Importance sampling for McKean-Vlasov SDEs

Importance sampling for McKean-Vlasov SDEs

Gonçalo dos Reis *

University of Edinburgh
and
Centro de Matemática e Aplicações
G.dosReis@ed.ac.uk

Greig Smith †

University of Edinburgh
Maxwell Institute for Mathematical Sciences
School of Mathematics
G.Smith-13@sms.ed.ac.uk

Peter Tankov‡

CREST–ENSAE
Palaiseau, France
peter.tankov@ensae.fr

01h02, 28/03/2018 (File: IS-MVSDEs_ArXiv.tex)

Abstract

This paper deals with the Monte-Carlo methods for evaluating expectations of functionals of solutions to McKean-Vlasov Stochastic Differential Equations (MV-SDE) with drifts of super-linear growth. We assume that the MV-SDE is approximated in the standard manner by means of an interacting particle system and propose two importance sampling (IS) techniques to reduce the variance of the resulting Monte Carlo estimator. In the *complete measure change* approach, the IS measure change is applied simultaneously in the coefficients and in the expectation to be evaluated. In the *decoupling* approach we first estimate the law of the solution in a first set of simulations without measure change and then perform a second set of simulations under the importance sampling measure using the approximate solution law computed in the first step.

For both approaches, we use large deviations techniques to identify an optimisation problem for the candidate measure change. The decoupling approach yields a far simpler optimisation problem than the complete measure change, however, we can reduce the complexity of the complete measure change through some symmetry arguments. We implement both algorithms for two examples coming from the Kuramoto model from statistical physics and show that the variance of the importance sampling schemes is up to 3 orders of magnitude smaller than that of the standard Monte Carlo. The computational cost is approximately the same as for standard Monte Carlo for the complete measure change and only increases by a factor of 2–3 for the decoupled approach. We also estimate the propagation of chaos error and find that this is dominated by the statistical error by one order of magnitude.

Key Words: McKean-Vlasov Stochastic Differential Equation, interacting particle system, Monte Carlo simulation, importance sampling, large deviations

MSC2010: 65C05 (Monte Carlo methods), 65C30 (Stochastic differential and integral equations), 65C35 (Stochastic particle methods)

*G. dos Reis acknowledges support from the *Fundação para a Ciência e a Tecnologia* (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações CMA/FCT/UNL).

†G. Smith was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016508/01), the Scottish Funding Council, the University of Edinburgh and Heriot-Watt University.

‡The research of Peter Tankov was supported by the FIME Research Initiative www.fime-lab.org

1 Introduction

The aim of this paper is to develop efficient importance sampling algorithms for computing the expectations of functionals of solutions to McKean-Vlasov stochastic differential equations (MV-SDE). MV-SDE are stochastic differential equations where the coefficients depend on the law of the solution, typically written in the following form:

$$\mathrm{d}X_t = b(t, X_t, \mu_t)\mathrm{d}t + \sigma(t, X_t, \mu_t)\mathrm{d}W_t, \quad X_0 = x,$$

where μ_t denotes the law of the process X at time t, and W is a standard Brownian motion. MV-SDEs, also known as mean-field equations, were originally introduced in physics to describe the movement of an individual particle amongst a large number of indistinguishable particles interacting through their mean field. They are now used in a variety of other domains, such as finance, economics, biology, population dynamics etc.

Development of algorithms for the simulation of MV-SDEs is a very active area of research. One of the earliest works to consider the error and computational complexity involved in simulating a MV-SDE was [BT97]. More recently [GP15], [STT17] and [CM17] among others (see references therein) developed more efficient methods for simulating MV-SDEs under Lipschitz coefficients or stronger settings.

A common technique for the simulation of MV-SDEs is to use the interacting particle representation. Namely, we consider $i = 1, \dots, N$ particles, where each $X^{i,N}$ satisfies the SDE with $X_0^{i,N} = x_0$

$$\mathrm{d}X_t^{i,N} = b\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}t + \sigma\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}W_t^i, \quad \mu_t^{X,N}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^{N} \delta_{X_t^{j,N}}(\mathrm{d}x) \tag{1.1}$$

where $\delta_{X_t^{j,N}}$ is the Dirac measure at point $X_t^{j,N}$, and the Brownian motions $W^i, i = 1, \dots, N$ are independent. The so-called propagation of chaos result (see, e.g., [Car16]) states that under sufficient conditions, as $N \to \infty$, for every i, the process $X^{i,N}$ converges to X^i, the solution of the MV-SDE driven by the Brownian motion W^i.

The system (1.1) is a system of ordinary SDE and can be discretized with one of the many available methods such as the Euler scheme. Let $X_t^{i,N,n}$ be the i-th component of the solution of (1.1), discretized on $[0, T]$ over n steps. The quantity of interest, which, in our case is $\theta = \mathbb{E}[G(X)]$, will then be approximated by the Monte Carlo estimator

$$\hat{\theta}^{N,n} = \frac{1}{N}\sum_{i=1}^{N} G(X^{i,N,n}).$$

The precision of this approximation is affected by three sources of error.

- The statistical error, that is the difference between $\hat{\theta}^{N,n}$ and $\mathbb{E}[G(X^{i,N,n})]$.
- The discretization error, that is, the difference between $\mathbb{E}[G(X^{i,N,n})]$ and $\mathbb{E}[G(X^{i,N})]$.
- The propagation of chaos error of approximating the MV-SDE with the interacting particle system, that is, the difference between $\mathbb{E}[G(X^{i,N})]$ and $\mathbb{E}[G(X)]$.

The discretization error of ordinary SDEs has been analyzed by many authors, and it is well known that, e.g., under the Lipschitz assumptions the Euler scheme has weak convergence error of order $\frac{1}{n}$. It is of course well known, the standard deviation of the statistical error is of order of $\frac{1}{\sqrt{N}}$.

There has also been some work detailing the error from the propagation of chaos as a function of N, essentially for G and X nice enough the weak error is also of the order $\frac{1}{\sqrt{N}}$, see for example [KHO97] and

[Bos04] for further details. In spite of this relatively slow convergence, many MV-SDEs have a reasonably "nice" dependence on the law which makes the particle approximation a good technique. On the other hand, one often wants to consider *rare events* in the context of the MV-SDE, and in this realm the statistical error will dominate the propagation of chaos error. The focus of this paper is therefore on the statistical error of the Monte Carlo method. In view of the poor convergence of the standard Monte Carlo, it is typical to enhance the standard approach with a so-called *variance reduction* technique. Importance sampling, which is the focus of this paper, is one such technique. We will discuss the point of statistical against propagation of chaos error in more detail in Section 5.

Importance sampling is based on the following identity, valid for any probability measure $\mathbb{Q}$ (absolutely continuous with respect to $\mathbb{P}$)

$$\mathbb{E}[G(X)] = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}G(X)\right].$$

The variance of the Monte Carlo estimator obtained by simulating X under the measure $\mathbb{Q}$ and correcting by the corresponding Radon-Nikodym density is different from that of the standard estimator, and can be made much smaller by a judicious choice of the sampling measure $\mathbb{Q}$.

Importance sampling is most effective in the context of *rare event simulation*, e.g., when the probability $\mathbb{P}[G(X) > 0]$ is small. Since the theory of large deviations is concerned with the study of probabilities of rare events, it is natural to use measure changes appearing in or inspired by the large deviations theory for importance sampling. We refer, e.g., to [DW04] and references therein for a review of this approach and to [GHS99, GR08, Rob10] for specific applications to financial models. The large deviations theory, on the one hand, simplifies the computation of the candidate importance sampling measure, and on the other hand, allows to define its optimality in a rigorous asymptotic framework.

The main contribution of this paper is two-fold. Firstly we show how one can apply a change of measure to MV-SDEs, and propose two algorithms that can carry this out: the *complete measure change* algorithm and the *decoupling* algorithm. In the complete measure change approach, the IS measure change is applied simultaneously in the coefficients and in the expectation to be evaluated. In the decoupling approach we first estimate the law of the solution in a first set of simulations without measure change and then perform a second set of simulations under the importance sampling measure using the approximate solution law computed in the first step.

Secondly, for both approaches, we use large deviations techniques to obtain an optimisation problem for the candidate measure change. We focus on the class of Cameron-Martin transforms, under which the measure change is given by

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{F}_T} = \mathcal{E}(\int_0^T f_t \mathrm{d}W_t) := \exp\left(\int_0^T f_t \mathrm{d}W_t - \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\right), \tag{1.2}$$

where f_t is a deterministic function. Following earlier works on the subject, we use the large deviations theory to construct a tractable proxy for the variance of $G(X)$ under the new measure. Of course, the presence of the interacting particle approximation introduces additional complexity at this point. Moreover, unlike the work of [GR08] which considered a very restrictive class of SDEs (the geometric Brownian motion), here we deal with a general class of MV-SDE where the drifts are of super-linear growth and satisfy a monotonicity type condition. This is very important in practice since many MV-SDEs fall into this category.

We then minimise the large deviations proxy to obtain a candidate optimal measure change for the two approaches that we consider. We find that the decoupling approach yields an easier optimisation problem than the complete measure change, which results in a high dimensional problem. However, by using exchangeability arguments the latter problem can be transformed into a far simpler two dimensional one. We implement both algorithms for two examples coming from the Kuramoto model from

statistical physics and show that the variance of the importance sampling schemes is up to 3 orders of magnitude smaller than that of the standard Monte Carlo. Moreover, the computational only increases by a factor of 2–3 for the decoupling approach and is approximately the same as standard Monte Carlo for the complete measure change. We also estimate the propagation of chaos error and find that this is dominated by the statistical error by one order of magnitude. That being said, although the complete measure change appears to operate well in certain situations, it does rely on a change of measure which isn't too "large". We come back to this point throughout.

The manuscript is organized as follows. In Section 2 we gather the preliminary results. In Section 3 we discuss how importance sampling and measure changes can be carried out for MV-SDE, and in Section 4 we introduce our concept of optimality and identify the candidate optimal measure changes using the theory of large deviations. Section 5 illustrates numerically our results while proofs from Section 4 are carried out in Section 6.

Acknowledgements The authors would like to thank Daniel Lacker (Columbia University), for the helpful discussion.

2 Preliminaries

Throughout the paper we work on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ satisfying the usual conditions, where $\mathcal{F}_t$ is the augmented filtration of a standard multidimensional Brownian motion W.

We consider some finite terminal time $T < \infty$ and use the following notation for spaces, which are standard in the McKean-Vlasov literature (see [Car16]). We define $\mathbb{S}^p$ for $p \geq 1$, as the space of $\mathbb{R}^d$-valued, $\mathcal{F}_\cdot$-adapted processes Z, that satisfy, $\mathbb{E}[\sup_{0\leq t\leq T} |Z(t)|^p]^{1/p} < \infty$. Similarly, $L_t^p(\mathbb{R}^d)$, defines the space of $\mathbb{R}^d$-valued, $\mathcal{F}_t$-measurable random variables X, that satisfy, $\mathbb{E}[|X|^p]^{1/p} < \infty$.

We will work with $\mathbb{R}^d$, the d-dimensional Euclidean space of real numbers, and for $a = (a^1, \cdots, a^d) \in \mathbb{R}^d$ and $b = (b^1, \cdots, b^d) \in \mathbb{R}^d$ we denote by $|a|^2 = \sum_{i=1}^d a_i^2$ the usual Euclidean distance on $\mathbb{R}^d$ and by $\langle a, b\rangle = \sum_{i=1}^d a^i b^i$ the usual scalar product.

Given the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on this space, and write $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if $\mu \in \mathcal{P}(\mathbb{R}^d)$ and for some $x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} |x-y|^2 \mu(\mathrm{d}y) < \infty$. We then have the following metric on the space $\mathcal{P}_2(\mathbb{R}^d)$ (Wasserstein metric) for $\mu,\ \nu \in \mathcal{P}_2(\mathbb{R}^d)$ (see [dRST17]),

$$W^{(2)}(\mu,\nu) = \inf_{\pi} \left\{ \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2 \pi(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{2}} \ :\ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ with marginals } \mu \text{ and } \nu \right\}.$$

2.1 McKean-Vlasov stochastic differential equations

Let W be an l-dimensional Brownian motion and take the progressively measurable maps $b : [0,T]\times\Omega\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$ and $\sigma : [0,T]\times\Omega\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^{d\times l}$. MV-SDEs are typically written in the form,

$$\mathrm{d}X_t = b(t, X_t, \mu_t)\mathrm{d}t + \sigma(t, X_t, \mu_t)\mathrm{d}W_t, \quad X_0 = x_0, \tag{2.1}$$

where μ_t denotes the law of the process X at time t, i.e. $\mu_t = \mathbb{P}\circ X_t^{-1}$. Consider the following assumption on the coefficients.

Assumption 2.1. *Assume that σ is Lipschitz in the sense that there exists $L > 0$ such that for all $t \in [0,T]$ and all $x, x' \in \mathbb{R}^d$ and $\forall \mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$ we have that*

$$|\sigma(t,x,\mu) - \sigma(t,x',\mu')| \leq L(|x-x'| + W^{(2)}(\mu,\mu')), \qquad \mathbb{P}\text{-a.s.}$$

and let b satisfy

1. *Monotone growth condition in x and Lipschitz in law: there exists $L > 0$ such that for all $t \in [0,T]$, all $x, x' \in \mathbb{R}^d$ and all $\mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$ we have that*

$$\langle x - x', b(t,x,\mu) - b(t,x',\mu)\rangle \leq L|x-x'|^2 \quad \textit{and} \quad |b(t,x,\mu) - b(t,x,\mu')| \leq W^{(2)}(\mu,\mu')) \quad \mathbb{P}\textit{-a.s.}.$$

2. *Locally Lipschitz with polynomial growth in x: there exists $q \in \mathbb{N}$ with $q > 1$ such that for all $t \in [0,T]$, $\forall \mu \in \mathcal{P}_2(\mathbb{R}^d)$ and all $x,\ x' \in \mathbb{R}^d$ the following holds.*

$$|b(t,x,\mu) - b(t,x',\mu)| \leq L(1 + |x|^q + |x'|^q)|x - x'| \qquad \mathbb{P}\textit{-a.s.}.$$

Under these assumptions, an existence and uniqueness result for the solution of the MV-SDE is given in [dRST17]. Note that this can be generalised to include random initial conditions.

