t\right), \tag{4.96}$$

where $\mathbf{X}_t \in \mathbb{R}^{N \times p}$ denotes the design matrix associated to the quadratic model, $\bar{\mathbf{V}}_t = \left(\bar{V}_t(X_{t-1}^1, X_t^1), \ldots, \bar{V}_t(X_{t-1}^N, X_t^N)\right)^T$ and $\mathbf{d}_t = \left(d_t(X_{t-1}^1), \ldots, d_t(X_{t-1}^N)\right)^T$. Although this procedure does not enforce the positive definite constraints (4.38)-(4.39), in our numerical implementations, we find that these constraints are satisfied when the step sizes $\{\Delta_t\}_{t=1}^T$ are sufficiently small. Furthermore, one could also project onto the space of symmetric positive definite matrices, using for example the result

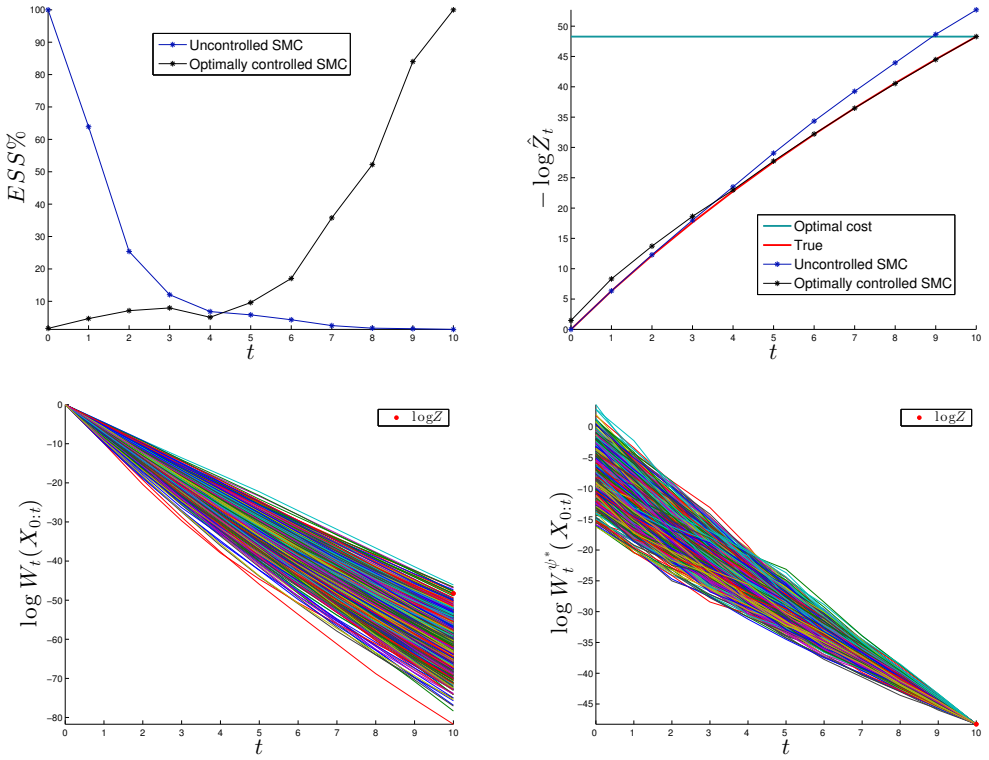

Figure 4.2: Comparison of uncontrolled and optimally controlled SMC samplers in terms of effective sample size (*top left*), normalizing constant estimation (*top right*) and variance of weights (*bottom row*). The problem setting here is $d = 4, \xi = 10, \rho = 0.8$ and the uncontrolled SMC sampler's configuration is $N = 100, T = 10, \Delta_t = 0.1, \lambda_t = t/T$.

by Higham (1988) for Frobenius norm. Although the computational complexity of (4.96) is $O(N)$, it scales quite costly in d as this requires inversion of a $p \times p$ matrix. For problems with large d, it might be worth considering the use of iterative linear solvers which offer reduced complexity. While it is tempting to also reduce complexity by restricting A_t to a diagonal matrix, we find that this approach performs poorly when correlation in the target distribution is significant.

Figure 4.3 shows the coefficients estimated by the controlled SMC sampler. It is striking that with just $N = 100$ particles, we are able to accurately estimate, in a single iteration, the true coefficients in dimension $d = 12$ (here $p = 91$). That said, we typically need to increase N with d to prevent the Gram matrix $\mathbf{X}_t^T \mathbf{X}_t$ from being ill-conditioned. Moreover, we see from Figure 4.4 that, for this particular example, it is unnecessary to iterate the algorithm as the twisting functions converge immediately to an invariant distribution that is very concentrated around the optimal.

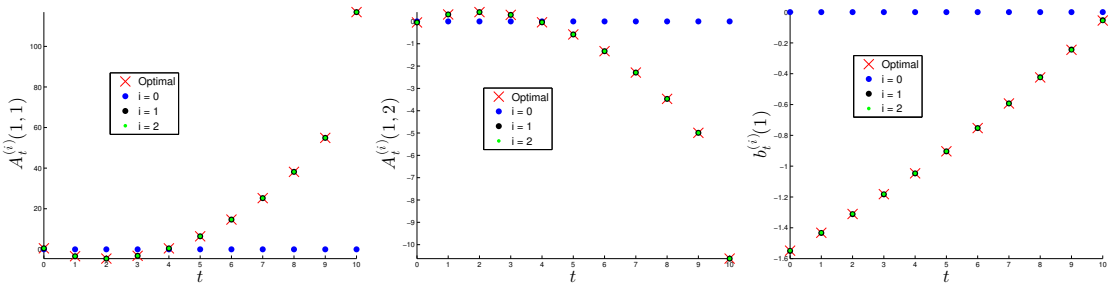

Figure 4.3: Comparing coefficients of value functions estimated by the controlled SMC sampler against true coefficients. The problem setting here is $d = 12, \xi = 24, \rho = 0.99$ and the uncontrolled SMC sampler's configuration is $I = 2, N = 100, T = 10, \Delta_t = 0.1, \lambda_t = t/T$.

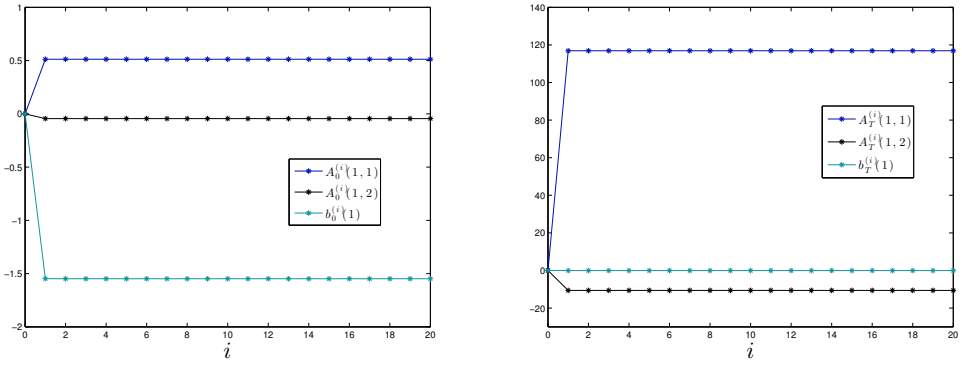

Figure 4.4: Coefficients of value functions estimated by the controlled SMC sampler over iterations. The problem setting here is $d = 12, \xi = 24, \rho = 0.99$ and the uncontrolled SMC sampler's configuration is $I = 20, N = 100, T = 10, \Delta_t = 0.1, \lambda_t = t/T$.

