Stochastic Approximation with Biased MCMC for Expectation Maximization

Samuel Gruffaz  Kyurae Kim  Alain Olivierio Durmus  Jacob R. Gardner
Centre Borelli, ENS Paris-Saclay   UPenn   CMAP, CNRS, École Polytechnique   UPenn

Abstract

The expectation maximization (EM) algorithm is a widespread method for empirical Bayesian inference, but its expectation step (E-step) is often intractable. Employing a stochastic approximation scheme with Markov chain Monte Carlo (MCMC) can circumvent this issue, resulting in an algorithm known as MCMC-SAEM. While theoretical guarantees for MCMC-SAEM have previously been established, these results are restricted to the case where asymptotically unbiased MCMC algorithms are used. In practice, MCMC-SAEM is often run with asymptotically biased MCMC, for which the consequences are theoretically less understood. In this work, we fill this gap by analyzing the asymptotics and non-asymptotics of SAEM with biased MCMC steps, particularly the effect of bias. We also provide numerical experiments comparing the Metropolis-adjusted Langevin algorithm (MALA), which is asymptotically unbiased, and the unadjusted Langevin algorithm (ULA), which is asymptotically biased, on synthetic and real datasets. Experimental results show that ULA is more stable with respect to the choice of Langevin stepsize and can sometimes result in faster convergence.

1 INTRODUCTION

Probabilistic modeling with latent variables is an essential tool for modeling observational data generated from complex latent structures. While eliciting priors for this class of models is often straightforward, e.g., data generated from unobserved groups can be modeled using mixtures (McLachlan et al., 2019), while group-level variabilities can be modeled with mixed effects (Kuhn and Lavielle, 2005), how we should set the hyper-parameters $\theta$ is not always self-evident. For this, the empirical Bayes paradigm applies the maximum likelihood principle (Robbins, 1956; Efron, 2019). That is, it infers the (hyper-)parameters $\theta$ from data by maximizing the marginal log-likelihood $l(\theta) \triangleq \log(p(y|\theta))$,

$$\arg\max_{\theta \in \Theta} l(\theta) = \arg\max_{\theta \in \Theta} \log \left( \int_{\mathcal{Z}} p(y, z|\theta) dz \right) \tag{1}$$

where $y \in \mathcal{Y} \subset \mathbb{R}^d$ are the observations, $z \in \mathcal{Z} \subset \mathbb{R}^d$ are the latent variables, and $\theta \in \Theta \subset \mathbb{R}^d$ denote the parameters.

Unfortunately, the marginal log-likelihood is often intractable, making the empirical Bayes problem hard. As a solution, Dempster et al. (1977) have proposed the expectation-maximization (EM) algorithm, which has subsequently been immensely successful in practice, and its convergence properties have been studied extensively (McLachlan and Krishnan, 2007). It is now known to converge to stationary points under mild conditions. However, the canonical EM algorithm may not apply immediately in many practical cases. For example, gradient ascent steps must be used when the maximization step (M-step) of EM is not available in closed form (Baey et al., 2023).

This work focuses on the setting where the expectation step (E-step) is intractable. In this case, the integral in Equation (1) needs to be numerically approximated using methods such as Monte Carlo (MC), importance sampling (IS), and Markov chain Monte Carlo (MCMC) (See the book by Robert and Casella (2004) for an overview of these methods). Furthermore, when the model belongs to the exponential family such that the M-step can be computed in closed-form, Delyon et al. (1999) show that the EM algorithm can be reduced to approximating sufficient statistics generated from the sampling algorithm. In particular, they leverage stochastic approximation (SA; Rob-
2 BACKGROUND

2.1 Expectation-Maximization

We assume that the joint distribution of the observations and latent variable belongs to the curved exponential family:

**H1.** For any \( y \in \mathcal{Y}, z \in \mathcal{Z} \) and \( \theta \in \Theta \),

\[
p(y, z|\theta) = h(y, z) \exp(S(y, z)^\top \phi(\theta) - \psi(\theta)) ,
\]

where \( \Theta, \mathcal{Z} \), are open, \( S: \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^d \) is continuous, and \( \phi: \Theta \to \mathbb{R}^d \) and \( \psi: \Theta \to \mathbb{R} \) are continuously differentiable.

In this paper, \( y \in \mathcal{Y} \) is fixed, and the parameter \( \theta \) is estimated through maximum marginal log-likelihood \( l: \Theta \to \log p(y|\theta) \). In most cases, \( l \) cannot be maximized directly and is not even tractable. EM solves this by employing the majorize-minimize principle: (See, e.g., Lange, 2016, Chapter 8.) An upper bound of \(-l\) is derived using Jensen’s inequality and a density \( q \) on \( \mathcal{Z} \), for any \( \theta \in \Theta \),

\[
- \log p(y|\theta) = \log \left( \int_\mathcal{Z} p(y, z|\theta) \frac{q(z)}{q(z)} dz \right) \\
\leq - \int_\mathcal{Z} \log \left( \frac{p(y, z|\theta)}{q(z)} \right) q(z) dz = Q(\theta, q) \\
= - \left( \mathbb{E}_{z \sim q} (\log(p(y, z|\theta))) dz + \text{Ent}(q) + \text{ cst}(y) \right) .
\]

Denoting \( D_\mathcal{Z} \triangleq \{ f \in L^1(\mathcal{Z}) : f \geq 0, \int_\mathcal{Z} f dz = 1 \} \), the function \( q \in D_\mathcal{Z} \mapsto Q(\theta, q) \) is minimized by \( q^*(z) \triangleq p(z|y, \theta) \) such that \( Q(\theta, q^*) = -l(\theta) \). Thus, by considering \( \theta^* = \arg\max_{\theta \in \Theta} Q(\theta, q^*) \), we have \( l(\theta^*) \leq l(\theta) \). This procedure offers a recipe to construct a maximizing sequence of \( l \). Moreover, under **H1.**

\[
\theta^* \triangleq \arg\max_{\theta \in \Theta} \mathbb{E}_{z \sim p(z|y, \theta)} (S(y, z))^\top \phi(\theta) - \psi(\theta) .
\]

We introduce the following assumption to maximize this last expression:

**H2.** Denoting \( L(s, \theta) \triangleq s \cdot \phi(\theta) - \psi(\theta) \), there exists a function \( \hat{\theta}: \mathbb{R}^d \to \Theta \), such that for any \( (s, \theta) \in \mathbb{R}^d \times \Theta \),

\[
L(s, \theta) \leq L(s, \hat{\theta}(s)) .
\]

If necessary, this assumption can also be met by “exponentializing” as shown by Debavelaere and Alllassonnière (2021).

The EM algorithm is defined as follows: Let \((s_k)_{k \geq 0}, (\theta_k)_{k \geq 0}\) be initialized from \(\theta_0 \in \Theta\) and follow the recursion for any \( k \geq 0 \):

1. **Expectation:** Denoting by \( \bar{s} : \Theta \to \mathbb{R}^d \) \( \bar{s} \triangleq \mathbb{E}_{z \sim p(z|y, \theta_0)} S(y, z)p(z|\theta_0, y) dz \), set \( s_k = \bar{s}(\theta_k) \).

2. **Maximization:** Set \( \theta_{k+1} = \hat{\theta}(s_k) \), which implies \( l(\theta_{k+1}) \geq l(\theta_k) \).

This algorithm quickly converges towards a local maximum of \( l \) under mild conditions. (See the review by McLachlan and Krishnan (2007).)
2.2 EM as a Root Finding Problem

The EM recursion can be seen as a fixed-point iterative scheme since, for any \( k \geq 0 \), \( \theta_{k+1} = \theta \circ \hat{s}(\theta_k) \). If the sequence \( (s_n) \) converges towards \( s^* \) and \( \hat{s} \circ \hat{\theta} \) is continuous, we then have

\[ h(s^*) = 0, \text{ where } h \triangleq \hat{s} \circ \hat{\theta} - \text{Id}. \]

Moreover, if \( s^* \) is a root of \( h \), then \( \hat{\theta}(s^*) \) is a root of \( \hat{s} \circ \hat{\theta} - \text{Id} \). And reciprocally, if \( s^* \) is a root of \( \hat{s} \circ \hat{\theta} - \text{Id} \), then \( \hat{s}(\hat{\theta}^*) \) is a root of \( h \). This suggests that EM can be reduced to a problem of finding the root \( h \) with respect to \( s \).

EM as Stochastic Approximation Finding the root is more general than gradient descent as \( h \) may not be the gradient of any known function. Instead, it is the descent direction for some Lyapunov function \( V: s \in \mathbb{R}^d \mapsto -l \circ \hat{\theta}(s) \). As such, the proposed scheme is part of the more general stochastic approximation (SA) framework (Dieuleveut et al., 2023). The most basic form of the SA algorithm is described as follows:

\[ s_{k+1} = s_k + \gamma_{k+1}H(s_k, Z_{k+1}), \quad Z_{k+1} \sim p(\cdot|\theta_k, y), \]

where \( H(s_k, Z_{k+1}) = S(y, Z_{k+1}) - s_k \) is a random oracle of \( h(s_k) \) and \( (\gamma_k) \) is a deterministic stepsize sequence. Solving the EM problem using these iterates is known as SAEM algorithm (Delyon et al., 1999).

While we have motivated the reduction of optimizing \( l \) to finding the root of \( h \), it is not apparent the solutions such that \( h = 0 \) and \( \nabla l = 0 \) are equivalent. The following lemma will clarify the relationship between \( h, l \), and the Lyapunov function \( V \) given assumptions on the Hessian of \( L \) at its maximums:

**H 3.** The functions \( \hat{\theta}, l(\cdot), \hat{s}(\cdot), \hat{\psi}(\cdot) \) are \( p \)-continuously differentiable with \( p > d \). Moreover, denote

\[ A(s) \triangleq \partial_s \hat{\theta}(s) \top \partial^2 L(s, \hat{\theta}(s)) \partial_s \hat{\theta}(s). \]

Then, there exist \( \lambda_m, \lambda_M > 0 \) such that for any \( s \in \mathbb{R}^d: \lambda_m |v|^2 \leq \langle A(s)v|v \rangle \leq \lambda_M |v|^2, \quad v \in \mathbb{R}^d. \)

**Lemma 1.** Under **H1, H2 and H3**, \( V \) is \( p \)-continuously differentiable and verifies for any \( s \in \mathbb{R}^d \),

\[ F(s) \triangleq (\nabla V(s))h(s) \leq -\lambda_m |h(s)|^2 \]

\[ |\nabla V(s)| \leq \lambda_M |h(s)| \]

\[ S = \{ s \in \mathbb{R}^d: F(s) = 0 \} = \{ s \in \mathbb{R}^d: \nabla V(s) = 0 \} \]

\[ \hat{\theta}(S) = \{ \theta \in \Theta: l(\theta) = 0 \}, \text{ int}(V(S)) = \emptyset \]

**Proof.** In the proof by Delyon et al. (1999, Lemma 2), they derive that for any \( s \in \mathbb{R}^d, \nabla V(s) = -A(s)h(s), \) the results are then straightforward with the regularity assumption on \( A \).

