Optimal Stochastic Gradient Langevin Dynamics for Approximate Bayesian Inference

Ayoub El hanchi

January 11, 2019

1 Introduction

In this report, we study approximate Bayesian inference using stochastic gradient Markov chain Monte Carlo (SG MCMC) algorithms for densities with an approximately quadratic logarithm. Our goal is to find the optimal algorithm in the subclass of stochastic gradient Langevin dynamics (SGLD) for this special case. This is motivated by two facts. First, heuristically, we do not expect to perform much better on other types of densities since this type of density is the simplest in many respects. Studying this special case gives us therefore a heuristic lower bound on the performance of these algorithms. Second, under some regularity conditions, the Bernstein-von Mises theorem asserts that the posterior distribution converges to a normal distribution in total variation as the number of observations becomes large. We can therefore expect the quadratic assumption to approximately hold when the size of the dataset is large, and we expect our algorithm to perform near optimally in that setup.

To motivative our definition of optimality, we start by recalling the core results in the theory of discrete-time Markov chains as it relates to MCMC methods. We then state recent similar results in the theory of continuous-time Markov processes, and use these results to construct stochastic gradient MCMC algorithms following the approach taken in [3]. Finally we consider the special case of densities with approximately quadratic logarithms, and derive an optimal algorithm in the subclass of stochastic gradient Langevin dynamics algorithms.

2 Relevant MCMC theory results

2.1 Monte Carlo integration and Bayesian inference

The major computational challenge in parametric Bayesian inference is the evaluation of the expectation corresponding to the posterior predictive distribution:

$$p(Z_{n+1}|Z_1^n = z_1^n) = \int_\Theta p(Z_{n+1}|\theta)p(\theta|Z_1^n = z_1^n)d\theta = \mathop{\mathbb{E}}_{\theta\sim P(\theta|Z_1^n=z_1^n)}[p(Z_{n+1}|\theta)] \tag{1}$$

Analytical solutions to this integral exist only in rare cases, when the prior and posterior distributions are conjugate for example. Deterministic numerical integration methods generally

require the evaluation of the integrand at an exponential number of points in the dimension of the space of parameters Θ, and are therefore infeasible for large dimensions.

An alternative to deterministic schemes are stochastic ones, usually referred to as Monte Carlo methods. The Monte Carlo estimator of (1) is given by:

$$\mathop{\mathbb{E}}_{\theta\sim P(\theta|Z_1^n=z_1^n)}[p(Z_{n+1}|\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} p(Z_{n+1}|\theta=\tilde{\theta}_i) =: \langle I\rangle \tag{2}$$

where $(\tilde{\theta}_i)_{i=1}^N$ is a finite sequence of independent and identically distributed random variables with distribution $P(\theta|Z_1^n = z_1^n)$. By the strong law of large numbers, the Monte Carlo estimator converges to the expectation almost surely as $N \to \infty$. Using the i.i.d assumption on $(\tilde{\theta}_i)_{i=1}^N$ we also have by the central limit theorem, assuming that the variance $\sigma := \mathop{\mathrm{Var}}_{\theta\sim P(\theta|Z_1^n=z_1^n)}[p(Z_{n+1}|\theta)]$ exists:

$$\langle I\rangle \to \mathop{\mathbb{E}}_{\theta\sim P(\theta|Z_1^n=z_1^n)}[p(Z_{n+1}|\theta)] + \mathcal{N}(0, \frac{1}{N}\sigma) \;\text{ as }\; N\to\infty \;\text{ in distribution}$$

If we use the standard deviation as a measure of error, then this tells us that the Monte Carlo estimator converges to the true value of the expectation with an error proportional to $\frac{1}{\sqrt{N}}$, which is independent of dimension !