4.4.1.3 Comparison of algorithmic performance

We now compare the controlled SMC sampler against AIS as correlation parameter ρ, location parameter ξ and dimension d vary one at a time (with default fixed at $\rho = 0.5, \xi = 10, d = 4$). We consider AIS in three different regimes: we increase either the number of MALA iterations M (abbreviated as `AIS-Iterations`), the number of particles N (`AIS-Particles`) or the number of time steps T (`AIS-Steps`), when comparisons are done at a fixed computational cost, measured in terms of run time. Otherwise, these tuning parameters are kept at $M = 1, N = 1000, T = 50$. As we found that a single iteration and time step is adequate for the controlled SMC sampler, N was increased to match the computational cost of other algorithms. For all algorithms, we used a linear tempering schedule, i.e. $\lambda_t = t/T$. The step sizes of

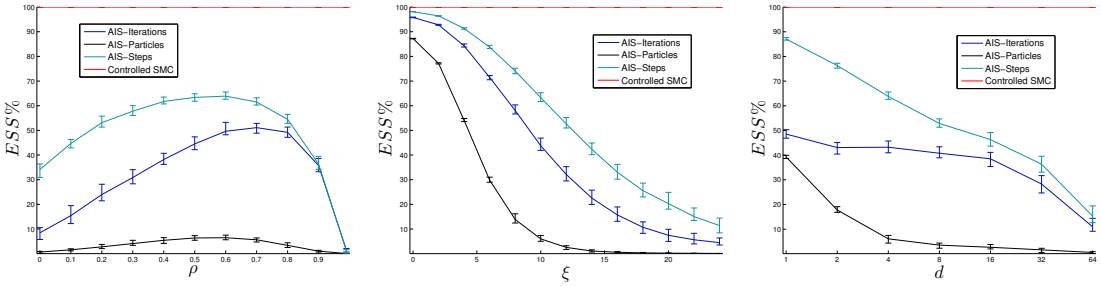

Figure 4.5: Comparison of algorithms in terms of ESS as correlation parameter ρ (*left*), location parameter ξ (*middle*) and dimension d (*right*) vary one at a time. Lines and error bars indicate median and interquartile range of 100 repetitions respectively. Note that for ease of illustration, we plotted the percentage instead of the actual ESS obtained.

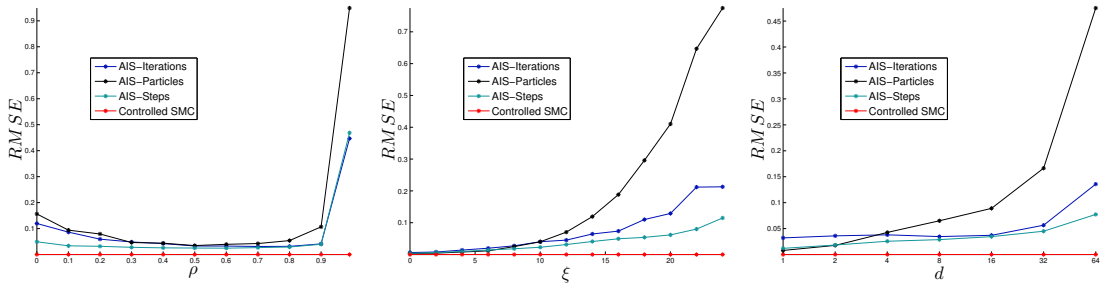

Figure 4.6: Comparison of algorithms in terms of RMSE in the estimation of $\log Z$ as correlation parameter ρ (*left*), location parameter ξ (*middle*) and dimension d (*right*) vary one at a time.

MALA moves were tuned to achieve suitable acceptance probabilities while that of the controlled SMC sampler were kept at 1. The results are summarized in Figure 4.5 and 4.6, which plots ESS and RMSE in the estimation of $\log Z$, respectively, for 100 independent repetitions of each algorithm.

4.4.2 Bayesian logistic regression

Consider a binary regression problem: each observation $y_i \in \{0, 1\}, i = 1, \ldots, n$ is modelled as independent Bernoulli random variable with probability of success $\eta(x^T \mathbf{X}_i)$, where $x \in \mathbb{R}^d$ denotes the unknown regression coefficients and $\mathbf{X}_i \in \mathbb{R}^d$ the i^{th} row of a model matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. The logistic regression model corresponds to taking the logistic function as link function, i.e. $\eta(u) = (1 + e^{-u})^{-1}$ for $u \in \mathbb{R}$,

hence the likelihood function is given by

$$L(x;y) = \exp\left(y^T \mathbf{X} x - \sum_{i=1}^{n} \log\left(1 + e^{x^T \mathbf{X}_i}\right)\right) \tag{4.97}$$

where $y = (y_1, \ldots, y_n)^T \in \{0,1\}^n$ denotes the data set. Following Hanson et al. (2014), we prescribe a Gaussian prior distribution $\pi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ of the form $\mu_0 = 0_d, \Sigma_0 = \frac{\pi^2 n}{3d}(\mathbf{X}^T\mathbf{X})^{-1}$. It is easy to check that

$$\nabla \log L(x;y) = \mathbf{X}^T y - \sum_{i=1}^{n} \frac{\mathbf{X}_i}{1 + e^{-x^T \mathbf{X}_i}}, \tag{4.98}$$

so the additive constant required in (4.44) is simply $\kappa_t = \lambda_t \mathbf{X}^T y$.

4.4.2.1 Implementation details

For this example, we used the same function classes $\{\mathsf{F}_t\}_{t=0}^T$ described in Section 4.4.1.2 and carried out the learning procedure in the same manner. The rationale for choosing the same parameterization is based on the following informal argument: as this choice results in twisting functions (4.36)-(4.37) that are log-concave and the distributions $\{\pi_t\}_{t=0}^T$ have log-concave densities, in view of Proposition 4.9, we expect the optimal sequence of value functions that we have to approximate at each iteration to be convex, and adequately approximated by a quadratic model. Moreover, in contrast to the Gaussian example where this choice provided the correct parameterization, this additional layer of approximation also explains why we observe that the resulting algorithm now exhibits sensitivity to the number of iterations and the choice of step size (see Figure 4.7), which we tuned using pilot runs.

In Figure 4.8, we plot the coefficients of twisting functions generated by the controlled SMC sampler over $I = 10,000$ iterations with a burn-in of 100. These plots reveal that the invariant distribution of these coefficients is likely to be Gaussian and consequently that at stationarity, the corresponding value functions are Gaussian processes.

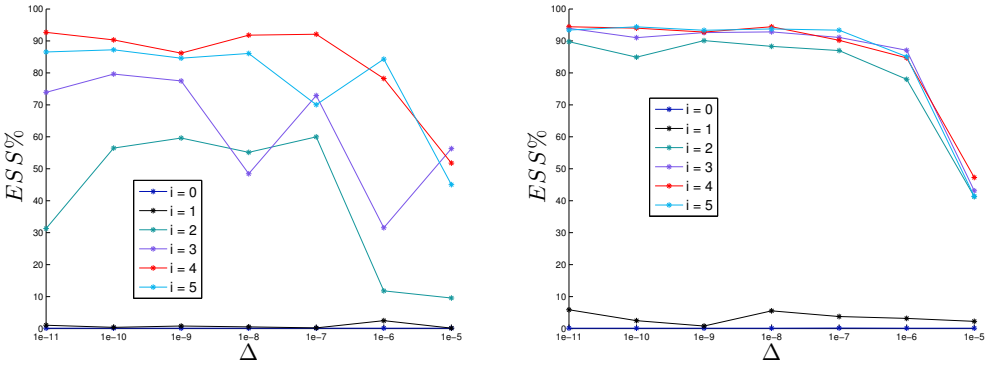

Figure 4.7: Illustrating sensitivity of ESS to the number of iterations and step size taken by the controlled SMC sampler on the Australian credit (*left*) and German credit (*right*) data sets.

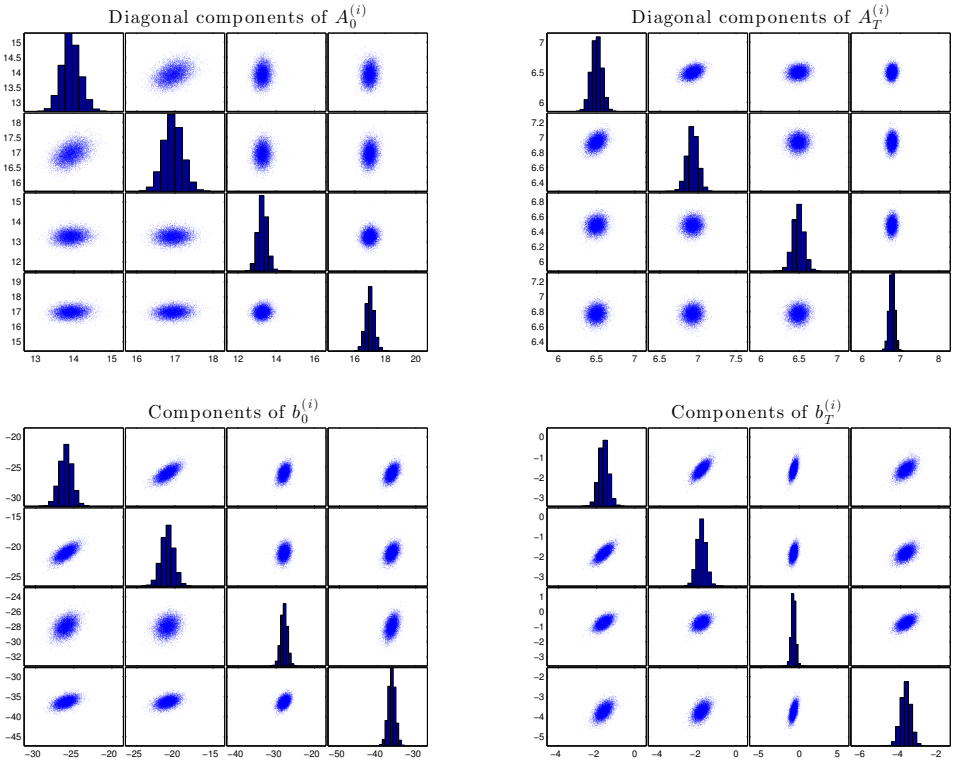

Figure 4.8: Some coefficients generated by the controlled SMC sampler over iterations when employed on the Heart disease data set.