In other words, \( h(s) \) is a proxy of \(-\nabla V(s)\) for any \( s \in \mathbb{R}^d \). This Lemma makes it clear that converging to \( S \) by working in the sufficient statistics space recovers the solutions of the original problem (1). From now on, we drop the dependence of \( y \) in \( S(y, z) \) such that \( S(z) \triangleq S(y, z) \) since we are interested only in \( z \).

**MCMC-SAEM** SAEM can be generalized to models with intractable likelihoods by leveraging Markov chain Monte Carlo (MCMC) as proposed by Kuhn and Lavielle (2004). An MCMC algorithm form a Markov kernel \( \Pi_{\theta} \), such that, for any \( z \in \mathcal{Z} \),

\[ \lim_{n \to \infty} ||\Pi_{\theta}(z^*) - \pi_\theta||_{TV} = 0, \text{ where } \pi_\theta \text{ is the target distribution related to } p(\cdot|\theta, y). \]

It means that asymptotically, sampling from \( \Pi_{\theta}(z, \cdot) \) is nearly equivalent to sampling from \( \pi_\theta \). Then, the Markov chain is said to be asymptotically unbiased, contrary to the case \( \lim_{n \to \infty} ||\Pi_{\theta}(z, \cdot)^{(n)} - \pi_\theta||_{TV} > 0 \), where we say that the chain is asymptotically biased.

Note that, in general, the chain is biased in finite time \( ||\Pi_{\theta}(z, \cdot)^{(n)} - \pi_\theta||_{TV} > 0 \) for any \( n \geq 0 \). Despite this, MCMC-SAEM, MCMC applied to SAEM, generates a sequence \( (\theta_k) \) that converges to a local maximum almost surely (Dieuleveut et al., 2023; Kuhn and Lavielle, 2004), where the E-step is replaced by

\[ s_{k+1} = s_k + \gamma_{k+1}(S(y, z_{k+1}) - s_k), \quad z_{k+1} \sim \Pi_{\theta_k}(z_k, \cdot), \]

for any \( k \geq 0 \), starting with \( (s_0, z_0) \in \mathbb{R}^d \times \mathcal{Z} \).

2.3 MCMC-SAEM with Approximate MCMC Algorithms

At the core of the practical performance of MCMC-SAEM is the choice of the MCMC algorithm. While MALA is often used, its asymptotically biased counterpart ULA has been shown to mix faster at high dimensions (Durmus and Moulines, 2017). This means that, in finite time, the total bias of ULA can be lower compared to the asymptotically unbiased MALA. Therefore, while we consider a general biased Markov chain to study the convergence of MCMC-SAEM in Section 3 and 4, we are interested in studying the difference between ULA and MALA. Our experiments in Section 5 will exclusively focus on this question.

**ULA and MALA** The ULA Markov chain \( (X_k)_{k \geq 0} \) is derived from the Euler–Maruyama discretization scheme associated with the Langevin diffusion related to the force \( U \triangleq -\nabla \log(\pi) \) if \( \pi \) is the target distribution, at iteration \( k \geq 0 \)

\[ X_{k+1} = X_k - \eta_{k+1} U(X_k) + \sqrt{2\eta_{k+1}} Z_{k+1} \]

where \( (Z_k)_{k \geq 0} \) is an i.i.d sequence of standard Gaussian \( d \)-dimensional random vectors and \( (\eta_k)_{k \geq 1} \) is a sequence of stepsize, which can be either constant or de-
crease to 0. The MALA Markov chain follows the same recursion by adding an acceptance/rejection step of $X_{k+1}$, making the chain asymptotically unbiased.

The properties of MALA chains have been studied in depth in (Roberts and Tweedie, 1996), where it is shown that the chain converges geometrically fast (geometric ergodicity) under mild assumptions on the tail of the distribution. This is key since geometric ergodicity is the main assumption of the MCMC-SAEM convergence theorem (Kuhn and Lavielle, 2004). Despite being geometrically ergodic, the adjustment step of MALA can become a curse in high dimensions. To maintain a sufficient level of acceptance, a smaller stepsize must be used, resulting in a slow mixing chain. (Mixing is a fundamentally non-asymptotic notion, unlike geometric ergodicity.)

On the other hand, for ULA, denoting its limiting distribution as $\pi$, the mixing rate improves with $\eta$ (Durmus and Moulines, 2017) at the cost of increasing $||\pi'' - \pi||_{TV}$ (de Bortoli et al., 2021). This means one can trade off asymptotic bias for a faster mixing rate. Considering this, we will expand the existing analysis, where $\eta$ is no longer a stepsize but a “knob” that controls bias, which can vary across the iterations.

### 2.4 Stochastic Approximation with Biased Dynamics

To incorporate biased MCMC chains into our analysis, we slightly modify the SA formalism (Dieuleveut et al., 2023) to allow some freedom on the bias parameter $\eta$.

Let $(\gamma_n)_n$ and $(\eta_n)_n$ be two monotone nonincreasing sequences with $\gamma_0, \eta_0 \in (0,1]^2$. Define the nonhomogeneous Markov chain $(\{Y_n = (Z_n, S_n)\})_n$ on $\mathbb{Z} \times \mathbb{R}^d$ as follows: Let $s_0 = \theta \in \mathbb{R}^d$, $z_0 = z \in \mathbb{R}^d$ and for $n \geq 0$,

$$
\begin{align*}
Z_{n+1} &\sim \Pi_{\eta_{n+1}}(Z_n, \cdot) \\
\gamma_{n+1} &\sim \gamma_n + I_H(s_n, Z_{n+1}),
\end{align*}
$$

where $\{\Pi_{\eta}(Y_n) : s, \eta \in \mathbb{R}^d \times (0, \eta_0)\}$ is a family of Markov kernel transition probabilities and $H : s, z \in \mathbb{R}^d \times \mathbb{Z} \mapsto S(z) - s$ is a field which satisfy the following conditions:

**H4.** For any $s \in \mathbb{R}^d$ and $\eta \in (0, \eta_0]$, the Markov kernel $\Pi_{\eta}$ has a single stationary distribution $\pi_{\hat{\theta}(s)}$, also denoted as $\pi_{s, \eta}$, such that $\pi_{s, \eta} \Pi_{\eta}(s, \cdot) = \pi_{s, \eta}$. In addition, $H : \mathbb{R}^d \times \mathbb{Z} \rightarrow \mathbb{R}^d$ is measurable for all $s \in \mathbb{R}^d$ and $\int_{\mathbb{Z}} |H(s, z)| \pi_{s, \eta}(dz) < \infty$.

By considering the filtration $\{\mathcal{F}_n = \sigma(s_0, Z_i, i \leq n)\}_n$, the following decomposition clarifies the deterministic and stochastic parts of the dynamic:

$$
H(s_n, Z_{n+1}) = h(s_n) + \xi_n,
$$

where $h(s_n)$ is the mean field drift, $\xi_n$ is the Markovian noise and $(\beta_n)_n$ the bias. In the MCMC-SA context, this becomes

$$
\begin{align*}
e_n &= S(z_{n+1}) - \tilde{s}_{\eta_{n+1}}(\hat{\theta}(s_n)) \\
\beta_n &= \tilde{s}_{\eta_{n+1}}(\hat{\theta}(s_n)) - \tilde{s}(\hat{\theta}(s_n)) \\
h(s) &= \tilde{s}(\hat{\theta}(s)) - s,
\end{align*}
$$

This way, for any $s \in \mathbb{R}^d$, $\tilde{s}_{\eta}(\hat{\theta}(s))$ is the biased approximation of $\tilde{s}(\hat{\theta}(s))$. Lastly, we assume that the asymptotic bias is finite:

**H5.** $\limsup_{n \rightarrow \infty} |\beta_n| = \beta < +\infty$

This assumption is reasonable if the growth of $z \mapsto S(z)$ can be mitigated by the tail decay of $\pi_{s, \eta}$ and $\pi_{\theta}$.

### 3 ASYMPTOTIC ANALYSIS

We estimate how the asymptotic bias of MCMC impacts the asymptotic convergence of $(s_n)$ generated by the recursion (2). The analysis takes most of its arguments from (Tadić and Doucet, 2017) where the framework is slightly different, $h = \nabla f$ where $f$ has the same regularity that in H3. We extend their results by using Lemma 1, i.e., the mean field is not a result but the proxy of a Lyapunov gradient. The results are not specific to the SA scheme but can be generalized to other SA schemes as Lemma 1 applies, but this is beyond the scope of this paper.

#### 3.1 Technical Assumptions

All the assumptions presented in this part are also used in (Tadić and Doucet, 2017). We take the following assumptions on $(\gamma_n), (\epsilon_n)$:

**A1.** $\limsup_{n \rightarrow \infty} |\gamma_{n+1} - \gamma_n| < \infty$, $\lim_{n \rightarrow \infty} \gamma_n = 0$ and $\sum_{n=0}^{\infty} \gamma_n = \infty$.

Denoting by $a(n, t) = \max \left\{ k \geq n : \sum_{i=n}^{t-1} \gamma_i \leq t \right\}$ for $n \geq 0$ and $t \in (0, \infty)$, $a(n, t)$ is well defined thank to A1. Furthermore, we assume:

**A2.** $(\epsilon_n)$ and $(\beta_n)$ are $\mathbb{R}^d$-valued stochastic processes satisfying,

$$
\lim_{n \rightarrow \infty} \max_{1 \leq k \leq a(n, t)} \left| \sum_{i=n}^{k} \gamma_i \epsilon_i \right| = 0
$$

almost surely on $\{ \sup_n |s_n| < \infty \}$ for any $t \in (0, \infty)$. 

A2 can be established from assumptions of geometric ergodicity on the Markov kernels \( \{\Pi_s, s \in \mathbb{R}^d, \eta \in (0, \eta_0]\} \) and by controlling the growth of the sufficient statistics. Due to its technicalities, this development is relegated to Appendix B.1.3.

In numerous cases, we can even make the following assumption,

**A3.** \( V(\cdot) \) is real analytic on \( \mathbb{R}^d \).

This implies that \( V \) can be locally represented by a power series.