The challenge then remains in generating a realization of the sequence $(\tilde{\theta}_i)_{i=1}^N$. Unfortunately, in most cases, the problem of generating this sequence is as difficult as the original integration problem. A natural question to ask therefore is whether we can preserve a convergence rate independent of the dimension of Θ if we replace the sequence $(\tilde{\theta}_i)_{i=1}^N$ with a "weakly dependent" sequence $(\theta_i)_{i=1}^N$ where each θ_i is distributed according to some distribution that is "close" to $P(\theta|Z_1^n = z_1^n)$. With conditions on how "close" these distributions are, and how "weak" the dependence between the θ_i's is, the answer turns out to be yes, and this is one important motivation for Markov chain Monte Carlo methods.

2.2 Discrete-time MCMC

A computationally convenient and well studied way of obtaining the sequence described in the previous paragraph is by simulating a Markov chain $(\theta_i)_{i\in\mathbb{N}}$ with the desired properties of "weak dependence" and "closeness" to the posterior distribution. A reasonable way to formally define the latter is by requiring that the distribution of the θ_i's approaches the posterior distribution as $n \to \infty$, independently of the starting value θ_0.

Definition 1 *A Markov chain $(\theta_i)_{i\in\mathbb{N}}$ with stationary distribution π is said to be Harris ergodic if:*

$$\forall\theta_0\in\Theta \quad \|P^n(\theta_0,\cdot)-\pi\|_{TV} \to 0 \text{ as } n\to\infty$$

Irreducible, aperiodic, Harris recurrent Markov chains fully characterize Harris ergodic chains [1].

One way to formalize our intuitive notion of "weak dependence" is by requiring that the random variables $(\theta_j)_{j=0}^{i}$ and $(\theta_j)_{j=i+n}^{\infty}$ become independent as $n \to \infty$ for all $i \in \mathbb{N}$. A chain that satisfies this property is said to be mixing, and the precise sense in which we mean independence gives rise to different notions of mixing. Here we restrict ourselves to the notion of strong mixing which is the most useful in our context.

Definition 2 *A Markov chain $(\theta_i)_{i\in\mathbb{N}}$ is said to be strongly mixing if:*

$$\alpha(n) := \sup\{|P(A \cap B) - P(A)P(B)| \mid A \in \sigma(\Theta_0^t), B \in \sigma(\Theta_{t+n}^{\infty}), t \in \mathbb{N}\} \to 0 \text{ as } n \to \infty$$

where $\sigma(\Theta_i^j)$ denotes the product sigma algebra on Θ^{j-i+1}.

Up until now we have been thinking about dependence and convergence separately, but in fact, it can be shown that Harris ergodicity implies strong mixing [1]. We might therefore expect that Harris ergodicity is enough to prove a central limit theorem, but this is in fact not the case. Necessary and sufficient conditions for the existence of a central limit theorem for Harris ergodic chains exist [2], but these conditions are somewhat difficult to check, so we consider here sufficient conditions based on the rate of convergence of the chain. We start by making the following definitions:

Definition 3 *A Markov chain $(\theta_i)_{i\in\mathbb{N}}$ with stationary distribution π satisfying:*

$$\forall \theta_0 \in \Theta \quad \|P^n(\theta_0, \cdot) - \pi\|_{TV} \leq M(\theta_0)\gamma(n)$$

for real valued functions M and γ is said to be:

- *polynomially ergodic of order m if $\gamma(n) \leq an^{-m}$ for some constants $a \in \mathbb{R}$ and $m \in \mathbb{N}$*
- *geometrically ergodic if $\gamma(n) \leq ar^n$ for some constants $a \in \mathbb{R}$ and $r \in [0, 1)$*
- *uniformly ergodic if it is geometrically ergodic and $M(\theta_0)$ is bounded*

Depending on the rate of convergence of the chain and the nature of the functional we would like to estimate the expectation of, different central limit theorems exist. The following theorem summarizes some of the well know results, and is taken from [2]:

Theorem 1 *Let $(\theta_i)_{i\in\mathbb{N}}$ be a Harris ergodic Markov chain with stationary distribution π, and let $f : \Theta \to \mathbb{R}^k$ be a measurable function. If any of the following conditions hold:*