Theorem 2.2 ([dRST17, Theorem 3.3])**.** *Suppose that b and σ satisfy Assumption 2.1 and are integrable in the sense that, for some $m \geq 2$,*

$$\mathbb{E}\Big[\Big(\int_0^T |b(t,0,\delta_0)|\mathrm{d}t\Big)^m\Big] < \infty \quad \textit{and} \quad \mathbb{E}\Big[\Big(\int_0^T |\sigma(t,0,\delta_0)|^2\mathrm{d}t\Big)^m\Big] < \infty.$$

Then there exists a unique solution for $X \in \mathcal{S}^m([0,T])$ to the MV-SDE (2.1)*. Moreover, for some positive constant C we have $\mathbb{E}\big[\sup_{t\in[0,T]} |X_t|^m\big] \leq C(1 + |x_0|^m)e^{CT}$.*

2.2 Large Deviation Principles

In this section, we state the main results from the large deviations theory that we use throughout, for a full exposition the reader can consult texts such as [DZ10] or [DE11]. The large deviation principle (LDP) characterizes the limiting behaviour, as $\epsilon \to 0$, of a family of probability measures $\{\mu_\epsilon\}$ in exponential scale on the space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, with $\mathcal{X}$ a topological space so that open and closed subsets of $\mathcal{X}$ are well-defined, and $\mathcal{B}_\mathcal{X}$ is the Borel σ-algebra on $\mathcal{X}$. The limiting behaviour is defined via a so-called rate function. We assume the probability spaces have been completed, consequently, $\mathcal{B}_\mathcal{X}$ is the complete Borel σ-algebra on $\mathcal{X}$. We have the following definition [DZ10, pg.4].

Definition 2.3 (Rate function)**.** *A rate function I is a lower semicontinuous mapping $I : \mathcal{X} \to [0,\infty]$ (such that for all $\alpha \in [0,\infty)$, the level set $\Psi_I(\alpha) := \{x : I(x) \leq \alpha\}$ is a closed subset of $\mathcal{X}$). A good rate function is a rate function for which all the level sets $\Psi_I(\alpha)$ are compact subsets of $\mathcal{X}$. The effective domain of I, denoted D_I, is the set of points in $\mathcal{X}$ of finite rate, namely, $D_I := \{x : I(x) < \infty\}$.*

We use the standard notation: for any set Γ, $\overline{\Gamma}$ denotes the closure, Γ^o denotes the interior and finally Γ^C denotes the complement of Γ. As is standard practice in LDP theory, the infimum of a function over an empty set is interpreted as ∞. We then define what it means for this sequence of measures to have an LDP [DZ10, pg.5].

Definition 2.4. *A family of probability measures, $\{\mu_\epsilon\}$ with $\epsilon > 0$ satisfies the large deviation principle with a rate function I if, for all $\Gamma \in \mathcal{B}$,*

$$-\inf_{x\in\Gamma^o} I(x) \leq \liminf_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) \leq \limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) \leq -\inf_{x\in\overline{\Gamma}} I(x)\,. \tag{2.2}$$

It is also typical to have LDP defined in terms of a sequence of random variables Z_ϵ, in which case one replaces $\mu_\epsilon(\Gamma)$ by $\mathbb{P}[Z_\epsilon \in \Gamma]$.

The following result can be viewed as a generalisation of Laplace's approximation of integrals to the infinite dimensional setting and transfers the LDP from probabilities to expectations (see [DZ10]).

Lemma 2.5 (Varadhan's Lemma). *Let $\{\mu_\epsilon\}$ be a family of measures that satisfies a large deviation principle with good rate function I. Furthermore, let Z_ϵ be a family of random variables in $\mathcal{X}$ such that Z_ϵ has law μ_ϵ and let $\varphi : \mathcal{X} \to \mathbb{R}$ be any continuous function that satisfies the following integrability (moments) condition for some $\gamma > 1$,*

$$\limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}\left[\exp\left(\frac{\gamma}{\epsilon}\varphi(Z_\epsilon)\right)\right] < \infty\,.$$

Then,

$$\lim_{\epsilon\to 0} \epsilon \log \mathbb{E}\left[\exp\left(\frac{1}{\epsilon}\varphi(Z_\epsilon)\right)\right] = \sup_{x\in\mathcal{X}}\{\varphi(x) - I(x)\}\,.$$

As is discussed in [GR08], one needs a slight extension to Varadhan's lemma to allow the function φ to take the value $-\infty$. The extension is proved in [GR08].

Lemma 2.6. *Let $\varphi : \mathcal{X} \to [-\infty, \infty)$ and assume the conditions in Lemma 2.5 are satisfied. Then the following bounds hold for any $\Gamma \in \mathcal{B}$*

$$\begin{aligned}\sup_{x\in\Gamma^0}\{\varphi(x) - I(x)\} &\le \liminf_{\epsilon\to 0} \epsilon \log\left(\int_{\Gamma^o} \exp\left(\frac{1}{\epsilon}\varphi(Z_\epsilon)\right) \mathrm{d}\mu_\epsilon\right)\\ &\le \limsup_{\epsilon\to 0} \epsilon \log\left(\int_{\overline{\Gamma}} \exp\left(\frac{1}{\epsilon}\varphi(Z_\epsilon)\right) \mathrm{d}\mu_\epsilon\right) \le \sup_{x\in\overline{\Gamma}}\{\varphi(x) - I(x)\}\,.\end{aligned}$$

The previous lemma allows us to control the $\liminf$ and $\limsup$ of the process even when they are not equal (as is the case in Varadhan's lemma).

2.3 Importance Sampling and large deviations

To motivate our approach we recall ideas from the pioneering works [GHS99], [GR08] and [Rob10] which establish a connection between large deviations and importance sampling. Importance sampling uses the following idea. Consider the problem of estimating $\mathbb{E}_{\mathbb{P}}[G(X)]$ where X is some random variable/process governed by probability $\mathbb{P}$. Through Radon-Nikodym theorem we can rewrite this expectation under a new measure $\mathbb{Q}$ weighted by the Radon-Nikodym derivative, thus $\mathbb{E}_{\mathbb{P}}[G(X)] = \mathbb{E}_{\mathbb{Q}}[G(X)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}]$. Although the expectations (first moments) are the same, the variance under $\mathbb{Q}$ is,

$$\mathrm{Var}_{\mathbb{Q}}\Big[G(X)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] = \mathbb{E}_{\mathbb{P}}\Big[G(X)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] - \mathbb{E}_{\mathbb{P}}\Big[G(X)\Big]^2\,. \tag{2.3}$$

As it turns out, if one can choose $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \frac{G}{\mathbb{E}_{\mathbb{P}}[G]}$, then the variance under $\mathbb{Q}$ is zero, i.e. we have no error in our Monte Carlo simulation. Unfortunately though, in order to choose such a change of measure one would need to a priori know the value of $\mathbb{E}_{\mathbb{P}}[G(X)]$ i.e. the value we wish to estimate in the first place.

Instead one typically chooses $\mathbb{Q}$ to minimise (2.3) over a set of equivalent probability measures, chosen to add only a small amount of extra computation and such that the process X is easy to simulate under the new measure. Specializing to the Brownian filtration, a common choice of $\mathbb{Q}$ is the Girsanov transform, (1.2) where f is often taken to be a deterministic function.

For example in [TFC16] the authors develop an importance sampling procedure in the context of Gaussian random vectors through a so-called "tilting" parameter, which corresponds to shifting the mean of the Gaussian random vector via a Girsanov transform. Although this method is intuitive, it still requires estimation of the Jacobian of G w.r.t. the tilting parameter and applying Newton's method to select the optimal parameter value. These steps can be computationally expensive, and it is difficult to obtain rigorous optimality results.

Even after one has reduced the set of measures $\mathbb{Q}$ to optimise over, in general the problem of minimizing (2.3) will not have a closed form solution. Thus we instead minimize a proxy for the variance obtained in the so-called small noise asymptotic regime as discussed in [GHS99] and [GR08]. Assuming that a Girsanov change of measure is used, we want to minimise

$$\mathbb{E}_{\mathbb{P}}\left[G(W)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right] = \mathbb{E}_{\mathbb{P}}\left[\exp\left(2F(W) - \int_0^T f_t \mathrm{d}W_t + \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\right)\right], \quad \text{with } F = \log(G). \tag{2.4}$$

Typically G is defined as a functional of the SDE, but here with a slight abuse of notation we have redefined it as the functional of the driving Brownian motion. It is important for this type of argument that we are able to write the solution of the SDE in terms of BM as well, i.e. we can write $X_t = H(t, W_\cdot)$. Finding the optimal f by minimizing (2.4) is in general intractable, hence an asymptotic approximation of the variance should be constructed. Let us consider,

$$\epsilon \log \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{1}{\epsilon}\left(2F(\sqrt{\epsilon}W) - \int_0^T \sqrt{\epsilon} f_t \mathrm{d}W_t + \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\right)\right)\right],$$

which equals $\log$ of (2.4) when $\epsilon = 1$, the *small noise asymptotic approximation* is then,

$$L(f) := \limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{1}{\epsilon}\left(2F(\sqrt{\epsilon}W) - \int_0^T \sqrt{\epsilon} f_t \mathrm{d}W_t + \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\right)\right)\right].$$

One then computes a candidate variance reduction parameter f^* by minimizing $L(f)$, which can be thought of as approximating $\mathbb{E}_{\mathbb{P}}\left[G(W)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right]$ by $\exp(L(f))$. Crucially, L is in a form that can be evaluated using the Varadhan's lemma, i.e., we can change L into a supremum depending on the rate function. The parameter f^*, which minimises L over some predefined space is known as *asymptotically optimal*, see [GR08]. We will give a precise definition of this concept later. It is important to note, these approximations are not approximations for the original problem (calculate $\mathbb{E}_{\mathbb{P}}[G(X)]$), they are only approximations to help choose the change of measure we want to apply.

3 Importance sampling for MV-SDEs

Leaving LDPs and the optimality of the IS (importance sampling) on the side, let us discuss how IS can be achieved for MV-SDEs with a given measure change.

Recall that MV-SDEs take the form (2.1). Because we change the measure we make explicit the dependence on the law of the solution process $\mu^X_{t,\mathbb{P}} = \mathbb{P}\circ X_t^{-1}$. If one knows the law μ^X beforehand , then one can treat the MV-SDE as a "standard" SDE and use IS as usual. However, typically one does not have access to the law, and the MV-SDE must be approximated by a so-called particle system approximation.

The interacting particle system approximation. We approximate (2.1) (driven by the $\mathbb{P}$-Brownian motion $W^{\mathbb{P}}$), using an N-dimensional system of interacting particles. Let $i = 1, \dots, N$ and consider N particles $X^{i,N}$ satisfying the SDE with $X_0^{i,N} = x_0$

$$\mathrm{d}X_t^{i,N} = b\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}t + \sigma\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}W_t^{i,\mathbb{P}}, \quad \mu_t^{X,N}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}(\mathrm{d}x) \tag{3.1}$$

where $\delta_{X_t^{j,N}}$ is the Dirac measure at point $X_t^{j,N}$, and the independent $\mathbb{P}$-Brownian motions $W^{i,\mathbb{P}}, i = 1, \dots, N$ (also independent of the BM $W^{\mathbb{P}}$ appearing in (2.1)). Due to the several changes of the measure throughout this section we keep track of which W we refer to.

Remark 3.1 (On the empirical measure $\mu_t^{X,N}$)**.** *Unlike standard measures, empirical measures do not have dependence on the underlying measure $\mathbb{P}$, namely empirical measures are maps that depend on a sequence of $\omega^i \in \Omega$, thus one should write $\mu_t^{X,N}$ instead of $\mu_{t,\mathbb{P}}^{X,N}$. Of course, this is a pathwise statement, since the ω^i are generated under $\mathbb{P}$, the* distribution *of the empirical measure does depend on $\mathbb{P}$.*

Propagation of chaos. In order to show that the particle approximation is of use, one shows a so-called propagation of chaos result. Although different types exist a common one is the pathwise convergence result where we consider the system of non interacting particles

$$\mathrm{d}X_t^i = b(t, X_t^i, \mu_{t,\mathbb{P}}^{X^i})\mathrm{d}t + \sigma(t, X_t^i, \mu_{t,\mathbb{P}}^{X^i})\mathrm{d}W_t^{i,\mathbb{P}}, \quad X_0^i = x_0\,, \quad t \in [0,T]\,, \tag{3.2}$$

which are of course just MV-SDEs and since the X^is are independent, then $\mu_{t,\mathbb{P}}^{X^i} = \mu_{t,\mathbb{P}}^{X}\ \forall i$. Under sufficiently nice conditions, one can then prove the following convergence result (see [Car16, Theorem 1.10] for example)

$$\lim_{N\to\infty} \sup_{1\le i\le N} \mathbb{E}_{\mathbb{P}}\left[\sup_{0\le t\le T} |X_t^{i,N} - X_t^i|^2\right] = 0\,.$$

Note that, for all SDEs appearing below, we have initial condition x_0 and work on the interval $[0,T]$.

Setup to change measures. When it comes to changing the measure under which we simulate we are also changing our approximation of the law. Since MV-SDEs depend explicitly on the law, this makes importance sampling more difficult. This will be one of the main points throughout this section.

