Applying standard Markov chain Monte Carlo (MCMC) algorithms to large data sets is computationally expensive. Both the calculation of the acceptance probability and the creation of informed proposals usually require an iteration through the whole data set. The recently proposed stochastic gradient Langevin dynamics (SGLD) method circumvents this problem in three ways: it generates proposals which are only based on a subset of the data, it skips the accept-reject step and it uses sequences of decreasing step-sizes. In [20], we provided the mathematical foundations for the decreasing step size SGLD for example, by establishing a central limit theorem. However, the step size of the SGLD and its extensions is in practice not decreased to zero. This is the focus of the present article. We characterise the asymptotic bias explicitly and its dependence on the step size and the variance of the stochastic gradient. On that basis we derive a modified SGLD which removes the bias due to the variance of the stochastic gradient up to first order in the step size. Moreover, we are able to obtain bounds on finite time bias and the MSE. The theory is illustrated with a Gaussian toy model for which the bias and the MSE for the estimation of moments can be obtained explicitly. For this toy model we study the gain of the SGLD over the standard Euler method in the limit of data items $N \to \infty$.

Keywords: Markov Chain Monte Carlo, Langevin Dynamics, Big Data, Fixed step size

1 Introduction

In many applications, for example in molecular dynamics [10] [12] and throughout statistics [12], one is interested in estimating expectations with respect to the invariant measure of a stochastic differential equation (SDE) by running a numerical scheme that approximates its time dynamics. In particular, numerical trajectories are used for the construction of an empirical measure either by averaging one single long trajectory or by averaging over many realisations to obtain a finite ensemble average (see for example [12]).

An alternative strategy, in order to avoid issues related to the bias of the numerical procedure, is to use it as a proposal [3] for Markov chain Monte Carlo (MCMC) algorithms.
However, for many modern machine learning applications such as topic modelling [8] standard MCMC algorithms become computationally expensive. This is because, both the calculation of the acceptance probability and the creation of informed proposals usually require an iteration through the whole data set.

Consider for example the following stochastic differential equation (SDE) for $\theta \in \mathbb{R}^d$

$$d\theta(t) = \frac{1}{2} \nabla \log \pi(\theta(t)) dt + dW_t,$$

where $W_t$ is a d-dimensional standard Brownian motion. Under appropriate assumptions on $\pi(\theta)$, it is possible to show that the dynamics generated by (1) are ergodic with respect to $\pi(\theta)$. We consider the case that $\pi$ arises in a Bayesian inference problem for $\theta$ with prior density $\pi_0(\theta)$ and i.i.d. data $X_i$ with likelihood $\pi(X_i|\theta)$. In this case, we have the following equation

$$\nabla \log \pi(\theta) = \nabla \log \pi_0(\theta) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log \pi(X_{\tau_i}|\theta).$$

Any of the traditional approaches described above would require an evaluation of the log-likelihood through the whole data set. The recently proposed stochastic gradient Langevin dynamics (SGLD) [21] circumvents this problem by generating proposals which are only based on a subset of the data, by skipping the accept-reject step and by using a decreasing step-size sequence $(\delta_k)_{k \geq 0}$. In particular one has

$$\theta_{k+1} = \theta_k + \delta_k \left( \nabla \log \pi_0(\theta_k) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log \pi(X_{\tau_i}|\theta_k) \right) + \sqrt{\delta_k} \xi_k,$$

where $\xi_k$ is a standard Gaussian random variable on $\mathbb{R}^d$, and $\tau = (\tau_1, \cdots, \tau_n)$ is a random subset of $[N] := \{1, \cdots, N\}$, generated, for example, by sampling with or without replacement from $[N]$. The idea behind this algorithm is that, since the quantity appearing in front of $\delta_k$ in (2) is an unbiased estimator of $\nabla \log \pi(\theta)$, the limiting dynamics ($k \to \infty$) of (2) should behave similarly to the case $n = N$. This case has been investigated in [20]. We have shown that in this case the step size weighted sample average is consistent and satisfies a CLT with rate depending on the decay of $\delta_m$. This rate is limited asymptotically by $m^{-\frac{3}{2}}$ and achieved by an asymptotic step size decay $\sim m^{-\frac{3}{2}}$.

In practice, the SGLD and its extensions, the Stochastic Gradient Hamiltonian Monte Carlo [4] and the Stochastic Gradient Thermostat Monte Carlo [6] use step sizes that are only decreased up to a point. Thus, we set $\delta_k = h$ and analyse the behaviour of (2) in the limit of $k \to \infty$. In particular, when $n = N$, the dynamics generated by (2) are ergodic with respect to the numerical invariant measure $\pi_h(\theta; N)$ which can in general be shown [11] to be $O(h)$ apart from the true invariant measure $\pi(\theta)$. We assume for now that the dynamics generated by (2) remain ergodic and we denote the invariant measure in this case by $\pi_h(\theta; n)$. The question that thus arises here is what can be said about the distance between $\pi_h(\theta; n)$ and $\pi(\theta)$. In particular, the interest lies in the difference of the expectation with respect to $\pi_h(\theta; n)$ and the posterior.

This is the first question we tackle in this paper. In particular, using the recent generalizations [1] of the approach by Talay and Tubaro [19] we are able to derive an expansion of the asymptotic bias in powers of the step size that allows us, additionally to the work in [18], to identify explicitly the contribution of the unbiased estimator. This is very important
as in turn allows us to construct a modification of the SGLD algorithm, the mSGLD, which asymptotically in $h$ is equivalent to the Euler method where all the points are taken into account in the calculation of the likelihood. This modification is different from the Stochastic Gradient Fisher Scoring, a modification of the additive noise of the discretisation of the Langevin SDE in order to match the Bernstein von Mises posterior [17]. In contrast, the mSGLD is a local modification based on the estimated variance of the stochastic gradient.

The expansion of the asymptotic bias is also complemented by bounds on the mean square error of the finite time average of the following form

$$
E\left(\frac{1}{K} \sum_{k=0}^{K-1} \phi(\theta_k) - \pi(\phi)\right) \leq C(n) \left( h^2 + \frac{1}{Kh} \right),
$$

where the RHS only depends on $n$ through the constant. This is achieved by extending the work in [11] from $\mathbb{T}^d$ to $\mathbb{R}^d$. The main difficulty in achieving this relates with combining the results of [15] and [20], in order to establish the existence of nice, well controlled solutions to the corresponding Poisson equation [11]. Having such an expression for the mean square error also allows us to establish a relation between decreasing and non decreasing step-size and the order of the numerical method and its convergence rate. In particular, for the SGLD we are able to show that in terms of the decay of the mean square error the fixed step size SGLD and the decreasing step size SGLD are equivalent.

Our theoretical findings are confirmed by our numerical simulations. More precisely, we start by studying a one dimensional Gaussian toy model both in terms of the asymptotic bias and the mean square error of time averages. The simplicity of the model allows us to obtain explicit expressions for these quantities and thus illustrate in a clear way the connection with the theory. More precisely, we confirm the scaling of the step size and number of steps for a prescribed MSE obtained from the upper bound in Equation (3) matches the scaling derived from the analytic expression for the MSE for this toy model. More importantly, this simplicity allows us to make significant analytic progress in the study of the asymptotic bias and mean square error of time averages in the limit of big data $N \to \infty$. In particular, we are able to show that the SGLD reduces the computational complexity by one order of magnitude for the estimation of the second moment in comparison with the Euler method if the error is quantified through the MSE. Having obtained all this useful insight from this simple problem we then proceed by studying a problem for a Bayesian regression model, observing similar quantitative behaviour for the SGLD and mSGLD.

The rest of the paper is organised as follows. In Section 2 we review some known results about the effect of the numerical discretisation of (1) in terms of the finite time weak error as well as in terms of the invariant measure approximation. In Section 3 we apply these results to analyse the finite and long time properties of the SGLD, as well as to construct a new method which asymptotically, in $h$, behaves exactly as the Euler method ($n = N$) while we still sub-sample the data set. Furthermore, in Section 4 we discuss the properties of the finite time sample averages, which we discuss further in terms of a Gaussian toy model in Section 5 and in terms of a Bayesian logistic regression model in Section 6. Finally, we conclude this paper by Section 7 with a discussion on some possible extensions of this work.

SJV and YWT gratefully acknowledge EPSRC for research funding through grant EP/K009362/1. Moreover, the authors thank Rémi Bardenet for fruitful discussions.
2 Review of Weak Order Results

In Section 2.1, we describe some preliminary results related to ergodicity of SDEs and their numerical approximations, using the standard framework of the Fokker-Planck and backward Kolmogorov equations. In Section 2.2, we discuss a global error expansion for the weak error, while Section 2.3 is devoted to a discussion of the bias of the numerical ergodic averages with respect to the invariant measure for a general numerical method.

2.1 Exact and Numerical Invariant Measures for SDEs

Let us denote by $\rho(\theta,t)$ the probability density of $\theta(t) \in \mathbb{E}$ defined by (1) with initial condition $\theta_0 = \theta$. Then, we have

$$
E(\phi(\theta(t))|\theta_0 = \theta) = \int_{\mathbb{E}} \phi(y)\rho(y,t)dy,
$$

where $\rho(y,t)$ is the solution of the Fokker-Planck equation

$$
\frac{\partial \rho}{\partial t} = L^* \rho,
\rho(y,0) = \rho_0(y),
$$

with $\rho_0(y) = \delta(y - \theta)$ (Dirac mass) for the deterministic initial condition of $\theta(t)$ and $L^*$ is given by

$$
L^* \rho = -\nabla_y \cdot (\nabla \log \pi(y) \rho) + \frac{1}{2} \nabla_y \cdot \nabla_y \cdot \rho.
$$

This operator is the $L^2$-adjoint of

$$
L := \nabla_{\theta} \log \pi(\theta) \cdot \nabla_{\theta} + \frac{1}{2} \Delta_{\theta},
$$

the generator of the Markov process $\theta(t)$ defined by (1). Moreover, we consider

$$
u(\theta,t) = E(\phi(\theta(t))|\theta_0 = \theta),
$$

where $\theta(t)$ is the solution of (1). We note that $u(\theta,t)$ is the solution of the backward Kolmogorov equation

$$
\frac{\partial u}{\partial t} = Lu,
u(\theta,0) = \phi(\theta).
$$

A formal Taylor series expansion for $u$ in terms of the generator operator $L$ of the Markov process has been derived in \[22\]. This has been made rigorous for $\theta \in \mathbb{T}^d$ in \[5\] and recently for $\theta \in \mathbb{R}^d$ in \[9\]. The Taylor series is of the following form

$$
u(\theta,h) - \phi(\theta) = \sum_{j=1}^{l} \frac{h^j}{j!} L^j \phi(\theta) + h^{l+1} r_l(\pi,\phi)(\theta),
$$

for all positive integers $l$, with the remainder satisfying a bound of the form $|r_l(\pi,\phi)(\theta)| \leq C_l(1 + |\theta|^\kappa_l)$ for some constants $C_l, \kappa_l$. 


Remark 2.1. One way to turn $u(\theta, h) = \phi(\theta) + h\mathcal{L}\phi + \frac{h^2}{2}\mathcal{L}^2\phi + \cdots$ into a rigorous expansion into $E = T^d$ as it has been done in [5]. Another way is to follow the approach in [19, Lemma 2] and to assume that $\log \pi$ is $C^\infty$ with bounded derivatives of any order. This fact, together with the assumption that $|\phi(\theta)| \leq C(1 + |\theta|^s)$ for some positive integer $s$ is enough to prove that the solution $u$ of (9) has derivatives of any order that have a polynomial growth of the form (11), with other constants $C, s$ that are independent of $t \in [0, T]$.

In turn, these regularity bounds establish that (10) holds.

Now assume that one solves (1) with a one step integrator of the form

$$\theta_{n+1} = \Psi(\theta_n, h, \xi_n),$$

(12)

where $h$ denotes the time step and $\xi_n$ are independent random variables. Then using this formulation we can define

$$U(\theta, h) = \mathbb{E}(\phi(\theta_1)|\theta_0 = \theta),$$

(13)

for the expectation of numerical solution at time $h$, where again for simplicity we assume that the initial condition $\theta_0$ is deterministic and the following regularity and consistency assumption on the integrator, which is easily satisfied by any reasonable numerical method.

Assumption 2.1. Let $\nabla \log \pi$ be $C^\infty$ with bounded derivatives of any order. We assume for all deterministic initial conditions $X_0$ that

$$|\mathbb{E}(\theta_1 - \theta_0)| \leq C(1 + |\theta_0|)h, \quad |\theta_1 - \theta_0| \leq M(1 + |\theta_0|)\sqrt{h},$$

(14)

where $C$ is independent of $h$ small enough and $M$ is a random variable that has bounded moments of all orders independent of $h$ and $\theta_0$. Let us also assume that (13) has a weak Taylor series expansion of the form

$$U(\theta, h) = \phi(\theta) + hA_0(\pi)\phi(\theta) + h^2A_1(\pi)\phi(\theta) + \ldots,$$

(15)

where $A_i(\pi), i = 0, 1, 2, \ldots$ are linear differential operators with coefficients depending smoothly on the drift function $\nabla \log \pi(\theta)$, and its derivatives (depending on the choice of the integrator). In addition, we assume that $A_0(\pi)$ coincides with the generator $\mathcal{L}$ given in (7), which means that the method giving rise to (13) has (at least) weak order one,

$$A_0(\pi) = \mathcal{L}.$$

(16)

Assumption 2.1 immediately implies that we have the rigorous expansion

$$U(\theta, h) = \phi(\theta) + \sum_{i=0}^{l} h^{i+1}A_i(\pi)\phi(\theta) + h^{l+2}R_l(\pi, \phi)(\theta)$$

(17)

for all positive integers $l$, with a remainder satisfying $|R_l(\pi, \phi)(\theta)| \leq C_l(1 + |\theta|^k)$.