4.4.2.2 Comparison of algorithmic performance

We now perform a comparison of algorithms on the analysis of three real data sets[2] with different characteristics. Like before, we first tune the controlled SMC sampler and compare it against AIS (in various regimes) at a fixed computing time. The configurations used for each sampler are given in Table 4.1. The results summarized in Table 4.2 show that the controlled SMC sampler obtained at least three order of magnitude gains over AIS and across all data sets. We also note that the comparison done here is conservative in the sense that we only included the output from the final iteration of Algorithm 5.

			Data set		
			Heart disease ($n = 270, d = 14$)	*Australian credit* ($n = 690, d = 15$)	*German credit* ($n = 1000, d = 25$)
Algorithm	`AIS-Iterations`	M	2	2	2
		N	1000	1000	1000
		T	20	20	20
		Δ	0.05	0.0275	0.01
	`AIS-Particles`	M	1	1	1
		N	1200	2000	1500
		T	20	20	20
		Δ	0.05	0.0275	0.01
	`AIS-Steps`	M	1	1	1
		N	1000	1000	1000
		T	25	40	35
		Δ	0.05	0.025	0.01
	`Controlled SMC`	I	3	4	4
		N	6500	10000	8000
		T	1	1	1
		Δ	10^{-10}	10^{-11}	10^{-11}

Table 4.1: Configurations of each sampler for each data set. Notationally, N refers to the number of particles, T the time steps and Δ the constant step size used in each sampler. For AIS algorithms, M denotes the number of MALA moves used at each time iteration and I the number of iterations taken by the controlled SMC sampler.

[2]Data sets were downloaded from the UCI Machine Learning Repository and standardized before analysis.

			Data set		
			Heart disease ($n = 270, d = 14$)	*Australian credit* ($n = 690, d = 15$)	*German credit* ($n = 1000, d = 25$)
Algorithm	`AIS-Iterations`	ESS (IQR)	5.5688 (6.5680)	1.7742 (1.5275)	1.3801 (1.1362)
		ESS%	0.56%	0.18%	0.14%
		$\log \hat{Z} \pm$SD	-118.4378 ± 0.8660	-254.9354 ± 2.3788	-532.4756 ± 3.7082
		RMSE	0.9837	4.8094	15.0063
	`AIS-Particles`	ESS (IQR)	3.2188 (3.5882)	1.5762 (1.2667)	1.0546 (0.3976)
		ESS%	0.27%	0.08%	0.07%
		$\log \hat{Z} \pm$SD	-118.9623 ± 1.4145	-257.1527 ± 2.5286	-539.9356 ± 4.9142
		RMSE	1.7259	6.8805	22.5424
	`AIS-Steps`	ESS (IQR)	3.5735 (3.7007)	2.0064 (1.8243)	1.2475 (0.8060)
		ESS%	0.36%	0.20%	0.12%
		$\log \hat{Z} \pm$SD	-118.5834 ± 1.1022	-254.3903 ± 2.0405	-532.8401 ± 3.7630
		RMSE	1.2598	4.1693	15.3731
	`Controlled SMC`	ESS (IQR)	6133 (72.8052)	9133 (123.0457)	7634 (62.8757)
		ESS%	94.35%	91.33%	95.43%
		$\log \hat{Z} \pm$SD	-117.9634 ± 0.0039	-250.7489 ± 0.0034	-517.9294 ± 0.0028
		RMSE	0.0039	0.0034	0.0028

Table 4.2: Summary of results obtained by each sampler in 100 independent repetitions. RMSE refers to the estimation error of $\log Z$, which we computed by taking reference to an estimate obtained using many repetitions of a SMC sampler with a large number of particles.

5 Conclusions

In this chapter, we give some concluding remarks and mention some future lines of research.

5.1 On transport methods

The use of transport theory to design more efficient Monte Carlo methods is still an early area of research that has received increased attention in recent years. As existing literature is spread across various subjects, the first objective of this thesis was to provide a review of these ideas in a coherent framework (Chapter 2 and Section 3.3).

In Chapter 3, we focused on the problem of constructing transport maps using flows. The main purpose of this work is as follows. Firstly, we provide a theoretical justification for the validity of these methods (Section 3.2) as such an understanding was lacking when this research was undertaken. As demonstrated in Example 3.8 and 3.11, these considerations are important as pathologies can occur and often manifest as numerical difficulties in practice. Future work in this direction could consider the use of Ambrosio et al. (2005; Theorem 8.2.1) to relax assumption A1 in Theorem 3.5. Secondly, we construct a solution to the flow transport problem (Section 3.3.7) and introduce the Gibbs flow as a computationally tractable approximation (Section

3.3.8). Future work could involve seeking other approximate solutions to the flow transport problem. Thirdly, we show how to implement a SMC sampler which uses approximate Gibbs flow and MCMC moves (Section 3.4). With minor changes, this can be seen as a more general framework to implement a sampler based on any approximate solution to the flow transport problem.

5.2 On optimal control methods

Exploiting the connection between smoothing for state space models and Kullback-Leibler optimal control is a subject of recent interest (Kappen and Ruiz 2016, Pieralberto et al. 2016, Ruiz and Kappen 2016). In Chapter 4, we leverage this connection to develop more efficient SMC samplers.

The methodology developed in Section 4.1 and 4.2 is novel and the resulting controlled SMC sampler can be thought of as a type of adaptive importance sampler that is trained using reinforcement learning. The performance of such control methods depends on the choice of function classes used to approximate the optimal sequence of twisting or value functions of the control problem. In Section 4.4, we have only implemented constant functions in the parameterization (4.35); future work will experiment with the full flexibility of this function class and the use of ADP algorithm (4.29) for twisting functions.

In contrast to the above references, we provide a theoretical analysis of our proposed methodology in Section 4.3. Most of these results can be applied or easily adapted to study the algorithm proposed by Pieralberto et al. (2016). Future work in this direction could involve relaxing the assumptions in Proposition 4.11 and 4.15, extending the limit theorems in Section 4.3.4 to the non-linear least squares setting and characterizing the invariant distribution of Theorem 4.18 in the regime where the number of samples is large. The latter does not follow straightforwardly from Theorem 4.16 and 4.17 as the dependence of (4.61) and (4.76) on the given sequence of twisting functions is complex. Lastly, we also hypothesize that the framework presented in Section 4.3.5 could be of broader interest in the ADP community to analyze value function iteration algorithms.