Fix a deterministic square-integrable function $\dot{h} \in L_0^2(\mathbb{R})$. Then one can define the probability measure $\mathbb{Q}$ via the Girsanov transform $\frac{d\mathbb{Q}}{d\mathbb{P}}|_{\mathcal{F}_T} := \mathcal{E}(\int_0^T \dot{h}_t \mathrm{d}W_t^{\mathbb{P}})$, see (1.2), so that $\mathrm{d}W_t^{\mathbb{Q}} = \mathrm{d}W_t^{\mathbb{P}} - \dot{h}_t\mathrm{d}t$ is a $\mathbb{Q}$-Brownian motion. We note that the Radon-Nikodym density $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}|_{\mathcal{F}_t} = \mathcal{E}(\int_0^\cdot \dot{h}_s \mathrm{d}W_s^{\mathbb{P}})_t =: \mathcal{E}_t$ is itself the solution of the SDE

$$\mathrm{d}\mathcal{E}_t = \dot{h}_t\mathcal{E}_t\mathrm{d}W_t^{\mathbb{P}}, \quad \mathcal{E}_0 = 1 \quad \Rightarrow \quad \mathcal{E}_t = \exp\Big\{\int_0^t \dot{h}_s \mathrm{d}W_s^{\mathbb{P}} - \frac{1}{2}\int_0^t |\dot{h}_s|^2\mathrm{d}s\Big\}.$$

Since $\mathbb{P}$ and $\mathbb{Q}$ are equivalent, one can also define $Z_t := \mathcal{E}_t^{-1} := \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}|_{\mathcal{F}_t}$. With our conditions on $\dot{h}$ it is also a straightforward task to show $\mathcal{E}_t$ and Z_t are in $\mathbb{S}^p$ for all $p \ge 1$.

Recall our goal: *estimate $\mathbb{E}_{\mathbb{P}}[G(X_T)] = \mathbb{E}_{\mathbb{Q}}[G(X_T)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}]$ for some function G by simulating X under $\mathbb{Q}$.* In the following paragraphs we present two alternative ways to achieve this goal.

A running example. We present our algorithm in general setting with (2.1). For the sake of clarity and easiness of presentation, we often recourse to a particular class of MV-SDEs (under $\mathbb{P}$),

$$\mathrm{d}X_t = \hat{b}\big(t, X_t, \mathbb{E}_{\mathbb{P}}[f(X_t)]\big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{P}}, \quad X_0 = x_0\,, \quad t \in [0,T]\,. \tag{3.3}$$

with $\sigma \in \mathbb{R}_+$ and $f, \hat{b}$ nice[1]. We believe many of the arguments that are used at this level can be extended to cover more general MV-SDEs (such as higher order interactions). However, obtaining analogous results to those of standard MV-SDEs, such as propagation of chaos, is made more challenging by the inclusion of the measure change. Therefore, these have to be considered on a case by case basis.

3.1 Fixing the Empirical Law - a decoupling argument

An obvious way to solve the problem of IS is to approximate the law of the MV-SDE under $\mathbb{P}$ and use that as a fixed input to a new equation which will be simulated under $\mathbb{Q}$. In this set up the McKean-Vlasov SDE turns into an SDE with random coefficients. The algorithm is as follows.

[1]We use $\hat{b}$ here since it takes the expectation rather than a measure input.

1. Use (3.1) with N particles to approximate (2.1). Use some numerical scheme (under $\mathbb{P}$, say Euler) to simulate the particles in time, calculating an empirical law over $[0,T]$. This gives an approximation for the empirical law μ_t^N which is then fixed.

   Define a new SDE, approximating the original MV-SDE (2.1), which is now a *standard SDE with random coefficients*

$$\mathrm{d}\bar{X}_t = b(t,\bar{X}_t,\mu_t^N)\mathrm{d}t + \sigma(t,\bar{X}_t,\mu_t^N)\mathrm{d}W_t^{\mathbb{P}}, \quad \bar{X}_0 = x_0, \tag{3.4}$$

   where $W^{\mathbb{P}}$ is a $\mathbb{P}$-Brownian motion independent of the $\{W^{i,\mathbb{P}}\}_{i=1,\cdots,N}$ appearing in (3.1). SDEs with random coefficients appear typically in optimal control, hence the reader can consult texts such as [YZ99, Chapter 1] for further details on existence uniqueness of such SDEs.

2. Change the probability measure to $\mathbb{Q}$, which is our importance sampling measure change. Simulate (3.4) under this new measure, i.e.

$$\mathrm{d}\bar{X}_t = \Big(b(t,\bar{X}_t,\mu_t^N) + \dot{h}_t\sigma(t,\bar{X}_t,\mu_t^N)\Big)\,\mathrm{d}t + \sigma(t,\bar{X}_t,\mu_t^N)\mathrm{d}W_t^{\mathbb{Q}}, \quad \bar{X}_0 = x_0\,.$$

3. This second run is therefore standard importance sampling, but the SDE has random coefficients i.e. the empirical law is random.

We will refer to algorithms of this form as *Decoupling Algorithms*. This scheme has the disadvantage in that it requires twice the amount of simulation and one will require a handle on the error coming from the original approximation of the law.

It is not a requirement to use interacting particles to approximate the law of the SDE, any approximation will work. The goal here is to make the SDEs independent.

3.2 Complete Measure Change

An alternative is to change the measure under which we are simulating in the coefficients *and* the Brownian motion. This is not a simple problem and as far as we are aware changing the measure of a MV-SDE and its particle approximation is not discussed elsewhere in the literature (for this purpose[2]), we therefore provide a discussion along with the pitfalls here. This is more complex than the decoupled case and for clarity we use (3.3) throughout.

The measure changed version of (3.3) takes the form,

$$\begin{aligned}\mathrm{d}X_t &= \Big(\hat{b}(t,X_t,\mathbb{E}_{\mathbb{P}}[f(X_t)]) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}}\\ &= \Big(\hat{b}(t,X_t,\mathbb{E}_{\mathbb{Q}}\Big[f(X_t)Z_t\Big]) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}}.\end{aligned}$$

where again $Z := \mathcal{E}^{-1}$.

In view of simulation, we re-write the measure changed MV-SDE as a system

$$\mathrm{d}X_t = \Big(\hat{b}\Big(t,X_t,\mathbb{E}_{\mathbb{Q}}\big[F(X_t,Z_t)\big]\Big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}}, \quad \text{and} \quad \mathrm{d}Z_t = \dot{h}_t Z_t\mathrm{d}W_t^{\mathbb{Q}}, \quad Z_0 = 1\,,$$

where $F(x,y) = f(x)y$. We now write the interacting particle system for the pair X, Z under $\mathbb{Q}$:

$$\begin{aligned}\mathrm{d}X_t^{i,N} &= \Big(\hat{b}\big(t,X_t^{i,N},\frac{1}{N}\sum_{j=1}^N F(X_t^{j,N},Z_t^{j,N})\big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{Q}},\\ \mathrm{d}Z_t^{i,N} &= \dot{h}_t Z^{i,N}\mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z_0^{i,N} = 1\,.\end{aligned} \tag{3.5}$$

[2]Measures changes for MV-SDE appear in methods requiring to remove the drift altogether, for instance in establishing weak solutions to MV-SDEs, see e.g. [DG87].

The importance sampling estimator of $\theta = \mathbb{E}^{\mathbb{P}}[G(X_T)]$ then takes the form

$$\hat{\theta}_h = \frac{1}{N}\sum_{i=1}^{N} Z_T^{i,N} G(X_T^{i,N}). \tag{3.6}$$

Remark 3.2. *One may be tempted to use a different approach, namely first apply an interacting particle approximation under $\mathbb{P}$ which yields*

$$\mathrm{d}X_t^{i,N} = \hat{b}\big(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} f(X_t^{j,N})\big)\mathrm{d}t + \sigma \mathrm{d}W_t^{i,\mathbb{P}},$$

and then change the measure for the particle system, writing

$$\mathrm{d}X_t^{i,N} = \Big(\hat{b}\big(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} f(X_t^{j,N})\big) + \sigma \dot{h}_t\Big)\mathrm{d}t + \sigma \mathrm{d}W_t^{i,\mathbb{Q}},$$

where we have taken the same $\dot{h}$ for every Brownian motion in order for all particles to have the same law. However, it is easy to see by the standard propagation of chaos result that as $N \to \infty$, this particle system converges to the solution of the MV-SDE

$$\mathrm{d}X_t = \Big(\hat{b}\big(t, X_t, \mathbb{E}^{\mathbb{Q}}[X_t]\big) + \sigma \dot{h}_t\Big)\mathrm{d}t + \sigma \mathrm{d}W_t^{\mathbb{Q}} = \hat{b}\big(t, X_t, \mathbb{E}^{\mathbb{Q}}[X_t]\big)\mathrm{d}t + \sigma \mathrm{d}W_t^{\mathbb{P}},$$

which is not what one is looking for.

To state a propagation of chaos result for the particle system (3.5) we introduce the auxiliary system of non-interacting particles,

$$\begin{aligned} \mathrm{d}X_t^i &= \Big(\hat{b}\Big(t, X_t^i, \mathbb{E}_{\mathbb{Q}}\big[F(X_t^i, Z_t^i)\big]\Big) + \sigma \dot{h}_t\Big)\mathrm{d}t + \sigma \mathrm{d}W_t^{i,\mathbb{Q}}, \\ \mathrm{d}Z^i &= \dot{h}_t Z^i \mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z^i = 1\,. \end{aligned} \tag{3.7}$$

Proposition 3.3. *Consider the following measure changed MV-SDE (see* (3.7)*),*

$$\mathrm{d}X_t^i = \Big(\hat{b}\Big(t, X_t^i, \mathbb{E}_{\mathbb{Q}}\Big[f(X_t^i) Z_t^i\Big]\Big) + \sigma \dot{h}_t\Big)\mathrm{d}t + \sigma \mathrm{d}W_t^{i,\mathbb{Q}}, \quad \mathrm{d}Z_t^i = \dot{h}_t Z_t^i \mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z_0^i = 1\,, \tag{3.8}$$

where $\hat{b}$ is continuous in time, $\hat{b}$ and f are Lipschitz in space, and $\hat{b}$ is a bounded Lipschitz function in its third variable. Let $X_t^{i,N}$, denote the corresponding particle approximation (see (3.5)*). Then the following pathwise propagation of chaos result holds,*

$$\lim_{N\to\infty} \sup_{1\le i\le N} \mathbb{E}_{\mathbb{Q}}\left[\sup_{0\le t\le T} |X_t^{i,N} - X_t^i|^2\right] = 0\,.$$

This proposition may be used to analyze the convergence of the Monte Carlo estimator (3.6). Indeed, due to the fact that there is no coupling (or law dependency) in $Z_t^{i,N}$, $Z^{i,N} = Z^i$ and $\hat{\theta}_h$ can be represented as follows.

$$\hat{\theta}_h = \frac{1}{N}\sum_{i=1}^{N} Z_T^i G(X_T^i) + \frac{1}{N}\sum_{i=1}^{N} Z_T^i (G(X_T^{i,N}) - G(X_T^i).)$$

The first term above converges to θ as $N \to \infty$ by the law of large numbers, and the second term can be shown, e.g., to converge to zero in probability using Proposition 3.3 if G is sufficiently regular.

Proof of Proposition 3.3. The idea of the proof is to appeal to a Gronwall type inequality, but this is made difficult due to the presence of Z term in (3.8). One can note, due to the assumptions on the coefficients of the SDE, all p-moments exist. Using our prescribed form of the MV-SDE we obtain,

$$|X_t^{i,N} - X_t^i|^2 \leq C \int_0^t \big|\hat{b}\Big(s, X_s^{i,N}, \frac{1}{N}\sum_{j=1}^N f(X_s^{j,N})Z_s^j\Big) - \hat{b}\left(s, X_s^i, \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\right)\big|^2 \mathrm{d}s\,.$$

Let $s \in [0,T]$, then introduce the terms, $\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^{j,N})Z_s^j\Big)$ and $\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j\Big)$, where the empirical measure in the second term is the one constructed from the i.i.d. SDEs in (3.8), hence each X^j corresponds to a independent realisation of the MV-SDE, namely it has the correct distribution. Splitting the original difference into three, we use the Lipshitz property in space for the first one, to obtain,

$$\big|\hat{b}\Big(s, X_s^{i,N}, \frac{1}{N}\sum_{j=1}^N f(X_s^{j,N})Z_s^j\Big) - \hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^{j,N})Z_s^j\Big)\big|^2 \leq C|X_s^{i,N} - X_s^i|^2\,.$$

For the second difference we use the fact that $\hat{b}$ is bounded along with the Lipschitz property in the third variable, which yields

$$\begin{aligned}
&\big|\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^{j,N})Z_s^j\Big) - \hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j\Big)\big|^2 \\
&\leq C\big|\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^{j,N})Z_s^j\Big) - \hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j\Big)\big| \leq C\frac{1}{N}\sum_{j=1}^N Z_s^j|X_s^{j,N} - X_s^j|\,.
\end{aligned}$$

Finally, again from the Lipschitz property we obtain,

$$\big|\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j\Big) - \hat{b}\left(s, X_s^i, \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\right)\big|^2 \leq C\big|\frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j - \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\big|^2\,.$$

Hence the following bound holds,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\Big[\sup_{0\leq t\leq T}|X_t^{i,N} - X_t^i|^2\Big] \\
&\leq C\int_0^T \mathbb{E}_{\mathbb{Q}}[|X_s^{i,N} - X_s^i|^2] + \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbb{Q}}\Big[Z_s^j|X_s^{j,N} - X_s^j|\Big] + \mathbb{E}_{\mathbb{Q}}\Big[\big|\frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j - \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\big|^2\Big]\mathrm{d}s\,.
\end{aligned}$$

One can use Cauchy-Schwarz along with the properties of Z to obtain,

$$\mathbb{E}_{\mathbb{Q}}\Big[Z_s^j|X_s^{j,N} - X_s^j|\Big] \leq C\mathbb{E}_{\mathbb{Q}}\Big[|X_s^{j,N} - X_s^j|^2\Big]^{\frac{1}{2}} \leq C\mathbb{E}_{\mathbb{Q}}\Big[\sup_{0\leq u\leq s}|X_u^{j,N} - X_u^j|^2\Big]^{\frac{1}{2}}\,.$$

Although at first it appears one cannot use Gronwall here, there is a nonlinear generalisation due to Perov (see [MPF12, Theorem 1, p360]) which we can use since the nonlinear term on the RHS is square root of the term on the left. Finally, take the supremum over i and using the fact that the variables $f(X_s^j)Z_s^j$ are i.i.d. and square integrable, we obtain,

$$\begin{aligned}
\sup_{1\leq i\leq N}\mathbb{E}_{\mathbb{Q}}\Big[\sup_{0\leq t\leq T}|X_t^{i,N} - X_t^i|^2\Big] &\leq Ce^C\int_0^T \mathbb{E}_{\mathbb{Q}}\Big[\big|\frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j - \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\big|^2\Big]\mathrm{d}s \\
&\leq \frac{Ce^C}{N}\int_0^T \mathbb{E}_{\mathbb{Q}}\Big[\big|f(X_s^1)Z_s^1 - \mathbb{E}_{\mathbb{Q}}[f(X_s^1)Z_s^1]\big|^2\Big]\mathrm{d}s \to 0
\end{aligned}$$

as $N \to \infty$, which concludes the proof. □

The Complete Measure Change Algorithm

We now describe the algorithm for simulating a general MV-SDE under a complete measure change.