In addition using Assumption 2.1 we have that if the numerical solution has local weak order $p$ then the following local error formula holds

$$\mathbb{E}(\phi(\theta(h))) - \mathbb{E}(\phi(\theta_1)) = h^{p+1} \left( \frac{\mathcal{L}^{p+1}}{(p+1)!} - A_p \right) \phi(\theta_0) + \mathcal{O}(h^{p+2}).$$

(18)
2.2 Global Weak Error Expansion

We now study how accurate the numerical invariant measure $\pi_h$ is compared to the true invariant measure $\pi$. The first step to show this is the establishment of a global error expansion for the weak error

$$E(\phi, h, T) = |\mathbb{E}(\phi(\theta(T))) - \mathbb{E}(\phi(\theta_M))|,$$

where $X_M$ denotes the numerical solution at the final time $T = Mh$ calculated with a time step $h$ with a numerical method of weak order $p$.

In the sequel, we assume that solutions $\theta(t)$ of (1) are ergodic. In Remark 2.2 some necessary conditions in order for $\theta(t)$ to be ergodic are stated.

Remark 2.2. Let $\theta(t)$ the solution of (1). The following assumptions imply that $\theta(t)$ is ergodic (see [7]):

1. $\nabla \log \pi$ is of class $C^\infty$, with bounded derivatives of any order;
2. there exists $\beta > 0$ and a compact set $K \subset \mathbb{R}^d$ such that $\forall \theta \in \mathbb{R}^d - K, \langle \theta, \nabla \log \pi(\theta) \rangle \leq -\beta \theta^2$.

Consider now the numerical approximation $\theta_n$ of $\theta(t)$ satisfying (1) at time $t = nh$ with time-step given by $h$. A natural question to ask though is now which conditions is the underlying discretization of (1) ergodic. This is a difficult one to answer in general, since there exist cases where the underlying numerical discretization is not ergodic, or it doesn’t converge exponentially fast [16]. This relates mainly to the properties of the drift coefficient $f$ and its behaviour as $x \to \infty$. For the Euler-Maruyama and the Milstein scheme this has been investigated in [19].

In what follows we assume that the Markov chain defined by our numerical solution is indeed ergodic. Under this assumption the following theorem combines results derived by Talay [19] and Milstein [13]. Precisely, the expression in Equation (20) has been proved in [19] for specific methods (e.g. the Euler-Maruyama or the Milstein methods), while the general procedure to infer the global weak order from the local weak order is due to Milstein [13] (see [14, Chap.ts 2.2]). The difference here is the formulation of the error function (21) in terms of the generator $L$ and the operators $A_i$ in Assumption 2.1. This formulation will be particularly useful for our main results.

Theorem 2.3. Let $\pi$ satisfying the assumption in Remark 2.2 and $\theta_1, \ldots, \theta_M$ in Equation (1) be generated by the one step integrator (12) for equation (1) on $[0, T]$ ($\mathbb{E} = \mathbb{R}^d$) satisfying Assumption 2.1. If the local weak order $p$ estimate in Equation (18) holds, then we have the following expansion of the global error in Equation (19), for all $\phi \in C_p^2(\mathbb{R}^d, \mathbb{R})$,

$$E(\phi, h, T) = h^p \int_0^T \mathbb{E}(\psi_e(\theta(s), s))ds + O(h^{p+1})$$

where $\psi_e(\theta, t)$ satisfies

$$\psi_e(\theta, t) = \left(\frac{1}{(p+1)!} \mathcal{L}^{p+1} - A_p\right) v(\theta, t),$$

with $v(\theta, t) = \mathbb{E}(\phi(\theta(T)))|\theta(t) = \theta$ satisfying

6
\[
\frac{\partial v}{\partial t} + \mathcal{L}v = 0,
\]
\[
v(\theta, T) = \phi(\theta).
\]

(22)

**Proof.** The proof can be found in [1]. □

### 2.3 Expansions of Invariant Measures

Using Theorem 2.3 one can obtain a similar expansion to (20) for the difference between the true and the numerical ergodic averages. In particular, we have the following theorem.

**Theorem 2.4.** Assume that the conditions 1., 2., 3. from Remark 2.2 hold, and let \( \phi : \mathbb{R}^d \to \mathbb{R} \) be a smooth function satisfying (11). Then, if a numerical method of weak order \( p \) is ergodic, it satisfies

\[
\lim_{K \to \infty} \frac{1}{K} \sum_{i=1}^{K} \phi(\theta_i) - \int_{\mathbb{R}^d} \phi(y) \pi(y) dy = -\lambda_p h^p + O(h^{p+1})
\]

(23)

for any deterministic initial condition with \( \lambda_p \) defined as

\[
\lambda_p = \int_0^{+\infty} \int_{\mathbb{R}^d} \left( \frac{1}{(p+1)!} \mathcal{L}^{p+1} - A_p \right) u(y, t) \pi(y) dy dt
\]

(24)

and \( u(y, t) \) satisfying (9).

**Proof.** The proof is similar to the one found in [19, Theorem 4], with the main difference (see [1] for more details) being that now (20) is used as the starting point of the proof instead of the specific formula for the Euler-Maruyama method used in [19]. □

Theorem 2.4 provides an explicit expression of the first term in the error \( e(\phi, h) \) for the invariant measure. It will thus be the key result in analysing the long time behaviour of the SGLD method. Intuitively equation (24) says that if we want to calculate the error between the numerical and the true ergodic averages, we need to take into account the discrepancy between the true and the numerical solution given by

\[
\left( \frac{1}{(p+1)!} \mathcal{L}^{p+1} - A_p \right) u(y, t)
\]

for long times and then average over all the possible initial conditions \( y \) with respect to invariant measure \( \pi(y) \). The following example illustrates this in the case where an explicit form is known for \( u(y, t) \)

**Example 2.5.** We consider the case where \( \theta \in \mathbb{R} \) and \( \pi(\theta) = e^{-(\theta - \mu)^2/2\sigma^2} \) for which equation (1) becomes

\[
d\theta(t) = -\frac{1}{2} \left( \frac{\theta(t) - \mu}{\sigma^2} \right) dt + dW_t
\]

(25)

corresponding to the one dimensional Ornstein-Uhlenbeck process. If we now choose \( \phi(\theta) = \theta^2 \), then a simple calculation reveals that the solution of equation (22) is

\[
u(\theta, t) = \sigma^2 \left( 1 - e^{-t/\sigma^2} \right) + \theta^2 e^{-t/\sigma^2} + \frac{\theta \mu}{\sigma^2} (1 - e^{-t/2\sigma^2}) e^{-t/2\sigma^2} + \frac{\mu^2}{4\sigma^4} (1 - e^{-t/2\sigma^2})^2
\]
If we now solve (25) using the Euler-Maruyama method, then we have that \( p = 1 \). In addition, one has

\[
\frac{1}{2} \mathcal{L}^2 - A_1 = \frac{1}{8\sigma^4}(\theta - \mu) \frac{d}{d\theta} - \frac{1}{4\sigma^2} \frac{d^2}{d\theta^2}.
\]

Using this together with the equation for \( u(\theta, t) \), we find

\[
\left( \frac{1}{2} \mathcal{L}^2 - A_1 \right) u(\theta, t) = \frac{e^{-t/2\sigma^2}}{2\sigma^2} - \frac{e^{-t/2\sigma^2}(\mu - \theta)}{4\sigma^4} \left[ \frac{1 - e^{-t/2\sigma^2}}{\mu + e^{-t/2\sigma^2} \theta} \right]
\]

Formula (24) now implies that

\[
\lambda_p = - \int_0^\infty \left[ \int_{-\infty}^{+\infty} \left( \frac{e^{-t/2\sigma^2}}{2\sigma^2} - \frac{e^{-t/2\sigma^2}(\mu - \theta)}{4\sigma^4} \left[ \frac{1 - e^{-t/2\sigma^2}}{\mu + e^{-t/2\sigma^2} \theta} \right] \right) \frac{e^{-(\theta - \mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma}} d\theta \right] dt = \frac{1}{4}.
\]

On the other hand for the case of the Ornstein-Uhlenbeck process, it is known [22] that

\[
\pi_h \sim N(\mu, \sigma_h^2) \quad \text{and} \quad \sigma_h^2 = \frac{\sigma^2}{1 - \frac{h}{4\sigma^2}} = \sigma^2 + \frac{1}{4} h + O(h^2).
\]

### 3 Weak Convergence Analysis

We study the weak convergence properties of the SGLD method in the light of Theorems 2.3 and 2.4. The analysis in Section 3.1 implies that at leading order term there is a cost associated with not calculating the likelihood over all points. Thus, we introduce in Section 3.2 a modification of the original algorithm which has an error that is, asymptotically in \( h \), identical to the error of the Euler method (all data points taken into account in the calculation of the likelihood).

#### 3.1 Stochastic Gradient Langevin Dynamics

Theorems 2.3 and 2.4 imply that in order to characterise the leading order error term both for the weak convergence and the invariant measure, we need to calculate the corresponding differential operators \( A_0, A_1, \ldots \) in (15). To simplify the presentation and to illustrate the main ideas, we present the calculations only in the case where \( \theta(t) \) is one dimensional. The calculations when \( \theta(t) \in \mathbb{R}^d \) are presented in the appendix for completeness.

We start our calculations by rewriting the SGLD method in the following form

\[
\theta_{j+1} = \theta_j + h \hat{f}(\theta_j) + \sqrt{h} \xi_j,
\]

where

\[
\hat{f}(\theta) = \frac{1}{2} \left( \nabla \log \pi_0(\theta) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log \pi(X_{\tau_i}|\theta) \right),
\]

and by noting that

\[
\mathbb{E}_\tau \hat{f}(\theta) = f(\theta) := \frac{1}{2} \nabla \log \pi(\theta), \quad \forall \quad n \leq N.
\]

\footnote{see [22] for more details}
Using this notation, we will now expand $\phi(\theta_1)$ in powers of $h$ and then take expectations with respect to random variable $\xi_0$. In particular, we obtain

$$\mathbb{E}_{\xi_0}(\phi(\theta_1)) = \phi(\theta_0) + h \left( \dot{f}(\theta_0) \phi'(\theta_0) + \frac{1}{2} \phi''(\theta_0) \right)$$

$$+ \frac{h^2}{2} \left( \ddot{f}(\theta_0) \phi''(\theta_0) + \dot{f}(\theta_0) \phi^{(3)}(\theta_0) + \frac{1}{4} \phi^{(4)}(\theta_0) \right) + O(h^3).$$

If we now take expectations with respect to the random sampling and use (28), we obtain

$$\mathbb{E}(\phi(\theta_1)) = \phi(\theta_0) + h \mathcal{L} \phi(\theta_0) + \frac{h^2}{2} \left( \mathbb{E}_r(\ddot{f}^2(\theta_0)) \phi''(\theta_0) + f(\theta_0) \phi^{(3)}(\theta_0) + \frac{1}{4} \phi^{(4)}(\theta_0) \right) + O(h^3),$$

where $\mathcal{L}$ is the generator of $[1]$. We thus see that the SGLD method is a first order weak method and

$$A_1(\pi)\phi = \frac{1}{2} \left( \mathbb{E}_r(\ddot{f}^2(\theta_0)) + f(\theta_0) \phi^{(3)} + \frac{1}{4} \phi^{(4)} \right).$$

Now using the exact expression [25] for $\mathcal{L}$, we obtain

$$\frac{1}{2} \mathcal{L}^2 - A_1 = \frac{1}{2} \left( f(\theta)f''(\theta) + \frac{1}{2} f''(\theta) \right) \partial_\theta + \frac{1}{2} \left( f'(\theta) + f^2(\theta) - \mathbb{E}_r(\ddot{f}^2(\theta_0)) \right) \partial^2_\theta$$

$$= \frac{1}{2} \left( f(\theta)f''(\theta) + \frac{1}{2} f''(\theta) \right) \partial_\theta + \frac{1}{2} \left( f'(\theta) - \text{Var}(\ddot{f}(\theta)) \right) \partial^2_\theta. \quad (30)$$

We thus see that in the case of SGLD the leading order error term contains an extra factor of $\text{Var}(\ddot{f}(\theta))\partial^2_\theta$ when compared to the Euler method ($n = N$), in which case $\text{Var}(\ddot{f}(\theta)) = 0$. This can be understood as the penalty associated with not using all the available points for calculating the likelihood at every time step, as it results in an extra term in the corresponding error expressions given in Theorems 2.3 and 2.4 when compared with the Euler method. More precisely, for $n \ll N$ the term $\text{Var}(\ddot{f}(\theta))$ is of size $\mathcal{O}(N^2)$ thus making the leading order error term $\mathcal{O}(hN^2)$ in (29) (here $p = 1$). We discuss this and its implications further in Section 5 in the context of a simple one dimensional Gaussian model.