### 3.2 Main Result

For any compact \( Q \subset \mathbb{R}^d, \Lambda_Q \) denotes the event \( \cup_{n=0}^{\infty} \cap_{k=n}^{\infty} \{\theta_k \in Q\} \), such that on \( \Lambda_Q, \{\theta_k\} \) is bounded and the bound is controlled by \( Q \). The main result on the asymptotic bias of the recursion (2) can be stated as follows:

**Theorem 1.** Suppose that \( H_{1-5}, A_{1-2} \). Let \( Q \subset \mathbb{R}^d \) be any compact set. Then, the following are true:

(I) There exists a (deterministic) non-decreasing function \( \psi_Q : [0, \infty) \rightarrow [0, \infty) \) (independent of \( \eta \) and depending only on \( V(\cdot) \)) such that \( \lim_{t \rightarrow 0} \psi_Q(t) = \psi_Q(0) = 0 \) and

\[
\limsup_{n \rightarrow \infty} d(s_n, S) \leq \psi_Q(\beta)
\]

almost surely on \( \Lambda_Q \).

(II) There exists a real number \( K_Q \in (0, \infty) \) (independent of \( \beta \) and depending only on \( V(\cdot) \)) such that

\[
\limsup_{n \rightarrow \infty} \|\nabla V(s_n)\| \leq K_Q \beta^{3/2} , \quad \limsup_{n \rightarrow \infty} V(s_n) - \liminf_{n \rightarrow \infty} V(s_n) \leq K_Q \beta^q
\]

almost surely on \( \Lambda_Q \), where \( q = (p - d)/(p - 1) \).

(III) If \( V(\cdot) \) satisfies **A3** there exist real numbers \( r_Q \in (0,1), L_Q \in (0, \infty) \) (independent of \( \beta \) and depending only on \( V(\cdot) \)) such that

\[
\limsup_{n \rightarrow \infty} \|\nabla V(s_n)\| \leq L_Q \beta^{1/2} , \quad \limsup_{n \rightarrow \infty} d(V(s_n), V(S)) \leq L_Q \beta , \quad \limsup_{n \rightarrow \infty} d(s_n, S) \leq L_Q \beta^{3/2} .
\]

almost surely on \( \Lambda_Q \).

Proof. See Appendix B.1.2 for the proof. \( \square \)

Theorem 1-(I) formalizes the intuition that \( \lim_{n \rightarrow \infty} d(s_n, S) = 0 \) if \( \beta = \limsup_{n \rightarrow \infty} |\beta_n| \rightarrow 0 \). If \( (\eta_n) \) encodes the stepsize of ULA, the last condition can be deduced from \( \lim_{n \rightarrow \infty} |\eta_n| = 0 \), since a smaller stepsize decreases bias. Even though it is the case that \( \beta > 0 \) in practice since \( \eta_n = \eta > 0 \) is fixed, we can quantify the impact of the bias on \( V \) in Theorem 1-(I). Note the impact of the regularity of \( V \) according to the dimension encoded in \( q = (p - d)/(p - 1) \): at the limit when \( p \rightarrow \infty \), we recover \( q = 1 \) as in Theorem 1-(III). Therefore, the impact of the bias \( \beta \) on \( \limsup_{n \rightarrow \infty} \|\nabla V(s_n)\| \) is smoothed by the regularity of \( V \). For the case of \( \beta = 0 \), we recover the convergence of the sequence towards stationary points \( S \), which has already been established by Kuhn and Lavielle (2004).

Remark that Theorem 1 is local in the sense that it holds on the event \( \Lambda_Q \) and not globally. (i.e., there exists a compact set \( Q \) such that \( \Lambda_Q \) holds almost surely.) Previous results have assumed that the sequence \( (s_n) \) is bounded to make their result global and have argued that we can bound it by design through “reinitialization” as studied by Kuhn and Lavielle (2004); Andrieu et al. (2005). Here, we did not use the recursion proposed (Andrieu et al., 2005, p.9) to more aligned with practical implementations.

The constants \( K_Q, L_Q \) depend explicitly on the bounds on \( V, |V'| \) (the maximal singular value of \( A \) to be precise) and the Yomdin and Lojasiewicz constants applied to \( V \). (See, e.g. Proposition 8.1 and 8.2 by Tadić and Doucet (2017), which are generalizations of Sard’s Theorem.) The explicit forms of \( K_Q, L_Q \) are given at the end of the proof in Appendix B.1.2.

### 4 NON ASYMPOTOTIC ANALYSIS

For the non-asymptotic analysis, we assume the bias parameters are fixed for any \( n \geq 0, \eta_n \in (0, \eta_0] \). Furthermore, we rely on assumptions typical for non-asymptotic analysis of SA. Our main result is a non-asymptotic high probability convergence guarantee for SAEM with biased MCMC.

#### 4.1 Technical Assumptions

First, we impose assumptions on the MCMC kernel, which are standard in the non-asymptotic analysis of stochastic approximation with state-dependent Markovian noise.

| N1 | The update is bounded by a constant \( 0 < \sigma < \infty \) as \( \sup_{(s,z) \in \mathbb{R}^d \times Z} |H(s,z) - h(s)| \leq \sigma \). |
| N2 | For any \( s \in \mathbb{R}^d \), there exists a solution \( \nu(s) : Z \rightarrow \mathbb{R} \). |

While this assumption is quite strong, it is necessary to control the properties of the solution to the Poisson equation as in the following assumption:

| N3 | \( \exists v(s), \forall s \in \mathbb{R}^d, \int_{Z} v(s,z) \mu_d(z)\,dz = 0 \). |
\( \mathbb{R}^d \) to the Poisson equation such that
\[
\nu^0 - \Pi^t_s \nu^0 = S(\cdot) - \hat{s}_n(\hat{\theta}(s)) ,
\]
where for any \( z \in \mathcal{Z} \), \( \Pi^t_s \nu^0(z) = \int_{\Delta} \nu^0(z) \Pi^t_s(z, \d z) \) and for any \( s \in \mathbb{R}^d \), \( \hat{s}_n(\hat{\theta}(s)) = \int_{\Delta} S(z) \d x_{s, n} \). Moreover, there exist some bound \( L^{(0)}_\nu, L^{(1)}_\nu > 0 \) such that
\[
\sup_{(s, z) \in \mathbb{R}^d \times \mathcal{Z}} \{|\nu^0(s)|, |\Pi^t_s \nu^0(s)|\} \leq L^{(0)}_\nu ,
\]
\[
\sup_{(s, z) \in \mathbb{R}^d \times \mathcal{Z}} |\Pi^t_s \nu^0(z) - \Pi^t_s \nu^0(z')| \leq L^{(1)}_\nu |s - s'| .
\]
Remark that \( \textbf{N1-2} \) implies \( \textbf{H4} \). Also, under \( \textbf{N2} \),
\[
e_n = S(Z_{n+1}) - \hat{s}_n(\hat{\theta}(s_{n})) = \nu^0(Z_{n+1}) - \Pi^t_s \nu^0(Z_{n+1}) ,
\]
which is crucial for the analysis in the Supplementary material Appendix B.2. This type of assumption has first been used by Karimi et al. (2019) and has since been standard in the analysis of stochastic approximation with state-dependent Markovian noise. See, e.g., the works of Alacaoglu and Lyu (2023, Assumption 3.7), Roy et al. (2022, Assumption 2.4) for some recent examples. If we assume that \( \textbf{(i)} \) both \( \Pi^t_s \), \( H(s, z) \) are uniformly Lipschitz with respect to \( s \) for any \( z \in \mathcal{Z} \), \( \textbf{(ii)} \) \( \Pi \) is uniformly Lipschitz, \( \text{(iii)} \) \( \textbf{N1} \) holds, Karimi et al. (2019, Lemma 7) establish \( L^{(0)}_\nu, L^{(1)}_\nu \) explicitly.

In this work, we are particularly interested in asymptotically biased MCMC algorithms:

\textbf{N3.} The asymptotic bias of the MCMC kernel is bounded for some \( 0 \leq \tau_0, \tau_1 < \infty \) as
\[
|\beta_k|^2 \leq \tau_0 + \tau_1 |h(s_k)|^2 .
\]
Moreover, under \( \textbf{H3} \), \( C_{b_1} \triangleq \lambda_M \left( \frac{1}{2} \sqrt{\tau_0} + \sqrt{\tau_1} \right) < \lambda_m \), where \( \lambda_m, \lambda_M > 0 \) are defined in \( \textbf{H3} \).

This assumption has been used by Dieuleveut et al. (2023), and encompasses both iterate-dependent and independent bias. It is a refinement of \( \textbf{H5} \). The condition about \( C_{b_1} \) is specific to the SAEM framework; the bias is bounded by the matrix conditioning of \( A \) given in \( \textbf{H3} \). This reveals crucial in the high probability bound.

The remaining assumptions are standard in the non-asymptotic analysis of stochastic approximation. (See H1-2 in the recent review by Dieuleveut et al. (2023).)

\textbf{N4.} The Lyapunov function \( V \) is smooth and bounded below such that
\[
|\nabla V(s) - \nabla V(s')| \leq L_V |s - s'| , \quad V(s) \geq V^* \quad \text{for } V^* = \inf_{s \in \mathbb{R}^d} V(s) > -\infty \text{ and some } 0 < L_V < \infty .
\]
For the EM setting, \( \textbf{N4} \) is problem-dependent and can not be recovered directly from \( \textbf{H3} \).

\subsection{4.2 Main Result}

\textbf{Theorem 2.} Assume \( \textbf{H1-3}, \textbf{N1-3} \) and \( \textbf{N4} \),

Then, given a stepsize satisfying with \( \alpha_1, \alpha_2 > 0 \),
\[
\gamma_k + 1 \leq \gamma_k , \quad \gamma_k \leq \alpha_1 \gamma_{k+1} + 1 , \quad \gamma_0 \leq \frac{1}{2} (\lambda_m - C_{b_1}) / C_{n_1} ,
\]
with probability at least \( 1 - \delta \), the MCMC-SAEM algorithm guarantees that
\[
\min_{k=0, \ldots, n} |h(s_k)|^2 \leq \frac{2}{\lambda_m - C_{b_1}} \times \left( V(\theta_0) - V^* + C_0 + \log \left( \frac{1}{\delta} \right) + C_{n_2} \sum_{k=0}^{n} \gamma_{k+1}^2 + C_{b_2} \right) ,
\]
where the constants are
\[
C_0 = L^{(0)}_\nu (\gamma_0 + 2 \lambda_M) , \quad C_{b_1} = \lambda_M \left( \frac{2}{5} \sqrt{\tau_0} + \sqrt{\tau_1} \right) , \quad C_{b_2} = \frac{1}{2} \lambda_M \lambda_m , \quad C_{n_1} = L^{(1)}_\nu \lambda_M \sigma + L_V L^{(0)}_\nu \lambda_M (1 + \sigma) + L_V \sigma^2 , \quad C_{n_2} = \left( 2 \lambda_M L^{(0)}_\nu \right)^2 + L^{(1)}_\nu \lambda_M \left( \frac{1}{2} + \alpha_1 \sigma + \alpha_1 \frac{1}{2} \right) + L^{(0)}_\nu \lambda_M (L_V + \alpha_2 + 1) + L_V .
\]

\textbf{Proof.} See Appendix B.2 for the proof.