1. Simulate the following particle system for the MV-SDE after the measure change:

$$\begin{aligned}\mathrm{d}X_t^{i,N} &= \left(b\left(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^N Z_t^j \delta_{X_t^{j,N}}\right) + \dot{h}_t \sigma\left(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^N Z_t^j \delta_{X_t^{j,N}}\right)\right)\mathrm{d}t \\ &\quad + \sigma\left(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^N Z_t^j \delta_{X_t^{j,N}}\right)\mathrm{d}W_t^{i,\mathbb{Q}}, \\ \mathrm{d}Z_t^i &= \dot{h}_t Z_t^i \mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z_0^i = 1\,.\end{aligned}$$

2. Compute the importance sampling estimator using the following formula:

$$\hat{\theta}_h = \frac{1}{N}\sum_{i=1}^N Z_T^{i,N} G(X_T^{i,N}).$$

We will refer to algorithms of this form as *Complete Measure Change Algorithms*. An advantage one can immediately see is that one simulates the particles only once. A key disadvantage is that the importance sampling to estimate the object of interest $\mathbb{E}[G(X_T)]$, may yield a poorer estimation of the original law μ and the term $\mathbb{E}_{\mathbb{Q}}[f(X_t)Z_t]$ in (3.7). We will discuss this in Section 5.

4 Optimal Importance Sampling for McKean-Vlasov SDEs

The previous section detailed algorithms for simulating MV-SDEs under an arbitrary change of measure. We now want to use the theory of large deviations to determine, in a certain optimal way, a measure change which will reduce the variance of the estimate.

An important point here is that we will be using the LDP for Brownian motion, rather than the MV-SDEs. There are several works dealing with Large Deviations for MV-SDEs and their associated interacting particles systems, see [BDF12], [Fis14], [dRST17] but such results are not of use to us here since we must be able to (cheaply) simulate the SDE after the change of measure. We therefore restrict to the Girsanov measure change since we know how the SDE changes under the measure change.

In this section we first show how the LDP framework can be applied to both algorithms to yield a simplified optimisation problem for finging the asymptotically optimal measure change (Theorems 4.9 and 4.6) and then demonstrate how these simplified optimization problems may be solved in practice.

4.1 Preliminaries

We recall some of the main concepts for importance sampling with LDP, see [GR08] for further discussion. We denote by $\mathbb{W}_T^d$ the standard d-dimensional Wiener space of continuous functions over the time interval $[0,T]$ which are zero at time zero and in the one-dimensional case we simply write $\mathbb{W}_T$ instead of $\mathbb{W}_T^1$. This space is endowed with the topology of uniform convergence and with the usual Wiener measure $\mathbb{P}$, defined on the completed filtration $\mathcal{F}_T$, which makes the process $\mathbf{W}_t(x) = x_t$ with $x \in \mathbb{W}_T^d$ a standard d-dimensional Brownian motion.

The goal is to estimate the expected value of some functional $\tilde{G} : \mathbb{W}_T^d \to \mathbb{R}_+$ continuous in the uniform topology ($\tilde{G}$ is explained later). For the change of measure, one considers a Girsanov transform

where the allowed functions are from the Cameron-Martin space of absolutely continuous functions with square integrable derivative, i.e. (if $d = 1$ we just write $\mathbb{H}_T = \mathbb{H}_T^1$)

$$\mathbb{H}_T^d = \left\{ h : [0,T] \mapsto \mathbb{R}^d : \ h_0 = 0 \,, \ h_\cdot = \int_0^\cdot \dot{h}_t dt \,, \ \int_0^T |\dot{h}_t|^2 \, \mathrm{d}t < \infty \ \text{ i.e. } \ \dot{h}_t \in L_t^2(\mathbb{R}^d) \right\}.$$

For any deterministic drift $h \in \mathbb{H}_T^d$, the stochastic exponential defines the Radon-Nikodym derivative for an equivalent measure $\mathbb{Q}$ namely, ($W^{\mathbb{P}}$ is a standard $\mathbb{P}$-Brownian motion)

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^T \dot{h}_t \mathrm{d}W_t^{\mathbb{P}} - \frac{1}{2} \int_0^T (\dot{h}_t)^2 \, \mathrm{d}t \right).$$

Under this new measure $\mathbb{Q}$, the process $W_\cdot^{\mathbb{Q}} = W_\cdot^{\mathbb{P}} - h_\cdot$ is a standard d-dimensional $\mathbb{Q}$-Brownian motion.

Standing assumptions

We consider MV-SDEs with nonlinear interaction between the SDE and its law. In this section we concentrate on one-dimensional SDEs of the form,

$$\mathrm{d}X_t = b(t, X_t, \mu_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad X_0 = x_0. \tag{4.1}$$

Throughout this section we will refer to the following assumptions (similar to assumptions in Section 2), for functions $b : [0,T] \times \Omega \times \mathbb{R} \times \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ and $\sigma > 0$ constant.

Assumption 4.1. *Assume that b is Lipschitz in the sense that $\exists L > 0$ such that $\forall t \in [0,T]$, $\forall x, x' \in \mathbb{R}$ and $\forall \mu, \mu' \in \mathcal{P}_2(\mathbb{R})$ we have that*

$$|b(t,x,\mu) - b(t,x',\mu')| \leq L(|x - x'| + W^{(2)}(\mu, \mu')).$$

Moreover, $\forall\ x \in \mathbb{R}$ and $\mu \in \mathcal{P}_2(\mathbb{R})$, b is continuous over the interval $[0,T]$.

Assumption 4.2. *Assume b satisfies the monotone growth and local Lipschitz conditions in Assumption 2.1. Further, $\forall\ x \in \mathbb{R}$ and $\mu \in \mathcal{P}_2(\mathbb{R})$, let b be continuous in time over the interval $[0,T]$.*

In view of Section 2, either of these assumptions yield the existence of a unique strong solution to (4.1). We further use the following assumption for the terminal function G. Note that this assumption is on G as a function of the SDE, rather than the driving Brownian motion as is the case in [GR08].

Assumption 4.3. *The functional G is non-negative, continuous and satisfies the following growth condition*

$$\log(G(x)) \leq C_1 + C_2 \sup_{t \in [0,T]} |x_t|^\alpha \,,$$

for $x : [0,T] \mapsto \mathbb{R}$ a continuous function starting at x_0 where $C_1,\ C_2$ are positive constants and $\alpha \in [1,2)$.

4.2 The decoupling algorithm

We first consider the decoupling algorithm presented in Section 3.1. We build μ_t^N, from an independent N-particle system which is simulated under a numerical scheme, and then consider the following approximation of SDE[3] (4.1),

$$\mathrm{d}\overline{X}_t = b(t, \overline{X}_t, \mu_t^N)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad X_0 = x_0 \,. \tag{4.2}$$

[3]The measure, μ^N is a random measure, but is independent of the process $\overline{X}$ thus we have decoupled the SDE.

In order to distinguish the current SDE from the previous particle approximation we introduce a so-called copy space (see for example [BLPR17]) $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t)_{t\geq 0}, \tilde{\mathbb{P}})$ (with the usual conditions and $\tilde{\mathcal{F}}_t$ is the augmented filtration over the N-dimensional Brownian motion). The N-system SDEs used to approximate this measure is then defined on this space, hence (4.2) is defined on the product space $(\Omega, \mathcal{F}, \mathbb{P}) \otimes (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$.

Our aim is now to minimize over $h \in \mathbb{H}_T$ the variance conditional on the trajectory of μ^N:

$$\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\big[\, G(\overline{X}_T)^2 \mathcal{E}_T^{-1} \big| \tilde{\mathcal{F}}_T \big], \quad \mathrm{d}\mathcal{E}_t = \dot{h}_t \mathcal{E}_t \mathrm{d}W_t^{\mathbb{P}}, \quad \mathcal{E}_0 = 1,$$

and we make use of small noise asymptotics in order to write this variance in a "LDP" tractable form, hence we define, for $h \in \mathbb{H}_T$

$$L(h;\mu^N) := \limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}} \left[\exp\left(\frac{1}{\epsilon} \Big(2\log(\overline{G}(\sqrt{\epsilon}W)) - \int_0^T \sqrt{\epsilon} \dot{h}_t \mathrm{d}W_t + \frac{1}{2}\int_0^T \dot{h}_t^2 \mathrm{d}t \Big) \right) \Big| \tilde{\mathcal{F}}_T \right], \quad (4.3)$$

where $\overline{G}(W) := G(\overline{X}(W))$. One should also keep in mind that $\overline{G}$ also depends on μ^N, however, we suppress this notation for ease of presentation.

Remark 4.4. *In* (4.3)*, we have a conditional expectation, thus $L(h;\mu^N)$ is technically a random variable in $\tilde{\Omega}$. This is not typically the case when using Varadhan's lemma, however, because the random variable is independent of the Brownian motion and $\overline{G}$ is still $\tilde{\mathbb{P}}$-a.s. continuous w.r.t. the Brownian motion (Section 6.2), upon checking the moment condition, we are still able to use Varadhan's lemma, $\tilde{\mathbb{P}}$-a.s..*

This leads to the following natural definition of asymptotic optimality.

Definition 4.5 (Asymptotic optimality)**.** *Fix an $\tilde{\omega} \in \tilde{\Omega}$ (and hence a μ^N). A change of measure with parameter $h \in \mathbb{H}_T$ is said to be* asymptotically optimal w.r.t. μ^N *if it is a solution to the problem,*

$$\min_{h\in\mathbb{H}_T} L(h;\mu^N) \quad \tilde{\mathbb{P}}\text{-a.s.}\,.$$

One of the main advantages of arguing under this LDP framework is that with μ^N fixed the optimal change of drift is a deterministic function that is relatively cheap to compute and use. The idea of an (optimal) stochastic measure change, which is more computationally expensive to execute, is explored in [DW04].

Theorem 4.6. *Let Assumptions 4.3 and 4.2 hold and fix $\tilde{\omega} \in \tilde{\Omega}$. Furthermore assume that there exists $u \in \mathbb{H}_T$ such that $\overline{G}(u) > 0$. Then the following statements hold:*

i. Let $h \in \mathbb{H}_T$ such that $\dot{h}$ is of finite variation. Then Varadhan's lemma holds for the small noise asymptotics, namely we can rewrite (4.3) *as,*

$$L(h;\mu^N) = \sup_{u\in\mathbb{H}_T} \left\{ 2\log(\overline{G}(u)) - \int_0^T \dot{h}_t \dot{u}_t \mathrm{d}t + \frac{1}{2}\int_0^T \dot{h}_t^2 \mathrm{d}t - \frac{1}{2}\int_0^T \dot{u}_t^2 \mathrm{d}t \right\} \quad \tilde{\mathbb{P}}\text{-a.s.}\,. \quad (4.4)$$

ii. There exists an asymptotically optimal parameter h^* which minimizes (4.4).

iii. Consider a simplified optimization problem

$$\sup_{u\in\mathbb{H}_T} \left\{ 2\log(\overline{G}(u)) - \int_0^T \dot{u}_t^2 \mathrm{d}t \right\}. \quad (4.5)$$

There exists a maximizer h^{**} for this problem. If

$$L(h^{**};\mu^N) = 2\log(\overline{G}(h^{**})) - \int_0^T (\dot{h}_t^{**})^2 \mathrm{d}t\,, \quad (4.6)$$

then h^{**} is asymptotically optimal and is the unique maximizer of (4.5).

All of these results are $\tilde{\mathbb{P}}$-a.s. since the particle system yields a random measure from $\tilde{\Omega}$. The proof of this theorem requires several auxiliary results which we defer to Section 6.2. One should also note that the requirement for $\overline{G} > 0$ for some u is not restrictive, it is purely there for technical reasons since one cannot have a maximiser if $\log(\overline{G}(u)) = -\infty$ for all $u \in \mathbb{H}_T$. The assumption that $\dot{h}$ has finite variation is necessary to establish the continuity of the functional in Varadhan's lemma.