### 3.2 Modified SGLD

As we have seen in the previous section, the SGLD method introduces an extra term of order $\mathcal{O}(N^2)$ when $n \ll N$ in the leading order error term appearing in Theorems 2.3 and 2.4 relating to the weak error and the errors corresponding to the ergodic averages, respectively. Ideally, one would like to make this extra term disappear as this would imply a smaller error constant in front of the time step $h$. One way to achieve this would be to take $n = N$, that is at every time step we use the full amount of data available in the calculation of the log-posterior. However, this is exactly what one would like to avoid, since using the full amount of data at each time step is computationally expensive. In order to deal with this issue, we introduce the following modification of the SGLD method

$$\theta_{j+1} = \theta_j + h\ddot{f}(\theta_j) + \sqrt{h} \left( 1 - \frac{h}{2} \text{Var}(\ddot{f}(\theta_j)) \right) \xi_j. \quad (31)$$
A similar calculation for the case of this method yields that
\[ \frac{1}{2}L^2 - A_1 = \frac{1}{2} \left( f(\theta)f'(\theta) + \frac{1}{2} f''(\theta) \right) \partial_\theta + \frac{1}{2} f'(\theta) \partial_\theta^2 \]  
which implies that the leading order term in the weak error and the error for the ergodic averages is now the same as for the Euler method (using all data per step). Furthermore, following the calculations in the Appendix, we have the modification of the SGLD in higher dimensions
\[ \theta_{j+1} = \theta_j + h \tilde{f}(\theta_j) + \sqrt{h} \left( I - \frac{h}{2} \text{Cov} \tilde{f}(\theta_j) \right) \xi_j \]  
where
\[ \text{Cov} \tilde{f}(\theta)_{km} = E \left[ (\tilde{f}_k(\theta) - E(\tilde{f}_k(\theta))) (\tilde{f}_m(\theta) - E(\tilde{f}_m(\theta)))^T \right] \]
and \( \xi_j \) is now a \( d \)-dimensional standard random variable.

### 3.2.1 The Variance Estimator

Except for special cases \( \text{Var} f(\theta_j, n, \tau) \) does not have a closed form. The simplest possible way to proceed without it, is by replacing it by an unbiased estimator,
\[ \hat{\text{Var}} \tilde{f}(\theta) := \frac{N(N-n)}{n(n-1)} \sum_{i=1}^n \left( \nabla \log \pi(x_{\tau_i} | \theta) - \tilde{f}(\theta) \right)^2 N \]
This replacement does not change Equation (32) because the smallest order contribution to Equation (29) is of the form
\[ -h^2 E \hat{\text{Var}}^2 \xi_j = -h^2 \text{Var} \cdot 1. \]

### 4 Finite time Sample Averages

The mean square error (MSE) of the finite time sample average has two contributions - bias and variance. In the previous sections we have focused on the bias and have derived an analytic expression for the asymptotic bias in Section 3.1 and its dependence on the step size and the subsampling procedure. In Section 4.2 we bound both variance and bias for the finite time sample averages of the SGLD. This is an extension of [11] which addresses finite time sample averages of discretisations of diffusions
\[ d\theta_t = f(\theta_t) + g(\theta_t)dW_t \]  
on the torus. We continue with brief summary of methodology of [11] before adapting it to the SGLD.

### 4.1 Preliminaries

We summarise the connection between Poisson equation and time average of the diffusion. The Poisson equation is an elliptic PDE on basis of the generator associated with Equation (34) given by
\[ L\psi = \frac{1}{2} \nabla \psi \cdot \nabla f + \frac{1}{2} g(\theta)g(\theta)^* : D^2 \psi \]
where the $\star$ denotes the transpose. We consider the Poisson equation

$$L\psi = \phi - \bar{\phi} \text{ on } \mathbb{R}^d$$

(35)

where $\bar{\phi} = \int \phi(x) \pi(dx)$. For the SDE the time average of $\phi$ can be related to $\psi$ by the following application of Ito’s formula

$$\psi(\theta(t)) - \psi(\theta(0)) = \int_0^t \phi(\theta_s) ds + \int_0^t \nabla \psi(\theta_s) \cdot g(\theta_s)dW_s$$

$$\frac{1}{t} \int_0^t \phi(\theta_s) ds - \bar{\phi} = \frac{1}{t} (\psi(\theta(t)) - \psi(\theta(0))) - \frac{1}{t} \int_0^t \nabla \psi(\theta_s) \cdot g(\theta_s)dW_s$$

Notice that the second term on the RHS is a martingale and the first can typically be controlled. This allows us to derive an error bound for the ergodic average. For a more elaborate summary of this technique we point the reader to Section 4.2 of [11] and references therein.

We are interested in the time average of the Euler discretisation and the SGLD. We can build on the ideas of Section 5 of [11] which considers time discretisations of Equation (34). The article considers a time discretisation of Equation (34) of the following form

$$\theta_{k+1} = \theta_k + hf(\theta_k, h) + hg(\theta_k, h)\eta_k, \quad \eta \sim \mathcal{N}(0, I).$$

In [11] a Taylor expansion is used to express

$$\Delta\psi(\theta_{k+1}) = \psi(\theta_{k+1}) - \psi(\theta_k) = h (A_h \psi)(\theta_k) + R_k$$

where $R_k$ is the remainder term.

Using that $L\psi = \phi - \bar{\phi}$, summing over $k$ and dividing by $hK$ yields

$$\hat{\phi}_K := \frac{1}{K} \sum_{k=0}^{K-1} \phi(\theta_k) = \bar{\phi} + \frac{1}{Kh} (\psi(\theta_K) - \psi(\theta_0)) - \frac{1}{Kh} \sum_{k=0}^{K-1} h (A_h - L) \psi(\theta_k) - \frac{1}{Kh} \sum_{k=0}^{K-1} R_k.$$

Controlling $A_h - L$ and the remainder gives rise to Theorem 5.1 and 5.2 in [11] stating that

$$\text{Bias} \left| \mathbb{E}\hat{\phi}_K - \bar{\phi} \right| \leq C \left( h + \frac{1}{h \cdot K} \right)$$

$$\text{MSEE} \left( \hat{\phi}_K - \bar{\phi} \right)^2 \leq C \left( h^2 + \frac{1}{h \cdot K} \right).$$

(36)

In [11] these results have been derived for discretisations of SDEs on the torus. This simplifies the presentation because the derivates of $\psi$ are bounded on a compact set. However, the same arguments hold if

$$\sup_k \mathbb{E} \left\| D^{(i)} \psi(\theta_k) \right\| < \infty.$$
4.2 The Bias and the MSE of finite time SGLD Averages

We consider the SDE
\[
d\theta_t = f(\theta_t)dt + g(\theta_t)\,dW_t.
\] (37)

In the following we consider \(g = I\) but we keep \(g\) to make presentation clearer. The recursion of the corresponding SGLD reads as follows
\[
\Delta_{k+1} = \theta_{k+1} - \theta_k = \hat{f}_k h + h^\frac{3}{2} g_k \xi_{k+1}
\]
where \(\hat{f}\) is an unbiased estimate of \(f\). The focus of this section is to establish results similar to Equation (36) for the SGLD. This is formulated in Theorem 4.1 under assumptions on both the estimator and the associated continuous time diffusion. The proof of this result is contained in the Appendix 8. We end this section by formulating sufficient condition on \(\pi\) that ensure that the associated continuous time diffusion satisfies the assumption of Theorem 4.1. We believe that these sufficient condition can be weakened, but this requires improving the results of [15] which is out of the scope of this article.

Notice that we use a slightly different notation than [11] in order to match the notation of [20]. In particular, we want to remark that we use \(\Delta_{k+1} = \theta_{k+1} - \theta_k\) instead of \(\delta_n = X_{n+1} - X_n\). We use the abbreviations \(\phi_k = \phi(\theta_k), \, \hat{f}_k = \hat{f}(\theta_k, \tau_k, h)\) for the estimate of the drift, \(g_k = g(\theta_k, h) = I, \, \psi_k = \psi(\theta_k), \, V_k = V(\theta_k)\) and \(D^k \psi_k = D^k \psi(\theta_k)\). Moreover, we write
\[
A_h \phi = \nabla \phi \cdot \mathbb{E}_\tau \hat{f}(\theta, \tau, h) + \frac{1}{2} S(\cdot, h) : \nabla \nabla
\]
where \(S(x, h) = g(x, h)g(x, h)^T = I\). The 3rd order Taylor expansion with remainder of \(\psi\) yields
\[
\psi_{k+1} = \psi_k + D\psi_k [\Delta_{k+1}] + \frac{1}{2} D^2 \psi_k [\Delta_{k+1}^2] + \frac{1}{6} D^3 \psi_k [\Delta_{k+1}^3] + R_{k+1}
\]
\[
R_{k+1} = \frac{1}{5} \int_0^1 s^3 D^4 \psi (s \theta_k + (1-s) \theta_{k+1}) [\Delta_{k+1}^4] \, ds.
\]

Expressing the expansion in terms of \(A_h\) yields
\[
\psi_{k+1} = \psi_k + h A_h \psi + h^\frac{3}{2} D^2 \psi_k [g_k \xi_{k+1}] + h D \psi_k \left[ \hat{f}_k - \mathbb{E}_\tau \hat{f}(\theta_k, \tau, h) \right] / h_k
\]
\[
+ h^\frac{3}{2} D^2 \psi_k [\hat{f}_k, g_k \xi_{k+1}] + \frac{1}{2} h^2 D^2 \psi_k \left[ \hat{f}_k^2 \right] + \frac{1}{6} D^3 \psi_k [\Delta_{k+1}^3] + r_{k+1} + R_{k+1}
\]
with \(r_{k+1} = h^\frac{1}{2} \left( D^2 \psi_k [g_k \xi_{k+1}, g_k \xi_{k+1}] - S(x, h) \right)\). Summing the above over the first \(K\) terms
and dividing by $T = hK$ yields

\[
\frac{\psi_K - \psi_0}{Kh} = \frac{1}{K} \sum_{k=0}^{K-1} (\phi_k - \bar{\phi}) + \sum_{k=0}^{K-1} (A - A_h) \psi_k
\]

\[
= \frac{1}{T} \sum_{k=0}^{K-1} r_{k+1} + \frac{1}{T} h^\frac{1}{2} \sum_{k=0}^{K-1} D\psi_k [g_k \xi_{k+1}] + \frac{1}{T} h^\frac{3}{2} \sum_{k=0}^{K-1} D^2\psi_k \left[ \hat{f}_k, g_k \xi_{k+1} \right]
\]

\[
M_{1,K} + \frac{1}{T} h \sum_{k=0}^{K-1} D\psi_k \left[ \hat{f}_k - \mathbb{E}_\tau f(\theta_k, \tau, h) \right]
\]

\[
M_{4,K}
\]

\[
+ \frac{1}{T} h^\frac{3}{2} \sum_{k=0}^{K-1} D^2\psi_k \left[ \hat{f}_k, \hat{f}_k \right] + \frac{1}{T} \sum_{k=0}^{K-1} R_{k+1} + \frac{1}{T} \sum_{k=0}^{K-1} D^3\psi_k \left[ \Delta^3_{k+1} \right], \quad (38)
\]

\[
S_{1,K}
\]

\[
S_{2,K}
\]

\[
S_{3,K}
\]

Where $M_{i,k}$ indicate the Martingale terms and $S_{i,k}$ other remainder terms. We split

\[
S_{3,K} = M_{0,K} + \tilde{M}_{0,K} + S_{0,K} + \tilde{S}_{0,K}
\]

in terms of

\[
M_{0,K} = \frac{1}{6} h^\frac{3}{2} \sum_{k=0}^{K-1} \left( D^3\psi_k \left[ (g_k \eta_{k+1})^{3^\otimes} \right) \right)
\]

\[
\tilde{M}_{0,K} = \frac{1}{2} \sum_{k=0}^{K-1} h^\frac{5}{2} D^3\psi_k \left[ \hat{f}_k, \hat{f}_k, g_k \eta_{k+1} \right]
\]

\[
S_{0,K} = \frac{1}{6} \sum_{k=0}^{K-1} 3h^2 D^3\psi_k \left[ g_k \eta_{k+1}, g_k \eta_{k+1}, \hat{f}_k \right]
\]

\[
\tilde{S}_{0,K} = \frac{1}{6} \sum_{k=0}^{K-1} h^3 D^3\psi_k \left[ \hat{f}_k^{3^\otimes} \right].
\]

**Theorem 4.1.** Suppose that there is a function $V$ such that the following three assumptions hold

1. derivatives of the solution $\psi$ to the Poisson equation can be bounded in terms of $V$

\[
\|D^k\psi\| \lesssim V^{p_k}, \quad \text{for} \ k = 0, \ldots, 4. \quad (39)
\]