- *$(\theta_i)_{i\in\mathbb{N}}$ is polynomially ergodic of order $m > 1$, $\mathop{\mathbb{E}}_{\theta_0\sim\pi}[M(\theta_0)] < \infty$, and f is bounded π-almost surely.*
- *$(\theta_i)_{i\in\mathbb{N}}$ is geometrically ergodic, and $\exists\delta > 0$ such that $\mathop{\mathbb{E}}_{\theta\sim\pi}\left[f(\theta)^{2+\delta}\right]$ exists.*
- *$(\theta_i)_{i\in\mathbb{N}}$ is uniformly ergodic, and $\mathop{\mathbb{E}}_{\theta\sim\pi}[f(\theta)^2]$ exists.*

Then for any choice of distribution of θ_0:

$$\frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \to \mathop{\mathbb{E}}_{\theta\sim\pi}[f(\theta)] + \mathcal{N}(0, \frac{1}{N}\sigma_f) \text{ as } N \to \infty$$

where $\sigma_f := \mathop{\mathrm{Var}}_{\theta_0\sim\pi}[f(\theta_0)] + 2\sum_{i=1}^{\infty} \mathop{\mathrm{Cov}}_{\theta_0\sim\pi}[f(\theta_0), f(\theta_i)] = \mathop{\mathrm{Var}}_{\theta_0\sim\pi}[f(\theta_0)] + 2\gamma_f$

Remark 1 *It can seem a little counter-intuitive that the variance σ_f does not depend on the initial choice of θ_0, but it can be shown that for Harris ergodic chains, if a central limit theorem holds for some initial distribution of θ_0, then it holds for any other initial distribution [2]. In particular, we expect the same variance asymptotically whether the chain starts off in its stationary distribution or in some other distribution.*

Traditional MCMC methods such as the Metropolis-Hastings algorithm and the Gibbs sampler are based on yet another sufficient condition for the central limit theorem, namely reversibility of the Markov chain. Reversibility is easier to impose in the construction of Markov chains, but it is more restrictive than the above conditions.

2.3 Continuous-time MCMC

Looking at Theorem 1, we see that the asymptotic variance σ_f depends on the auto-covariance of the chain. A chain with a large auto-covariance can thus be very inefficient. Another practical concern in the application of MCMC methods is the rate of convergence of the distribution of the estimator to the normal distribution of the central limit theorem. Many methods have been developed to deal with these two problems, some of which are gradient-based MCMC methods. Probably the most well known such methods are Hamiltonian Monte Carlo (HMC) and the Metropolis-adjusted Langevin algorithm (MALA). As oppose to traditional MCMC algorithms, these algorithms are based on continuous time Markov processes.

In this section, we mention some important results in the development of continuous-time MCMC methods, the first of which is taken from [3], and is a characterization of the set of continuous-time Markov processes with a desired stationary distribution π.

Theorem 2 *Let $(q_t)_{t\in[0,\infty)} := ((\theta_t, r_t))_{t\in[0,\infty)}$ be a continuous-time Markov process with stationary distribution $\tilde{\pi}$ with density $\exp\{-H(z)\}$ such that the marginal stationary distribution of θ is π with density $\exp\{-U(\theta)\}$. Then there exists a matrix valued function $D(q)$ such that the Markov process $(q_t)_{t\in[0,\infty)}$ satisfies the stochastic differential equation (SDE):*

$$dq = \mu(q)dt + \sigma(q)dW_t \tag{3}$$

where:

$$\mu(q) := D(q)\nabla_q \log(H(q)) + \Gamma(q)$$
$$\sigma(q) := \sqrt{(D(q) + D(q)^T)}$$
$$\Gamma_i(q) = \sum_{j=1}^{d} \frac{\partial}{\partial q_j} D_{ij}(q)$$

The above theorem is particularly important from an optimization point of view since it allows us to explore the space of all continuous Markov process with the right invariant distribution by varying the function $D(q)$.