If we set \( \gamma_k = (\lambda_m - C_{b_1}) / 2 C_{n_1} \sqrt{k} \) for any \( k \geq 1 \), then the stepsize satisfies (3) with \( \alpha_1, \alpha_2 = \sqrt{2} (\sqrt{2} - 1) / \sqrt{2 \gamma_1} \), and thus the right hand in (4) becomes \( O(\log(n)/n + 2 C_{b_2} / (\lambda_m - C_{b_1})) \). The results are coherent with Theorem 1-(III), the bias is proportional to \( \lambda_M \sqrt{\beta} \) as \( n \to \infty \).

The key step in establishing the high-probability bound is to ensure the non-asymptotic transient bias of the Markov chain concentrates. This is done by constructing a Martingale following the strategy of Karimi et al. (2019), while the concentration inequality is by Li and Orabona (2020, Lemma 1). Although this Lemma relied on Gaussian tails, this automatically follows from \( \textbf{N2} \). Therefore, with this set of assumptions, SA with MCMC is well-behaved under a concentration perspective.

Overall, as in Section 3, we conclude that the regularity of the problem determines the limit of biased MCMC within SAEM.
5 EXPERIMENTS

We will now empirically compare the performance of MCMC-SAEM with approximated and asymptotically exact MCMC kernels. In particular, we compare ULA and MALA. While the computation cost is comparable—MALA is slightly more expensive as it requires an additional evaluation of the unnormalized target density—their practical behavior can be different, as we will see in the experiments. All experiments were implemented in the Julia language (Bezanson et al., 2017). For the stepsize, we use $\gamma_k = 1/\sqrt{k}$ for all experiments. Furthermore, in the E-step of each SAEM iteration, 4 MCMC steps are performed as burn-in unless stated otherwise. The source code used for the experiments is publicly available online.

5.1 Logistic Regression on a Synthetic Dataset

To illustrate the difference in the behavior of ULA and MALA, we first consider a toy problem.

**Model** The model is a typical logistic regression model with a Gaussian prior on the coefficients:

$$\beta \sim N(\mu_1, \sigma_1^2)$$

$$p_i = \text{logistic}(\beta^T x_i)$$

$$y_i \sim \text{Bernoulli}(p_i)$$

We optimize for the hyperparameters $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$. The number of datapoints is 1000, while the dimensionality of $\beta$ is 100. The regression matrix is randomly generated to have a condition number of $\kappa = 1000$. We initially run the respective MCMC kernel for 10 iterations as burn-in, and then run MCMC-SAEM for 100 iterations. The true parameter is $\theta^* = (1.0, 1)$, while we initialize the algorithm with $\theta_0 = (0, 1)$. The MCMC chain is initialized from a standard Gaussian.

**Results** For a small $\eta$ (ULA/MALA stepsize), ULA and MALA perform similarly since MALA reduces to ULA. On the other hand, they start to behave differently in the large $\eta$ regime, as illustrated in Figure 1: MALA starts to reject more proposals, which results in slower convergence. On the other hand, ULA does not reject anything, making constant progress. This illustrates that, in the large-$\eta$ regime, the asymptotic bias of ULA becomes less critical since MALA suffers from non-asymptotic transient bias from the rejections.

5.2 Pharmacokinetics

A popular application of MCMC-SAEM is nonlinear mixed modeling, which often arises in longitudinal studies with nonlinear models of progression. Here, we will consider modeling the pharmacokinetics of Theophylline, which is a drug for respiratory diseases (Davidian and Giltinan, 1995; Pinheiro and Bates, 1995).

**Model** Following the formulation of Kuhn and Lavielle (2005), we assume the concentration of the drug on the $i$th patient at the $j$th measurement can be modeled as

$$\log V_i \sim N(\mu_V, \sigma_V^2)$$

$$\log k_{ai} \sim N(\mu_{kai}, \sigma_{kai}^2)$$

$$\log Cl_i \sim N(\mu_{Cl}, \sigma_{Cl}^2)$$

$$y_{ij} \sim N(h(V_i, Cl_i, k_{ai}, t_{ij}), \sigma^2)$$

where $h$ is a first-order one-compartment model:

$$h(V_i, Cl_i, k_{ai}, t_{ij}) \triangleq \frac{d_i k_{ai}}{V_i} \left( e^{-\frac{Cl_i}{V_i} t_{ij}} - e^{-k_{ai} t_{ij}} \right),$$

and for each of the $i$th patient,

$$y_{ij}$$ is the concentration of the drug at $t_{ij}$ (mg/L),

$t_{ij}$ is the time of the $j$th measurement (hours),

$d_i$ is the administered dosage (mg/kg),

$k_{ai}$ is the drug absorption rate,

$Cl_i$ is the drug’s clearance, and

$V_i$ is the volume of the central compartment.

$z_i = (\log V_i, \log k_{ai}, \log Cl_i) \in \mathbb{R}^3$ for $i = 1, \ldots, n$ are the latent variables local to each patient sampled using MCMC, while the hyperparameters $\theta = (\mu_{kai}, \mu_V, \mu_{Cl}, \sigma_{kai}, \sigma_V, \sigma_{Cl}, \sigma) \in \mathbb{R}^3 \times \mathbb{R}_{>0}^4$ are inferred by maximizing the marginal likelihood.

**Dataset** We use the Theophylline dataset preprocessed and distributed by the saemix package (Comets et al., 2017), which is based on the one originally distributed by NONMEM (Boeckmann et al., 1994). This dataset contains 12 patients who were administered

![Figure 1: Trajectory of the MCMC-SAEM iterates for $\theta_2$ with a large MALA/ULA stepsize of $\eta = 5 \times 10^{-3}$. MALA only makes “occasional” progress due to rejections, while ULA makes progress nonetheless, albeit with some asymptotic bias. The dotted line marks the true value $\theta_2$.](image-url)
We initialize at $\theta_0 = (-1, 0, 0, 1, 1, 1)$, run MCMC-SAEM for 1000 iterations after 100 initial burn-in MCMC steps. We randomly split the patients into a train and test set of 9 : 3. Then, we estimate the marginal log-predictive density of the test patients resulting from the hyper-parameters found by SAEM. For estimating the test marginal LPD, we use the average of 100 importance weights drawn using annealed importance sampling (Neal, 2001), each using 100 annealing steps with a quadratic schedule.

### Results

The results are shown in Figure 2. We can see that ULA converges for the widest range of stepsizes. In this example, MALA struggles the most because the likelihood is highly non-smooth. Misspecifications of the hyper-parameters result in a sudden increase in the rejection rate. Since ULA is immune to this problem, it makes constant progress as long as it doesn’t diverge.

### 5.3 Robust Poisson Regression

Our first realistic experiment is a generalized linear model (GLM) with a Poisson likelihood. In particular, we consider a robustified, or “localized” (Wang and Blei, 2018), Poisson regression model, also known as the Poisson-log-normal model (Cameron and Trivedi, 2013, §4.2.4). The model is described as follows:

$$
\eta_i \sim \mathcal{N} (\beta^\top x_i + \beta_0, \sigma)
$$

### 5.4 Logistic Regression with Automatic Relevance Determination

Automatic relevance determination (ARD; Neal, 1996; MacKay, 1996) is prior on regression coefficients, where each regressor $\beta_i$ is assigned its own scale parameter $\gamma_i$. When maximizing the marginal likelihood with respect to the relevance parameters $\gamma_i$, the ARD prior is known to have a sparsifying effect, where irrelevant features are pruned as $\gamma_i \to \infty$. This “shrinkage” effect is lost if one does fully Bayesian inference. Therefore the empirical Bayes version of the problem is especially relevant.

Unfortunately, solving the maximum marginal likelihood problem is challenging, even for linear regression models (Tipping, 2001; Wipf and Nagarajan, 2007). Here, we demonstrate that MCMC-SAEM can be used to solve the ARD problem for logistic regression with...
Figure 4: Test average log-predictive density (LPD) for logistic regression with automatic relevance determination versus MALA/ULA stepsize $\eta$. The colored bands are 80% bootstrap confidence intervals of the mean computed from 32 independent train-test splits of a ratio of 8 : 1.

Table 2: LOGISTIC REGRESSION DATASETS

<table>
<thead>
<tr>
<th>NAME</th>
<th># data</th>
<th>dim(z)</th>
<th>dim(\theta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>phishing</td>
<td>11054</td>
<td>68</td>
<td>69</td>
</tr>
<tr>
<td>german</td>
<td>1000</td>
<td>217</td>
<td>216</td>
</tr>
<tr>
<td>caravan</td>
<td>9822</td>
<td>620</td>
<td>619</td>
</tr>
</tbody>
</table>

Bernoulli likelihoods.

The model is described as:

$$
\beta_0 \sim \mathcal{N}(0, 10) \quad \beta \sim \mathcal{N}(0, \gamma^{-1})
$$

$$
p_i = \text{logistic}(\beta^T x_i + \beta_0)
$$

$$
y_i \sim \text{Bernoulli}(p_i),
$$

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_d) \in \mathbb{R}_{+}^d$.

We optimize the hyperparameters $\theta = \gamma$ using MCMC-SAEM for 2000 iterations, after 100 burn-in steps, starting from an initial point of $\theta_0 = (1, \ldots, 1)$. On this problem, the result was quite sensitive to the initial point. After running MCMC-SAEM, we test the quality of the hyperparameters by estimating the log-predictive density (LPD) on a held-out test dataset using samples from the posterior. The samples are separately drawn using 2000 steps of MALA after 2000 adaptation/burn-in steps. MALA is automatically tuned to target an acceptance rate of 0.57 (Roberts and Rosenthal, 2001) using Nesterov’s dual averaging procedure (Nesterov, 2009). We replicate this over 32 independent train-test splits. The datasets were obtained from the UCI repository (Dua and Graff, 2017) and are shown in Table 2. The categorical variables were one-hot encoded, while the continuous features were z-standardized.

Furthermore, for this problem, preconditioning MALA and ULA is crucial since the scale of the posterior greatly varies depending on whether a feature is pruned or not. We use a diagonal preconditioner $P$ where the diagonal is set as $P_{ii} = 1/(\gamma_i^2 + 0.01) + \delta$ for $\beta$ and 1 for $\beta_0$, $\delta = 2 \times 10^{-16} > 0$ is necessary to ensure that the MCMC chain is not reducible even when a feature is pruned by $\gamma_i \to \infty$.