2. The drift $f$ and the error from the estimate $H := \hat{f}(\theta) - \mathbb{E}_\tau \hat{f}(\theta, \tau, h)$ are supposed to satisfy

\[
\mathbb{E}_\tau H(\theta, \tau)^{2p} \lesssim V(\theta)^p \quad \forall p \leq p^*
\]

\[
\|f\|^2 \lesssim V.
\]
for $p^* = \max \{2p_{\psi,2} + 2, 2p_{\psi,4} + 4, 2p_{\psi,3} + 1, 2p_{\psi,3} + 3\}$. Moreover, we suppose that the $EV^p(\theta_k)$ is bounded from above and that this bound is independent of $k$, that is

$$\sup_k EV^p(\theta_k) < \infty, \quad \forall p \leq p^*. \tag{40}$$

3. Suppose that $V$ additionally satisfies

$$\sup_s V(s\theta + (1-s)Y)^p \lesssim V(\theta)^p + V(Y)^p, \quad \forall \theta, Y, p \leq p^*. \tag{41}$$

Under these assumptions we have

$$\text{Bias}(\hat{\phi}_K) = \left| \mathbb{E}\hat{\phi}_K - \bar{\phi} \right| = O \left( h + \frac{1}{K}h \right) \tag{42}$$

$$\mathbb{E}(\hat{\phi}_K - \bar{\phi})^2 = C \left( h^2 + \frac{1}{K}h \right) \tag{43}$$

where

$$\hat{\phi}_K = \frac{1}{K} \sum_{k=1}^{K} \phi(\theta_k), \quad \bar{\phi} = \mathbb{E}_\mu \phi.$$ 

**Remark 4.2.** At first sight, fixed step size SGLD seems advantageous compared to decreasing step size SGLD because its variance is of order $K^{-1}$. However, optimising over the step size $h$ results in bound of the MSE of order $K^{-\frac{3}{2}}$ in Equation (43). This agrees with rate of $K^{-\frac{3}{4}}$ in CLT for decreasing step size SGLD that we obtained in [20]. This analysis is based on the upper bound in Equation (43) but it is confirmed for the analytic expression for the MSE of a toy model in Section 5.2.2.

Theorem 4.1 and results for decreasing step size SGLD [20] hold under assumptions formulated in terms of the solution $\psi$ of the Poisson equation. In the following, we show how these can be verified in terms of $\pi$. The assumption of Theorem 4.1 can be summarised as follows

$$\sup_i EV^p(\theta_i) < \infty \tag{40}$$

$$\sup_s V(s\theta + (1-s)Y)^p \lesssim V(\theta)^p + V(Y)^p, \quad \forall \theta, Y, p \tag{41}$$

$$\|D^k\psi\| \lesssim V_{p_{\psi,k}}, \quad \text{for } k = 0 \ldots 4. \tag{39}$$

**Proof.** See Appendix 8. \hfill \Box

Equation (40) is established for all $p \leq p_H$ by Lemma 5 in [20] if we impose the following assumption.

**Assumption 4.1.** The drift term $\theta \mapsto \frac{1}{2} \nabla \log \pi(\theta)$ is continuous. The function $V : \mathbb{R}^d \rightarrow [1, \infty)$ tends to infinity as $\|\theta\| \rightarrow \infty$, is twice differentiable with bounded second derivatives and satisfies the following conditions:
1. $V$ is a Lyapunov function for the SDE, i.e. there are constants $\alpha, \beta > 0$ such that for every $\theta \in \mathbb{R}^d$ we have
\[
\left( \nabla V(\theta), \frac{1}{2} \nabla \log \pi(\theta) \right) \leq -\alpha V(\theta) + \beta.
\] (44)

2. The following bounds hold
- There exists an exponent $p_H \geq 2$ such that
\[
E[\|H(\theta, U)\|^{2p_H}] \lesssim V^{p_H}(\theta).
\] (45)
Moreover, this implies that $E[\|H(\theta, U)\|^{2p}] \lesssim V^p(\theta)$ for any exponent $0 \leq p \leq p_H$.
- For every $\theta \in \mathbb{R}^d$ we have
\[
\|\nabla V(\theta)\|^2 + \|\nabla \log \pi(\theta)\|^2 \lesssim V(\theta).
\] (46)

Notice that for $\hat{f}$ based on subsampling we obtain that $p_H = \infty$ if
\[
\|\nabla \log p(y | \theta)\|^2 \leq C(y)V.
\]

Notice that Assumption 4.1 also implies
\[
E\|\Delta_i\|^{2p} \leq V_i^p
\]
\[
E\|\hat{f}_i\|^{2p} \leq V_i^p
\] (47)
if for $p \leq p_H$. Equation (41) could now simply be formulated as an additional assumption, however currently we need even stronger assumptions to verify Equation (39). In Appendix 8.1 we show how the results of [15] can be used to establish Equation (39) if Equations (44) and (46) hold for $V = \|\theta\|^2 + 1$.

5 A one-dimensional Gaussian Toy Model

We consider a simple linear Gaussian model, that is
\[
\theta \sim \mathcal{N}(0, \sigma_\theta^2), \quad x_i \sim \mathcal{N}(\theta, \sigma_x^2).
\] (48)

In this case the posterior is given by
\[
\pi = \mathcal{N}(\mu_p, \sigma_p^2) = \mathcal{N}\left( \frac{\sum_{i=1}^N x_i}{\sigma_x^2}, \left( \frac{1}{\sigma_\theta^2} + \frac{N}{\sigma_x^2} \right)^{-1} \right).
\] (49)

For this choice of $\pi$ Equation (1) becomes
\[
d\theta(t) = -\frac{1}{2} \left( \frac{\theta(t)}{\sigma_p^2} \right) dt + dW_t,
\] (50)
and its numerical discretisation by the SGLD reads as follows

\[
\theta_{j+1} = \theta_j - Ah\theta_j + B_j h + \sqrt{h}\xi_k, \tag{51}
\]

where \(\xi_k \sim \mathcal{N}(0, 1)\) and

\[
A = \frac{1}{2} \left( \frac{1}{\sigma_0^2} + \frac{N}{\sigma_x^2} \right),
\]

\[
B_j = \frac{N}{n} \sum_{i=1}^{n} X_{\tau_i} \sigma_x^2.
\tag{52}
\]

By \(\tau = (\tau_1, \ldots, \tau_n)\) we denote a random subset of \([N] = \{1, \ldots, N\}\), generated, for example, by sampling with or without replacement from \([N]\).

This toy model illustrates a clear connection with the theory. We obtain an analytic expression for the asymptotic bias in Section 5.1 that matches the analysis of the bias presented in Section 3. Moreover, we confirm for this toy model that the upper bound on the MSE derived in Section 4.2 suggests the correct scaling for number of steps and step size for minimal computational effort, see Section 5.2.2. The toy model does not only provide analytic expression for bias and MSE that match the theory but is also amenable in limit of data items \(N \to \infty\) which is the focus of Section 5.2.3.

### 5.1 Analysis of the Bias

We start by calculating the posterior mean. In particular, using Equation (51) and taking expectations with respect to \(\xi_k\), we have

\[
\mathbb{E}(\theta_{j+1}|B) = (1 - Ah)\mathbb{E}(\theta_j|B) + B_j h, \tag{53}
\]

which can be solved to obtain

\[
\mathbb{E}(\theta_j|B) = (1 - Ah)^j\mathbb{E}(\theta_0) + \sum_{k=0}^{j-1} h(1 - Ah)^k B_{j-k-1}.
\]

If we now take the expectation with respect to \(B_j\), and use the fact that \(\mathbb{E}(B_j) = \mathbb{E}(B)\) and take the limit of \(j \to \infty\), we have

\[
\mathbb{E}(\theta_\infty) = \sum_{k=0}^{\infty} (1 - Ah)^k h\mathbb{E}(B),
\]

which implies

\[
\mathbb{E}(\theta_\infty) = h/(1 - (1 - Ah))\mathbb{E}(B) = \frac{\mathbb{E}(B)}{\frac{\sum_{i=1}^{N} X_i}{\sigma_x^2} + N}.
\]

We thus see that the SGLD is capturing the correct limiting expectation of the posterior independently of the choice of the time-step \(h\).

We now investigate the behaviour of the limiting variance under the SGLD. In doing this we will be using the law of total variance

\[
\text{Var}[Y] = \mathbb{E}(\text{Var}[Y | X]) + \text{Var}(\mathbb{E}[Y | X]).
\]
We thus have
\[ \text{Var}[\theta_{j+1}] = \mathbb{E}(\text{Var}[\theta_{j+1} | B]) + \text{Var}(\mathbb{E}[\theta_{j+1} | B]). \]
A simple calculation now shows that
\[ \text{Var}[\theta_{j+1} | B] = (1 - Ah)^2 \text{Var}[\theta_{j+1} | B] + h \]
and
\[ \text{Var}(\mathbb{E}[\theta_{j+1} | B]) = (1 - Ah)^2 \text{Var}(\mathbb{E}[\theta_j | B]) + h^2 \text{Var}(B). \]
Combining these two results, we see that
\[ \text{Var}(\theta_{j+1}) = (1 - Ah)^2 \text{Var}(\theta_j) + h + h^2 \text{Var}(B). \]
If we now take the limit of \( j \to \infty \), we have that
\[ \text{Var}(\theta_{\infty}) = \frac{1}{2A - A^2 h} + \frac{h \text{Var}(B)}{2A - A^2 h}. \] (54)
We note here that the fact that we are evaluating the likelihood in a subset of the available points leads to an extra error term for the limiting variance, compared to the case when all available data points \( (n = N, \text{Var}(B) = 0) \) are taken into account. In this case the result we find coincides with [22]. Furthermore, the additional first order error term for the SGLD is equal to
\[ \frac{\text{Var}(B)}{2A} \]
which coincides with what one would find using [30] and the expression for \( \lambda_p \) in [44]. In particular, the extra term in [30] when compared with the Euler method is now given by
\[ -\frac{1}{2} \text{Var}(B) \partial_\theta^2 u(\theta, t) = -e^{-t/\sigma_p^2}, \]
where we have used the expression for \( u(\theta, t) \) from Example 2.5 replacing \( \mu \) and \( \sigma \) by \( \mu_p \) and \( \sigma_p \) respectively, as well as the fact that for this simple toy model \( \text{Var}(\hat{f}(\theta)) = \text{Var}(B) \). A simple integration now of this term according to the formula [24] gives that the overall contribution of the extra term is equal to
\[ \sigma_p \frac{\text{Var}(B)}{2A}, \]
and thus agreeing with our analysis.

We now consider the modified SGLD discussed in Section 3.2. In this case the numerical discretisation of Equation (50) becomes
\[ \theta_{j+1} = \theta_j - Ah \theta_j + B_j h + \sqrt{h} \left( 1 - \frac{h}{2} \text{Var}(B) \right) \xi_j \] (55)
where
\[ \text{Var}(B) = \frac{1}{4 \sigma_x^4 n} \sum_{j=1}^{N} \left( X_i^2 - \left( \frac{1}{N} \sum_{i=1}^{N} X_i \right)^2 \right)^2 \]
for sampling with replacement and

$$\text{Var}(B) = \frac{1}{4\sigma_x^2} \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} \left( X_i - \frac{1}{N} \sum_{i=1}^{N} X_i \right)^2$$

$$= \frac{1}{4\sigma_x^2} \frac{N(N-n)}{n} \text{Var} X$$ (56)

for sampling without replacement. A similar calculation to the case of SGLD now shows that

$$E(\theta_{\infty}) = \sum_{i=1}^{N} X_i \quad \text{and} \quad \text{Var}(\theta_{\infty}) = \frac{1}{2A - A^2 h} + \frac{h^2 \text{Var}^2(B)}{4(2A - A^2 h)}$$ (57)

We thus see, as expected from the discussion in Section 3.2, that the leading order error term for the modified SGLD is now the same as the one would obtain by using all points. This is illustrated further in Figure 1 where we plot the different errors for the invariant measure variance in the case of the SGLD, the modified SGLD and the Euler method (n=N) for a dataset containing $10^3$ points. We observe that if we choose $n = 10$ points for evaluating our

![Figure 1: Comparison for the invariant measure error for the SGLD, the mSGLD and the Euler method. We have used $N = 10^3$. We used $n = 10$ on the LHS and $n = 200$ on the RHS.](image)

likelihood function for large values of our timestep $h$, the SGLD is superior to the mSGLD. However, as $h$ is reduced, this is no longer the case. Furthermore, when we take $n = 200$, we see that the mSGLD outperforms the SGLD for all the time-steps used, but more importantly its error is now directly comparable with the one of the Euler method where all the data points are used for evaluating the likelihood.