In order to build continuous-time Markov processes for our purposes, we need at least ergodicity of the process. Various results on ergodicity of continuous time Markov processes are given in [4, 5], but they are quite technical. Assuming the existence of a stationary distribution, a sufficient condition for ergodicity of a process of the type (3) is global Lipschitzness of $\mu(q)$ and $\sigma(q)$, although this might be too restrictive since some densities of interest do not have a Lipschitz gradient. Another sufficient condition is local Lipschitzness of $\mu(q)$ and $\sigma(q)$ and existence of a solution to the stochastic differential equation (3) for all times $t \in [0, \infty)$.

A complete treatment of the construction of MCMC methods using continuous-time Markov processes would require the equivalent of a central limit theorem in continuous-time, and a reliable way to check the conditions of such a theorem. In practice however, we are only able to simulate the process (3) by discretizing it. The simplest discretization scheme is the Euler–Maruyama method which replaces the SDE (3) by the discretized version:

$$\Delta q = \epsilon\mu(q) + \sqrt{\epsilon}\sigma(q)w_t \text{ where } w_t \sim \mathcal{N}(0, I) \text{ for some } 0 < \epsilon \ll 1 \tag{4}$$

which describes a discrete-time Markov chain and for which the results in section 2.3 hold. Therefore, one strategy for giving a rigorous theoretical justification for the use of a process satisfying (4) in an MCMC algorithm would be to show the following, under conditions on $\mu(q)$, $\sigma(q)$, and ϵ:

- The existence of a stationary distribution $\tilde{\pi}'$ for the process described by (4).
- The applicability of the central limit theorem on (4).
- The closeness of $\tilde{\pi}'$ to the stationary distribution $\tilde{\pi}$ of the continuous-time process described by (3) in a total variation sense.

To my knowledge, no such results exist for reasonably general conditions on $\mu(q)$ and $\sigma(q)$, and it is, I think, an interesting problem to consider for future investigations.

3 Optimal SGLD for approximately quadratic log densities

3.1 Stochastic gradient Langevin dynamics

For large datasets, simulating the process (4) becomes too computationally expensive since the gradient is a sum over the contributions from each observation. One way to overcome this problem is by replacing the full gradient with a stochastic gradient. For large enough batch sizes, we can use the central limit theorem to quantify the added noise and preserve the dynamics of the process. We give here a formal implementation of this idea in the context

of Bayesian inference following the approach taken in [3]. The original SGLD algorithm was first given in [8].

For the rest of this report, we will restrict our attention to the case where $z := \theta$ and $D(\theta) := D$ is constant in (3). The resulting MCMC algorithms from SDEs of this type are usually referred to as Langevin type algorithms. The discretized version is given by:

$$\Delta\theta = -\epsilon D \nabla_\theta U(\theta) + \mathcal{N}(0, \epsilon(D + D^T)) \tag{5}$$

In Bayesian inference and under the i.i.d. assumption, we have:

$$\begin{aligned} -\nabla_\theta \log p(\theta|Z_1^N) \propto \nabla_\theta U(\theta) &= -\nabla_\theta \log p(Z_1^N|\theta) - \nabla_\theta \log p(\theta) \\ &= \left(\sum_{i=1}^N -\nabla_\theta \log p(Z_i|\theta)\right) - \nabla_\theta \log p(\theta) \\ &=: \left(\sum_{i=1}^N g(Z_i, \theta)\right) + h(\theta) \end{aligned}$$

The idea behind a stochastic gradient is to replace $\sum_{i=1}^N g(Z_i, \theta)$ with an unbiased estimate that is easier to compute. Let $S := \{S_i\}_{i=1}^m$ where the S_i's are independent uniform random variable on $[N] := \{1, \ldots, N\}$ and where $|S| := m \ll N$. Then we define the stochastic gradient to be the random variable:

$$\nabla_\theta^S U(\theta) := N\left(\frac{1}{|S|}\sum_{S_i \in S} g(Z_{S_i}, \theta)\right) + h(\theta)$$