Remark 4.7 (Concavity of $\log(\overline{G})$ and asymptotic optimality)**.** *Consider the problem of minimizing* (4.4) *and assume that one can interchange the inf and the sup. Then,*

$$\begin{aligned}\inf_{h\in\mathbb{H}_T} L(h;\mu^N) &= \sup_{u\in\mathbb{H}_T}\inf_{h\in\mathbb{H}_T}\left\{2\log(\overline{G}(u)) - \int_0^T \dot{h}_t\dot{u}_t\mathrm{d}t + \frac{1}{2}\int_0^T \dot{h}_t^2\mathrm{d}t - \frac{1}{2}\int_0^T \dot{u}_t^2\mathrm{d}t\right\}\\ &= \sup_{u\in\mathbb{H}_T}\left\{2\log(\overline{G}(u)) - \int_0^T \dot{u}_t^2\mathrm{d}t\right\}\end{aligned}$$

because the inner problem is solved by $h = u$. Therefore, a sufficient condition for asymptotic optimality of h^{**} is the exchangeability of inf and sup above. Since L is a convex function in h, and the integral terms in (4.4) *are concave in u, a sufficient condition for such exchangeability is that $\log(\overline{G})$ is concave. Indeed, in the case of convex-concave functions we can invoke the minimax principle to swap infimum and supremum, see* [ET99, pg. 175] *for example.*

In [GR08]*, the process X was a geometric Brownian Motion and the authors were able to explicitly link the concavity of $\log(\overline{G})$ with the properties of the function G. Here the dependence of $\overline{G}$ on the Brownian motion is more complex, and it appears to be difficult to check concavity. Hence, in general one has to check numerically whether* (4.6) *holds. However, even if* (4.6) *fails, one can still use h^{**} to construct a candidate importance sampling measure if this is justified by superior numerical performance.*

4.3 The complete measure change algorithm

Here we focus on the algorithm discussed in Section 3.2. Recall that we are interested in evaluating, $\mathbb{E}_{\mathbb{P}}[G(X)]$. We now change the measure to $\mathbb{Q}$ and calculate the variance,

$$\mathrm{Var}_{\mathbb{Q}}\Big[G(X)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] = \mathbb{E}_{\mathbb{P}}\Big[G(X)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] - \mathbb{E}_{\mathbb{P}}\Big[G(X)\Big]^2.$$

Minimising the variance is equivalent to minimize the first term in the RHS. As a first step to constructing a tractable proxy for this variance we consider a particle approximation of X:

$$\mathrm{d}X_t^{i,N} = b\left(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}\right)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{P}}, \quad X_0^{i,N} = x_0\,, \tag{4.7}$$

$$\mathrm{d}\mathcal{E}_t^i = \dot{h}_t\mathcal{E}_t^i\mathrm{d}W_t^{i,\mathbb{P}}, \quad \mathcal{E}_0^i = 1, \tag{4.8}$$

where $W^{i,\mathbb{P}}$ denotes the driving $\mathbb{P}$-Brownian motion of particle i, and all $W^{i,\mathbb{P}}$s are independent of each other. We approximate $\mathbb{E}_{\mathbb{P}}[G^2(X)(\mathcal{E}_T)^{-1}]$ with $\mathbb{E}_{\mathbb{P}}[G^2(X^{i,N})(\mathcal{E}_T^{i,N})^{-1}]$. Since $\mathcal{E}^i = \mathcal{E}^{i,N}$ (due to the absence of cross dependency), one can equivalently minimize

$$\mathbb{E}_{\mathbb{P}}\big[\, G^2(X^{i,N})(\mathcal{E}_T^i)^{-1}\,\big], \text{ over all } h \in \mathbb{H}_T. \tag{4.9}$$

In order to use the LDP theory to minimize (4.9), we define $\tilde{G}$ as the functional dependent on the underlying $\mathbb{P}$-Brownian motions, i.e., for all $i \in \{1, \dots, N\}$, $\tilde{G}_i : \mathbb{W}_T^N \mapsto \mathbb{R}$, where, $\tilde{G}_i(W^1, \dots, W^N) := G(X^{i,N}(W^1, \dots, W^N))$. The corresponding small noise asymptotics takes the following form:

$$\bar{L}(h) := \limsup_{\epsilon \to 0} \epsilon \log \left(\mathbb{E}_{\mathbb{P}} \left[\exp \left(\frac{1}{\epsilon} \Big(2 \log \big(\tilde{G}_i \big(\sqrt{\epsilon} W^1, \dots, \sqrt{\epsilon} W^N \big) \big) \right. \right. \right. \\ \left. \left. \left. - \int_0^T \sqrt{\epsilon} \dot{h}_t \mathrm{d}W_t^i + \frac{1}{2} \int_0^T (\dot{h}_t)^2 \mathrm{d}t \Big) \right) \right] \right), \quad h \in \mathbb{H}_T \tag{4.10}$$

where we remark that the value of this expression does not depend on the choice of i. We may also define an optimality result here

Definition 4.8 (Asymptotic optimality)**.** *A change of measure with parameter $h \in \mathbb{H}_T$ is said to be* asymptotically optimal *if it is a solution to the problem,*

$$\min_{h \in \mathbb{H}_T} \bar{L}(h) \, .$$

We then obtain the following result for $\bar{L}$ (compare with Theorem 4.6).

Theorem 4.9. *Fix $N \in \mathbb{N}$ and let Assumptions 4.3 and 4.1 hold. Assume that there exists $(u^1, \hat{u}) \in \mathbb{H}_T^2$ such that $\tilde{G}_1(u^1, \hat{u}, \dots, \hat{u}) > 0$. Then the following statements hold*

i. Let $h \in \mathbb{H}_T$ such that $\dot{h}$ is of finite variation. Then Varadhan's lemma holds for the small noise asymptotics and we can rewrite (4.10) *as*

$$\bar{L}(h) = \sup_{u \in \mathbb{H}_T^N} \left\{ 2 \log(\tilde{G}_1(u^1, \dots, u^N)) - \int_0^T \dot{h}_t \dot{u}_t^1 \mathrm{d}t + \frac{1}{2} \int_0^T (\dot{h}_t)^2 \mathrm{d}t - \frac{1}{2} \int_0^T |\dot{u}_t|^2 \mathrm{d}t \right\} , \tag{4.11}$$

ii. There exists an asymptotically optimal parameter h^* which minimizes (4.11).

- *Consider a simplified optimization problem*

$$\sup_{u^1 \in \mathbb{H}_T, \hat{u} \in \mathbb{H}_T} \left\{ 2 \log(\tilde{G}_1(u^1, \hat{u}, \dots, \hat{u})) - \int_0^T (\dot{u}_t^1)^2 \mathrm{d}t - \frac{N-1}{2} \int_0^T \dot{\hat{u}}_t^2 \mathrm{d}t \right\} . \tag{4.12}$$

There exists a maximizer (h^{**}, u^{**}) for this problem. If

$$\bar{L}(h^{**}) = 2 \log \left(\tilde{G}_1(h^{**}, u^{**}, \dots, u^{**}) \right) - \int_0^T (\dot{h}_t^{**})^2 \mathrm{d}t - \frac{N-1}{2} \int_0^T (\dot{u}_t^{**})^2 \mathrm{d}t \, . \tag{4.13}$$

then h^{**} is asymptotically optimal and is the unique maximizer of (4.12)*, where we have taken $i = 1$ without loss of generality.*

The proof of this theorem is deferred to Section 6.1. Similarly to the previous discussion if $\log(\tilde{G}_1)$ is a concave function in $u \in \mathbb{H}_T^N$, then we know that (4.13) holds (this is discussed at the end of Section 6.1). However, in general (4.13) is difficult to check since, even with h^* fixed, $\bar{L}$ is still an N-dimensional optimisation problem.

There is also a difficulty in quantifying how the measure change affects the propagation of chaos error i.e. a measure change that is good for the statistical error may be damaging to the propagation of chaos error. We discuss this point further in Section 5.

4.4 Computing the optimal measure change

The exponential form of the SDEs (the log-normal class) considered in [GR08] and [Rob10] allows the maximisation to be written in the form of an Euler-Lagrange equation (calculus of variations approach). Due to the more general coefficients here, we obtain a more complex interaction between the Brownian motion and the value of the SDE. Consequently we need to look towards the more general theory of optimal control to calculate the change of measure[4]. Deterministic optimal control is a large subject area and one can consult [FR75] or [YZ99] for example. We recall that we are working under the $\mathbb{P}$-measure.

One of the most important results from optimal control is Pontryagin's maximum principle. Roughly speaking, Pontryagin's maximum principle gives a set of differential equations that the optimal control must satisfy. Let us recall the main ideas following [YZ99, p.102]. We start with the controlled dynamical system $x(t)$ which takes the following form:

$$\begin{cases} \dot{x}(t) = b(t, x(t), u(t)), \quad \text{a.e. } t \in [0, T] \\ x(0) = x_0 \,, \end{cases} \tag{4.14}$$

where u is our "control", which is defined in a metric space (U, d) and associated to this we have a *cost functional*

$$J(u(\cdot)) = \int_0^T f(t, x(t), u(t))\mathrm{d}t + h(x(T)) \,, \tag{4.15}$$

f is typically referred to as the *running cost* and h the *terminal cost*. We then have the following assumption.

Assumption 4.10. *For ease of writing we denote by $\varphi(t, x, u)$ to be any of the functions $b(t, x, u)$, $f(t, x, u)$ or $h(x)$. We then assume the following,*

- *(U, d) is a separable metric space and $T > 0$.*
- *The maps $b : [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}^n$, $f : [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are measurable and there exists a constant $L > 0$ and a modulus of continuity $\eta : [0, \infty) \to [0, \infty)$ such that,*

$$\begin{cases} |\varphi(t, x, u) - \varphi(t, \hat{x}, \hat{u})| \le L|x - \hat{x}| + \eta(d(u, \hat{u})) & \forall t \in [0, T] \; x, \hat{x} \in \mathbb{R}^n, \; u, \hat{u} \in U \,, \\ |\varphi(t, 0, u)| \le L & \forall (t, u) \in [0, T] \times U \,. \end{cases}$$

- *The maps b, f and h are C^1 in x and there exists a modulus of continuity $\eta : [0, \infty) \to [0, \infty)$ such that,*

$$|\partial_x \varphi(t, x, u) - \partial_x \varphi(t, \hat{x}, \hat{u})| \le \eta\Big(|x - \hat{x}| + d(u, \hat{u})\Big) \quad \forall t \in [0, T] \; x, \hat{x} \in \mathbb{R}^n, \; u, \hat{u} \in U \,.$$

As discussed in [YZ99, p.102], Assumption 4.10 implies that (4.14) admits a unique solution and (4.15) is well defined. Let us denote by $\mathcal{U}[0, T] := \{u(\cdot) : [0, T] \to U \mid u \text{ is measurable}\}$, then optimal control problem is to find $u^* \in \mathcal{U}[0, T]$ that satisfies,

$$J(u^*) = \inf_{u \in \mathcal{U}[0,T]} J(u) \,. \tag{4.16}$$

Such u^* is referred to as an *optimal control*, and the corresponding $x^*(\cdot) := x(\cdot; u^*)$ the *optimal state trajectory*. We can then state the deterministic version of Pontryagin's maximum principle as [YZ99, p.103].

[4] Even though we are initially dealing with SDEs, in the large deviations asymptotics, the trajectory of the Brownian motion becomes a deterministic control.

Theorem 4.11. *[Pontryagin's Maximum Principle] Let Assumption 4.10 hold and let (x^*, u^*) be the optimal pair to* (4.16)*. Then, there exists a function $p : [0,T] \to \mathbb{R}^n$ satisfying the following,*

$$\begin{cases} \dot{p}(t) = -\partial_x b(t, x^*(t), u^*(t))^\intercal p(t) + \partial_x f(t, x^*(t), u^*(t)), & \text{a.e. } t \in [0,T] \\ p(T) = -\partial_x h(x^*(T))\,, \end{cases} \tag{4.17}$$

and

$$H(t, x^*(t), u^*(t), p(t)) = \max_{u \in U}\{H(t, x^*(t), u, p(t))\} \quad \text{a.e. } t \in [0,T]\,,$$

where

$$H(t,x,u,p) := \langle p, b(t,x,u)\rangle - f(t,x,u) \quad (t,x,u,p) \in [0,T] \times \mathbb{R}^n \times U \times \mathbb{R}^n\,.$$

Typically p is referred to as the *adjoint function* and (4.17) the *adjoint equation*, and the function H is called the *Hamiltonian*.

Remark 4.12 (An alternative approach)**.** *The maximum principle is not the only way one can use to solve this problem. An alternative is by solving the so-called Hamilton-Jacobi-Bellman (HJB) equation. This approach is typically more difficult since the HJB is a* partial differential equation*.*

Maximum principle for Theorems 4.6 and 4.9. The maximum principle allows to translate the simplified optimization problems of Theorems 4.6 and 4.9 into boundary value problems for ODE. One can observe that we are actually interested in $\dot{u}$ rather than u, that is, in the decoupled case we can write the controlled dynamics as

$$X_t(\dot{u}) = x_0 + \int_0^t b(s, X_s(\dot{u}), \mu_s^N)\mathrm{d}s + \int_0^t \sigma \dot{u}_s \mathrm{d}s\,.$$

The theory above is for infimum while we are interested in supremum, therefore we use the fact that $\sup\{f\} = -\inf\{-f\}$.

▷ *For the decoupling algorithm* Theorem 4.11 yields the following equations for the adjoint function and trajectory under optimal control $\dot{u}^*$ (for a given μ^N),

$$\text{(Decoupled)} \quad \begin{cases} \dot{p}_t = -\partial_x b(t, X_t(\dot{u}^*), \mu_t^N) p_t\,, & p_T = \dfrac{2G'(X(\dot{u}^*))}{G(X(\dot{u}^*))}\,, \\ \dot{X}_t = b(t, X_t, \mu_t^N) + \frac{1}{2}\sigma^2 p_t\,, & X_0 = x_0\,, \end{cases} \tag{4.18}$$

that is, the optimal control is related to p through, $\dot{u}_t^* = \frac{1}{2}\sigma p_t$.

▷ *For the complete measure change algorithm* the argument is similar argument to the above one, but here we also need to deal with the measure term. Noting that we have two controls to optimise over (recall Theorem 4.9) we obtain more complex expressions. Theorem 4.11 yields the following system of ODEs,

$$\text{(Complete)} \quad \begin{cases} \dot{p}_t^1 = -\partial_{X^1} b(t, X_t^1, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) p_t^1 - \partial_{X^1} b(t, \hat{X}_t, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) p_t^2\,, & p_T^1 = \frac{2G'(X^1)}{G(X^1)}\,, \\ \dot{p}_t^2 = -\partial_{\hat{X}} b(t, X_t^1, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) p_t^1 - \partial_{\hat{X}} b(t, \hat{X}_t, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) p_t^2\,, & p_T^2 = 0\,, \\ \dot{X}_t^1 = b(t, X_t^1, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) + \frac{1}{2}\sigma^2 p_t^1\,, & X_0^1 = x_0\,, \\ \dot{\hat{X}}_t = b(t, \hat{X}_t, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) + \frac{1}{2(N-1)}\sigma^2 p_t^2\,, & \hat{X}_0 = x_0\,, \end{cases} \tag{4.19}$$

similarly we obtained, $\dot{u}_t^* = \frac{1}{2}\sigma p_t$ and $\dot{\hat{u}}_t^* = 0$ as the optimal controls. From Theorem 4.9 we obtain the measure change as $\dot{h} = \dot{u}$.