These numerical observations indicate that there is a sensitive dependence with $n$ for the bias of the invariant measure variance in the case of mSGLD. In particular, because the asymptotic mean is correct for both the SGLD and the mSGLD the bias for the variance of the invariant measure takes the form

$$\sigma_p^2 - \text{Var} \sigma_{\infty}.$$ 

From Equations (54) and (55), we see that the bias is sum of two terms the bias due the step size and an additive term due to subsampling which takes the form
\[ h^2 \frac{\text{Var}(B)^2}{4(2A - A^2h)} = \frac{h^2 (\frac{N-n}{n})^2 N^2 \text{Var}(X)^2}{4(2A - A^2h)} \]

Because \( A \sim N \), we see that for the choice \( h = \frac{r}{A} \) the above stays bounded if and only if \( n \gtrsim N^{\frac{1}{2}} \). In contrast, the same consideration for SGLD shows that bias due to subsampling vanishes as long as \( n \to \infty \) as \( N \to \infty \). We can also identify the regime in which the mSGLD has a smaller bias than the SGLD using Equations (54) and (55):

\[ \frac{h^2 \text{Var}^2(B)}{4(2A - A^2h)} \leq \frac{h \text{Var}(B)}{2A - A^2h}. \]

Setting \( h = \frac{r}{A} \) for some \( r < 1 \) for the recursion in Equation (51) to be stable and rearranging yields

\[ 1 \geq \frac{h}{4} \frac{1}{16\sigma_x^2} \frac{N(N - n)}{N} \text{Var} X \]

Letting \( c = \frac{n}{N} \) be the proportion of the subset and using Equation yields

\[ c \geq \frac{2rN\text{Var} X}{16\sigma_x^2 \left( \frac{1}{\sigma_x^2} + \frac{N}{\sigma_z^2} \right) + 2rN\text{Var} X}, \]

which in the limit of \( N \to \infty \) gives

\[ c \geq \frac{2r\text{Var} X}{16 + 2r\text{Var} X}. \]

For fixed \( r \), we see that the mSGLD has a smaller bias than the SGLD if \( c \geq c_0 \) where \( c_0 \) is given by RHS of the above equation. This restriction on \( n \) is not a very desirable feature of the mSGLD. However, suppose that \( n < N \) is fixed and the aim is to approximate posterior expectations to higher precision. In this case we need to decrease \( h \) to reduce the bias and increase \( M \) (in order to have approximation to the SDE over the same time interval). Thus, the results of Section 5.1 imply that the bias of the mSGLD becomes closer to that of the Euler method while the SGLD has an additional term. We suspect that the additional term in general increases the bias. In this case our arguments suggest that there is a regime of accuracy below which the mSGLD is efficient. Actually, this is exactly what we find in Section 5.2.2, where the computational efficiency of the two methods is analysed for a fixed accuracy.

### 5.2 Analytic Investigation of the Toy Model

We will now further investigate the performance of SGLD and mSGLD in terms of the mean square error in the estimation of the second moment

\[ \text{MSE} = \mathbb{E} \left( \frac{1}{M} \sum_{k=0}^{M-1} \theta_j^2 - (\mu_p^2 + \sigma_p^2)^2 \right)^2 \quad (58) \]

as well as the relative error in the estimation of the posterior variance

\[ \text{ERE} = \left| \mathbb{E} \left( \frac{1}{M} \sum_{i=0}^{M-1} \theta_i^2 - \left( \frac{1}{M} \sum_{i=0}^{M-1} \theta_i \right)^2 \right) \frac{\sigma_p^2}{\sigma_z^2} - 1 \right|. \quad (59) \]
The advantage of studying these error quantities for the simplistic model given in Equation (50) is that for this one, we can calculate analytic expressions of the form $E[^p_{\theta^j}]$, depending on the data set $X$, the total number of steps $M$, the subset size $n$, as well as the step size parameter $r = hA$. In Appendix 10.1 we derive analytic expressions of the moments, in order to obtain analytical expressions for the MSE and the ERE.

We use this expressions to study both the MSE and the ERE for fixed and increasing $N$ in Section 5.2.1 by generating an artificial data using Equation (48). For $n < N$ the SGLD and the mSGLD can perform more steps with the same computational budget. That can be visualised by plotting the MSE and ERE against the (effective) iterations through the data set, that is likelihood evaluations divided by $N$. Notice that this quantity is proportional to the computational budget. Using this approach we study the SGLD and mSGLD for fixed and increasing $N$. This comparison is refined in Section 5.2.2 by comparing the computational cost of the SGLD and the mSGLD for fixed level accuracy specified in terms of MSE for optimally chosen $h$, $n$ and $M$. A numerical solution to the resulting optimisation problem demonstrates that the SGLD is advantageous in the lower accuracy regime while the SGLD degenerates to $n = N$ in the high accuracy regime. The mSGLD does not degenerate and seems to maintain a constant speed up factor compared to the Euler method.

In Section 5.2.3 we take the expectation with respect to the realisation of the data for fixed $\theta^\dagger$ and study the limit $N \to \infty$. The bottom line is that the SGLD

- reduces the computational cost from $N^{2+\epsilon}$ to $N^{1+\epsilon}$ for the estimation of the second moment with the MSE of its estimate going to zero.

- is not able to reduce the computational cost for estimation of the variance in terms of an algebraic power in $N$ with the ERE of its estimate going to zero.

### 5.2.1 The MSE and the ERE for fixed and increasing $N$

We now investigate the behaviour of both the MSE and the ERE for fixed and increasing $N$ using the explicit expression for a data set $X$ obtained using the provided Mathematica file. We start by considering $N = 10^3$ and show the effect of varying $n$ before considering $n$ as a function of $N$ and considering the behaviour for growing $N$.

In our first test we generate an artificial data set $X$ of size $N = 10^3$, choose $r = \frac{1}{20}$ and plot the analytical expressions of the MSE against the total number of steps $M$, for different values of subset size $n$, both for the SGLD and the mSGLD in Figure 2. As we observe, the MSE asymptotes for both the SGLD and the mSGLD as $M \to \infty$ due to the bias introduced by the fixed step size (see discussion in Section 5.1). This bias also depends on the choice of $n$ and as we can see for the values of $n$ chosen, it is smaller for the case of the mSGLD.

We now consider the behaviour for growing $N$, where in each instance we generate a new data set $X$. Figure 3 depicts the MSE for $N = 10^i$ with $i = 1, \ldots, 4$ for the subset choices $n = N^{0.1}, N^{0.5}$ and $N^{0.9}$ each compared to the Euler method corresponding to $n = N$. In this plot we notice that the SGLD outperforms the mSGLD for $n = N^{0.1}$ and $n = N^{0.5}$. This of course is not surprising since as we have already discussed in Section 5.1 in the case of mSGKD one needs to choose $n \gtrsim N^{\frac{1}{2}}$ in order for the bias not to blow up in the limit of large $N$ and $n \gtrsim N$ to have the bias of the same size as SGLD or smaller.

In the following we repeat the same analytic experiments we did for the MSE for the ERE. As $N \to \infty$, the denominator goes to zero making this problem more difficult than the MSE of estimating the second moment. We start by fixing $N = 10^3$ and generate an artificial
Figure 2: Expected MSE of the sample average for the SGLD and the mSGLD for the second moment of the posterior

data set \(X\) of size \(N\), while we set \(r = 10^{-2}\) and consider the impact of changing \(n\) for both the SGLD and the mSGLD. The results are depicted in Figure 4 where we note that the bias increases for smaller values of \(n\). Similarly to the case of the MSE, we now depict in Figure 5 the behaviour of the ERE as \(N\) grows, again for the choice \(r = 10^{-2}\). We notice that the number of iterations through the data set required for the ERE < 1 increases with the size of the data set quite significantly.

5.2.2 Minimising Computational Effort for Constrained MSE

In Section 5.2.1 we compared the Euler method, the SGLD and the mSGLD for the same choice of \(r = \frac{h}{A}\). In the following we try to minimise the computational effort with respect to the condition \(\text{MSE} \leq \alpha\). We assume that the computational effort is proportional to \(M \cdot n\) which leads to the problem

\[
\min_F \quad M \cdot n \\
\text{subject to} \quad \text{MSE}(r; M, n) \leq \epsilon^2 \\
w.r.t. \quad r, M, n.
\]

Even though we have an analytic expressions for the MSE, the solution to the optimisation problem does not have closed form. To conclude our analysis, we illustrated the numerical solution to this problem for \(N = 1000\) for the Euler method, the SGLD and the mSGLD. Note that in case of the Euler method \(n\) is fixed to be 1000. The results, depicted in Figure 6, can be summarised as follows:

1. as \(\epsilon\) becomes smaller the gain of the SGLD over the Euler method in terms of computational effort decreases, the reason being that \(n\) increases;

2. as \(\epsilon\) becomes smaller the mSGLD gains efficiency over the SGLD, the reason being that \(n\) does not increase as much as \(\alpha\) decreases.

The upper bound obtained in Equation (43) suggests a scaling of \(M \sim \epsilon^{-3}\) and \(r \sim \epsilon\) to obtain an MSE of order \(\epsilon^2\) with minimal computational effort. The numerical minimisation
Figure 3: Expected MSE of the time average for the SGLD and the mSGLD for the second moment of the posterior as \( N \to \infty \).
of $M$ with respect to $r$ and $M$ subject to $\text{MSE}(r, M) \leq \epsilon^2$ confirms this scaling by empirically, see Figure 7.

### 5.2.3 Limit of the MSE and ERE for well-specified Data as $N \to \infty$

We now perform a frequentist’s analysis by fixing $\theta^\dagger = 1$ and generating data according to the model. We note that in this setting, the expectations for the MSE and the ERE are with respect to the realisation of the data and the noise driving the algorithm. This results in analytic expressions\[^2\] for the MSE and the ERE depending only on $M, n, r$ and $N$. We then choose $M, n$ and $r$ as functions of $N$ and study the limit $N \to \infty$.

For the Euler method ($n = N$) we need to take $h < \frac{1}{A} \approx \frac{1}{N}$ in order to make Equation (53) stable. Moreover, we need the number of steps $M$ to be of order $N$ to approximate the diffusion to a time of order $O(1)$. Because we evaluate $N$ data points per step, the complexity is of order $O(N^2)$. Furthermore, we verify, using Mathematica, that for the Euler method ($n = N$)

$$\lim_{N \to \infty} \text{MSE} = 0, \quad \lim_{N \to \infty} \text{ERE} = 0 \quad (61)$$

for the choice $M = N^{1+2\epsilon}$ and $r = N^{-\epsilon}$ for any $\epsilon > 0$. The computational cost for fixed $N$ is $M \cdot n = N^{2+2\epsilon}$, thus, confirming the heuristics we used for the Euler method in Section 5.2.1.

A natural question thus to ask in terms of the SGLD is if one can have Equation (61) to hold but for smaller computational complexity than the Euler method. Using Mathematica, we obtain the following theorem for the MSE.

**Theorem 5.1.** For any $\epsilon > 0$ there exist $h = N^{-1-\epsilon}$, $M = N^{1+2\epsilon}$ and $n = 1$ satisfying

$$\lim_{N \to \infty} \mathbb{E}_{\theta, X} \left( \frac{1}{M} \sum_{k=0}^{M-1} \theta_j^2 - \left( \mu_p^2 + \sigma_p^2 \right) \right)^2 = 0.$$

\[^2\]see Appendix 10.2 for a sketch of the derivation for the MSE (the derivation for ERE is similar)
Figure 5: Expected relative error of the SGLD and the mSGLD for estimating the variance of the posterior as $N \to \infty$. 
(a) Minimised Work for Euler, SGLD and mS-(b) Subset size \( n \) for the minimum for SGLD and GLD

\[ M \cdot n \]

subject to \( MSE \leq \alpha \)

Figure 6: Minimisation of Computational Work \( \propto M \cdot n \) subject to \( MSE \leq \alpha \)

(a) Line fit yields \( M(\epsilon) \sim \epsilon^{-3.03} \) as \( \epsilon \to 0 \)

(b) Line fit yields \( r(\epsilon) \sim \epsilon^{0.95} \) as \( \epsilon \to 0 \)

Figure 7: Scaling of \( r(\epsilon) \) and \( M(\epsilon) \) for minimal computational effort subject to \( MSE(r(\epsilon), M(\epsilon)) \leq \epsilon^2 \)
This constitutes a substantial gain compared to the Euler method because it reduces the computational complexity from almost quadratic to almost linear.

Such computational improvement is not observed for the ERE. In particular, we now choose \( r = N^{-\alpha} \), \( n = N^\beta \) and \( M = N^\gamma \), hence the computational cost is \( N^{\beta+\gamma} \). The step size \( h = \frac{1}{N} \) satisfies \( h \sim N^{-1-\alpha} \) because \( A \sim N \). Because the algorithm performs \( M = N^\gamma \) steps, we expect it to approximate the diffusion on the time interval \( h \cdot M = N^{-1-\alpha+\gamma} \). Therefore it reasonable to require that \( \gamma > 1 + \alpha \). Under this assumption, Mathematica reduces the limit above to

\[
\lim_{N \to \infty} \text{ERE} = \lim_{N \to \infty} \left( \frac{2 \cdot N^{-\alpha-\beta+3}}{(N+1)^2(N-2)^2} - \frac{N^{-2a-\beta+3}}{(N+1)^2(N-2)^2} \right)
\]

\[
= \begin{cases} 
0 & \text{if } \alpha + \beta > 1 \\
\infty & \text{if } \alpha + \beta < 1.
\end{cases}
\]

Ensuring that \( \lim_{N \to \infty} \text{ERE} = 0 \) requires \( \alpha + \beta > 1 \), this in turn implies that the computational complexity satisfies

\[
\text{steps} \times \text{cost per step} = N^{1+\alpha} N^\beta = N^{1+\alpha+\beta} \gtrsim N^2.
\]

Thus, there is no computational gain for the ERE in the limit \( N \to \infty \) on the algebraic scale in \( N \).