It is clear by linearity of expectation and the definition of S that the stochastic gradient is an unbiased estimator of the full gradient. By the central limit theorem for an independent sequence we also have:

$$\nabla_\theta^S U(\theta) \xrightarrow{d} \nabla_\theta U(\theta) + \mathcal{N}\left(0, \frac{N^2}{S} C(\theta)\right) \text{ as } |S| \to \infty$$

where:

$$C(\theta) := \operatorname*{Var}_{i \sim \mathcal{U}([N])}[g(Z_i, \theta)] = \frac{1}{N}\sum_{i=1}^N \left[g(Z_i, \theta) - \frac{1}{N}\sum_{j=1}^N g(Z_j, \theta)\right]\left[g(Z_i, \theta) - \frac{1}{N}\sum_{j=1}^N g(Z_j, \theta)\right]^T$$

Assuming that the central limit approximation holds, we replace the full gradient with the stochastic gradient in (5) and remove the added noise to get the discretized stochastic gradient Langevin dynamics equation:

$$\Delta\theta = -\epsilon D \nabla_\theta^S U(\theta) + \mathcal{N}\left(0, \epsilon(D + D^T) - \epsilon^2 \frac{N^2}{S} D C(\theta) D\right) \tag{6}$$

Assuming that the conditions outlined at the end of section 2.3 hold for this choice of $\mu(\theta)$, $\sigma(\theta)$, and a choice of ϵ, the chain described by this update equation converges to a distribution close to the posterior. An interesting question to ask is how can we choose the matrix D so that we maximize the efficiency of our sampler ? This is the problem we discuss in the next sections.

Note that $C(\theta)$ is in general a function of θ, and a full computation of it at each step defeats the purpose of using a stochastic gradient since it requires a sum over all observations. One could try to estimate $C(\theta)$ at each step using only the current batch, but in general we have $|S| \ll N$, and such an estimate is not precise enough to preserve the dynamics of the process.

3.2 Discretized Langevin dynamics for approximately quadratic log densities

For the rest of this report, we will assume that the posterior $\pi(\theta) \propto \exp\{-U(\theta)\}$ satisfies:

$$U(\theta) \approx \frac{1}{2}(\theta - \overline{\theta})^T H(\theta - \overline{\theta})$$

where $\overline{\theta}$ is the maximum a posteriori estimate, and where H is the Hessian matrix at $\overline{\theta}$. We will assume without loss of generality that $\overline{\theta} = 0$ so that the posterior is approximately $\mathcal{N}(0, H^{-1})$.

We write $H = H' + G$ where H' is the Hessian of the likelihood and G is the Hessian of the prior, both of which are assumed to be approximately constant and evaluated at $\overline{\theta}$. By the assumption that H' is approximately constant, we have the relationship $H' \approx NC(\theta)$, and therefore $C(\theta) \approx C$ is approximately constant and can be evaluated only once at $\overline{\theta}$.

With these assumptions, the discretized Langevin dynamics equation (5) becomes:

$$\theta_{i+1} = (I - \epsilon DH)\theta_i + \mathcal{N}(0, \epsilon(D + D^T)) \tag{7}$$

Processes of this form are known as Gaussian VAR(1) processes. It is easy to verify that as $\epsilon \to 0$, this process has the posterior as its stationary distribution, but this does not hold for $\epsilon > 0$. Our next task will be to find the family of Gaussian VAR(1) chains that are Harris ergodic whose stationary distribution is the posterior.

3.3 Harris ergodic Gaussian VAR(1) processes converging to the posterior

We consider processes of the form:

$$\theta_i = (I - \epsilon AH)\theta_{i-1} + \mathcal{N}(0, \epsilon B)$$

where it is assumed that $B \succeq 0$. Expanding the right-hand side and using the affine property of the normal distribution we get:

$$\theta_i \sim \mathcal{N}\left((I - \epsilon AH)^i \theta_0, \sum_{j=0}^{i-1} (I - \epsilon AH)^j \epsilon B (I - \epsilon HA^T)^j\right)$$

Therefore a necessary condition for the convergence of the mean to 0 is:

$$0 \prec \epsilon\lambda(AH) \prec 2$$

where $\lambda(AH)$ is the vector of eigenvalues of AH. This condition also guarantees the convergence of the series $S = \sum_{j=0}^{\infty}(I - \epsilon AH)^j \epsilon B(I - \epsilon HA^T)^j$ representing the asymptotic variance.