Appendices

A

Jacobian of Gibbs velocity field

Contents

A.1 Expression of Jacobian

Consider the form of the Gibbs velocity field $\tilde{f}$ given in (3.76). For notational ease, we write $\tilde{f}_i(t,x) = N_i(t,x)/\gamma_t(x)$ where N_i is the numerator and $\gamma_t(x) := \pi_0(x)L(x)^{\lambda(t)}$ is the unnormalized density at time $t \in [0,1]$. Now for $(t,x) \in [0,1] \times \mathbb{R}^d$, consider the $(i,j)^{th}$ element of the Jacobian matrix $\nabla \tilde{f}(t,x)$:

$$\begin{aligned} \partial_{x_j} \tilde{f}_i(t,x) &= \tilde{f}_i(t,x) \partial_{x_j} \log \tilde{f}_i(t,x) \\ &= \tilde{f}_i(t,x) \left(\partial_{x_j} \log N_i(t,x) - \partial_{x_j} \log \gamma_t(x) \right) \\ &= \frac{\partial_{x_j} N_i(t,x)}{\gamma_t(x)} - \tilde{f}_i(t,x) \partial_{x_j} \log \gamma_t(x). \end{aligned} \tag{A.1}$$

Note first that

$$\partial_{x_j} \log \gamma_t(x) = \partial_{x_j} \log \pi_0(x) + \lambda(t) \partial_{x_j} \log L(x). \tag{A.2}$$

The tricky term to compute is

$$\begin{aligned}\frac{\partial_{x_j} N_i(t,x)}{\gamma_t(x)} = \lambda'(t)\Bigg\{&\partial_{x_j} F_t(x_i|x_{-i}) \int_{-\infty}^{\infty} \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\,\mathrm{d}y_i \\ &+ F_t(x_i|x_{-i})\partial_{x_j}\left(\int_{-\infty}^{\infty} \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\,\mathrm{d}y_i\right) \\ &- \partial_{x_j}\left(\int_{-\infty}^{x_i} \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\,\mathrm{d}y_i\right)\Bigg\}\Bigg/\gamma_t(x).\end{aligned} \tag{A.3}$$

For diagonal entries, i.e. $i = j$, this is

$$\frac{\partial_{x_i} N_i(t,x)}{\gamma_t(x)} = \lambda'(t)\left\{\frac{\int_{-\infty}^{\infty} \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\,\mathrm{d}y_i}{\int_{-\infty}^{\infty} \gamma_t(y_i, x_{-i})\,\mathrm{d}y_i} - \log L(x)\right\}. \tag{A.4}$$

The terms needed in off-diagonal entries are

$$\begin{aligned}\partial_{x_j} F_t(x_i|x_{-i}) = &\int_{-\infty}^{x_i} \partial_{x_j} \log \pi_t(y_i, x_{-i})\pi_t(y_i|x_{-i})\,\mathrm{d}y_i \\ &- F_t(x_i|x_{-i}) \int_{-\infty}^{\infty} \partial_{x_j} \log \pi_t(y_i, x_{-i})\pi_t(y_i|x_{-i})\,\mathrm{d}y_i,\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}\partial_{x_j}\left(\int \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\,\mathrm{d}y_i\right) = &\int \partial_{x_j} \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\,\mathrm{d}y_i \\ &+ \int \log L(y_i, x_{-i})\gamma_t(y_i, x_{-i})\partial_{x_j} \log \gamma_t(y_i, x_{-i})\,\mathrm{d}y_i\end{aligned} \tag{A.6}$$

with appropriate limits.

A.2 Expression for truncated Gaussians application

For notational ease, we write the Gibbs velocity field (3.101) as

$$\tilde{f}_i(t,x) = \frac{N_i(t,x)}{\pi_0(x)\int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i})\,\mathrm{d}y_i} \tag{A.7}$$

where N_i denotes the numerator. Now for $(t,x) \in [0,1] \times \text{supp}(\pi_t)$, consider the $(i,j)^{th}$ element of the Jacobian matrix $\nabla \tilde{f}(t,x)$:

$$\begin{aligned}
\partial_{x_j} \tilde{f}_i(t,x) &= \tilde{f}_i(t,x) \partial_{x_j} \log \tilde{f}_i(t,x) \qquad \text{(A.8)} \\
&= \tilde{f}_i(t,x) \left(\partial_{x_j} \log N_i(t,x) - \partial_{x_j} \log \pi_0(x) - \partial_{x_j} \log \left(\int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i \right) \right) \\
&= \frac{\partial_{x_j} N_i(t,x)}{\pi_0(x) \int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i} - \tilde{f}_i(t,x) \partial_{x_j} \log \pi_0(x) \\
&\quad - \tilde{f}_i(t,x) \partial_{x_j} \log \left(\int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i \right).
\end{aligned}$$

For diagonal entries, i.e. $i = j$, we have

$$\partial_{x_i} \tilde{f}_i(t,x) = \frac{\beta_i'(t) \pi_0(\beta_i(t), x_{-i}) - \alpha_i'(t) \pi_0(\alpha_i(t), x_{-i})}{\int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i} - \tilde{f}_i(t,x) \partial_{x_i} \log \pi_0(x). \quad \text{(A.9)}$$

The terms needed in off-diagonal entries are

$$\begin{aligned}
\partial_{x_j} N_i(t,x) &= \alpha_i'(t) \pi_0(\alpha_i(t), x_{-i}) \partial_{x_j} \log \pi_0(\alpha_i(t), x_{-i}) \int_{x_i}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i \qquad \text{(A.10)} \\
&\quad + \alpha_i'(t) \pi_0(\alpha_i(t), x_{-i}) \int_{x_i}^{\beta_i(t)} \partial_{x_j} \log \pi_0(y_i, x_{-i}) \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i \\
&\quad + \beta_i'(t) \pi_0(\beta_i(t), x_{-i}) \partial_{x_j} \log \pi_0(\beta_i(t), x_{-i}) \int_{\alpha_i(t)}^{x_i} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i \\
&\quad + \beta_i'(t) \pi_0(\beta_i(t), x_{-i}) \int_{\alpha_i(t)}^{x_i} \partial_{x_j} \log \pi_0(y_i, x_{-i}) \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i
\end{aligned}$$

and

$$\partial_{x_j} \log \left(\int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i \right) = \frac{\int_{\alpha_i(t)}^{\beta_i(t)} \partial_{x_j} \log \pi_0(y_i, x_{-i}) \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i}{\int_{\alpha_i(t)}^{\beta_i(t)} \pi_0(y_i, x_{-i}) \, \mathrm{d}y_i}. \quad \text{(A.11)}$$

Lastly, note that for multivariate Gaussian $\pi_0 = \mathcal{N}(\mu, \Sigma)$, we have

$$\nabla \log \pi_0(x) = \Sigma^{-1}(\mu - x). \quad \text{(A.12)}$$

B

Least squares approximations

Contents

B.1 Approximate projection operators

In this section, we give some results which characterize the error of approximate projection operators on an arbitrary probability space $(\Omega, \mathcal{F}, \mu)$. In the following, $\{X^n\}_{n\in\mathbb{N}}$ denotes a sequence of iid random variables with distribution μ, defined on a common underlying probability space with probability measure $\mathbb{P}$. We will write $\mathbb{E}_\mu$ and $\mathbb{E}$ to denote expectation with respect to μ and $\mathbb{P}$ respectively, and $\mu^N := N^{-1}\sum_{n=1}^N \delta_{X^n}$ to refer to a random probability measure based on $N \in \mathbb{N}$ samples.

B.1.1 Linear least squares

We first consider the linear least squares case where the function class of interest is

$$\mathsf{F} := \left\{\Phi(x)^T\beta : \beta \in \mathbb{R}^M\right\}, \tag{B.1}$$

where $\Phi := (\varphi_1, \ldots, \varphi_M)^T$ is a vector of $M \in \mathbb{N}$ pre-specified basis functions satisfying $\|\varphi_m\|_\infty \leq U$ for all $m = 1, \ldots, M$. It is easy to see that F is a linear subspace of $L^2(\mu)$ and is closed since every finite dimensional subspace of a normed vector space is closed. Let $\xi \in L^4(\mu)$ be the function to be approximated, write the squared loss function as $(\beta, x) \mapsto \ell_\beta(x) := \left(\Phi(x)^T\beta - \xi(x)\right)^2$ and risk function as $\beta \mapsto \mu(\ell_\beta)$. Assuming that the matrices

$$A_N := \frac{1}{N}\sum_{n=1}^{N}(\Phi\Phi^T)(X^n), \quad A := \mathbb{E}_\mu\left[(\Phi\Phi^T)(X)\right], \tag{B.2}$$

are invertible, we have explicit expressions of the least squares estimator and the orthogonal projection

$$\begin{aligned}\beta^N &:= \arg\min_{\beta\in\mathbb{R}^M} \|\Phi^T\beta - \xi\|^2_{L^2(\mu^N)} = A_N^{-1}b_N, \\ \beta^* &:= \arg\min_{\beta\in\mathbb{R}^M} \|\Phi^T\beta - \xi\|^2_{L^2(\mu)} = A^{-1}b,\end{aligned} \tag{B.3}$$

where $b_N := N^{-1}\sum_{n=1}^{N}(\xi\Phi)(X^n)$ and $b := \mathbb{E}_\mu[(\xi\Phi)(X)]$. To simplify notation involving central moments, for any function $\varphi \in L^1(\mu)$, we write $\bar{\varphi} := \varphi - \mu(\varphi)$ to denote the centered function.