The difference between (4.18) and (4.19) comes from the fact that for the complete measure change we have a higher dimensional problem. That is, we have two controls and two "SDEs" thus we have more terms to optimise. Recall, when one wishes to assess asymptotic optimality, (4.11) is still an N-dimensional problem.

Remark 4.13 (Accuracy of Change of Measure)**.** *In* [GR08]*, they were able to obtain explicit solutions in certain situations, but here, due to the increase in complexity, we expect this to rarely be the case. We therefore need to set reasonable tolerances in checking whether asymptotic optimality holds.*

5 Example: Kuramoto model

The Kuramoto model is a special case of a so-called system of coupled oscillators. Such models are of particular interest in physics and are used to study many different phenomena such as active rotator systems, charge density waves and complex biological systems amongst other things, see [KŁSG02] for further details. The corresponding SDE for the Kuramoto model is

$$\mathrm{d}X_t = \left(K \int_{\mathbb{R}} \sin(y - X_t)\mu_{t,\mathbb{P}}^X(\mathrm{d}y) - \sin(X_t) \right) \mathrm{d}t + \sigma \mathrm{d}W_t^{\mathbb{P}}\,, \quad t \in [0,T], \quad X_0 = x_0\,,$$

where K is the coupling strength and σ has the physical interpretation of the temperature in the system. We consider a terminal condition $G(x) = a\exp(bx)$ (satisfying Assumption 4.3). Our goal is to obtain the asymptotically optimal change of measure that improves the estimation of $\mathbb{E}_{\mathbb{P}}[G(\bar{X}_T)]$.

One can see that such a model easily satisfies the assumptions required in the paper. Let us now apply the theory from the previous section to calculate the optimal change of measure. We should point out here that we do not have the concavity required for asymptotic optimality to hold automatically, therefore we need to check this condition.

By our previous discussion, to apply the decoupling algorithm here we would generate a set of N weakly interacting SDEs which we denote by $Y^{i,N}$ and approximate the original SDE by,

$$\mathrm{d}\bar{X}_t = \left(\frac{K}{N} \sum_{i=1}^{N} \sin(Y_t^{i,N} - \bar{X}_t) - \sin(\bar{X}_t) \right) \mathrm{d}t + \sigma \mathrm{d}W_t^{\mathbb{P}}\,, \quad t \in [0,T], \quad \bar{X}_0 = x_0\,.$$

Let us now apply the theory from the previous section to calculate the optimal change of measure. Our optimal control argument implies solving $\tilde{\mathbb{P}}$-a.s.

$$\text{(Decoupled)} \quad \begin{cases} \dot{p}_t = \left(\frac{K}{N} \sum_{i=1}^{N} \cos(Y_t^{i,N} - X_t) + \cos(X_t) \right) p_t\,, & p_T = 2b\,, \\ \dot{X}_t = \left(\frac{K}{N} \sum_{i=1}^{N} \sin(Y_t^{i,N} - X_t) - \sin(X_t) \right) + \frac{1}{2}\sigma^2 p_t\,, & X_0 = x_0\,. \end{cases}$$

The complete measure change algorithm yields the following system,

$$\text{(Complete)} \quad \begin{cases} \dot{p}_t^1 = K\big(\frac{N-1}{N}\cos(\hat{X}_t - X_t^1) - \cos(X_t^1)\big)p_t^1 - \frac{K}{N}\cos(X_t^1 - \hat{X}_t)p_t^2\,, & p_T^1 = 2b\,, \\ \dot{p}_t^2 = -K\frac{N-1}{N}\cos(\hat{X}_t - X_t^1)p_t^1 + K\big(\frac{1}{N}\cos(X_t^1 - \hat{X}_t) + \cos(\hat{X}_t)\big)p_t^2\,, & p_T^2 = 0\,, \\ \dot{X}_t^1 = K\big(\frac{N-1}{N}\sin(\hat{X}_t - X_t^1) - \sin(X_t^1)\big) + \frac{1}{2}\sigma^2 p_t^1\,, & X_0^1 = x_0\,, \\ \dot{\hat{X}}_t = K\big(\frac{1}{N}\sin(X_t^1 - \hat{X}_t) - \sin(\hat{X}_t)\big) + \frac{1}{2(N-1)}\sigma^2 p_t^2\,, & \hat{X}_0 = x_0\,, \end{cases}$$

To show the numerical advantages one can achieve by using importance sampling we consider how the time taken and the estimate given by the algorithms change with the number of particles N.

For this example we use, $T = 1$, $\bar{X}_0 = 0$, $K = 1$, $\sigma = 0.3$, $a = 0.5$ and $b = 10$. For the numerics we use an Euler scheme with step size of $\Delta t = 0.02$. The systems of equations are solved using MATLAB's `bvp4c` function. For the importance sampling, we use the particle positions from the first Monte Carlo simulation as the empirical law.

	Monte Carlo			Decoupled			Complete		
N	Payoff	Error	Time	Payoff	Error	Time	Payoff	Error	Time
1×10^3	1.5066	0.1490	3	1.5729	0.0028	9	1.5419	0.0024	3
5×10^3	1.5895	0.0626	27	1.5840	0.0013	54	1.5710	0.0013	28
1×10^4	1.6813	0.0693	76	1.5728	0.0009	153	1.5860	0.0009	75
5×10^4	1.5899	0.0200	1 025	1.5820	0.0004	2 052	1.5738	0.0004	1 062
1×10^5	1.5807	0.0176	3 433	1.5731	0.0003	6 935	1.5882	0.0003	3 644

Table 5.1: Results from standard Monte Carlo and the importance sampling algorithms. Time is measured in seconds and error refers to square root of the variance.

We recall that the decoupling importance sampling requires two runs, here we use the same N for both of these. The first note one can make is how the time scales when increasing the number of particles, namely one can truly observe the N^2 complexity[5]. As expected the decoupling algorithm takes approximately twice as long as the standard Monte Carlo (computing the change of measure is not time consuming). Following this point we also observe that the complete measure change has roughly the same computational complexity as standard Monte Carlo. The other key point is the reduction in variance (standard error) one obtains with importance sampling. For this example we see that both importance sampling schemes reduce the variance by several orders of magnitude. Further, if one is interested in the decoupling algorithm it may be more efficient to take less simulations in the second importance sampled run. Finally, we checked the asymptotic optimality (for the decoupling) numerically and there is only a small difference between the two sides in (4.13), we therefore believe we are close to the optimal. Table 5.1 does show that the use of importance sampling in MV-SDEs is both viable and worthwhile.

▷ *Estimating the propagation of chaos error.* As was mentioned in the introduction, theoretically the statistical error and the propagation of chaos error converge to zero at the same rate. We now use this example to show that the statistical error dominates. Since the Euler scheme is the same in all examples we can neglect the bias caused by that. We can then decompose the error as

$$\frac{1}{N}\sum_{i=1}^{N} G(\bar{X}^{i,N}) - \mathbb{E}_{\mathbb{P}}[G(\bar{X}^1)] = \frac{1}{N}\sum_{i=1}^{N} G(\bar{X}^{i,N}) - \mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1,N})] + \mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1,N})] - \mathbb{E}_{\mathbb{P}}[G(\bar{X}^1)]\,.$$

The first difference on the RHS is the statistical error, and the second one is the propagation of chaos error. It is then clear that if one considers M realisations of $\frac{1}{N}\sum_{i=1}^{N} G(\bar{X}^{i,N})$ and takes the average this approximates $\mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1,N})]$ but does not change the propagation of chaos error. Hence for large M the error reduces to the propagation of chaos error. To show the propagation of chaos error is negligible compared to the statistical error here, we repeat the simulation for $N = 5 \times 10^3$ particles, $M = 10^3$ times and we obtain an average terminal value of 1.5772 (with an average standard error of 0.06533, which agrees with the result in Table 5.1). Comparing this to the 10^5 decoupled entry (which has almost no statistical error) in Table 5.1, we can conclude the propagation of chaos error at least an order of magnitude smaller than the statistical error.

[5]Even if one is able to optimize the code somewhat, with this method we cannot escape the extra complexity arising from the particle interaction.

Another example: a terminal condition function with steep slope

Let us consider the terminal condition $G(x) = \big(\tanh(a(x-b))+1\big)/2$, for a large (G can be understood as a mollified indicator function). Then $\mathbb{E}_{\mathbb{P}}[G(X_T)] \approx \mathbb{P}(X_T \geq b)$. We take the same set up as before but with $a = 15$ and $b = 1$ and note that the terminal condition for adjoint takes the form,

$$p_T = 2a\Big(1 - \tanh\big(a\big(X_T(\dot{u}^*) - b\big)\big)\Big).$$

We obtain the following table (we omit the times here since they are similar).

	Monte Carlo		Decoupled		Complete	
N	Payoff (10^{-9})	Error (10^{-9})	Payoff (10^{-9})	Error (10^{-9})	Payoff (10^{-9})	Error (10^{-9})
1×10^3	1.015	0.671	3.864	0.0250	8.456	0.101
5×10^3	1.093	0.752	3.952	0.0112	5.564	0.0185
1×10^4	8.829	7.071	3.910	0.0077	32.956	0.1520
5×10^4	1.106	0.271	3.970	0.0035	2.101	0.0024
1×10^5	5.158	1.990	3.901	0.0024	16.781	0.019

Table 5.2: Results from standard Monte Carlo and the importance sampling algorithms. Note that for ease of presentation the payoff and error are all scaled to be 10^{-9} of the values presented.

The results in Table 5.2 highlight the key differences in the algorithms. Clearly this is a difficult problem for standard Monte Carlo to solve. The reason of course being that although G is mollified it still changes value quickly over a small interval. For example G at 0.25 is approximately 10^{-10}, but $G(0.5) \approx 10^{-7}$ and $G(0.75) \approx 10^{-4}$, hence a reasonably small change in the value of the SDE can influence the outcome significantly. However, for the standard Monte Carlo run, only 60 of the $100,000$ were $> 1/2$ at the terminal time and none were above $3/4$. Hence standard Monte Carlo is not giving much information about the most important region of the function.

The importance sampling schemes again give reduced errors, however, this example highlights the differences between them. Although the complete measure change does have a smaller error than standard Monte Carlo the payoff oscillates around and hence the decoupled algorithm appears to be superior since the payoffs are consistent and the error decreases in the expected manner.

▷ *Robustness of complete measure change.* The above table shows why one has to consider the effect of the measure change on the propagation of chaos error. The reason this is more prominent here than in the previous example is because the magnitude of the optimal measure change is far larger. Hence, even when we use a large number of particles they may provide a poor approximation of the law, this is where this algorithm lacks robustness.

Remark 5.1 (Requirement for improved simulation)**.** *It is clear from these examples that combining importance sampling with MV-SDEs can provide a major reduction in the required computational cost, namely we can achieve a smaller variance with far less simulations (and hence time). When using decoupling, unfortunately one has to approximate the law first, which is computationally expensive to do using a particle approximation. Hence, one may look towards more sophisticated simulation techniques to speed up the first run, for example* [GP15] *or towards multilevel Monte Carlo such as* [STT17]*. However, with the ability to almost eliminate the variance one should always keep in mind the benefits from importance sampling.*

6 Proof of Main Results

We now provide the proofs of our two main theorems. Throughout we work under the $\mathbb{P}$-measure and we omit it as a superscript in our Brownian motions. Some arguments align with those of [GR08] and we quote them where appropriate.

6.1 Proofs for Theorem 4.9

Continuity of the SDE w.r.t. Brownian motion is key as it allows to apply directly the contraction principle transferring Schilder's LDP for the Brownian motion to an LDP for the solution of the SDE; otherwise difficulties would arise when using Varadhan's lemma. Unlike the decoupled case, we will stick to Lipschitz coefficients here, the reason for this is that Lemma 6.3 does not generalise well for SDEs of the type (4.7).