6 Logistic Regression

In the following we present numerical simulations for a Bayesian logistic regression model. We assume the data \( y_i \in \{-1,1\} \) is modelled by

\[
p(y_i|x_i, \beta) = f(y_i^\beta x_i) \quad (62)
\]

where \( f(z) = \frac{1}{1+\exp(-z)} \in [0,1] \) and \( x_i \in \mathbb{R}^d \) are fixed covariates. We put a Gaussian prior \( \mathcal{N}(0, C_0) \) on \( \beta \), for simplicity we use \( C_0 = I \) subsequently. By Bayes’ rule the posterior \( \pi \) satisfies

\[
\pi(\beta) \propto \exp \left(-\frac{1}{2} \|\beta\|^2_{C_0} \right) \prod_{i=1}^{N} f(y_i^\beta x_i).
\]

We consider \( d = 3 \) and \( N = 1000 \) data points and choose the covariate to be

\[
x = \begin{pmatrix}
x_{1,1} & x_{1,2} & 1 \\
x_{2,1} & x_{2,2} & 1 \\
\vdots & \vdots & \vdots \\
x_{1000,1} & x_{1000,2} & 1
\end{pmatrix}
\]

for a fixed sample of \( x_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \) for \( i = 1, \ldots 1000 \) and \( j = 1, 2 \). We use a long run of a RWM to estimate the posterior mean. On that basis we estimate the MSE of the SGLD based mean estimate using 100 runs of the algorithm with step size \( h = 0.002 \) for various subset sizes. Figure 8 depicts the MSE as function of the iterations and effective iterations through the data set. We note that the mSGLD is superior for \( n = 150 \), inferior for \( n = 50 \) and for \( n = 10 \) the MSE does not drop below 1.
7 Conclusion

This article lays the mathematical foundations for stochastic gradient methods with fixed step size for posterior sampling. We derived an error expansion of the asymptotic bias in powers of the step size and identified how the constant in the leading order term depends on the unbiased estimator of the gradient. We construct the mSGLD to match the Euler method in this asymptotic expansion. These asymptotic results are complemented by upper bounds on the bias and the MSE over a finite time horizon. Minimising the MSE with respect to the step size yields that the error decays at the same rate as the decreasing step size SGLD, see Remark 4.2. These theoretical findings are completed with extensive analytic investigations of a one dimensional toy model that allows derivation of analytic expressions for the sample average and its moments allowing an exact quantification of expected errors. The results of this investigation are as follows

- In the high accuracy regime the SGLD deteriorates to the Euler method while the mSGLD prevails.
- For small data batches the bias of the mSGLD is larger than for the SGLD.
- In the limit as the number of data items goes to infinity the SGLD reduces computational complexity by one power of the number of data items

This recommends the constructing of new and a study of existing modifications of the SGLD such as the Stochastic Gradient Hamiltonian Monte Carlo [4] and the Stochastic Gradient Thermostat Monte Carlo [6].

References

[1] A. Abdulle, G. Vilmart, and K. Zygalakis. High order numerical approximation of the invariant measure of ergodic sdes. SIAM Journal on Numerical Analysis, 52(4):1600–1622, 2014.
[2] Nawaf Bou-Rabee and Eric Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for SDEs. *Communications on Pure and Applied Mathematics*, 63(5):655–696, 2010.

[3] S. Brooks, A. Gelman, G. Jones, and X.L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2011.

[4] Tianqi Chen, EB Fox, and Carlos Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. *arXiv Prepr. arXiv1402.4102*, pages 1–23, 2014.

[5] Arnaud Debussche and Erwan Faou. Weak backward error analysis for SDEs. *SIAM J. Numer. Anal.*, 50(3):1735–1752, 2012.

[6] N. Ding, Y. Fang, R. Babbush, C. Chen, R. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. To Appear in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014.

[7] Rafail Hasminskii. *Stochastic stability of differential equations*. Sijthoff and Noordhoff, The Netherlands, 1980.

[8] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[9] Marie Kopec. Weak backward error analysis for overdamped langevin processes. *IMA Journal of Numerical Analysis*, 2014.

[10] Benedict Leimkuhler and Charles Matthews. Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, 2013.

[11] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Convergence of Numerical Time-Averaging and Stationary Measures via Poisson Equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

[12] G. N. Milstein and M. V. Tretyakov. Computing ergodic limits for Langevin equations. *Phys. D*, 229(1):81–95, 2007.

[13] G.N. Milstein. Weak approximation of solutions of systems of stochastic differential equations. *Theory Probab. Appl.*, 30(4):750–766, 1986.

[14] Grigori N Milstein and M.V. Tretyakov. *Stochastic numerics for mathematical physics*. Scientific Computing. Springer-Verlag, Berlin and New York, 2004.

[15] E. Pardoux and A. Yu. Veretennikov. On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, 29(3):1061–1085, 2001.

[16] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):pp. 341–363, 1996.

[17] A. Koratticara S. Ahn and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *ICML*, 2012.
8 Appendix A: Finite time Ergodic Average based Poisson Equation

Proof of Theorem 4.1. Rearranging Equation (38) for \( \frac{1}{K} \sum_{k=0}^{K-1} (\phi_k - \bar{\phi}) \) the bias and the MSE can be controlled as follows:

\[
\mathbb{E} \frac{1}{Kh} \left( \psi_K - \psi_0 \right) \lesssim \frac{1}{T} \sup_i \mathbb{E} V_{\psi,0}^{\psi,0} \lesssim \frac{1}{T}.
\]

Because \( A_0 = \mathcal{L} \) for the SGLD it is left to bound to bound \( \mathbb{E} \frac{1}{T} S_{i,K} \) for \( i = 0 \ldots 4 \). First we consider \( i = 1 \) and use Equation (47)

\[
\frac{1}{T} \mathbb{E} S_{1,K} \lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} h^2 V_{\psi,2}^{\psi,2} \mathbb{E} r_k \| \hat{f}_k \|^2 \lesssim \frac{1}{T} \sum_{k=0}^{K-1} h^2 \sup_i \mathbb{E} V_i^{\psi,2+1} \lesssim \frac{1}{T} h^2 K \lesssim h.
\]

This procedure will be used over and over again. It can be summarised as follows:

1. bounding the terms in the sum by a power of \( V_\psi \), using Equation (47) and the assumption on derivates of \( \psi \);
2. then derive the bound using \( \sup_i \mathbb{E} V_i^p < \infty \).

For \( i = 2 \) we additionally use that Equation (41) implies

\[
\int_0^1 s^3 D^4 \psi (s \theta_k + (1 - s) \theta_{k+1}) \left[ \Delta_{k+1}^{4} \right] ds \lesssim (V_k^{\psi,4} + V_{k+1}^{\psi,4}) \| \Delta_k \|^4.
\]
which allows us to follow the general procedure

\[
\frac{1}{T} \mathbb{E} S_{2,K} \lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} R_{k+1}
\]

\[
\lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} h^2 \left( V_k^{p,4} + V_{k+1}^{p,4} \right) \mathbb{E}_T \| \Delta_{k+1} \|^4
\]

\[
\lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} h^2 \sup_i \mathbb{E} V_i^{p,4+2} (\theta_i)
\]

\[
\lesssim \frac{1}{T} h^2 \lesssim h.
\]

We apply the general procedure to \( \mathbb{E} S_{0,K} \)

\[
\frac{1}{T} \mathbb{E} S_{0,K} \lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} h^2 V_k^{p,3,3} \mathbb{E}_T \| g_k \eta_{k+1} \|^2 \| \hat{f}_k \|
\]

\[
\lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} h^2 \sup_i \mathbb{E} V_i^{p,3+\frac{3}{2}} (\theta_i)
\]

\[
\lesssim \frac{1}{T} h^2 \lesssim h.
\]

\[
\frac{1}{T} \mathbb{E} \tilde{S}_{0,K} \lesssim \frac{1}{T} \mathbb{E} \sum_{k=0}^{K-1} h^3 V_k^{p,3,3} \mathbb{E}_T \| \hat{f}_k \|^3
\]

\[
\lesssim \frac{1}{T} \sum_{k=0}^{K-1} h^3 \sup_i \mathbb{E} V_i^{p,3+\frac{3}{2}} (\theta_i)
\]

\[
\lesssim \frac{1}{T} K h^3 \lesssim h^2.
\]

Thus, we have established the bound on the bias given by Equation (42).

In order to establish the bound one the MSE in Equation (43) we note that Equation (38) yields

\[
\mathbb{E} \left( \frac{1}{K} \sum_{k=0}^{K-1} (\phi - \tilde{\phi}) \right)^2 \lesssim \mathbb{E} \left( \psi_K - \psi \right)^2 + \frac{1}{K^2} \mathbb{E} \left( \sum_{k=0}^{K-1} (A_n - A) \psi_k \right)
\]

\[
+ \frac{1}{T^2} \sum_{i=0}^{2} \mathbb{E} S_{i,K}^2 + \frac{1}{T^2} \sum_{i=0}^{2} \mathbb{E} M_{i,K}^2
\]

First we note that

\[
\mathbb{E} \left( \psi_K - \psi \right)^2 \lesssim \frac{1}{T^2} \sup_i \mathbb{E} V_i^{2p,0} \lesssim \frac{1}{T^2}.
\]

The \( S_{i,K}^2 \) terms can be bound in similar way as above with the additional use of Cauchy-Schwartz inequality to break the correlation between \( V_i^p \) and \( V_j^p \).
\[
\frac{1}{T^2} ES_{1,K}^2 \lesssim \frac{1}{T^2} \mathbb{E} \sum_{i,j=0}^{K-1} h^4 D^2 \psi_i \left[ \hat{f}_i, \hat{f}_i \right] D^2 \psi_j \left[ \hat{f}_j, \hat{f}_j \right]
\]
\[
\lesssim \frac{1}{T^2} \mathbb{E} \sum_{i,j=0}^{K-1} h^4 \| D^2 \psi_i \|_r \left\| \hat{f}_i \right\|^2 \| D^2 \psi_j \|_r \left\| \hat{f}_j \right\|^2
\]
\[
\lesssim \frac{1}{T^2} \sum_{i,j=0}^{K-1} h^4 E V_{i+1}^2 V_{j+1}^2
\]
\[
\lesssim \frac{1}{T^2} \sum_{i,j=0}^{K-1} h^4 \left( E V_{i+2}^4 \right)^{\frac{1}{2}} \left( E V_{j+2}^4 \right)^{\frac{1}{2}}
\]
\[
\lesssim \frac{1}{T^2} \sum_{i,j=0}^{K-1} h^4 \left( \sup_i E V_{i+2}^4 \right)
\]
\[
\lesssim K^2 h^4 \lesssim h^2.
\]

Similarly, we bound

\[
\frac{1}{T^2} ES_{2,K}^2 \lesssim \frac{1}{T^2} \mathbb{E} \sum_{i,j=0}^{K-1} R_{i+1} R_{j+1}
\]
\[
\lesssim \frac{1}{T^2} \mathbb{E} \sum_{i,j=0}^{K-1} \left( V_{i+1}^4 + V_{j+1}^4 \right) \left( V_{i+1}^4 + V_{j+1}^4 \right) E_r \| \Delta_{i+1} \|^4 E_r \| \Delta_{j+1} \|^4
\]
\[
\lesssim \frac{h^4}{T^2} \mathbb{E} \sum_{i,j=0}^{K-1} \left( V_{i+1}^4 + V_{j+1}^4 \right) \left( V_{i+1}^4 + V_{j+1}^4 \right) V_i^2 V_j^2
\]
\[
\lesssim \frac{h^4}{T^2} \mathbb{E} \sum_{i,j=0}^{K-1} \left( V_{i+1}^4 + V_{j+1}^4 \right) \left( V_{i+1}^4 + V_{j+1}^4 \right) V_i^2 V_j^2
\]
\[
\lesssim \frac{h^4}{T^2} \sum_{i,j=0}^{K-1} \left( \sup_i E V_{i+4}^4 \right)
\]
\[
\lesssim K^2 h^4 \lesssim h^2.
\]
Similar bounds can also be obtained for $\frac{1}{T^2}E S_{0,K}^2$ and $\frac{1}{T^2}E S_{0,K}^2$:

$$\frac{1}{T^2}E S_{0,K}^2 \lesssim \frac{h^4}{T^2} \sum_{i,j=0}^{K-1} \mathbb{E} V_{i}^{p_{0,3}} \left\| \hat{f}_i \right\| \left\| V_j^{p_{0,3}} \hat{f}_j \right\|$$

$$\lesssim \frac{h^4 K^2}{T^2} \sup_i \mathbb{E} V_{i}^{2p_{0,3} + 1} \lesssim h^2$$

$$\frac{1}{T^2}E S_{0,K}^2 \lesssim \frac{h^6}{T^2} \sum_{i,j=0}^{K-1} \mathbb{E} V_{i}^{p_{0,3}} \left\| \hat{f}_i \right\|^3 \left\| V_j^{p_{0,3}} \hat{f}_j \right\|^3$$

$$\lesssim \frac{h^6 K^2}{T^2} \sup_i \mathbb{E} V_{i}^{2p_{0,3} + 2} \lesssim h^4$$