**Results**

The results are shown in Figure 4. We can see that ULA converges to a high quality solution for the widest range of stepsizes. On german, only ULA achieves the highest level of accuracy. Notably, in the large $\eta$ regime, the MALA chain tended to reduce to a state with an acceptance rate close to 0. ULA, on the other hand, is immune to this issue since it always makes progress as long as it does not diverge. Furthermore, when $\eta$ was too large, ULA immediately diverged at the initial SAEM iterations, which is easier to diagnose and correct.

6 DISCUSSIONS

In this work, we theoretically and empirically studied the impact of approximate MCMC algorithms. The theory suggests that they are feasible for SAEM in high dimensions as soon as the marginal log-likelihood is smooth enough. That is, the asymptotic bias will have a minimal effect on the found solution. We empirically confirmed this fact on multiple statistical problems. Furthermore, in our experiments, we observe that ULA versus MALA represents a trade-off between asymptotic bias versus non-asymptotic bias, where the latter can be more significant on high-dimensional and poorly conditioned problems. That is, with large stepsizes, MALA converges slower than ULA due to rejections, incurring a large non-asymptotic transient bias. As a result, ULA converges faster on these problems.

On a different note, this work provides a clear use case of approximate MCMC algorithms in statistics. While recent works (Akyildiz et al., 2023; Kuntz et al., 2023; de Bortoli et al., 2021) in empirical Bayes estimation leveraging approximate MCMC didn’t explore the benefits of approximate MCMC methods over exact MCMC, we showed here that being approximate can, in fact, be better.
Acknowledgements

The authors would like to thank the anonymous reviewers for their critical and constructive feedback.

K. Kim was supported by a gift from AWS AI to Penn Engineering’s ASSET Center for Trustworthy AI. J. R. Gardner was supported by NSF award [IIS-2145644]. Part of the work of A. O. Durmus is funded by the European Union (ERC, Ocean, 101071601). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References


Dheeru Dua and Casey Graff. UCI machine learning repository. 2017.


Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
       Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
       No. But we present a non-asymptotic convergence guarantee.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
       Yes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
       Yes.
   (b) Complete proofs of all theoretical results.
       [Yes/No/Not Applicable]
       Yes.
   (c) Clear explanations of any assumptions.
       [Yes/No/Not Applicable]
       Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
       Yes. The link to the repository is disclosed in Section 5.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
       [Yes/No/Not Applicable]
       Yes.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
       Yes. See the main text.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
       Yes. In Appendix C.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
       Yes.
   (b) The license information of the assets, if applicable.
       Not applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable.
       Not applicable.
   (d) Information about consent from data providers/curators.
       Not applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
       Not applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
       Not applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
       Not applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
       Not applicable.
# Table of Contents

1 INTRODUCTION

2 BACKGROUND
   2.1 Expectation-Maximization
   2.2 EM as a Root Finding Problem
   2.3 MCMC-SAEM with Approximate MCMC Algorithms
   2.4 Stochastic Approximation with Biased Dynamics

3 ASYMPTOTIC ANALYSIS
   3.1 Technical Assumptions
   3.2 Main Result

4 NON ASYMPTOTIC ANALYSIS
   4.1 Technical Assumptions
   4.2 Main Result

5 EXPERIMENTS
   5.1 Logistic Regression on a Synthetic Dataset
   5.2 Pharmacokinetics
   5.3 Robust Poisson Regression
   5.4 Logistic Regression with Automatic Relevance Determination

6 DISCUSSIONS

A TECHNICAL ASSUMPTIONS ON THE MARKOV KERNEL

B PROOFS
   B.1 Asymptotic Convergence
      B.1.1 Auxiliary Lemma
      B.1.2 Proof of Theorem 1
      B.1.3 Proof of Theorem 3
   B.2 Non-Asymptotic Convergence
      B.2.1 Auxiliary Lemmas
      B.2.2 Proof of Lemma 7
      B.2.3 Proof of Lemma 4
      B.2.4 Proof of Theorem 2

C COMPUTATIONAL RESOURCES
A TECHNICAL ASSUMPTIONS ON THE MARKOV KERNEL

In this section, we give conditions that imply A2 in terms of a bound from below of the Markov kernel on a small set and a drift condition toward this small set (see (Nummelin, 1991) for the definitions and main results). It offers some insights regarding what properties are essential for the state-dependent Markov kernels to have convergence guarantees. These conditions are verified for ULA (de Bortoli et al., 2021).

Define, for convergence guarantees. These conditions are verified for ULA (de Bortoli et al., 2021). It offers some insights regarding what properties are essential for the state-dependent Markov kernels to have small set and a drift condition toward this small set (see (Nummelin, 1991) for the definitions and main results).

For any \( M_1 \).

\[ \text{constant} \ l \]

weakened by considering subgeometric ergodicity conditions (Debavelaere et al., 2021). (DRI2) is just a control kernels \( s, \eta \) which are helpful to control the markovian noise \( \bar{\nu} \). For any \( \sqrt{c} \) expressed in \( s \).

Note that the constants \( C_{s,n} \) may be bounded over the compact sets of \( \mathbb{R}^d \), i.e. for each \( \mathcal{K} \subset \mathbb{R}^d \), there exists \( C \) such that \( C_{s,n} \in C \) and \( \cup \mathcal{K} \subset [0,1] \) such that for any function \( f \in \mathcal{L}_{\nu_{s,\eta}^c} \) and any integer \( k > 0 \)

\[ \| (\Pi_s)^k f - \pi_{s,\eta}(f) \|_{\nu_{s,\eta}^c} \leq C_{s,\eta} \gamma_{s,\eta}^k \| f \|_{\nu_{s,\eta}^c}. \]

We stress here the fact that the function \( W \) is global but that the bounds in the previous equations depend on the particular compact \( \mathcal{K} \) under consideration.

\(^2\)We use in this article the standard terminology and the notations introduced in (Nummelin, 1991, Chapter 4,5)
Lemma 2. Assume $M1$. Then $H4$, $M2$ are satisfied.

It is (Andrieu et al., 2005, Proposition 6.1) adapted to our framework here $\Theta = \mathbb{R}^d \times [0, \eta_0]$, $\beta = 1$. We removed the implications depending on the context of (Andrieu et al., 2005).

Furthermore, if we control the learning steps $(\gamma_n)$, the bias parameter $(\eta_n)$ and momentums of the latent variables $(Z_n)$, the desired result can be established.

M 3. The sequence $(\gamma_n)$ and $(\eta_n)$ are nonincreasing, positive and satisfy $\sum_{k=0}^{\infty} \gamma_k = \infty$, $\lim_{k \to \infty} \gamma_k = 0$, $\limsup_{k \to \infty} |\gamma_k^{-1} - \gamma_{k+1}^{-1}| = 0$ and

$$\sum_{k=1}^{\infty} \{\gamma_k^2 + \gamma_k|\eta_{k+1} - \eta_k|^\alpha + \gamma_{k+1}^{1+\alpha}\} < \infty$$

where $\alpha$ is defined in $M2$.

M 4. For any compact $K \subset \mathbb{R}^d$, there exists a constant $C > 0$ such that for any $z \in Z$, $\sup_{s \in K} \sup_{k \geq 0} \mathbb{E}_{z,s} [W^{lc}(Z_k) \mathbb{1}_{\{\sigma(K) \geq k\}}] \leq CW^{lc}(z)$.

$$\sigma(K) = \inf\{k \geq 0 : s_k \notin K\} \cup \{+\infty\},$$

$W$ is defined in $M2$.

Theorem 3. Assume $M1$, $M3$ and $M4$ then $A2$ is satisfied.

The proof is postponed to Appendix B.1.3.

The assumption M4 is essential and cannot be recovery directly from M1, if we want this hypothesis, we have to change a little the recursion (2) to ensure that $(s_n)$ stays in a compact by design and that $|s_{n+1} - s_n|$ is controlled by a non increasing sequence $(\epsilon_n)$ (Andrieu et al., 2005, p.9). We did not choose this framework because it gives rise to technicalities that are rarely implemented in practice. Even if it offers a guarantee of convergence, there are some drawbacks regarding convergence speed.
B PROOFS

B.1 Asymptotic Convergence

In this section, we will prove Theorem 1 in Section 3 and Theorem 3 in Appendix A.

B.1.1 Auxiliary Lemma

Before to prove the main theorem, we introduce another technical Lemma similar to Lemma 1,

**Lemma 3.** Assume $\textbf{H}1$-$\textbf{H}3$, then for any $s \in \mathbb{R}^d$,

$$\nabla V(s) = -A(s)h(s) \quad \text{and} \quad |\nabla V(s)|^2/\lambda_M \leq \langle \nabla V(s)\vert h(s) \rangle \leq |\nabla V(s)|^2/\lambda_m . \tag{5}$$

**Proof.** In the proof of (Delyon et al., 1999, Lemma 2), the authors derive that for any $s \in \mathbb{R}^d$, $\nabla V(s) = -A(s)h(s)$. By $\textbf{H}3$, $A(s)$ is invertible and $A^{-1}(s)$ has respectively for minimal and maximal eigen values $1/\lambda_M$ and $1/\lambda_m$, thus, writing for any $s \in \mathbb{R}^d$, $h(s) = -A^{-1}(s)\nabla V(s)$ yields the inequality (5). \qed
B.1.2 Proof of Theorem 1

Except for some inequalities and constants, the proof is nearly the same as in (Tadić and Doucet, 2017). The notations of (Tadić and Doucet, 2017) and ours are not the same: the function $V$ and the sequences $(s_n), (\gamma_n), (\beta_n)$ in this paper are replaced by the function $f$ and the sequences $(\theta_n), (\alpha_1), (\eta_n)$ in their paper. In the following, we detail all the proof modifications in (Tadić and Doucet, 2017) to get the desired theorem. We expose the sketch of proof related to each part of the proof before to details the changes.

Proof of (I) First remark that $S = \{ h(s) = 0 : s \in \mathbb{R}^d \}$ by

$$
\frac{d\theta(t)}{dt} \in F_\beta(\theta(t))^{\text{adic}}, \quad F_\beta(\theta(t))^{\text{adic}} = \{ -\nabla\theta(s(t)) + v, v \in \mathbb{R}^{d_s} : |v| \leq \beta \}. \quad (6)
$$

Then, by (Benaïm, 2006, Proposition 4.1, Theorem 5.7), on the event $\Lambda_Q$, all limit points of $(\theta_n)$ are in a set $R_{Q,2\beta}$ called "recurrent set" which is related to $(6)[\beta]$: there is a link between the asymptotics of the discretized and the continuous flow. By (Benaïm et al., 2012, Theorem 3.1), working with $(6)$ instead of the discretization, there exists $\psi_Q : [0,\infty) \to [0,\infty)$ (depending only on $f(\cdot)$) such that $\lim_{t \to 0} \psi_Q(t) = \psi_Q(0) = 0$ and $R_{Q,\beta} \subset \{ x \in \mathbb{R}^d : d(x, R_{Q,0}) \leq \psi_Q(\beta/2) \}$ for any $\beta \geq 0$. Finally, by (Hurley, 1995, Proposition 4), we can show that $R_{Q,0} = \{ x \in \mathbb{R}^d : \nabla f(x) = 0 \}$ provided that $f$ is $p$-continuously differentiable with $p > d$ by using Sard’s theorem.