Notice that:

$$S = \epsilon B + (I - \epsilon AH)S(I - \epsilon HA^T)$$

Setting $S = H^{-1}$ we get the condition:

$$B = 2A - \epsilon AHA^T$$

in addition, since $AHA^T \succeq 0$, we have $2A = B + \epsilon AHA^T \succeq 0$ and therefore A must be positive semi-definite, and in particular symmetric.

All Gaussian VAR(1) processes converging to the posterior in distribution are therefore given by:

$$\theta_i = (I - \epsilon AH)\theta_{i-1} + \mathcal{N}(0, 2\epsilon A - \epsilon^2 AHA) \tag{8}$$

with the conditions:

$$2A - \epsilon AHA \succeq 0$$
$$0 \prec \epsilon\lambda(AH) \prec 2$$

and to allow for the use of a stochastic gradient, we see from (6) that we need to replace the first condition with the stronger condition:

$$2A - \epsilon AHA - \epsilon\frac{N^2}{S}ACA \succeq 0$$

Convergence in distribution is, of course, not enough for our purposes. It can be shown that geometric ergodicity also holds for all the processes above [7]. With the appropriate moment assumption on the likelihood function, Theorem 1 applies for the estimation of the predictive distribution (1) using chains of this family.

3.4 Formulation of the optimization problem

The result of the previous section gives us a family of Markov chains satisfying Theorem 1 for approximately quadratic log densities, assuming the existence of the $(2 + \delta)$ moment of f for some $\delta > 0$. We make this assumption for the rest of this report. In this section and the next, we search for the optimal Markov chain in this family.

Looking at Theorem 1, we see that the asymptotic variance σ_f depends on the Markov chain through the term γ_f. For a Markov chain P and a given function f, we write $\sigma_f = \nu(f, P)$ to make this dependence explicit. A natural way therefore to define an optimal Markov chain P given f is by requiring $\nu(f, P)$ to be minimal. Restricting our attention to

chains P of the form (8) which are indexed by a matrix $A \in \mathbb{S}_+^n$, we can write $\nu(f, P) = \nu(f, A)$.

One could ask if the solution of this optimization problem depends on f when restricted to the family (8) of Markov chains. In other words, is there a feasible $A \in \mathbb{S}_+^n$ such that for all feasible $B \in \mathbb{S}_+^n$ and for all choice of f, $\nu(f, A) \leq \nu(f, B)$?

One path to the answer of this question is the following. For Markov chains P and Q, we say that P is at least as efficient as Q and write $P \succeq_E Q$ if for all f, $\nu(f, P) \leq \nu(f, Q)$. Clearly, $\succeq_E$ defines a total preorder on the set of Markov Chains. Therefore, the question formulated in the last paragraph is equivalent to the question of whether $\succeq_E$ has a maximum element in the family of Markov chains defined by (8). More accessible necessary and sufficient conditions for establishing $P \succeq_E Q$ are given in [6]. I have attempted to use these to answer this question, but I have not succeeded yet, although I suspect that the answer is no.

Working under the assumption that the answer of the above question is indeed no, we are left with the conclusion that the optimization $\min_{A \in \mathbb{S}_+^n} \nu(f, A)$ over feasible A depends on the function f. We will therefore focus on the case $f(\theta) = \theta$, with the hope that minimizing for this choice of f will yield reasonably good performance on other choices of f. If $f(\theta)$ is highly non-linear, which is admittedly the case for most likelihood functions, this might not hold. The alternative would be to consider the most commonly used likelihood functions and find the optimal chain for each one of them, but we postpone such treatment for future investigations.