For Martingale terms the cross terms vanish which allows us to obtain the following bounds:

$$\frac{1}{T^2}EM_{1,K}^2 = \frac{1}{T^2} \sum_{i=0}^{K-1} \left( E \sum_{i,j=0}^{K-1} D^{2p_{0,3}} \left( g_i, h_i, g_{i+1}, h_{i+1} - g_i, g_{i+1} \right) \right)^2$$

$$\lesssim \frac{1}{T^2} \sum_{i=0}^{K-1} \mathbb{E} V_{i}^{2p_{0,3} + 1} \lesssim \frac{1}{T}.$$
The additional part for the SGLD is the term corresponding to the Martingale $M_{4,K}$

$$\frac{1}{T^2} M_{4,K}^2 \lesssim \frac{1}{T^2} \mathbb{E} h^2 \sum_{k=0}^{K-1} (D\psi_k(H_k))^2$$

$$\lesssim \frac{1}{T^2} \mathbb{E} h^2 \sum_{k=0}^{K-1} V^{2p,1} \mathbb{E}_T \|H_k\|^2$$

$$\lesssim \frac{1}{T^2} h^2 \sum_{k=0}^{K-1} \mathbb{E} V^{2p,1} + 1$$

$$\lesssim \frac{h}{T}$$

For all these calculations to go through need $\sup_i EV_i^{p^*}$ to be bounded. Collecting the orders present we see that

$$p^* = \max \{2p_{\psi,2} + 2, 2p_{\psi,4} + 4, 2p_{\psi,3} + 1, 2p_{\psi,3} + 3\}$$

is sufficient.

\[\square\]

### 8.1 Regularity of Poisson Equation

Theorem 1 and 2 of [15] characterise the smoothness and growth of the solution to the Poisson equation associated with Equation (37). The Assumptions needed to apply the results are

$$\left\langle f(\theta), \frac{\theta}{\|\theta\|} \right\rangle \leq -r \|\theta\|, \quad \|\theta\| \geq M_0 \quad (A.1)$$

$$0 < \lambda_- \leq \left\langle g(\theta) g^*(\theta) \frac{\theta}{\|\theta\|}, \frac{\theta}{\|\theta\|} \right\rangle \leq \lambda_+ < \infty \quad (A.2)$$

This holds if Assumption 4.1 is satisfied with $V(\theta) = \|\theta\|^2 + 1$.

**Theorem 8.1.** [15] Let $\|f(\theta)\| \leq C_1 + C_2 \|\theta\|^2$ and $\bar{f} = 0$ and Equations (A.1) and (A.2) are satisfied. Then there exists a solution $\psi \in W^2_{\text{loc}}$ to the Poisson equation

$$A\psi = \phi - \bar{\phi}$$

1. If there is a $C$ such that

$$|f(\theta)| \leq C (1 + \|\theta\|)^\beta$$

for some $\beta < 0$, then $u$ is bounded. Moreover

$$\sup_{\theta} |\psi(\theta)| \leq C \sup_{\theta} |f(1 + \|\theta\|)^{-\beta}$$

and

$$\|\nabla \psi\| \leq C$$
2. if there exists a constant $C$ and some $\beta > 0$ such that

$$|f(\theta)| \leq C(1 + \|\theta\|)^\beta$$

then there exist a constant such that

$$|\psi(\theta)| \leq C'(1 + \|\theta\|)^\beta.$$

Finally there exists $C$ such that

$$\|\nabla\psi\| \leq C \left(1 + |\theta|^\beta\right).$$

**Remark 8.2.** We believe that assumption Theorem 8.1 can be weakened to be of the form of Equation (44) but it is out of the scope of this article to explore this direction.

In order to iterate Theorem 8.1 we note that the derivatives $\psi$ can be expressed as solution to Poisson equations with different RHSs.

\[ A\psi = \phi - \bar{\phi} \quad \text{(A.3)} \]
\[ A\partial_i \psi = \partial_i \phi - \frac{1}{2}\nabla \psi \cdot \partial_i f \quad \text{(A.4)} \]
\[ A\partial_{ij} \psi = \partial_{ij} \phi - \frac{1}{2}\nabla \partial_{ij} \psi \cdot \partial_i f - \frac{1}{2}\nabla \psi \cdot \partial_{ij} f \quad \text{(A.5)} \]
\[ A\partial_{ijk} \psi = \partial_{ijk} \phi - \frac{1}{2}\nabla \partial_{ijk} \psi \cdot \partial_i f - \frac{1}{2}\nabla \partial_{ij} \psi \cdot \partial_k f - \frac{1}{2}\nabla \psi \cdot \partial_{ijk} f \quad \text{(A.6)} \]

We will denote by $\beta_{\psi,i}$ numbers that satisfy

$$\sup_{|\alpha|=i} \|\partial^\alpha \psi\| \lesssim \left(1 + |\theta|^{\beta_{\psi,i}}\right)$$

where we used multi-index notation for derivatives. We use a similar notation for the derivatives of $f$, that is $\beta_{f,i}$ and assume that these bounds are a priori given.

Using Theorem 8.1 we can obtain $p_{\psi,i}$ to satisfy Equation (39) in terms of the $\beta$’s which we formulate as the following lemma.

**Lemma 8.3.** Suppose that $\phi$ and its derivatives are bounded and Assumption 4.1 and Equations Equations (A.1) and (A.2) hold. Then the choice

\[ p_{\psi,0} = 0 \]
\[ p_{\psi,1} = 0 \]
\[ p_{\psi,2} = \frac{\beta_{f,1}}{2} \]
\[ p_{\psi,3} = \beta_{f,1} \vee \frac{\beta_{f,2}}{2} \]
\[ p_{\psi,4} = \frac{1}{2} \left(3 \beta_{f,1} \vee (\beta_{f,1} + \beta_{f,2}) \vee \beta_{f,3}\right) \]

satisfies Equation (39).
Proof. Assumption 4.1 yields that $\beta_{f,0} = 1$ is a valid choice. Applying Theorem 8.1 to Equation (A.3) implies that $\beta_{\psi,0} := 0$ satisfies Equation (A.7). Applying Theorem 8.1 to Equation (A.4) yields that $\beta_{\psi,2} \leq \beta_{f,1}$.

Applying Theorem 8.1 to Equation (A.5) yields that $\beta_{\psi,3} \leq 2\beta_{f,1} \lor \beta_{f,2}$.

Applying Theorem 8.1 to Equation (A.6) yields that $\beta_{\psi,4} \leq (\beta_{f,1} + (2\beta_{f,1} \lor \beta_{f,2}) \lor (\beta_{f,1} + \beta_{f,2}) \lor \beta_{f,3}$



Thus we have established Equation (39). □

9 Modified SGLD in more than one dimensions

We present some calculations needed for the derivation the form of the mSGLD in higher than one dimensions. In particular, we first calculate $L^2_0$ assuming a general diffusion matrix in front of the noise

$$L^2_0 \phi = f(\theta) \cdot \nabla \left( f(\theta) \cdot \nabla \phi + \frac{1}{2} \Sigma \Sigma^T : \nabla \nabla \phi \right)$$

$$= f_k \partial_k f_i \partial_i \phi + \frac{1}{2} \Sigma \Sigma^T : \nabla \nabla \left( f(\theta) \cdot \nabla \phi + \frac{1}{2} \Sigma \Sigma^T : \nabla \nabla \phi \right)$$

$$= f_k \partial_k f_i \partial_i \phi + \frac{1}{2} \Sigma \Sigma^T : \nabla \nabla \left( f(\theta) \cdot \nabla \phi + \frac{1}{2} \Sigma \Sigma^T : \nabla \nabla \phi \right)$$

Now since for the noise we have $\Sigma = I$ we find that

$$L^2_0 \phi = f_k \partial_k f_i \partial_i \phi + f_k \partial_k \partial_k \partial_i \phi + \frac{1}{2} f_k \partial_k \partial^2 \phi$$

$$= \frac{1}{2} \left[ \partial^2_k f_i \partial_i \phi + 2 f_k \partial_k \partial_i \phi + f_i \partial^2_k \partial_i \phi \right] + \frac{1}{4} \partial^4 \phi$$

On the other hand a simple calculation (similar to the one dimensional one) shows that for SGLD one has

$$A_1 \phi = \text{E}(\hat{f}_i \hat{f}_j) \partial_i \partial_j \phi + \frac{1}{2} \text{E}(\hat{f}_i) \partial^2_k \partial_i \phi + \frac{1}{8} \partial^4 \phi$$

We thus obtain that

$$\frac{1}{2} L^2_0 - A_1 = \frac{1}{2} f_k \partial_k f_i \partial_i \phi + \frac{1}{2} f_k \partial_k \partial_i \phi + \frac{1}{4} \partial^2_k f_i \partial_i \phi + \frac{1}{2} \partial_k \partial_i \phi$$

35
where we have used the fact that $\mathbb{E}(\hat{f}_i) = f_i$. We are thus in exactly the same situation as in one dimension where in comparison to the Euler method we have an extra term of the form

$$\frac{1}{2}(f_k f_i - \mathbb{E}(\hat{f}_k \hat{f}_i))\partial_k \theta_j \phi$$

Thus one can obtain a modified SGLD in high dimensions by introducing an appropriate perturbation to the noise as in one dimension.

## 10 Gaussian Toy Model

### 10.1 Expected MSE for fixed Data Sets

We outline how an analytic expression for the expected MSE of the ergodic average can be derived. The following method generalises to any polynomial test function but we concentrate on the ergodic average for the second moment of the posterior given by

$$S_2 = \frac{1}{M} \sum_{k=0}^{M-1} \theta_j^2.$$  

Its MSE can be expressed using Equation (49)

$$MSE = \mathbb{E}\left(\frac{1}{M} \sum_{k=0}^{M-1} \theta_j^2 - (\mu_p^2 + \sigma_p^2)\right)^2 = \mathbb{E}S_2^2 - 2\mathbb{E}S_2 (\mu_p^2 + \sigma_p^2) + (\mu_p^2 + \sigma_p^2)^2. \quad (C.1)$$

In order to express $\mathbb{E}S_2^2$ in Equation (C.1) we derive the recurrence equations for $\mathbb{E}\theta_j^i$ for $i = 1, \ldots, 4$ by taking the expectations of

$$\begin{align*}
\theta_{j+1} &= (1 - Ah)\theta_j + hB_j + \sqrt{h}\eta_j \\
\theta_{j+1}^2 &= (1 - Ah)\theta_j^2 + h^2B_j^2 + h\eta_j^2 + 2(1 - Ah)\theta_j hB_j + 2(1 - Ah)\theta_j \sqrt{h}\eta_j + 2hB_j \sqrt{h}\eta_j 2 \\
\theta_{j+1}^3 &= \left((1 - Ah)\theta_j + hB_j + \sqrt{h}\eta_j\right)^3 \\
\theta_{j+1}^4 &= (1 - Ah)^4\theta_j^4 + 4(1 - Ah)^3\theta_j^3 hB_j + 6(1 - Ah)^2\theta_j^2 h^2B_j^2 + 4(1 - Ah)\theta_j^2 h^3B_j^3 + h^4B_j^4 \\
&+ 4(1 - Ah)\theta_j^2 h^3B_j^3 + h^4B_j^4 + 4\eta_j^3 (\ldots) + 4\eta_j^3 (\ldots)  \\
&+ 6h\eta_j^2 ((1 - Ah)^2\theta_j^2 + 2(1 - Ah)\theta_j hB_j + h^2B_j^2) + \eta_j^4 h^2. \quad (C.3) \\
\end{align*}$$

The recurrent equation for $\mathbb{E}\theta_j$ is linear and first order and can therefore be solved explicitly. Plugging the result into the equation for $\mathbb{E}\theta_j^2$ turns it into a first order linear equation as well. Repeating this processes yields explicit expressions for $\mathbb{E}\theta_j^i$ $i = 1, \ldots, 4$. The sums can be carried out explicitly because the terms are of the form of a geometric sum or a geometric term with a polynomial factor. This allows us to obtain an analytic expression for $\mathbb{E}S_2^2$ by reducing it to $\mathbb{E}\theta_j^2$ as follows

$$\mathbb{E}S_2^2 = \frac{1}{M^2} \mathbb{E}\left(\sum_{i=0}^{M-1} \theta_j^i + 2 \sum_{i=0}^{M-1} \theta_j^i \sum_{j=i+1}^{M-1} \theta_j^i \right). \quad (C.6)$$
The cross terms can be removed using Equation (C.2) so that
\[ \theta_j^2 = (1 - Ah)^{2(j-i)} \theta_i^2 + \sum_{k=0}^{j-i-1} (1 - Ah)^{2k} \]
\[ \left[ h^2 B_{j-1-k}^2 + h^2 \eta_{j-1-k}^2 + 2 (1 - Ah) B_{j-1-k} h \theta_{j-1-k} + 2 (1 - Ah) \eta_{j-1-k} \theta_{j-1-k} \right]. \]