To adapt the reasoning, we take $h$ instead of $-\nabla f$ such that we consider $F_\beta(s(t)) = \{ h(s) + v, v \in \mathbb{R}^d : |v| \leq \gamma \}$ for any $t, \beta \geq 0$. (Benaïm et al., 2012, Theorem 3.1) and (Benaïm, 2006, Proposition 4.1, Theorem 5.7) remain usable as long as $h$ is continuous, which is given by H3. (Hurley, 1995, Proposition 4) can’t be applied directly. This Proposition is for gradient flow, i.e. $dx/dt = -\nabla f(x)$ with $f$ continuously differentiable. However, the proof of (Hurley, 1995, Proposition 4) can still be done by replacing $-\nabla f$ by $h$ and $f$ by $V$. Indeed, for any $s \in \mathbb{R}^d$ and $K \subset \mathbb{R}^d$ a compact such that $\phi([0,1], s) \in K$,

$$
V(\phi(0,s)) = V(\phi(1,s)) - \int_0^1 \frac{dV(\phi(t,s))}{dt} dt = \int_0^1 \langle \nabla V(\phi(t,s)) | h(\phi(t,s)) \rangle dt = F(\phi(t,s)) \geq \frac{1}{\lambda_M} \int_0^1 |\nabla V(\phi(t,s))|^2 dt
$$

where we use (5). This shows that the following flow $dx/dt = h(x)$ decreases an energy $V$. The others arguments in the proof of (Hurley, 1995, Proposition 4) remains the same as long as we remark that the regular points of $V$ are dense by Sard’s theorem and H3, and by (5) we have

$$
S = \{ s \in \mathbb{R}^d : \nabla h(s) = 0 \} = \{ s \in \mathbb{R}^d : \nabla h(s) = 0 \}. \quad (7)
$$

The proof of (I) is complete.

Proof of (II)-(III) We work on the event $\Lambda_Q$, i.e., the sequence $(s_n)$ is in the compact $Q$ for $n$ large enough.

In (Tadić and Doucet, 2017, Proposition 8.2), the authors show that for any $s \in \mathbb{R}^d$, the distances $d(s,S), d(f(s), S)$ can be bounded from above by a term of the form $a_1|\nabla f(s)|a_2$ with $a_1, a_2 > 0$ using geometric and regularity arguments. The remaining work is to bound $|\nabla f(s)|$, which is performed in (Tadić and Doucet, 2017, Proposition 8.3) using previous results (Tadić and Doucet, 2017, Lemma 8.1-3) all deriving from a Taylor expansion of $f(s_n(n,t)) - f(s_n)$ in $s_n$ given in (Tadić and Doucet, 2017, equation (32) p.17). The idea is to measure the influence of the bias $\beta$ and the noise $\epsilon_\eta$ compared to the influence of the velocity field $\nabla f$: if $\nabla f$ has more influence than the other terms, $f$ will decrease, and its gradient as well till the point where $\nabla f$ has less influence than the noise and the bias. They show that the norm of the gradient at the asymptotic is ruled only by the bias since the accumulated noise vanishes by A2.

In order to adapt the proof, we should only change the Taylor expansion (Tadić and Doucet, 2017, equation (32) p.17) by replacing $-\nabla f$ by $h$ and then propagating the changes in the analysis related to the norm of the gradient in (Tadić and Doucet, 2017, Lemma 8.1-3) by using (5).
By doing a Taylor expansion, for any $n \geq 0$, $t \in (0, \infty)$,
\[
V(s_{a(n,t)}) - V(s_n) = \sum_{i=n}^{a(n,t)-1} \gamma_i (\nabla V(s_n) h(s_n)) - \left( \nabla V(s_n) \right) \sum_{i=n}^{a(n,t)-1} \gamma_i \xi_i + |\phi_n(t)|
\]
where $\phi_n$ is defined in (Tadić and Doucet, 2017, p.17). Then, by using (5),
\[
V(s_{a(n,t)}) - V(s_n) \leq -|\nabla V(s_n)| \left( \frac{1}{\lambda M} |\nabla V(s_n)| \sum_{i=n}^{a(n,t)-1} \gamma_i - \left| \sum_{i=n}^{a(n,t)-1} \gamma_i \xi_i \right| + |\phi_n(t)| \right)
\]

We denote by $\phi = \limsup_{n \to \infty} |\nabla V(s_n)|$ and $C_{1,Q} = \sup_{s \in Q} |\nabla V(s)|$. The second change are in equations (38),(39) in (Tadić and Doucet, 2017, Lemma 8.1, p.18) which are replaced for any $t \in (0, \infty)$ by

\[
\limsup_{n \to \infty} \max_{n \leq k < a(n,t)} |V(s_k) - V(s_n)| \leq C_{1,Q} t(\phi + \beta) \max \left( \frac{1}{\lambda_M}, 1 \right)
\]
\[
\limsup_{n \to \infty} |\phi_n(t)| \leq C_{1,Q} t^2(\phi + \beta)^2 \max \left( \frac{1}{\lambda_M}, 1 \right)^2
\]
since the last equation in (Tadić and Doucet, 2017, p.18) has to be replaced by,
\[
|s_k - s_n| \leq \sum_{i=n}^{k-1} \gamma_i |h(s_i)| + \left| \sum_{i=n}^{k_1} \gamma_i \xi_i \right| \leq t(\phi + \epsilon) \max \left( \frac{1}{\lambda_M}, 1 \right) + \max_{n \leq i < a(n,t)} \left| \sum_{i=n}^{k_1} \gamma_i \xi_i \right|
\]
where we used again (5) and the fact that $|\nabla V(s_n)| \leq \phi$ for $n$ large enough as in (Tadić and Doucet, 2017).

From these changes, the constants used in the proof are modified. The constant $\gamma$ is replaced p.20 by $\gamma = 2\lambda_M(\epsilon + \beta)$ and $C_{2,Q} = 4\lambda_M M_Q$. The (Tadić and Doucet, 2017, Lemma 8.3) holds on $(\Lambda_Q \setminus N_0) \cap (\phi > 2\beta \lambda_M)$. It impacts the final constant, we have $K_Q = \lambda_M \max\left(2, \tilde{C}_Q, C_{2,Q} \right)$ instead of $\max\left(2, \tilde{C}_Q, C_{2,Q} \right)$, where $M_Q$ is a constant related to the Yomdin theorem (Tadić and Doucet, 2017, Proposition 8.1, 8.2) and $\tilde{C}_Q = \sup_Q |\nabla V(s)|$. 

B.1.3 Proof of Theorem 3

We show now the convergence a.e of a constant which depends only upon the compact set \(E\). Using the existence of solution to the Poisson equation, for any \(k \geq 1\), we have,

\[ e_{k-1} = S(z_k) - 1_{\eta_k} (s_{k-1}) = (I - \Pi_{s_{k-1}}^{\eta_k}) \nu_{s_{k-1}}^{\eta_k} (z_k), \]

and we denote by,

\[ T_n = \sum_{k=1}^n \gamma_k e_{k-1} 1_{\gamma_k \geq k}. \]

We want to show that \(T_n\) converges almost surely on the event \(\tilde{A}_K\). Using for all \(k \geq 0\):

\[ 1_{\gamma_k \geq k} = 1_{\gamma_k \geq k+1} + 1_{\gamma_k = k}, \]

we may write \(T_n = \sum_{i=1}^5 T_n^{(i)}\) where,

\[
\begin{align*}
(A.1) & \quad T_n^{(1)} = \sum_{k=1}^n \gamma_k \left( \nu_{s_{k-1}}^{\eta_k} (Z_k) - \Pi_{s_{k-1}}^{\eta_k} \nu_{s_{k-1}}^{\eta_k} (Z_{k-1}) \right) 1_{\gamma_k \geq k}, \\
(A.2) & \quad T_n^{(2)} = \sum_{k=1}^{n-1} \gamma_k^{k+1} \left( \Pi_{s_k}^{\eta_k+1} \nu_{s_k}^{\eta_k+1} (Z_k) - \Pi_{s_k}^{\eta_k} \nu_{s_k}^{\eta_k} (Z_k) \right) 1_{\gamma_k \geq k+1}, \\
(A.3) & \quad T_n^{(3)} = \sum_{k=1}^{n-1} (\gamma_k^{k+1} - \gamma_k) \Pi_{s_k}^{\eta_k} \nu_{s_k}^{\eta_k} (Z_k) 1_{\gamma_k \geq k+1}, \\
(A.4) & \quad T_n^{(4)} = \sum_{k=1}^{n-1} \gamma_k \Pi_{s_k}^{\eta_n} \nu_{s_k}^{\eta_n} (Z_n) 1_{\gamma_k \geq n}, \\
(A.5) & \quad T_n^{(5)} = -\sum_{k=1}^{n-1} \gamma_k \Pi_{s_k}^{\eta_k} \nu_{s_k}^{\eta_k} (Z_k) 1_{\gamma_k = k}.
\end{align*}
\]