Proposition B.1. *Assume that* $\lambda_{\min}(A) > 0$ *and let* $\Phi^T\beta^* := \mathsf{P}^\mu\xi$ *be the orthogonal projection of* ξ *unto* F. *Denote its approximation by* $\Phi^T\beta^N := \mathsf{P}^{\mu,N}\xi$ *for a given* $N \in \mathbb{N}$ *such that* $\mathbb{E}\left[\lambda_{\min}^{-4}(A_N)\right] < \infty$. *We have*

$$\mathbb{E}\left[\mu(\ell_{\beta^N})\right] \leq \mu(\ell_{\beta^*}) + \left(\varrho(N) + UM^{1/2}\lambda_{\min}^{-1}(A)\|\xi\|_{L^2(\mu)}\theta(N)\right)^2 U^2M\mathbb{E}\left[\lambda_{\min}^{-4}(A_N)\right]^{1/2} \tag{B.4}$$

where

$$\begin{aligned}\varrho^2(N) &:= \sum_{m=1}^{M}\sqrt{3\left(N^{-2}-N^{-3}\right)\mu^2\left(\left(\overline{\xi\varphi_m}\right)^2\right) + N^{-3}\mu\left(\left(\overline{\xi\varphi_m}\right)^4\right)}, \\ \theta^2(N) &:= \sum_{n,m=1}^{M}\sqrt{3\left(N^{-2}-N^{-3}\right)\mu^2\left(\left(\overline{\varphi_n\varphi_m}\right)^2\right) + N^{-3}\mu\left(\left(\overline{\varphi_n\varphi_m}\right)^4\right)}.\end{aligned} \tag{B.5}$$

Proof. Since F is linear, $\Phi^T\beta^N - \Phi^T\beta^*$ lies in F and is orthogonal to the residual $\Phi^T\beta^* - \xi$. By Pythagorean theorem, $\|\Phi^T\beta^N - \xi\|^2_{L^2(\mu)} = \|\Phi^T\beta^N - \Phi^T\beta^*\|^2_{L^2(\mu)} + \|\Phi^T\beta^* - \xi\|^2_{L^2(\mu)}$ hence

$$\mathbb{E}\left[\mu(\ell_{\beta^N})\right] = \mathbb{E}\|\Phi^T\beta^N - \Phi^T\beta^*\|^2_{L^2(\mu)} + \mu(\ell_{\beta^*}). \tag{B.6}$$

By Cauchy-Schwarz inequality,

$$\|\Phi^T\beta^N - \Phi^T\beta^*\|^2_{L^2(\mu)} = \|\Phi^T(\beta^N - \beta^*)\|^2_{L^2(\mu)} \leq U^2 M|\beta^N - \beta^*|^2. \tag{B.7}$$

By Young's inequality,

$$\begin{aligned} |\beta^N - \beta^*|^2 &= \left|A_N^{-1}(b_N - b) + (A_N^{-1} - A^{-1})b\right|^2 \\ &\leq (1+\varepsilon)\left|A_N^{-1}(b_N - b)\right|^2 + (1+\varepsilon^{-1})\left|(A_N^{-1} - A^{-1})b\right|^2 \end{aligned} \tag{B.8}$$

for any $\varepsilon > 0$. We first consider

$$\begin{aligned} \left|(A_N^{-1} - A^{-1})b\right|^2 &= \left|A_N^{-1}(A - A_N)A^{-1}b\right|^2 \\ &\leq \|A_N^{-1}\|_2^2 \, \|A - A_N\|_2^2 \, \|A^{-1}\|_2^2 \, |b|^2 \\ &\leq \lambda_{\min}^{-2}(A_N) \, \|A - A_N\|_F^2 \, \lambda_{\min}^{-2}(A) \, |b|^2 \\ &\leq \lambda_{\min}^{-2}(A_N) \, \|A - A_N\|_F^2 \, \lambda_{\min}^{-2}(A) \, U^2 M \|\xi\|^2_{L^2(\mu)}. \end{aligned} \tag{B.9}$$

In the first line, we used the identity $G^{-1} - H^{-1} = G^{-1}(H - G)H^{-1}$ for any invertible $G, H \in \mathbb{R}^{M\times M}$; the first inequality follows from sub-multiplicativity of the spectral matrix norm $\|\cdot\|_2$; the second inequality exploits its relationship to the Frobenius norm $\|\cdot\|_F$; the last line uses Jensen's inequality. Similarly, we have

$$\left|A_N^{-1}(b_N - b)\right|^2 \leq \|A_N^{-1}\|_2^2 |b_N - b|^2 = \lambda_{\min}^{-2}(A_N)|b_N - b|^2. \tag{B.10}$$

Combining (B.8), (B.9) & (B.10), taking expectations and applying Cauchy-Schwarz inequality gives

$$\begin{aligned} \mathbb{E}|\beta^N - \beta^*|^2 &\leq (1+\varepsilon)\mathbb{E}\left[\lambda_{\min}^{-4}(A_N)\right]^{1/2} \sum_{m=1}^{M} \mathbb{E}\left[\mu^N\left(\overline{\xi\varphi_m}\right)^4\right]^{1/2} \\ &+ (1+\varepsilon^{-1})\,\lambda_{\min}^{-2}(A)\,U^2 M\,\|\xi\|^2_{L^2(\mu)}\,\mathbb{E}\left[\lambda_{\min}^{-4}(A_N)\right]^{1/2} \sum_{n,m=1}^{N} \mathbb{E}\left[\mu^N\left(\overline{\varphi_n\varphi_m}\right)^4\right]^{1/2}. \end{aligned} \tag{B.11}$$

Noting that we have $\mathbb{E}\left[\mu^N(\bar{\varphi})^4\right] = 3(N^{-2} - N^{-3})\mu^2(\bar{\varphi}^2) + \mu(\bar{\varphi}^4)N^{-3}$ for any function $\varphi \in L^4(\mu)$ and minimizing over $\varepsilon > 0$ gives (B.4). □

Existence of inverse moments of $\lambda_{\min}(A_N)$ is an assumption on the rate of decay of its distribution around zero. Such a condition is necessary, otherwise

the Gram matrix A_N could be ill-conditioned and consequently the least squares estimator β^N could be unbounded.

Bounds similar to (B.4) have been studied under various settings. In Györfi et al. (2006; Theorem 11.3), ξ was assumed to be bounded and the analysis was done under truncation of the approximation $\Phi^T\beta^N$ whenever it exceeds $\|\xi\|_\infty$; Belomestny et al. (2010; Theorem 4.5) also considered a similar truncation under the condition $\lambda_{\min}(A_N) \leq \lambda_{\min}(A)/2$. We find that these existing results are not easily applicable here as one would need to obtain these a priori estimates and enforce them in practice. Lastly, we note that a probabilistic bound with weaker conditions is given in Oliveira (2016; Theorem 1.2).

B.1.1.1 Proof of Corollary 4.12

Proof. We apply Proposition B.1 for each $t = 0, \ldots, T$. Note first that by Jensen's inequality, we have $\|\xi\|_{L^p(Q^\psi_{t-1,t})} \leq \|w_t^\psi\|_\infty \|\mathsf{F}_{t+1}\|_{L^p(Q^\psi_{t,t+1})}$ for any $\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}$ and $p = 2, 4$. Additionally, under boundedness of the basis functions and Holder's inequality