With $f(\theta) = \theta$, we have:

$$
\begin{aligned}
\sigma_f &= \operatorname*{Var}_{\theta_0 \sim \pi}[\theta_0] + 2 \sum_{i=1}^{\infty} \operatorname*{Cov}_{\theta_0 \sim \pi}[\theta_0, \theta_i] \\
&= \mathop{\mathbb{E}}_{\theta_0 \sim \pi}\left[\theta_0 \theta_0^T\right] + 2 \sum_{i=1}^{\infty} \mathop{\mathbb{E}}_{\theta_0 \sim \pi}\left[\theta_0 \theta_0^T (I - \epsilon A H)^T + \theta_0 \sum_{j=1}^{i-1} w_i^T)\right] \\
&= \mathop{\mathbb{E}}_{\theta_0 \sim \pi}\left[\theta_0 \theta_0^T\right] + 2 \sum_{i=1}^{\infty} \mathop{\mathbb{E}}_{\theta_0 \sim \pi}\left[\theta_0 \theta_0^T\right] (I - \epsilon A H)^T \\
&= H^{-1} + 2 H^{-1} \sum_{i=1}^{\infty} (I - \epsilon A H)^T \\
&= H^{-1} + \frac{2}{\epsilon} H^{-1} A^{-1} H^{-1} - 2 H^{-1} \\
&= \frac{2}{\epsilon} H^{-1} A^{-1} H^{-1} - H^{-1}
\end{aligned}
$$

which is positive semi-definite by the condition on the noise in (8) as expected. Taking the Frobenius norm as a measure of the distance between σ_f and 0, we can formulate our

optimization problem as follows:

$$\min_{A \in S_+^n} \quad \left\| \frac{2}{\epsilon} H^{-1} A^{-1} H^{-1} - H^{-1} \right\|_F \tag{9}$$

$$\text{subject to} \quad 2A - \epsilon AHA - \epsilon \frac{N^2}{|S|} ACA \succeq 0 \tag{9a}$$

$$0 \prec \epsilon \lambda(AH) \prec 2 \tag{9b}$$

3.5 Solution of the optimization problem

To solve the optimization problem of the previous section, let $X := 2A^{-1} - \epsilon H - \epsilon \frac{N^2}{|S|} C$. Then condition (9a) can be rewritten as $AXA \succeq 0$, which is equivalent to $X \succeq 0$ since $A \succeq 0$. The objective function then becomes :

$$\begin{aligned} &\left\| \frac{1}{\epsilon} H^{-1} (X + \epsilon H + \epsilon \frac{N^2}{|S|} C) H^{-1} - H^{-1} \right\|_F \\ &= \left\| \frac{2}{\epsilon} H^{-1} X H^{-1} + \frac{N^2}{|S|} H^{-1} C H^{-1} \right\|_F \end{aligned}$$

Ignoring condition (9b) for the moment, $X = 0$ is clearly the minimizer of this objective. We now show that $X = 0$ satisfies (9b). Solving for A we get:

$$A = \frac{2}{\epsilon} \left(H + \frac{N^2}{|S|} C \right)^{-1} \tag{10}$$

We now have:

$$\begin{aligned} \epsilon \lambda(AH) \succ 0 &\iff \frac{1}{\epsilon} \lambda(H^{-1} A^{-1}) \succ 0 \\ &\iff 2\lambda \left(H^{-1} (H + \frac{N^2}{S} C) \right) \succ 0 \\ &\iff \lambda(I + \frac{N^2}{S} H^{-1} C) \succ 0 \end{aligned}$$

and for the other inequality:

$$\begin{aligned} \epsilon \lambda(AH) \prec 2 &\iff \frac{1}{\epsilon} \lambda(H^{-1} A^{-1}) \succ 2 \\ &\iff 2\lambda \left(H^{-1} (H + \frac{N^2}{S} C) \right) \succ 2 \\ &\iff \lambda(I + \frac{N^2}{S} H^{-1} C) \succ 1 \end{aligned}$$

Now both H^{-1} and C are positive definite, so their product $H^{-1}G$ has positive real eigenvalues, and is diagonalizable. Hence we have:

$$\begin{aligned}\lambda(I + \frac{N^2}{S}H^{-1}C) &= \lambda(I + \frac{N^2}{S}P\Lambda P^{-1}) \\ &= \lambda(P^{-1}(I + \Lambda)P) \\ &\succ 1\end{aligned}$$

Therefore (10) is feasible, and is a solution to (9).