Lemma 6.1. *Fix $N \in \mathbb{N}$, let Assumption 4.1 hold and let $X \in \mathbb{S}^p$ for $p \geq 2$ denote the N-dimensional strong solution to the SDE system defined in* (4.7)*. Then X is continuous w.r.t. the set of N Brownian motions in the uniform topology.*

Proof. To show continuity in the uniform topology we consider two sets of iid Brownian motions, $W_t = (W_t^1, \dots, W_t^N)$ and $\tilde{W}_t = (\tilde{W}_t^1, \dots, \tilde{W}_t^N)$ and show continuity by analyzing the difference between, $\tilde{X}_t^i := X_t^i(\tilde{W}_t^1, \dots, \tilde{W}_t^N)$ and X_t^i with $i \in \{1, \cdots, N\}$. We have,

$$|\tilde{X}_t^{i,N} - X_t^{i,N}| \leq \int_0^t |b(s, \tilde{X}_s^{i,N}, \frac{1}{N}\sum_{j=1}^N \delta_{\tilde{X}_s^{j,N}}) - b(s, X_s^{i,N}, \frac{1}{N}\sum_{j=1}^N \delta_{X_s^{j,N}})|\mathrm{d}s + |\int_0^t \sigma \mathrm{d}\tilde{W}_s^i - \int_0^t \sigma \mathrm{d}W_s^i|\,.$$

Considering the time integral first, we can bound as follows,

$$\Big|b(s, \tilde{X}_s^{i,N}, \frac{1}{N}\sum_{j=1}^N \delta_{\tilde{X}_s^{j,N}}) - b(s, X_s^{i,N}, \frac{1}{N}\sum_{j=1}^N \delta_{X_s^{j,N}})\Big| \leq C\Big(|\tilde{X}_s^{i,N} - X_s^{i,N}| + \Big(\frac{1}{N}\sum_{j=1}^N (\tilde{X}_s^{j,N} - X_s^{j,N})^2\Big)^{\frac{1}{2}}\Big)\,,$$

where we used the Lipschitz property and the definition of the Wasserstein-2 metric for empirical distributions (see [BJGR17], for example). Noting that for the second term,

$$\Big(\frac{1}{N}\sum_{j=1}^N (\tilde{X}_s^{j,N} - X_s^{j,N})^2\Big)^{\frac{1}{2}} \leq \max_{j\in\{1,\dots,N\}} |\tilde{X}_s^{j,N} - X_s^{j,N}| \leq \sum_{j=1}^N |\tilde{X}_s^{j,N} - X_s^{j,N}|\,.$$

Hence we can bound the drift by terms of the form $|\tilde{X}_s^{j,N} - X_s^{j,N}|$. This yields the following,

$$|\tilde{X}_t^{i,N} - X_t^{i,N}| \leq \int_0^t C\Big(|\tilde{X}_s^{i,N} - X_s^{i,N}| + \sum_{j=1}^N |\tilde{X}_s^{j,N} - X_s^{j,N}|\Big)\mathrm{d}s + C\sup_{0\leq s\leq t}|\tilde{W}_s^i - W_s^i|\,.$$

Taking supremums and summing over i on both sides yields,

$$\begin{aligned}\sum_{i=1}^N \sup_{0\leq t\leq T} |\tilde{X}_t^{i,N} - X_t^{i,N}| \leq& \int_0^T C\sum_{i=1}^N \sup_{0\leq t\leq s} |\tilde{X}_t^{i,N} - X_t^{i,N}|\mathrm{d}s + C\sum_{i=1}^N \sup_{0\leq s\leq T}|\tilde{W}_s^i - W_s^i| \\ \leq& Ce^{CT}\sum_{i=1}^N \sup_{0\leq s\leq T}|\tilde{W}_s^i - W_s^i|\,,\end{aligned}$$

where the final step follows from Gronwall's inequality. It is then clear that $\sum_{i=1}^N \sup_{0\leq s\leq T}|\tilde{W}_s^i - W_s^i| \to 0$ implies $\sum_{i=1}^N \sup_{0\leq t\leq T}|\tilde{X}_t^{i,N} - X_t^{i,N}| \to 0$, hence we obtain the required continuity. □

We next show that one can use Varadhan's lemma in this case.

Lemma 6.2. *Fix $N \in \mathbb{N}$, let $h \in \mathbb{H}_T$ and let Assumptions 4.3 and 4.1 hold.*
Then the integrability condition in Varadhan's lemma holds for (4.10)*. Namely for $\gamma > 1$*

$$\limsup_{\epsilon\to 0} \epsilon \log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\gamma}{\epsilon}\left(2\log\left(\tilde{G}_1(\sqrt{\epsilon}W^1, \dots, \sqrt{\epsilon}W^N)\right)\int_0^T \sqrt{\epsilon}\dot{h}_t \mathrm{d}W_t^1 + \frac{1}{2}\int_0^T (\dot{h}_t)^2\mathrm{d}t\right)\right)\right]\right) < \infty.$$

Proof. Using that $h \in \mathbb{H}_T$ is deterministic, $\dot{h} \in L^2([0,T],\mathbb{R}^N)$ and Cauchy-Schwarz we obtain,

$$\begin{aligned}\epsilon \log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\gamma}{\epsilon}\Big(2\log(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)) - \int_0^T \sqrt{\epsilon}\dot{h}_t \mathrm{d}W_t^1 + \frac{1}{2}\int_0^T (\dot{h}_t)^2 \mathrm{d}t\Big)\right)\right]\right) \\ \leq \frac{\gamma}{2}\int_0^T (\dot{h}_t)^2 \mathrm{d}t + \frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{4\gamma}{\epsilon}\log(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N))\right)\right]\right) \\ + \frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(-\frac{2\gamma}{\epsilon}\int_0^T \sqrt{\epsilon}\dot{h}_t \mathrm{d}W_t^1\right)\right]\right).\end{aligned}$$

It is then sufficient to show that the three terms are finite when we take $\limsup_{\epsilon\to 0}$. The first term is clearly finite by the conditions on h. Finiteness of the third term follows from [GR08, pg.16], namely $\forall\ i \in \{1,\dots,N\}$ the stochastic integral has the distribution $\int_0^T \dot{h}_t \mathrm{d}W_t^i \sim \mathcal{N}(0, \int_0^T (\dot{h}_t)^2 \mathrm{d}t)$. Thus we obtain,

$$\limsup_{\epsilon\to 0} \frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(-\frac{2\gamma}{\epsilon}\int_0^T \sqrt{\epsilon}\dot{h}_t \mathrm{d}W_t^1\right)\right]\right) = \gamma^2 \int_0^T (\dot{h}_t)^2 \mathrm{d}t < \infty\,.$$

The final term to consider is the terminal term, $\log(\tilde{G}_1)$. By definition of $\tilde{G}_1$ and Assumption 4.3 we have,

$$\log\left(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)\right) \leq C_1 + C_2 \sup_{0\leq t\leq T} |X^{1,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|^\alpha\,.$$

Applying similar arguments as in Lemma 6.1 we obtain

$$\begin{aligned}&|X_t^{1,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)| \\ &\quad\leq C + \int_0^t C\Big(|X_s^{1,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)| + \sum_{j=1}^N |X_s^{j,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|\Big)\mathrm{d}s + C\sqrt{\epsilon}\sup_{0\leq s\leq t}|W_s^1|\,.\end{aligned}$$

Noting that for $\alpha \geq 1$ and a_i nonnegative, $(\sum_{i=1}^N a_i)^\alpha \leq C^\alpha \sum_{i=1}^N a_i^\alpha$ and that the above estimate is true for any $X^{i,N}$, then taking supremums and summing over i yields,

$$\begin{aligned}&\sum_{i=1}^N \sup_{0\leq t\leq T} |X_t^{i,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|^\alpha \\ &\qquad\leq C^\alpha + \int_0^T C^\alpha \sum_{i=1}^N \sup_{0\leq t\leq s} |X_t^{i,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|^\alpha \mathrm{d}s + C^\alpha \sqrt{\epsilon}^\alpha \sum_{i=1}^N \sup_{0\leq s\leq T}|W_s^i|^\alpha \\ &\qquad\leq C^\alpha e^{C^\alpha}\sqrt{\epsilon}^\alpha \sum_{i=1}^N \sup_{0\leq s\leq T}|W_s^i|^\alpha\,,\end{aligned}$$

where the final line comes from Gronwall's inequality. It is useful for us to note this yields the bound

$$\log\left(\tilde{G}_1(W^1,\dots,W^N)\right) \leq C_1 + C_2 \sum_{i=1}^N \sup_{0\leq s\leq T}|W_s^i|^\alpha\,. \tag{6.1}$$

Using the previous results we have the following bound,

$$\begin{aligned}&\frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{4\gamma}{\epsilon}\log(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N))\right)\right]\right) \\ &\qquad\leq C + \sum_{i=1}^N \frac{\epsilon C}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{4\gamma C^\alpha}{\epsilon^{1-\alpha/2}}\sup_{0\leq s\leq T}|W_s^i|^\alpha\right)\right]\right),\end{aligned}$$

where we have used the independence of the Brownian motions to obtain the sum over i.

Finiteness of this term then follows by arguments similar to those in Lemma 7.6 and 7.7 in [GR08]. To conclude, we have shown that all terms are finite and the result follows. □

Before finishing the proof of Theorem 4.9, we note that the LDP for Brownian motion in pathspace is given by Schilder's theorem, which states that for a d-dimensional Brownian motion W, then $\sqrt{\epsilon}W$ satisfies a LDP with good rate function (see [DZ10]),

$$I(y) = \begin{cases} \frac{1}{2}\int_0^T |\dot{y}_t|^2 \mathrm{d}t & , \text{ if } y \in \mathbb{H}_T^d\,, \\ \infty & , \text{ if } y \in \mathbb{W}_T^d \backslash \mathbb{H}_T^d\,. \end{cases}$$

Proof of Theorem 4.9. The continuity of the SDE from Lemma 6.1 along with existence of a unique strong solution under Assumptions 4.1, ensure $\tilde{G}_1$ is a continuous function under Assumption 4.3. By assumption, there exists a point $(u^1, \hat{u}) \in \mathbb{H}_T^2$ such that $\tilde{G}(u^1, \hat{u}, \dots, \hat{u}) > 0$, this along with (6.1) and recalling $\alpha < 2$ we obtain the existence of maximisers by Lemma 7.1 of [GR08]. Similarly the $+\dot{h}^2$ yields existence of a minimising h for $\bar{L}$.

Moreover, continuity of $\tilde{G}$ w.r.t. the Brownian motion and finite variation of $\dot{h}$ implies the exponential term in (4.10) is continuous. Thus to use Varadhan's lemma we only need to check the integrability condition, which is given in Lemma 6.1, hence relation (4.11) follows.

The remaining part to be proved is that (4.13) implies asymptotic optimality. This essentially relies on showing that (4.12) provides a lower bound for L. To see this note that

$$\begin{aligned} \inf_{h\in\mathbb{H}_T} \bar{L}(h) &= \inf_{h\in\mathbb{H}_T} \sup_{u\in\mathbb{H}_T^N} \left\{ 2\log(\tilde{G}_1(u^1,\dots,u^N)) + \frac{1}{2}\int_0^T (\dot{h}_t - \dot{u}_t^1)^2 \mathrm{d}t - \frac{1}{2}\int_0^T (\dot{u}_t^1)^2 \mathrm{d}t - \frac{1}{2}\int_0^T |\dot{u}_t|^2 \mathrm{d}t \right\} \\ &\geq \sup_{u\in\mathbb{H}_T^N} \inf_{h\in\mathbb{H}_T} \left\{ 2\log(\tilde{G}_1(u^1,\dots,u^N)) + \frac{1}{2}\int_0^T (\dot{h}_t - \dot{u}_t^1)^2 \mathrm{d}t - \frac{1}{2}\int_0^T (\dot{u}_t^1)^2 \mathrm{d}t - \frac{1}{2}\int_0^T |\dot{u}_t|^2 \mathrm{d}t \right\}, \end{aligned}$$

where the inequality follows from the relation $\inf_a \sup_b\{\cdot\} \geq \sup_b \inf_a\{\cdot\}$. We have equality when the minimax result holds. Clearly the minimising h is $\dot{h}_\cdot = \dot{u}_\cdot$. It is then clear that the above supremum is bounded below by the case $u^2 = \cdots = u^N$, which yields the expression (4.12).

Strict convexity along with arguments on page 18 in [GR08] yields the uniqueness which completes the proof. □

6.2 Proofs for Theorem 4.6

We recall, that due to the independence of the original particle system from the SDE in question, we work on the product of two probability spaces, consequently (since μ^N will be a "realisation" coming from the space $\tilde{\Omega}$) our results are all $\tilde{\mathbb{P}}$-a.s..

As before we need to prove that the SDE is a continuous map of the Brownian motions. We were unable to find any results for the monotone (one-sided Lipschitz), locally Lipschitz case, we therefore provide a proof of this result here (Lemma 6.4). The proof of this relies on the following lemma.

Lemma 6.3. *Let Assumption 4.2 hold and let $\bar{X}$ be the solution to* (4.2)*. Then consider the following stochastic processes*

$$\begin{aligned} X_t^+ &:= x_0 \mathbb{1}_{\{x_0\geq 0\}} + \int_0^t C(|X_s^+| + 1)\mathrm{d}s + \sigma\left(\sup_{0\leq s\leq t} W_s - \inf_{0\leq s\leq t} W_s\right), \\ X_t^- &:= x_0 \mathbb{1}_{\{x_0\leq 0\}} - \int_0^t C(|X_s^-| + 1)\mathrm{d}s + \sigma\left(\inf_{0\leq s\leq t} W_s - \sup_{0\leq s\leq t} W_s\right), \end{aligned}$$

where C is the constant in the monotone condition of b.

Then, $\forall\, t \geq 0$, $X_t^- \leq \bar{X}_t \leq X_t^+$, $\mathbb{P} \otimes \tilde{\mathbb{P}}$-a.s..

Proof. Firstly, one can easily show through a standard Picard iteration argument that both $X^{\pm}$ have unique, progressively measurable solutions in $\mathbb{S}^2$. We argue by contradiction and show the upper bound $\bar{X} \leq X^+$, the lower bound follows by the same argument in the opposite direction. Since b is monotone (Assumption 4.2), we can derive the following bounds $\forall\, s \in [0,T]$ and $\mu \in \mathcal{P}_2(\mathbb{R})$,

$$b(s,x,\mu) \leq C(|x|+1) \quad \text{for } x \geq 0 \qquad \text{and} \qquad b(s,x,\mu) \geq -C(|x|+1) \quad \text{for } x \leq 0\,.$$

Assume that there exists a time t_2 such that $\bar{X}_{t_2} > X_{t_2}^+$. If $\bar{X}_t \geq 0$ for all $t \in [0,t_2]$, then,

$$X_{t_2}^+ - \bar{X}_{t_2} = x_0 \mathbb{1}_{\{x_0 \geq 0\}} - x_0 + \int_0^{t_2} C(|X_s^+|+1) - b(s,\bar{X}_s,\mu_s^N)\mathrm{d}s + \sigma\left(\sup_{0\leq s\leq t_2} W_s - \inf_{0\leq s\leq t_2} W_s\right) - \sigma W_{t_2} \geq 0,$$

which yields a contradiction. Alternatively, let $t_1 := \max\{t \leq t_2 : \bar{X}_t = 0\}$. By continuity, $\bar{X}_{t_1} = 0$ and so

$$\begin{aligned} X_{t_2}^+ - \bar{X}_{t_2} = x_0 \mathbb{1}_{\{x_0\geq 0\}} + \int_0^{t_2} C(|X_s^+|+1)\mathrm{d}s - \int_{t_1}^{t_2} b(s,\bar{X}_s,\mu_s^N)\mathrm{d}s \\ + \sigma\left(\sup_{0\leq s\leq t_2} W_s - \inf_{0\leq s\leq t_2} W_s\right) - \sigma\left(W_{t_2} - W_{t_1}\right) \geq 0, \end{aligned}$$

which contradicts $\bar{X}_{t_2} > X_{t_2}^+$ and thus proves the result. □

One can now use this lemma to prove the following result.