Plugging this into Equation (C.6) yields
\[ \mathbb{E}S_2^2 = \mathbb{E} \frac{1}{M^2} \sum_{i=0}^{M-1} \theta_i^4 \left( 1 + 2 \sum_{j=i+1}^{M-1} (1 - Ah)^{2(j-i)} \right) \]
+ \[ \mathbb{E} \frac{1}{M^2} \sum_{i=0}^{M-1} \theta_i^2 \sum_{j=i+1}^{M-1} \sum_{k=0}^{j-i-1} (1 - Ah)^{2k} \]
\[ \left[ h^2 B_{j-1-k}^2 + h \eta_{j-1-k}^2 + 2 (1 - Ah) B_{j-1-k} h \theta_{j-1-k} + 2 (1 - Ah) \eta_{j-1-k} \theta_{j-1-k} \right]. \]

Using the recurrence Equation to express \( \theta_{j-1-k} \) in terms of \( \theta_i \) we conclude that \( \mathbb{E}S_2^2 \) is equal to
\[ \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E} \theta_i^4 \left( 1 + 2 \sum_{j=i+1}^{M-1} (1 - Ah)^{j-i} \right) \]
+ \[ \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E} \theta_i^2 \sum_{j=i+1}^{M-1} \sum_{k=0}^{j-i-1} (1 - Ah)^{2k} \left[ h^2 \mathbb{E} B^2 + h \right]. \]
+ \[ \mathbb{E} \frac{1}{M^2} \sum_{i=0}^{M-1} \theta_i^2 \sum_{j=i+1}^{M-1} \sum_{k=0}^{j-i-1} (1 - Ah)^{2k} 2 (1 - Ah) B_{j-1-k} h \]
\[ \left( (1 - Ah)^{(j-1-k)-i} \theta_j + \sum_{l=0}^{(j-1-k)-i-1} (1 - Ah)^l \left( h B_{j-1-k-l-1} + \sqrt{h} \eta_{j-1-k-l-1} \right) \right) \]
Taking the expectations into the sum yields

\[
\mathbb{E} S_2^2 = \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E} \theta_i^4 \left( 1 + 2 \sum_{j=i+1}^{M-1} (1 - Ah)^{j-i} \right)
\]

\[
+ \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E} \theta_i^2 \sum_{j=i+1}^{M-1} \sum_{k=0}^{j-1-i} (1 - Ah)^{2k} \left[ h^2 \mathbb{E} B^2 + h \right].
\]

\[
+ \mathbb{E} \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E} \theta_i^3 \sum_{j=i+1}^{M-1} \sum_{k=0}^{j-1-i} (1 - Ah)^{2k} 2 (1 - Ah) \mathbb{E} Bh (1 - Ah)^{(j-1-k)-i}
\]

\[
+ \mathbb{E} \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E} \theta_i^4 \sum_{j=i+1}^{M-1} \sum_{k=0}^{j-1-i} (1 - Ah)^{2k} 2 (1 - Ah) \mathbb{E} Bh \sum_{l=0}^{(j-1-k)-i-1} (1 - Ah)^l h \mathbb{E} B.
\]

We have an expressions for \( \mathbb{E} B \) but in the following we derive the expressions for \( \mathbb{E} B^2 \) required to express \( \mathbb{E} S_2^2 \). The terms \( \mathbb{E} B^3 \) and \( \mathbb{E} B^4 \) are needed for the derivation of \( \mathbb{E} \theta_j^4 \). In order to derive expressions for \( \mathbb{E} B^p \), we introduce the power sums

\[
p_k = \sum_{i=1}^N X_i^k
\]

and the elementary symmetric polynomials

\[
e_0 = 1,
\]

\[
e_1 = X_1 + X_2 + \cdots + X_N
\]

\[
e_2 = \sum_{1 \leq i < j \leq N} X_i X_j,
\]

\[
\vdots = \vdots
\]

\[
e_N = \prod_{i=1}^N X_i.
\]

Computing \( e_i \) naively has complexity of order \( O \left( N^i \right) \) for \( i \ll N \). Using Newton’s identities

\[
e_1 = p_1
\]

\[
e_2 = \frac{1}{2} (e_1 p_1 - p_2)
\]

\[
e_3 = \frac{1}{3} (-e_1 p_2 + e_2 p_1 + p_3)
\]

\[
e_4 = \frac{1}{4} (e_1 p_3 - e_2 p_2 + e_3 p_1 - p_4),
\]

\( e_i \) can be expressed in terms of \( p_k, k \leq i \) each of which can be computed with complexity of order \( O(N) \).

We consider the term \( B = \frac{N}{n} \sum_{i=1}^n \frac{X_{\tau_i}}{2 \sigma^2} \) where \( \tau_i \) are sampled with replacement from a fixed
data set \{1, \ldots, N\}. The second moment can be calculated as follows

\[
\mathbb{E}B^2 = \left( \frac{N}{n^2\sigma_x^2} \right)^2 \sum_{i,j} \mathbb{E}X_{\tau_i}X_{\tau_j}
\]

\[
= \left( \frac{N}{n^2\sigma_x^2} \right)^2 \left( n(n-1) \mathbb{E}X_{\tau_1}X_{\tau_2} + n \mathbb{E}X_{\tau_1}X_{\tau_3} \right) / \text{Mom}_{2,1,2}.
\]

We use Newton’s identities to express \text{Mom}_{2,1} and \text{Mom}_{2,2}

\[
\text{Mom}_{2,1} = \frac{2e_2}{N(N-1)}
\]

\[
\text{Mom}_{2,2} = \frac{p_2}{N}.
\]

Similarly, we obtain

\[
\mathbb{E}B^3 = \left( \frac{N}{n^2\sigma_x^2} \right)^3 \sum_{i,j,k} \mathbb{E}X_{\tau_i}X_{\tau_j}X_{\tau_k}
\]

\[
= \left( \frac{N}{n^2\sigma_x^2} \right)^3 \left( n(n-1)(n-2) \mathbb{E}X_{\tau_1}X_{\tau_2}X_{\tau_3} + 3n(n-1) \mathbb{E}X_{\tau_1}X_{\tau_2}^2 + n \mathbb{E}X_{\tau_1}X_{\tau_3}^2 \right) / \text{Mom}_{3,1,2,3}.
\]

\[
\text{Mom}_{3,1} = \frac{\sum_{i \neq j \neq l} X_iX_jX_l}{N(N-1)(N-2)} = \frac{6e_3}{N(N-1)(N-2)}
\]

\[
\text{Mom}_{3,2} = \frac{p_1p_2 - p_3}{N(N-1)}
\]

\[
\text{Mom}_{3,3} = \frac{p_3}{N}.
\]

A similar calculation yields
\[ E \theta_j^4 = \left( \frac{N}{n2\sigma_x^2} \right)^4 \sum_{i,j,k,l} \mathbb{E} X_{r_i} X_{r_j} X_{r_k} X_{r_l} \]

\[ = \left( \frac{N}{n2\sigma_x^2} \right)^4 \left( n(n-1)(n-2)(n-3) \mathbb{E} X_{r_1} X_{r_2} X_{r_3} X_{r_4} + \left( \frac{4}{2} \right) n(n-1)(n-2) \mathbb{E} X_{r_1}^2 X_{r_2} X_{r_3} \right) \left( \frac{Mom_{4,1}}{Mom_{4,2}} \right) \]

\[ + \left( \frac{N}{n2\sigma_x^2} \right)^4 \left( \frac{1}{2} \left( \frac{4}{2} \right) n(n-1) \mathbb{E} X_{r_1}^2 X_{r_2} + \left( \frac{4}{3} \right) n(n-1) \mathbb{E} X_{r_1}^2 X_{r_3} + n \mathbb{E} X_{r_1}^4 \right) \left( \frac{Mom_{4,3}}{Mom_{4,4}} \right) \]

\[ \frac{Mom_{4,1}}{Mom_{4,2}} = \frac{N(N-1)(N-2)(N-3)}{N(N-1)(N-2)(N-3)} = \frac{24e_4}{N(N-1)(N-2)(N-3)} \]

\[ Mom_{4,2} = \frac{2e_2p_2 - 2p_3p_1 + 2p_4}{N(N-1)(N-2)} \]

\[ Mom_{4,3} = \frac{p_2p_2 - p_4}{N(N-1)} \]

\[ Mom_{4,4} = \frac{p_3p_1 - p_4}{N(N-1)} \]

\[ Mom_{4,5} = \frac{p_4}{N}. \]

### 10.2 Expected MSE for Random Data

We sketch the derivation of the expected MSE

\[ \mathbb{E}_{\theta, X} \left( \frac{1}{M} \sum_{k=0}^{M-1} \theta_j^2 - \left( \mu_p^2 + \sigma_p^2 \right) \right)^2 \]

where we take expectation with respect to \( X_i \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma_X^2) \) for \( i = 1, \ldots, N \) and the randomness in the recursion for \( \theta_j \). We obtain an analytic expression for the MSE by deriving expressions for \( \mathbb{E}_{X, \theta} \theta_j^p \). We illustrate the computation for \( p = 2 \), noting that we assume \( \theta_0 = 0 \) a.s. We know that

\[ \theta_j^2 = \sum_{k=0}^{j-1} (1 - Ah)^{2k} \left[ h^2 B_{j-1-k}^2 + h \eta_{j-1-k} + 2 (1 - Ah) \eta_{j-1-k} \theta_{j-1-k} \sqrt{h} \right] \]

\[ + 2h^2 B_{j-1-k} \eta_{j-1-k} + 2 (1 - Ah) B_{j-1-k} \theta_{j-1-k} \theta_{j-1-k} \] (C.8)

\[ \sum_{k=0}^{j-1} (1 - Ah)^{2k} \left[ h^2 B_{j-1-k}^2 + 2 (1 - Ah) \eta_{j-1-k} \theta_{j-1-k} \sqrt{h} + 2h^2 B_{j-1-k} \eta_{j-1-k} \right] \] (C.9)

\[ h \eta_{j-1-k}^2 + 2 (1 - Ah) B_{j-1-k} h \sum_{l=0}^{j-1-k-1} (1 - Ah)^l \left( h B_{j-2-l} + \sqrt{h} \eta_{j-2-l} \right) \] (C.10)

The expectation \( \mathbb{E}_{X, \theta} \theta_j \) therefore boils down to calculating \( E B^2 \), \( EB B' \) and \( EB \) where \( B \) and \( B' \) are independent samples of Equation (52). We start by calculating
\[
E_{BB'} = \frac{N^2}{n^24\sigma_x^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}X_iX_j
= \frac{N^2}{n^24\sigma_x^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{1}{N}\mathbb{E}X^2 + \frac{N-1}{N}\mathbb{E}X\bar{X} \right)
= \frac{N^2}{n^24\sigma_x^2} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{N}(\theta^{12} + \sigma_x^2) + \frac{N-1}{N}\theta^{12} \right).
\]

Similarly, we obtain
\[
E_{B^2} = \frac{N^2}{n^24\sigma_x^4} \frac{n(\theta^{12} + \sigma_x^2) + n(n-1)\theta^{12}}{1}.
\]

Deriving \(\mathbb{E}\theta_j^p\) for \(p = 1, 2, 3, 4\) requires the calculation of \(\mathbb{E}B_1^{\alpha_1}B_2^{\alpha_2}B_3^{\alpha_3}B_4^{\alpha_4}\) where \(B_i\) are i.i.d. following the distribution of Equation (52) for \(\alpha_i \geq 0\) and \(\sum_i \alpha_i \leq 4\). The arguments so far allow us to derive \(T_1\) and \(T_3\) in

\[
MSE = \mathbb{E} \left( S_2^2 - 2S_2(\mu_p^2 + \sigma_p^2) + (\mu_p^2 + \sigma_p^2)^2 \right)
\]
\[
= \frac{\mathbb{E}S_2^2}{T_1} - 2\mathbb{E}S_2\mu_p^2 - 2\mathbb{E}S_2\sigma_p^2 + \mathbb{E} \left( \mu_p^4 + 2\mu_p^2\sigma_p^2 + \sigma_p^4 \right) \frac{1}{T_3} + \mathbb{E} \left( \mu_p^4 + 2\mu_p^2\sigma_p^2 + \sigma_p^4 \right) \frac{1}{T_4}.
\]

Recall that the posterior for this toy model is given by
\[
N(\mu_p, \sigma_p^2) = N \left( \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} X_i^2 + N}, \left( \frac{1}{\sigma_x^2} + \frac{N}{\sigma_x^2} \right)^{-1} \right)
\]
and hence \(T_4\) can be computed explicitly. The summands of \(T_2\) can be derived similarly to Equation (C.8) in terms of the quantities \(\mathbb{E}B\mu_p^2, \mathbb{E}B^2\mu_p^2\) and \(\mathbb{E}BB'\mu_p^2\). The explicit expression can be obtained from the supplemented Mathematica file.