We show now the convergence a.e of \(T_n^{(i)}, i = 1, \ldots, 5\) on the event \(\tilde{A}_K\). First remark that \(T_n^{(5)} = 0\) on \(\tilde{A}_K\). In the sequel \(C\) denotes a constant which depends only upon the compact set \(K\) through the quantities defined in the assumptions and whose value may change upon each appearance. Denoting by

\[ D(\gamma, K, z) = \sup_{s \in K} \sup_{K_{\gamma}} \mathbb{E}_{z, s} \left[ W^l (Z_k) 1_{\gamma_k \geq k} \right], \]

we have for any \(s \in \mathbb{R}^d\) and \(z \in Z\)

\[
\begin{align*}
(A.6) & \quad \mathbb{E}_{s, z} \left[ \sum_{k=1}^n \gamma_k^{k+1} \nu_{s_{k-1}}^{\eta_k} (Z_k) - \Pi_{s_{k-1}}^{\eta_k} \nu_{s_{k-1}}^{\eta_k} (Z_{k-1}) \right]^2 1_{\gamma_k \geq k} \\
& \quad \leq C \left( \sum_{k=0}^n \gamma_k^2 \right)^{l_c/2} D(\gamma, K, z), \\
(A.7) & \quad \mathbb{E}_{s, z} \left[ \sum_{k=1}^n \gamma_k^{k+1} \left( \Pi_{s_k}^{\eta_k+1} \nu_{s_k}^{\eta_k+1} (Z_k) - \Pi_{s_k}^{\eta_k} \nu_{s_k}^{\eta_k} (Z_k) \right) 1_{\gamma_k \geq k+1} \right] \\
& \quad \leq C \left( \sum_{k=1}^n \gamma_k \right)^{l_c} + \left( \sum_{k=1}^n \gamma_k \eta_k \right)^{l_c} D(\gamma, K, z)^{1+\alpha}, \\
(A.8) & \quad \mathbb{E}_{s, z} \left[ \sum_{k=1}^{n-1} \gamma_k^{k+1} \left| \Pi_{s_k}^{\eta_k} \nu_{s_k}^{\eta_k} (Z_k) \right| 1_{\gamma_k \geq k+1} \right] \\
& \quad \leq C \left( \sum_{k=1}^n \gamma_k \right)^{l_c} D(\gamma, K, z), \\
(A.9) & \quad \mathbb{E}_{s, z} \left[ \gamma_k^{k+1} \left| \Pi_{s_k}^{\eta_k} \nu_{s_k}^{\eta_k} (Z_n) \right| 1_{\gamma_k \geq n} \right] \leq C \gamma_k^{l_c} D(\gamma, K, z).
\end{align*}
\]
We will show these inequalities at the end of the proof. For any \( k \geq 1 \),
\[
\mathbb{E}_{s,z} \left[ (\nu_{s_{k+1}}^{\eta_k+1} (Z_{k+1}) - \Pi_{s_{k+1}}^{\eta_k+1} (Z_k)) \mathbb{I}_{\{\sigma(K) \geq k+1\}} \mid \mathcal{F}_k \right] = \\
(\Pi_{s_{k+1}}^{\eta_k+1} (Z_k) - \Pi_{s_{k+1}}^{\eta_k+1} (Z_k)) \mathbb{I}_{\{\sigma(K) \geq (k+1)\}} = 0
\]
\( T_n^{(i)} \) is a \((\mathbb{R}^d, \mathbb{F})\)-valued martingale. Applying Doob's theorem with (A.6), we have the convergence of \( (T_n^{(i)}) \) almost surely on \( \Lambda_K \). Since \( T_n^{(i)} \mathbb{I}_{\{\sigma(K) \geq n\}} = 0 \), we have
\[
T_n \mathbb{I}_{\{\sigma(K) \geq n\}} = \sum_{i=1}^4 T_n^{(i)} \mathbb{I}_{\{\sigma(K) \geq n\}}.
\]
Using the inequalities (A.7)-(A.9) with A1 gives the convergence of \( T_n^{(i)} \) for \( i = 2, 3, 4 \) on the event \( \Lambda_K \). On the event \( \{\sup_n |s_n| < \infty\} \), we can work in \( \Lambda_K \) for some compact set \( K \subset \mathbb{R}^d \), then \( \sum_i \gamma_i \epsilon_i \) converges almost surely, which yields the assumption A2 is given.

**Proof of (A.6)** Under M2, we have for any \( k \geq 1 \),
\[
\left| \nu^{\eta_k}_{s_{k-1}} (Z_k) - \Pi^{\eta_k}_{s_{k-1}} (Z_k) \right| \mathbb{I}_{\{\sigma(K) \geq k+1\}} \leq W(Z_k) \mathbb{I}_{\{\sigma(K) \geq k+1\}} + W(Z_{k-1}) \mathbb{I}_{\{\sigma(K) \geq k\}}
\]
and then,
\[
\mathbb{E}_{s,z} \left[ \sum_{k=1}^{\infty} \gamma_k^2 \left| \nu^{\eta_k}_{s_{k-1}} (Z_k) - \Pi^{\eta_k}_{s_{k-1}} (Z_{k-1}) \right|^2 \mathbb{I}_{\{\sigma(K) \geq k\}} \right] \leq C \sum_{k=1}^{\infty} \gamma_k^2 W(Z_k)^2 \mathbb{I}_{\{\sigma(K) \geq k\}}
\]
\[
\leq C \sum_{k=1}^{\infty} \gamma_k^2 \| D(\gamma, K, z) \|
\]
where we used that \( \{\gamma_i\} \) is decreasing and a Minkowski inequality.

**Proof of (A.7)** Under M2, we have For any \( k \geq 1 \),
\[
\gamma_{k+1} \left| \Pi^{\eta_k}_{s_{k+1}} (Z_k) - \Pi^{\eta_k}_{s_{k-1}} (Z_k) \right| \mathbb{I}_{\{\sigma(K) \geq k+1\}} \leq C \gamma_{k+1} W(Z_k) \mathbb{I}_{\{\sigma(K) \geq k+1\}} \]
then, using the bound on sufficient statistics by M2 and that \( s_k \) is in a compact, for any \( k \geq 1 \),
\[
|s_k - s_{k-1}| \mathbb{I}_{\{\sigma(K) \geq k+1\}} \leq \gamma_k \left| S(Z_{k+1}) - s_k \right| \mathbb{I}_{\{\sigma(K) \geq k+1\}} \leq \gamma_k W(Z_k) + C \mathbb{I}_{\{\sigma(K) \geq k+1\}}
\]
Thus, we have,
\[
\sum_{k=1}^{\infty} \gamma_{k+1} \left| \Pi^{\eta_k}_{s_{k+1}} (Z_k) - \Pi^{\eta_k}_{s_{k-1}} (Z_k) \right| \mathbb{I}_{\{\sigma(K) \geq k+1\}} \]
\[
\leq C \sum_{k=0}^{\infty} \gamma_{k+1} \left| \eta_{k+1} - \eta_k \right| \mathbb{I}_{\{\sigma(K) \geq k+1\}} \leq C \sum_{k=0}^{\infty} \gamma_{k+1} \left| \eta_{k+1} - \eta_k \right| W(Z_k) + \gamma_k^2 \left| W(Z_k) + W(Z_k) \mathbb{I}_{\{\sigma(K) \geq k+1\}} \right|
\]
in the last inequality, we used that \( \{\beta_k\} \) is bounded and that \( \gamma_k \) is decreasing. We conclude using the Minkowski’s inequality.

Remark that if \( \beta > 0 \), we have \( \epsilon_k = |s_k - s_{k-1}| = \Theta(\gamma_k) \) which makes it impossible to apply the proof of (Delyon et al., 1999) to show that we can have \( \{s_n\} \) in a compact by design. Indeed, it is needed to have \( \sum_{n} (\frac{\epsilon_n}{\epsilon_n})^\beta < \infty \), but here we have \( \epsilon_n \sim \gamma_n \) because of the bias.
Proof of (A.8) Under $M_2$, 
\[ \sum_{k=1}^{n-1} |\gamma_{k+1} - \gamma_k| \left| \Pi_{\gamma_{k+1}} \mu_{\gamma_{k+1}} (Z_k) \right| \mathbb{1}_{\{\sigma(K) \geq k + 1\}} \leq C \sum_{k=1}^{\infty} |\gamma_k - \gamma_{k+1}| W(Z_k) \mathbb{1}_{\{\sigma(K) \geq k + 1\}}, \]
and the proof follows from Minkowski's inequality.

Proof of (A.9) Under $M_2$, 
\[ \mathbb{E}_{s,z} \left[ \left| \Pi_{s_{n-1}^{-1} s_{n-1}^{-1}} (Z_n) \mathbb{1}_{\{\sigma(K) \geq n\}} \right|^{l_c} \right] \leq C \mathbb{E}_{s,z} \left[ W^{l_c} (Z_n) \mathbb{1}_{\{\sigma(K) \geq n\}} \right] \]
B.2 Non-Asymptotic Convergence

In this section, we will prove Theorem 2 in Section 4.

B.2.1 Auxiliary Lemmas

Before proving the theorem, we need some preliminary results. First, we derive a Robbins-Siegmund Lemma. Then, we control the Markov stochasticity of the process \((e_n)\) by using the Poisson solutions given by \(N2\). It brings out a weighted sum of Martingale difference, which can be bounded by using a concentration inequality given in (Li and Orabona, 2020, Lemma 1). Finally, we conclude by reorganizing the terms. We first state the Robbins-Siegmund Lemma before giving others technical results.

**Lemma 4. (Robbins-Siegmund Lemma)** Assume \(H1-3, N1, N3, \) and \(N4\). Then,

\[
V(s_{k+1}) \leq V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) \rangle + \gamma_{k+1} \langle \nabla V(s_k) | e_k \rangle + a_k + b_k | h(s_k) |^2 ,
\]

where the constants are

\[
a_k = \frac{1}{2} \lambda_M \sqrt{\rho} \gamma_{k+1} + L_V \sigma^2 \gamma_{k+1}^2 ,
\]

\[
b_k = \left( \frac{1}{2} \lambda_M \sqrt{\rho} + \lambda_M \sqrt{\gamma} \right) \gamma_{k+1} + L_V \gamma_{k+1}^2 .
\]

The proof is delayed after the small technical results that followed.

Then, we state the concentration inequality.

**Lemma 5 (Adaptation of Lemma 1 by Li and Orabona 2020).** Let \((E_n)_{n \geq 0}\) be a martingale difference sequence adapted to the filtration \((\mathcal{F}_{n+1} = \sigma(Z_i, i \leq n + 1))_{n \geq 0}\) and a stochastic process \((v_n)_{n \geq 0}\) adapted to the filtration \((\mathcal{F}_n)_{n \geq 0}\), such that for any \(k \geq 0\), \(\mathbb{E}^{\mathcal{F}_{k+1}} \exp \frac{E_k}{v_k^2} \leq \exp(1)\). Then, for any fixed \(\varrho > 0\) and \(\delta \in (0, 1)\), with probability at least \(1 - \delta\), it holds that for any \(n \geq 1\),

\[
n \sum_{k=1}^{n} E_k \leq \frac{3}{\varrho} \sum_{k=1}^{n} v_k^2 + \frac{1}{\varrho} \log \frac{1}{\delta} .
\]

This lemma will be used with the following:

**Lemma 6.** Assume \(H1-3 \) and \(N2\). Then, there exists \(c \in [0, 2L_v^{(0)}]\) such that the MCMC kernel has sub-Gaussian tails for \(\nu^\eta\), i.e., for any \(s \in \mathbb{R}^d\),

\[
\mathbb{E}^{\mathcal{F}_{k+1}} \exp \left( |\nu^\eta(Z_{k+1}) - \Pi^\eta \nu^\eta(Z_k) |^2 / c^2 \right) \leq \exp(1) .
\]

The proof is straightforward: by \(N2\), for any \(k \geq 0\) and \(s \in \mathbb{R}^d\),

\[
|\nu^\eta(Z_{k+1}) - \Pi^\eta \nu^\eta(Z_k) | / 2L_v^{(0)} \leq 1 ,
\]

which implies that (7) holds with at least \(c = 2L_v^{(0)}\).