$$
\begin{aligned}
&\sup_{\xi \in \mathsf{B}_t \mathsf{F}_{t+1}} \Big\{ 3(N^{-2} - N^{-3}) Q^\psi_{t-1,t} \left(\left[\xi\varphi_m - Q^\psi_{t-1,t}(\xi\varphi_m) \right]^2 \right)^2 \\
&\qquad + N^{-3} Q^\psi_{t-1,t} \left(\left[\xi\varphi_m - Q^\psi_{t-1,t}(\xi\varphi_m) \right]^4 \right) \Big\} \\
&\leq \sup_{\xi \in \mathsf{B}_t \mathsf{F}_{t+1}} \Big\{ 3(N^{-2} - N^{-3}) U_t^4 \|\xi\|^4_{L^2(Q^\psi_{t-1,t})} + N^{-3} U_t^4 \|\xi\|^4_{L^4(Q^\psi_{t-1,t})} \\
&\qquad + 4N^{-3} U_t^4 \|\xi\|_{L^1(Q^\psi_{t-1,t})} \|\xi\|^3_{L^3(Q^\psi_{t-1,t})} + 6N^{-3} U_t^4 \|\xi\|^2_{L^1(Q^\psi_{t-1,t})} \|\xi\|^2_{L^2(Q^\psi_{t-1,t})} \Big\} \\
&\leq U_t^4 \sup_{\xi \in \mathsf{B}_t \mathsf{F}_{t+1}} \Big\{ 3(N^{-2} + N^{-3}) \|\xi\|^4_{L^2(Q^\psi_{t-1,t})} + 5N^{-3} \|\xi\|^4_{L^4(Q^\psi_{t-1,t})} \Big\} \\
&\leq U_t^4 \|w_t^\psi\|^4_\infty \left(3(N^{-2} + N^{-3}) \|\mathsf{F}_{t+1}\|^4_{L^2(Q^\psi_{t,t+1})} + 5N^{-3} \|\mathsf{F}_{t+1}\|^4_{L^4(Q^\psi_{t,t+1})} \right)
\end{aligned}
\tag{B.12}
$$

for any $m = 1, \ldots, M_t$. □

B.1.2 Non-linear least squares

We now move to the non-linear least squares case where we only assume that F is a closed linear subspace of $L^2(\mu)$. We write $\xi \in L^\infty(\Omega)$ as the function to be

approximated by regressors $\varphi \in \mathsf{F}$ and denote the corresponding loss function as $(\varphi, x) \mapsto \ell_\varphi(x) := (\varphi(x) - \xi(x))^2$ and risk function as $\varphi \mapsto \mu(\ell_\varphi)$. We define the set of all loss functions as $\mathsf{L} := \{\ell_\varphi : \varphi \in \mathsf{F}\}$, the envelope $E(x) := \sup_{\ell \in \mathsf{L}} \ell(x)$ for a fixed $x \in \Omega$ and assume that $U := \sup_{\ell \in \mathsf{L}} \|\ell\|_\infty < \infty$. Due to measurability issues when taking supremum over an uncountable set, we shall additionally assume that L is permissible (Pollard 1984; Appendix C, Definition 1). To get quantitative bounds, we will rely on results from empirical processes to control the rate at which empirical averages converge to their expectations uniformly over the set L.

Proposition B.2. *Let $\varphi^* := \mathsf{P}^\mu \xi$ be the orthogonal projection of ξ unto F and its approximation $\varphi^N := \mathsf{P}^{\mu,N} \xi$ for a given $N \in \mathbb{N}$. Suppose that L is permissible and is a Vapnik-Chervonenkis (VC) class of functions, i.e. there exists $v, c > 0$ such that for every probability measure ν on $(\Omega, \mathcal{F})$ and $\alpha > 0$, we have*

$$N\left(\mathsf{L}, L^2(\nu), \alpha \|E\|_{L^2(\nu)}\right) \leq \left(\frac{c}{\alpha}\right)^v \tag{B.13}$$

where $N(S, m, \delta)$ denotes the δ-covering number of the metric space (S, m), defined as the minimum number of (open) balls of radius δ needed to cover S. We have

$$\mathbb{E}\left[\mu(\ell_{\varphi^N})\right] \leq \mu(\ell_{\varphi^*}) + \left(\sqrt{\frac{\pi}{2}} + C(v, c) U^{-(v+1)/2}\right) U N^{-1/2} \tag{B.14}$$

where $C(v, c) > 0$ is a constant that depends only on the VC characteristics of L. Moreover, if $\sup_{\ell \in \mathsf{L}} \mu((\ell - \mu(\ell))^2) \leq \sigma^2$ and $0 < \sigma \leq U$, then

$$\mathbb{E}\left[\mu(\ell_{\varphi^N})\right] \leq \mu(\ell_{\varphi^*}) + \varrho \sqrt{v \sigma^2 \log\left(\frac{cU}{\sigma}\right)} N^{-1/2} + \varrho v U \log\left(\frac{cU}{\sigma}\right) N^{-1} \tag{B.15}$$

where $\varrho > 0$ is a universal constant.

Proof. We begin by defining the events

$$\begin{aligned} A_\varepsilon &:= \{\mu^N(\ell_{\varphi^*}) - \mu(\ell_{\varphi^*}) < \varepsilon\}, \\ B_\varepsilon &:= \{\mu(\ell_{\varphi^N}) - \mu^N(\ell_{\varphi^N}) < \varepsilon\}, \\ C_\varepsilon &:= \{\mu(\ell_{\varphi^N}) - \mu(\ell_{\varphi^*}) < 2\varepsilon\}, \end{aligned} \tag{B.16}$$

for $\varepsilon > 0$. Noting that $\mu^N(\ell_{\varphi^N}) \leq \mu^N(\ell_{\varphi^*})$ as φ^N minimizes the empirical risk function by construction, it follows that $A_\varepsilon \cap B_\varepsilon \subseteq C_\varepsilon$. Hence we have

$$\begin{aligned}\mathbb{P}\left(\mu(\ell_{\varphi^N}) - \mu(\ell_{\varphi^*}) \geq 2\varepsilon\right) &= \mathbb{P}\left(C_\varepsilon^c\right) \\ &\leq \mathbb{P}\left(A_\varepsilon^c \cup B_\varepsilon^c\right) \\ &\leq \mathbb{P}\left(A_\varepsilon^c\right) + \mathbb{P}\left(B_\varepsilon^c\right).\end{aligned} \tag{B.17}$$

We now consider bounds on the tail probabilities $\mathbb{P}\left(A_\varepsilon^c\right)$ and $\mathbb{P}\left(B_\varepsilon^c\right)$. By Hoeffding's inequality

$$\mathbb{P}\left(A_\varepsilon^c\right) \leq \exp\left(-2\varepsilon^2 U^{-2} N\right), \tag{B.18}$$

and by Talagrand (1994; Theorem 1.3)

$$\mathbb{P}\left(B_\varepsilon^c\right) \leq \tilde{C}(v,c) U^{-v} N^{v/2} \varepsilon^v \exp\left(-2\varepsilon^2 U^{-2} N\right), \tag{B.19}$$

where $\tilde{C}(v,c) > 0$ is a constant that depends only on the VC characteristics of L. Noting that $\mu(\ell_{\varphi^N}) > \mu(\ell_{\varphi^*})$ as φ^* minimizes the risk function by definition, using (B.17) and bounds on the tail probabilities gives

$$\begin{aligned}&\mathbb{E}\left[\mu(\ell_{\varphi^N}) - \mu(\ell_{\varphi^*})\right] = \int_0^\infty \mathbb{P}\left(\mu(\ell_{\varphi^N}) - \mu(\ell_{\varphi^*}) > u\right) \mathrm{d}u \\ &\leq 2\int_0^\infty \exp\left(-2\varepsilon^2 U^{-2} N\right) \mathrm{d}\varepsilon + 2\int_0^\infty \tilde{C}(v,c) U^{-v} N^{v/2} \varepsilon^v \exp\left(-2\varepsilon^2 U^{-2} N\right) \mathrm{d}\varepsilon \\ &= \sqrt{\frac{\pi}{2}} U N^{-1/2} + C(v,c) U^{(1-v)/2} N^{-1/2}\end{aligned} \tag{B.20}$$

where $C(v,c) > 0$ is another constant that depends only on the VC characteristics of L. This establishes (B.14), to obtain (B.15) we consider a similar decomposition as before

$$\begin{aligned}\mathbb{E}\left[\mu(\ell_{\varphi^N}) - \mu(\ell_{\varphi^*})\right] = \mathbb{E}\left[\mu(\ell_{\varphi^N}) - \mu^N(\ell_{\varphi^N})\right] + \mathbb{E}\left[\mu^N(\ell_{\varphi^N}) - \mu^N(\ell_{\varphi^*})\right] \\ + \mathbb{E}\left[\mu^N(\ell_{\varphi^*}) - \mu(\ell_{\varphi^*})\right].\end{aligned} \tag{B.21}$$

Note that the second term is at most zero as φ^N minimizes the empirical risk function and the third term is zero by unbiasedness. To deal with the first term, we apply Giné and Guillou (2002; Theorem 2.1)

$$\mathbb{E}\left[\mu(\ell_{\varphi^N}) - \mu^N(\ell_{\varphi^N})\right] \leq \varrho\sqrt{v\sigma^2 \log\left(\frac{cU}{\sigma}\right)} N^{-1/2} + \varrho v U \log\left(\frac{cU}{\sigma}\right) N^{-1}. \tag{B.22}$$

□

The bounds in Proposition B.2 decompose the approximate projection error into a bias term given by the norm of the residual $\varphi^* - \xi \in \mathsf{F}^\perp$ and a variance term that depends on the sample size N and how well parameterized the chosen function class F is as an approximation of ξ, described by U and the VC characteristics of L (B.13). When we also have control over the variance σ^2, (B.15) provides a more informative upper bound: in the regime where N is large and σ^2 is small, observe that the constant $\varrho\sqrt{v\sigma^2 \log cU - v\sigma^2 \log \sigma}$ of the order $N^{-1/2}$ term would be small since $\lim_{\sigma\to 0} \sigma^2 \log \sigma = 0$. Like in the linear case, a similar bound was obtained in Györfi et al. (2006; Theorem 11.5) under truncation of the approximation φ^N whenever it exceeds $\|\xi\|_\infty$. We now apply Proposition B.2 to provide a more precise description of the approximate projection error in (4.47).