Lemma 6.4. *Let $\bar{X}$ be defined as in* (4.2)*, with coefficients satisfying Assumption 4.2, then $\bar{X}$ is a $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s. continuous map of Brownian motion in the uniform norm.*

Proof. To prove this result we require that, if $\sup_{0\leq s\leq t}|\tilde{W}_s - W_s| \to 0$, then $\sup_{0\leq s\leq t}|\bar{X}_s(\tilde{W}) - \bar{X}_s(W)| \to 0$. We note that we work on the uniform topology and hence we may assume that all (a finite number of) Brownian motions are uniformly bounded on $[0,T]$. Lemma 6.3, implies that we can bound the value $\bar{X}$ takes by the processes $X_\cdot^{\pm}$. It is a straightforward application of Gronwall's Lemma to deduce,

$$\begin{aligned} X_t^+ &\leq \left(x_0\mathbb{1}_{\{x_0\geq 0\}} + Ct + \sigma\left(\sup_{0\leq s\leq t} W_s - \inf_{0\leq s\leq t} W_s\right)\right)e^{Ct}\,, \\ X_t^- &\geq -\left(|x_0\mathbb{1}_{\{x_0\leq 0\}}| + Ct + \sigma\Big|\inf_{0\leq s\leq t} W_s - \sup_{0\leq s\leq t} W_s\Big|\right)e^{Ct}\,. \end{aligned}$$

Hence we can bound the value $\bar{X}$ can take as a function of its Brownian motion (which itself is bounded by the uniform topology). Let us now consider the difference in the SDEs driven by the different Brownian motions,

$$|\bar{X}_t(\tilde{W}) - \bar{X}_t(W)| \leq \int_0^t |b(s,\bar{X}_s(\tilde{W}),\mu_s^N) - b(s,\bar{X}_s(W),\mu_s^N)|\mathrm{d}s + \sigma|\tilde{W}_t - W_t|\,.$$

By Assumption 4.2, b is locally Lipschitz, hence,

$$|b(s,\bar{X}_s(\tilde{W}),\mu_s^N) - b(s,\bar{X}_s(W),\mu_s^N)| \leq C(\tilde{W},W)|\bar{X}_s(\tilde{W}) - \bar{X}_s(W)|\,.$$

Noting further that $\sigma|\tilde{W}_t - W_t| \leq \sigma \sup_{0\leq s\leq t}|\tilde{W}_s - W_s|$, then by Gronwall's inequality we obtain,

$$|\bar{X}_t(\tilde{W}) - \bar{X}_t(W)| \leq \sigma \left(\sup_{0\leq s\leq t}|\tilde{W}_s - W_s|\right) e^{C(\tilde{W},W)t}\,.$$

Again, by the uniform topology, we must have $\tilde{W}$ and W bounded, thus $C(\tilde{W},W) < \infty$ and hence, $\sup_{0\leq s\leq t}|\bar{X}_s(\tilde{W}) - \bar{X}_s(W)| \to 0$ when $\sup_{0\leq s\leq t}|\tilde{W}_s - W_s| \to 0$.

□

We now prove that the uniform integrability condition still holds, namely that we can still apply Varadhan's Lemma, in both settings.

Lemma 6.5. *Let $h \in \mathbb{H}_T$, then under Assumption 4.3 and 4.2 the integrability condition in Varadhan's lemma holds for* (4.3)*. Namely, for some $\gamma > 1$*

$$\limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{\gamma}{\epsilon}\left(2\log(\overline{G}(\sqrt{\epsilon}W)) - \int_0^T \sqrt{\epsilon}\dot{h}_t \mathrm{d}W_t + \frac{1}{2}\int_0^T \dot{h}_t^2 \mathrm{d}t\right)\right)\Big|\tilde{\mathcal{F}}\right] < \infty \quad \tilde{\mathbb{P}}\text{-a.s.}.$$

Proof. The h terms can be dealt with using the same arguments as before. The term we are interested in is the G term. Using arguments as in the proof of Lemma 6.2, we only need to prove the following holds,

$$\limsup_{\epsilon\to 0} \frac{\epsilon}{2} \log\left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{4\gamma}{\epsilon}\log\left(G(\bar{X}(\sqrt{\epsilon}W))\right)\right)\Big|\tilde{\mathcal{F}}\right]\right) < \infty\,.$$

Recall that Lemma 6.3, yields the bound, $X_t^- \leq \bar{X}_t \leq X_t^+$, $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s.. Hence, for $\alpha \in [1,2)$ we have the following bound $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s.,

$$\sup_{0\leq t\leq T}|\bar{X}_t|^\alpha \leq \sup_{0\leq t\leq T}|X_t^+|^\alpha + \sup_{0\leq t\leq T}|X_t^-|^\alpha = |X_T^+|^\alpha + |X_T^-|^\alpha\,,$$

where the final equality comes from the fact $|X^\pm|$ are nondecreasing processes. Due to the dependence on the external measure μ^N, all of these results are $\tilde{\mathbb{P}}$-a.s., but for ease of presentation we will omit it here. Further recall that by Gronwall's lemma (or see proof of Lemma 6.4), we can bound the processes $|X^\pm|$, thus,

$$|X_T^+|^\alpha \leq C^\alpha\left(x_0^\alpha \mathbb{1}_{\{x_0\geq 0\}} + C^\alpha + \sigma^\alpha\left(\sup_{0\leq s\leq T} W_s - \inf_{0\leq s\leq T} W_s\right)^\alpha\right)e^{C\alpha}\,,$$
$$|X_T^-|^\alpha \leq C^\alpha\left(|x_0\mathbb{1}_{\{x_0\leq 0\}}|^\alpha + C^\alpha + \sigma^\alpha\Big|\inf_{0\leq s\leq T} W_s - \sup_{0\leq s\leq T} W_s\Big|^\alpha\right)e^{C\alpha}\,.$$

Due to the fact that $\alpha \geq 1$, and $-\inf_{0\leq s\leq T} W_s = \sup_{0\leq s\leq T} -W_s \geq 0$, we have,

$$\Big|\inf_{0\leq s\leq T} W_s - \sup_{0\leq s\leq T} W_s\Big|^\alpha = \left(\sup_{0\leq s\leq T} W_s - \inf_{0\leq s\leq T} W_s\right)^\alpha \leq C^\alpha\left(\left(\sup_{0\leq s\leq T} W_s\right)^\alpha + \left(\sup_{0\leq s\leq T} -W_s\right)^\alpha\right)\,.$$

We express the bound w.r.t. the driving Brownian motion $\sqrt{\epsilon}W$ and obtain,

$$\sup_{0\leq t\leq T}|\bar{X}_t(\sqrt{\epsilon}W)|^\alpha \leq C^\alpha\left(|x_0|^\alpha + C^\alpha + C^\alpha\sigma^\alpha\sqrt{\epsilon}^\alpha\left(\left(\sup_{0\leq s\leq T} W_s\right)^\alpha + \left(\sup_{0\leq s\leq T} -W_s\right)^\alpha\right)\right)e^{C\alpha}\,.$$

We can simplify this further by noting,

$$\left(\sup_{0\leq s\leq T} W_s\right)^\alpha + \left(\sup_{0\leq s\leq T} -W_s\right)^\alpha \leq C^\alpha \sup_{0\leq s\leq T} |W_s|^\alpha .$$

Using these inequalities we obtain,

$$\begin{aligned}
&\frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{4\gamma}{\epsilon}\log(G(\bar{X}(\sqrt{\epsilon}W)))\right)\Big|\tilde{\mathcal{F}}\right]\right)\\
&\quad\leq \frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{4\gamma}{\epsilon}C_1 + \frac{4\gamma}{\epsilon}C_2\left(C^\alpha\left(|x_0|^\alpha + C^\alpha + C^\alpha\sigma^\alpha\sqrt{\epsilon}^\alpha \sup_{0\leq s\leq T}|W_s|^\alpha\right)e^{C\alpha}\right)\right)\Big|\tilde{\mathcal{F}}\right]\right) .
\end{aligned}$$

By splitting up the terms in the exponential this then reduces to the problem of considering,

$$\frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{4\gamma}{\epsilon^{1-\alpha/2}}C_2C^\alpha\sigma^\alpha \sup_{0\leq s\leq T}|W_s|^\alpha\right)\Big|\tilde{\mathcal{F}}\right]\right) .$$

One can show that this quantity is finite by following the same arguments as [GR08, pg.16]. □

We can now prove the second main theorem, the arguments follow similar lines to those we used to conclude the proof of Theorem 4.9.

Proof of Theorem 4.6. The continuity of the SDE from Lemma 6.4 along with existence of a unique strong solution under Assumption 4.2, ensure $\overline{G}$ is a $\tilde{\mathbb{P}}$-a.s. continuous function under Assumption 4.3. We then obtain the existence of the maximiser by Lemma 7.1 of [GR08].

Moreover, the $\tilde{\mathbb{P}}$-a.s. continuity of $\overline{G}$ w.r.t. the Brownian motion and finite variation of $\dot{h}$ implies that to use Varadhan's lemma we only need to check the integrability condition, which is given in Lemma 6.5. This with Lemma 7.6 in [GR08] is enough to complete the proof by arguments on page 18 in [GR08]. □

References

[BDF12] A. Budhiraja, P. Dupuis, and M. Fischer, *Large deviation properties of weakly interacting processes via weak convergence methods*, Ann. Probab. **40** (2012), no. 1, 74–102.

[BJGR17] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, *Inference in generative models using the Wasserstein distance*, arXiv:1701.05146 (2017).

[BLPR17] R. Buckdahn, J. Li, S. Peng, and C. Rainer, *Mean-field stochastic differential equations and associated PDEs*, The Annals of Probability **45** (2017), no. 2, 824–878.

[Bos04] M. Bossy, *Optimal rate of convergence of a stochastic particle method to solutions of 1d viscous scalar conservation laws*, Mathematics of computation **73** (2004), no. 246, 777–812.

[BT97] M. Bossy and D. Talay, *A stochastic particle method for the Mckean-Vlasov and the Burgers equation*, Mathematics of Computation of the American Mathematical Society **66** (1997), no. 217, 157–192.

[Car16] R. Carmona, *Lectures of BSDEs, stochastic control, and stochastic differential games with financial applications*, SIAM, 2016.

[CM17] D. Crisan and E. McMurray, *Cubature on Wiener space for McKean–Vlasov SDEs with smooth scalar interaction*, 2017. arXiv:1703.04177.

[DE11] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, Vol. 902, John Wiley & Sons, 2011.

[DG87] D. A. Dawson and J. Gärtner, *Large deviations from the McKean-Vlasov limit for weakly interacting diffusions*, Stochastics **20** (1987), no. 4, 247–308.

[dRST17] G. dos Reis, W. Salkeld, and J. Tugaut, *Freidlin-Wentzell LDPs in path space for McKean-Vlasov equations and the functional iterated logarithm law*, 2017. arXiv:1708.04961.

[DW04] P. Dupuis and H. Wang, *Importance sampling, large deviations, and differential games*, Stochastics: An International Journal of Probability and Stochastic Processes **76** (2004), no. 6, 481–508.

[DZ10] A. Dembo and O. Zeitouni, *Large deviations techniques and applications, volume 38 of stochastic modelling and applied probability*, Springer-Verlag, Berlin, 2010.

[ET99] I. Ekeland and R. Temam, *Convex analysis and variational problems*, SIAM, 1999.

[Fis14] M. Fischer, *On the form of the large deviation rate function for the empirical measures of weakly interacting systems*, Bernoulli **20** (2014), no. 4, 1765–1801.

[FR75] W. H. Fleming and R. W. Rishel, *Deterministic and stochastic optimal control*, Springer-Verlag, Berlin-New York, 1975. Applications of Mathematics, No. 1.

[GHS99] P. Glasserman, P. Heidelberger, and P. Shahabuddin, *Asymptotically optimal importance sampling and stratification for pricing path-dependent options*, Mathematical finance **9** (1999), no. 2, 117–152.

[GP15] E. Gobet and S. Pagliarani, *Analytical approximations of BSDEs with nonsmooth driver*, SIAM Journal on Financial Mathematics **6** (2015), no. 1, 919–958.

[GR08] P. Guasoni and S. Robertson, *Optimal importance sampling with explicit formulas in continuous time*, Finance and Stochastics **12** (2008), no. 1, 1–19.

[KHO97] A. Kohatsu-Higa and S. Ogawa, *Weak rate of convergence for an Euler scheme of nonlinear SDE's*, Monte Carlo Methods and Applications **3** (1997), 327–345.

[KŁSG02] M. Kostur, J. Łuczka, and L. Schimansky-Geier, *Nonequilibrium coupled Brownian phase oscillators*, Physical Review E **65** (2002), no. 5, 051115.

[MPF12] D. S. Mitrinovic, J. Pecaric, and A. M. Fink, *Inequalities involving functions and their integrals and derivatives*, Vol. 53, Springer Science & Business Media, 2012.

[Rob10] S. Robertson, *Sample path large deviations and optimal importance sampling for stochastic volatility models*, Stochastic Processes and their applications **120** (2010), no. 1, 66–83.

[STT17] L. Szpruch, S. Tan, and A. Tse, *Iterative particle approximation for Mckean-Vlasov SDEs with application to multilevel Monte Carlo estimation*, 2017. ArXiv:1706.00907.

[TFC16] H.-W. Teng, C.-D. Fuh, and C.-C. Chen, *On an automatic and optimal importance sampling approach with applications in finance*, Quantitative Finance **16** (2016), no. 8, 1259–1271.

[YZ99] J. Yong and X. Y. Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*, Vol. 43, Springer Science & Business Media, 1999.