3.6 Applicability to arbitrary posteriors

Notice that the optimal A has a $\frac{1}{\epsilon}$ factor in front of it, so that the final update equation does not depend on ϵ. Redefine A as:

$$A = 2\left(H + \frac{N^2}{S}C\right)^{-1}$$

Then what we have shown is that under the normal assumption on the posterior, the optimal Markov chain in the family described in (8) is the one given by the following equation:

$$\theta_{i+1} = \theta_i - A\nabla_\theta^S U(\theta_i) + \mathcal{N}(0, 2A - AHA - \frac{N^2}{S}ACA) \tag{11}$$

Note that the covariance of the noise vanishes. When the posterior is exactly a Gaussian, this update equation is not really useful since to evaluate A one needs to compute H, and if one has access to H, then one can either do the integration analytically, or use a more standard sampler for the Gaussian distribution. The real usefulness of (11) comes from interpreting it as the discretized stochastic gradient Langevin equation (6) with step size $\epsilon = 1$ and $D = A$. Removing the normality assumption on the posterior, and reintroducing a step size to the update equation (11) we get the following update equation:

$$\theta_{i+1} = \theta_i - \epsilon A\nabla_\theta^S U(\theta_i) + \mathcal{N}(0, 2\epsilon A - \epsilon^2 AHA - \epsilon^2\frac{N^2}{S}ACA) \tag{12}$$

which we refer to as the optimized stochastic gradient Langevin dynamics (OSGLD) equation. This equation has 3 key properties:

(i) For any $\epsilon > 0$, and under the normal assumption on the posterior, the chain will sample exactly from the posterior.

(ii) For $\epsilon = 1$, and under the normal assumption on the posterior, the chain will sample optimally from the posterior.

(iii) As $\epsilon \to 0$, the chain samples from an arbitrary posterior exactly.

Therefore, we can expect that if the posterior is close to a Gaussian, in the sense that its log density has a significant quadratic component, then by properties (i) and (iii), we can use a large step size while preserving good accuracy. By property (ii), we can also expect that the sampling will be near optimal.

4 Conclusion

In this report, we reviewed some important results of Markov Chain Monte Carlo theory for discrete time Markov chains, and mentioned some new results for continuous time Markov processes, which give rise to stochastic gradient MCMC algorithms. We then considered stochastic gradient Langevin dynamics algorithms and derived the optimal algorithm for this subclass for posteriors with approximately quadratic log densities. Many interesting problems remain unsolved in this area to my knowledge, and I hope to investigate some of them in the near future:

- Does there exist a discrete equivalent to the continuous-time Markov processes described in (3) ? In other words, is there a family of discrete-time Markov chains that uses only the gradient of the density and a normal random variable to sample from any given distribution ?
- Finding the optimal $D(q)$ in (3) for a given density. The notion of optimality here will also have to take into account the computational cost of simulating the resulting process.

References

[1] George Casella Christian P. Robert. *Monte Carlo Statistical Methods*. 2004.

[2] Galin L Jones. On the Markov chain central limit. 2004.

[3] Yi-An Ma, Tianqi Chen, and Emily B. Fox. A Complete Recipe for Stochastic Gradient MCMC. 2015.

[4] Sean P Meyn and R L Tweedie. Stability of Markovian Processes II : Continuous-Time Processes and Sampled Chains. 2008.

[5] Sean P Meyn and R L Tweedie. Stability of Markovian Processes III : Foster-Lyapunov Criteria for Continuous-Time Processes. 2008.

[6] Antonietta Mira and Fabrizio Leisen. Covariance ordering for discrete and continuous time Markov chains. 2009.

[7] Dag Tjostheim. Non-linear Time Series and Markov Chains. 2014.

[8] Max Welling and Yee-Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. 2011.