Corollary B.3. *(Non-linear least squares) For each $t = 0, \ldots, T$ denote the set of all squared loss functions associated to the approximation of $\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}$ by $\mathsf{L}_t(\xi) := \{(\varphi - \xi)^2 : \varphi \in \mathsf{F}_t\}$ and the set of all possible squared loss functions at time t by $\mathsf{L}_t := \{(\varphi - \xi)^2 : \varphi \in \mathsf{F}_t, \xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}\}$. Define the corresponding envelope functions as $E_t(\xi)(x) := \sup_{\ell \in \mathsf{L}_t(\xi)} \ell(x)$ and $E_t(x) := \sup_{\ell \in \mathsf{L}_t} \ell(x)$ for each $x \in \mathcal{X}^2$. Suppose that $U_t := \sup_{\ell \in \mathsf{L}_t} \|\ell\|_\infty < \infty$, L_t is permissible and is a VC class of functions, i.e. there exists $v_t, c_t > 0$ such that for every probability measure ν on $(\mathcal{X}^2, \mathcal{B}(\mathcal{X})^{\otimes 2})$ and $\alpha > 0$, we have*

$$N\left(\mathsf{L}_t, L^2(\nu), \alpha \inf_{\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}} \|E_t(\xi)\|_{L^2(\nu)}\right) \leq \left(\frac{c_t}{\alpha}\right)^{v_t}. \tag{B.23}$$

For $t = 0, \ldots, T$, we have

$$\varepsilon_t^2(N) = \sup_{\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}} \|\mathsf{P}_t^\psi \xi - \xi\|^2_{L^2(Q^\psi_{t-1,t})} + \left(\sqrt{\frac{\pi}{2}} + C(v_t, c_t) U_t^{-(v_t+1)/2}\right) U_t N^{-1/2} \tag{B.24}$$

where $C(v_t, c_t)$ is a constant that depends only on the VC characteristics of L_t. Moreover, if $\sup_{\ell \in \mathsf{L}_t} Q^\psi_{t-1,t}\left(\left[\ell - Q^\psi_{t-1,t}(\ell)\right]^2\right) \leq \sigma_t^2$ and $0 < \sigma_t \leq U_t$, then

$$\begin{aligned}\varepsilon_t^2(N) = \sup_{\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}} \|\mathsf{P}_t^\psi \xi - \xi\|^2_{L^2(Q^\psi_{t-1,t})} + \varrho\sqrt{v_t \sigma_t^2 \log\left(\frac{c_t U_t}{\sigma_t}\right)} N^{-1/2} \\ + \varrho v_t U_t \log\left(\frac{c_t U_t}{\sigma_t}\right) N^{-1}\end{aligned} \tag{B.25}$$

where $\varrho > 0$ is a universal constant.

Proof. Noting that for any $\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}, \alpha > 0$ and $\nu \in \mathcal{P}(\mathcal{X}^2)$, we have

$$N\left(\mathsf{L}_t(\xi), L^2(\nu), \alpha\|E_t(\xi)\|_{L^2(\nu)}\right) \leq N\left(\mathsf{L}_t(\xi), L^2(\nu), \alpha \inf_{\zeta \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}} \|E_t(\zeta)\|_{L^2(\nu)}\right) \quad \text{(B.26)}$$
$$\leq N\left(\mathsf{L}_t, L^2(\nu), \alpha \inf_{\zeta \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}} \|E_t(\zeta)\|_{L^2(\nu)}\right),$$

therefore assumption (B.23) implies that $\mathsf{L}_t(\xi)$ is a VC class with characteristics v_t, c_t for all $\xi \in \mathsf{B}_t^\psi \mathsf{F}_{t+1}$. Hence applying Proposition B.2 for each $t = 0, \ldots, T$ completes the proof. □

Bibliography

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart. Importance sampling: computational complexity and intrinsic dimension. *arXiv preprint arXiv:1511.06196*, 2015.

L. Ambrosio. *Lecture notes on optimal transport problems*. Springer, 2003.

L. Ambrosio. Transport equation and Cauchy problem for BV vector fields. *Inventiones mathematicae*, 158(2):227–260, 2004.

L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows in metric spaces and in the space of probability measures, Lect. Math. *ETH Zürich, Birkhäuser Verlag, Basel*, 2005.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

A. Barker. Monte Carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

A. R. Barron and X. Luo. Adaptive annealing. In *Proceedings of the Allerton Conference on Communications, Computation, and Control*, pages 665–673, 2007.

D. Belomestny, A. Kolodko, and J. Schoenmakers. Regression methods for stochastic control problems and their convergence analysis. *SIAM Journal on Control and Optimization*, 48(5):3562–3588, 2010.

J. D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3): 375–393, 2000.

J. Bérard, P. Del Moral, and A. Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *Electronic Journal of Probability*, 19(94):1–28, 2014.

K. Bergemann and S. Reich. An ensemble Kalman-Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 21(3):213–219, 2012.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic Programming*. Athena Scientific, 1996.

J. Besag. Comments on "Representations of knowledge in complex systems" by U. Grenander and M. Miller. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(4):591–592, 1994.

A. Beskos, A. Jasra, N. Kantas, and A. Thiery. On the convergence of adaptive sequential Monte Carlo methods. *Annals of Applied Probability*, 26(2):1111–1146, 2016.

M. J. Betancourt. Adiabatic Monte Carlo. *arXiv preprint arXiv:1405.3489*, 2014.

A. Beurling. An automorphism of product measures. *Annals of Mathematics*, 72(1): 189–200, 1960.

J. Bierkens and H. J. Kappen. Explicit solution of relative entropy weighted control. *Systems & Control Letters*, 72:36–43, 2014.

O. Bokanowski and B. Grébert. Deformations of density functions in molecular quantum chemistry. *Journal of Mathematical Physics*, 37(4):1553–1573, 1996.

M. Briers, A. Doucet, and S. Maskell. Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61–89, 2010.

R. A. Brualdi. *Combinatorial matrix classes*. Cambridge University Press, 2006.

P. Bunch and S. Godsill. Approximations of the optimal importance density using Gaussian particle flow importance sampling. *Journal of the American Statistical Association*, 111(514):748–762, 2016.

L. Caffarelli, M. Feldman, and R. McCann. Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *Journal of the American Mathematical Society*, 15(1):1–26, 2002.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer Science & Business Media, 2006.

G. Carlier, A. Galichon, and F. Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *arXiv preprint arXiv:1511.01437*, 2015.

S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85 (2):347–361, 1998.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3): 539–552, 2002.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics*, 32(6):2385–2411, 2004.

D. L. Cohn. *Measure theory*. Springer, 2013.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

D. Crisan and J. Xiong. Approximate Mckean–Vlasov representations for a class of SPDEs. *Stochastics An International Journal of Probability and Stochastics Processes*, 82(1):53–68, 2010.

G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics*, 90(5-6):1481–1487, 1998.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

B. Dacorogna and J. Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l'IHP Analyse non linéaire*, 7(1):1–26, 1990.

P. Dai Pra, L. Meneghini, and W. J. Runggaldier. Connections between stochastic control and dynamic games. *Mathematics of Control, Signals and Systems*, 9(4): 303–326, 1996.

F. Daum and J. Huang. Particle flow for nonlinear filters with log-homotopy. In *SPIE Defense and Security Symposium*, pages 696918–696918. International Society for Optics and Photonics, 2008.

F. Daum and J. Huang. Nonlinear filters with particle flow induced by log-homotopy. In *SPIE Defense, Security, and Sensing*, pages 733603–733603. International Society for Optics and Photonics, 2009.

F. Daum, J. Huang, and A. Noushin. Coulomb's law particle flow for nonlinear filters. In *SPIE Optical Engineering+ Applications*, pages 81370B–81370B. International Society for Optics and Photonics, 2011.

P. Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

P. Del Moral. *Mean field simulation for Monte Carlo integration*. CRC Press, 2013.

P. Del Moral and A. Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. *Annales de l'IHP Probabilités et statistiques*, 37(2):155–194, 2001.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5): 1009–1020, 2012a.

P. Del Moral, A. Doucet, and A. Jasra. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278, 2012b.

P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41(1): 45–76, 1999.

R. J. DiPerna and P. L. Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones mathematicae*, 98(3):511–547, 1989.

R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

W. Doeblin. Sur deux problèmes de M. Kolmogoroff concernant les chaînes dénombrables. *Bulletin de la Société Mathématique de France*, 66:210–220, 1938.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69. IEEE, 2005.

R. Douc, E. Moulines, and J. Olsson. Long-term stability of sequential Monte Carlo methods under verifiable conditions. *Annals of Applied Probability*, 24(5): 1767–1802, 2014a.

R. Douc, E. Moulines, and D. Stoffer. *Nonlinear time series: theory, methods and applications with R examples*. CRC Press, 2014b.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2009.

A. Doucet, N. de Freitas, and N. J. Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, 2011.

L. C. Evans and W. Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem*. American Mathematical Society, 1999.

L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. CRC Press, 2015.

P. Fearnhead and P. Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2001.

B. D. Froese and A. M. Oberman. Fast finite difference solvers for singular solutions of the elliptic Monge–Ampère equation. *Journal of Computational Physics*, 230 (3):818–834, 2011.

C. W. Gardiner. *Handbook of stochastic methods*. Springer, second edition, 1985.

A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. 1991.

A. L. Gibbs. Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stochastic Models*, 20(4): 473–492, 2004.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'IHP Probabilités et statistiques*, 38(6):907–921, 2002.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

E. Gobet. *Monte-Carlo Methods and Stochastic Processes: From Linear to Non-Linear*. CRC Press, 2016.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F - Radar and Signal Processing*, 140(2):107–113, 1993.

U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(4): 549–603, 1994.

A. Griffin, P. A. Jenkins, G. O. Roberts, and S. E. F. Spencer. Simulation from quasi-stationary distributions on reducible state spaces. *arXiv preprint arXiv:1612.01872*, 2016.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

M. Hairer and J. C. Mattingly. Spectral gaps in Wasserstein distances and the 2D stochastic Navier–Stokes equations. *Annals of Probability*, 36(6):2050–2091, 2008.

M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63, pages 109–117. Springer, 2011.

T. E. Hanson, A. J. Branscum, and W. O. Johnson. Informative g-priors for logistic regression. *Bayesian Analysis*, 9(3):597–612, 2014.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

J. D. Hol, T. B. Schön, and F. Gustafsson. On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop*, pages 79–82. IEEE, 2006.

A. Iserles. *A first course in the numerical analysis of differential equations*. Cambridge University Press, 2009.

P. E. Jacob, F. Lindsten, and T. B. Schön. Coupling of particle filters. *arXiv preprint arXiv:1606.01156*, 2016.

B. Jamison. Reciprocal processes. *Probability Theory and Related Fields*, 30(1): 65–86, 1974.

B. Jamison. The Markov processes of Schrödinger. *Probability Theory and Related Fields*, 32(4):323–331, 1975.

S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications*, 85(2):341–361, 2000.

C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997.

A. Jasra, D. A. Stephens, and C. C. Holmes. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.

A. M. Johansen, P. Del Moral, and A. Doucet. Sequential Monte Carlo samplers for rare events. In *Proceedings of the 6th International Workshop on Rare Event Simulation*, pages 256–267, Bamberg, Germany, 2006.

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.

L. V. Kantorovich. On a problem of Monge. *Journal of Mathematical Sciences*, 133 (4):1383–1383, 2006.

H. J. Kappen and H. C. Ruiz. Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, 162(5):1244–1266, 2016.

H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, 2012.

S. Kim, R. Ma, D. Mesa, and T. P. Coleman. Efficient Bayesian inference methods via convex optimization and optimal transport. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 2259–2263. IEEE, 2013.

C. Kipnis and S. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

H. Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.

A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

H. R. Künsch. Recursive Monte Carlo filters: Algorithms and theoretical analysis. *Annals of Statistics*, 33(5):1983–2021, 2005.

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.

A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.

B. Leimkuhler and C. Matthews. *Molecular Dynamics*. Springer, 2015.

F. Liang and W. H. Wong. Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666, 2001.

J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, pages 2370–2378, 2016.

G. Loeper and F. Rapetti. Numerical solution of the Monge–Ampère equation by a Newton's algorithm. *Comptes Rendus Mathematique*, 340(4):319–324, 2005.

N. Madras and D. Sezer. Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16(3):882–908, 2010.

Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.

J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 1993.

T. Mikami. Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probability Theory and Related Fields*, 129(2):245–260, 2004.

G. Moffa and J. Kuipers. Sequential Monte Carlo EM for multivariate probit models. *Computational Statistics & Data Analysis*, 72:252–272, 2014.

T. Moselhy and Y. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

J. Moser. On the volume elements on a manifold. *Transactions of the American Mathematical Society*, 120(2):286–294, 1965.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2): 125–139, 2001.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, 1994.

E. Nummelin. *General irreducible Markov chains and non-negative operators*. Cambridge University Press, 1984.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.

R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180 (3):378–384, 1981.

M. Parno and Y. Marzouk. Transport map accelerated Markov chain Monte Carlo. *arXiv preprint arXiv:1412.5492*, 2014.

M. Parno, T. Moselhy, and Y. Marzouk. A multiscale strategy for Bayesian inference using transport maps. *arXiv preprint arXiv:1507.07024*, 2015.

O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

G. Pieralberto, A. Lee, and A. M. Johansen. The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 2016.

D. Pollard. *Convergence of Stochastic Processes*. Springer, 1984.

P. Rebeschini and R. Van Handel. Can local particle filters beat the curse of dimensionality? *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. A Gaussian-mixture ensemble transform filter. *Quarterly Journal of the Royal Meteorological Society*, 138(662):222–233, 2012.

S. Reich. A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

J. Ridgway. Computation of Gaussian orthant probabilities in high dimension. *Statistics and Computing*, 26(4):899–916, 2016.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996a.

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996b.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7 (1):110–120, 1997.

M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472, 1952.

H. C. Ruiz and H. J. Kappen. Particle smoothing for hidden diffusion processes: Adaptive path integral smoother. *arXiv preprint arXiv:1605.00278*, 2016.

C. Schäfer and N. Chopin. Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing*, 23(2):163–184, 2013.

E. Schrödinger. Über die Umkehrung der Naturgesetze. *Angewandte Chemie*, 44 (30):636–636, 1931.

R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly*, 74(4):402–405, 1967.

R. L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

V. N. Sudakov. *Geometric problems in the theory of infinite-dimensional probability distributions*. American Mathematical Society, 1979.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22(1):28–76, 1994.

A. Talhouk, A. Doucet, and K. Murphy. Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(3):739–757, 2012.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1728, 1994.

N. S. Trudinger and X. J. Wang. On the Monge mass transfer problem. *Calculus of Variations and Partial Differential Equations*, 13(1):19–31, 2001.

J. N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, 2001.

S. Vaikuntanathan and C. Jarzynski. Escorted free energy simulations: Improving convergence by reducing dissipation. *Physical Review Letters*, 100(19):190601, 2008.

S. Vaikuntanathan and C. Jarzynski. Escorted free energy simulations. *Journal of Chemical Physics*, 134(5):054107, 2011.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

C. Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

W. Walter. *Ordinary Differential Equations*. Springer, 1998.

N. Whiteley. Stability properties of some particle filters. *Annals of Applied Probability*, 23(6):2500–2537, 2013.

D. Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85, 1994.

D. Xiu and G. E. Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

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.

S. L. Zeger and M. R. Karim. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86 (413):79–86, 1991.

Y. Zhou, A. M. Johansen, and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.