We introduce the Martingale difference related to Lemma 5. By using the solutions of the Poisson equation by \(N2\), we have for any \(k \geq 0\),

\[
D_k \triangleq \nu^\eta_{s_k}(Z_{k+1}) - \Pi^\eta \nu^\eta_{s_k}(Z_k) , \ E(D_k | \mathcal{F}_k) = 0
\]

where use the Markov property related to \((Z_k)\) to remark that \((D_k)_k\) is a sequence of Martingale difference adapted to the filtration \((\mathcal{F}_{k+1})_k\). This yields the following Lemma.

**Lemma 7.** Assume \(H1-3\) and \(N2\). Then, with probability at least \(1 - \delta\), for any \(n \geq 1\),

\[
- \sum_{k=1}^{n} \gamma_{k+1} \langle \nabla V(s_k) | D_k \rangle \leq \varrho \lambda_M^2 \rho^2 \sum_{k=1}^{n} \gamma_{k+1}^2 | h(s_k) |^2 + \frac{1}{\varrho} \log \frac{1}{\delta} ,
\]

where \(\varrho > 0\) is a free variable.
B.2.2 Proof of Lemma 7.

For any $k \geq 1$, denoting by

$$E_k \triangleq -\gamma_{k+1} \langle \nabla V(s_k) | D_k \rangle,$$

we have,

$$|E_k|^2 = \gamma_{k+1}^2 |\langle \nabla V(s_k) | D_k \rangle|^2 \\
\leq \gamma_{k+1}^2 |\nabla V(s_k)|^2 |D_k|^2 \quad \text{(Cauchy-Schwarz)} \\
\leq \lambda_M^2 \gamma_{k+1}^2 |h(s_k)|^2 |D_k|^2. \quad \text{(Lemma 1)}$$

Since $D_k$ is sub-Gaussian by Lemma 6, $(E_k)$ is also sub-gaussian such that

$$\mathbb{E}^\mathcal{F}_{k+1} \exp \left( |E_k|^2 / \left( \lambda_M^2 \gamma_{k+1}^2 |h(s_k)|^2 c^2 \right) \right) \leq \exp (1).$$

Thus, Lemma 5 applies after setting $v_k \triangleq \lambda_M^2 \gamma_{k+1}^2 |h(s_k)|^2 c^2$, which is adapted to the filtration $\mathcal{F}_k$. Then, with probability at least $1 - \delta$, for any $n \geq 1$,

$$\sum_{k=1}^{n} E_k \leq \frac{3}{4} \theta \sum_{k=1}^{n} v_k^2 + \frac{1}{\theta} \log \frac{1}{\delta} \\
= \frac{3}{4} \theta \sum_{k=1}^{n} \lambda_M^2 \gamma_{k+1}^2 |h(s_k)|^2 c^2 + \frac{1}{\theta} \log \frac{1}{\delta} \\
\leq \theta \lambda_M^2 c^2 \sum_{k=1}^{n} \gamma_{k+1}^2 |h(s_k)|^2 + \frac{1}{\theta} \log \frac{1}{\delta},$$

where we have only organized the constants in the last inequality.
B.2.3 Proof of Lemma 4

We now prove the Robbins-Siegmund Lemma. From the $L_2$-smoothness of the Lyapunov function by N4, we have for any $h \geq 0$,

$$V(s_{k+1}) \leq V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k, Z_{k+1}) \rangle + \frac{L_V \gamma_{k+1}^2}{2} |h(s_k, Z_{k+1})|^2$$

$$= V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) + \xi_k \rangle + \frac{L_V \gamma_{k+1}^2}{2} |h(s_k) + H(s_k, Z_{k+1}) - h(s_k)|^2,$$

applying the inequality $|a + b|^2 \leq 2|a|^2 + 2|b|^2$,

$$\leq V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) + \xi_k \rangle + L_V \gamma_{k+1}^2 \left(|h(s_k)|^2 + |H(s_k, Z_{k+1}) - h(s_k)|^2 \right),$$

and applying N1,

$$\leq V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) \rangle + \gamma_{k+1} \langle \nabla V(s_k) | \xi_k \rangle + L_V \gamma_{k+1}^2 |h(s_k)|^2 + L_V \gamma_{k+1}^2 \sigma^2$$

$$= V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) \rangle + \gamma_{k+1} \langle \nabla V(s_k) | \xi_k \rangle + L_V \gamma_{k+1}^2 \left(|h(s_k)|^2 + \sigma^2 \right).$$

Mean-field dynamics \quad MCMC bias dynamics

For the bias dynamics, recall that the bias can be decomposed as

$$\xi_k = \underbrace{e_k}_{\text{non-asymptotic bias}} + \underbrace{\beta_k}_{\text{asymptotic bias}}.$$

Considering this, the bias dynamics can be decomposed as

$$\langle \nabla V(s_k) | \xi_k \rangle = \langle \nabla V(s_k) | e_k + \beta_k \rangle = \langle \nabla V(s_k) | e_k \rangle + \langle \nabla V(s_k) | \beta_k \rangle.$$

For the asymptotic bias, we have

$$\langle \nabla V(s_k) | \beta_k \rangle \leq |\nabla V(s_k)| |\beta_k|,$$

applying Lemma 1 and N3,

$$\leq \lambda_M |h(s_k)| \sqrt{\tau_0 + \tau_1 |h(s_k)|},$$

applying the inequality $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$,

$$\leq \lambda_M \left(\sqrt{\tau_0} |h(s_k)| + \sqrt{\tau_1 |h(s_k)|^2} \right),$$

and the inequality $a \leq \frac{1}{2} (1 + a^2)$ for $a \geq 0$,

$$\leq \lambda_M \left(\frac{1}{2} \sqrt{\tau_0} + \left(\frac{1}{2} \sqrt{\tau_0} + \sqrt{\tau_1} \right) |h(s_k)|^2 \right) .$$

Combining the results, we have

$$V(s_{k+1}) \leq V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) \rangle + L_V \gamma_{k+1}^2 \left(|h(s_k)|^2 + \sigma^2 \right)$$

$$+ \gamma_{k+1} \langle \nabla V(s_k) | e_k \rangle + \lambda_M \gamma_{k+1} \left(\frac{1}{2} \sqrt{\tau_0} + \left(\frac{1}{2} \sqrt{\tau_0} + \sqrt{\tau_1} \right) |h(s_k)|^2 \right)$$

$$= V(s_k) + \gamma_{k+1} \langle \nabla V(s_k) | h(s_k) \rangle + \gamma_{k+1} \langle \nabla V(s_k) | e_k \rangle$$

$$+ \left(\frac{1}{2} \lambda_M \sqrt{\tau_0} \gamma_{k+1} + L_V \sigma^2 \gamma_{k+1}^2 \right) + \left(\frac{1}{2} \lambda_M \sqrt{\tau_0} \gamma_{k+1} + \lambda_M \sqrt{\tau_1} \gamma_{k+1} + L_V \gamma_{k+1}^2 \right) |h(s_k)|^2 .$$
B.2.4 Proof of Theorem 2

From Lemma 4, we have for any \( k \geq 0 \),

\[-\gamma_{k+1} \langle \nabla V (s_k) | h (s_k) \rangle \leq V (s_k) - V (s_{k+1}) + \gamma_{k+1} \langle \nabla V (s_k) | e_k \rangle + a_k + b_k | h (s_k) |^2 .\]

Applying Lemma 1,

\[\gamma_{k+1} \lambda_m | h (s_k) |_2^2 \leq V (s_k) - V (s_{k+1}) + \gamma_{k+1} \langle \nabla V (s_k) | e_k \rangle + a_k + b_k | h (s_k) |^2 .\]

Summing this bound for \( k = 0, \ldots, n \) forms a telescoping sum as

\[\lambda_m \sum_{k=0}^n \gamma_{k+1} | h (s_k) |_2^2 \leq V (s_0) - V (s_{n+1}) + \sum_{k=0}^n \gamma_{k+1} \langle \nabla V (s_k) | e_k \rangle + \sum_{k=0}^n a_k + \sum_{k=0}^n b_k | h (s_k) |^2 \]

\[\leq V (s_0) - V^* + \sum_{k=0}^n \gamma_{k+1} \langle \nabla V (s_k) | e_k \rangle + \sum_{k=0}^n a_k + \sum_{k=0}^n b_k | h (s_k) |^2 .\]

Now, we focus on the non-asymptotic bias dynamics term. Karimi et al. (Karimi et al., 2019, Theorem 2) show that the sum of inner products can be decomposed as

\[-\sum_{k=0}^n \gamma_{k+1} \langle \nabla V (s_k) | e_k \rangle = A_1 + A_2 + A_3 + A_4 + A_5 ,\]

where the terms are

\[A_1 = -\sum_{k=1}^n \gamma_{k+1} \langle \nabla V (s_k) | \nu^\eta_{\theta_k} (Z_{k+1}) - \Pi^\eta_{\theta_k} \nu^\eta_{\theta_k} (Z_k) \rangle ,\]

\[A_2 = -\sum_{k=1}^n \gamma_{k+1} \langle \nabla V (s_k) | \Pi^\eta_{\theta_k} \nu^\eta_{\theta_k} (Z_k) - \Pi^\eta_{\theta_{k-1}} \nu^\eta_{\theta_{k-1}} (Z_k) \rangle ,\]

\[A_3 = -\sum_{k=1}^n \gamma_{k+1} \langle \nabla V (s_k) - \nabla V (s_{k-1}) | \Pi^\eta_{\theta_{k-1}} \nu^\eta_{\theta_{k-1}} (Z_k) \rangle ,\]

\[A_4 = -\sum_{k=1}^n (\gamma_{k+1} - \gamma_k) \langle \nabla V (s_{k-1}) | \Pi^\eta_{\theta_{k-1}} \nu^\eta_{\theta_{k-1}} (Z_k) \rangle ,\]

\[A_5 = -\gamma_1 \langle \nabla V (s_0) | \Pi^\eta_{\theta_0} \nu^\eta_{\theta_0} (z_1) \rangle + \gamma_{n-1} \langle \nabla V (s_{n-1}) | \Pi^\eta_{\theta_{n-1}} \nu^\eta_{\theta_{n-1}} (z_n) \rangle .\]

\( A_1 \) is given by Lemma 7 where we set \( \rho = 1 \) and use that \( c \leq 2L^{(0)}_\nu \). On the other hand, for \( A_2, A_3, A_4, A_5 \), we can use the results in the proof of (Karimi et al., 2019, Theorem 2) by setting

\[c_0 = 0, \quad c_1 = \lambda_m, \quad d_0 = 0, \quad d_1 = \lambda_M, \quad L = L_V, \quad L^{(0)}_{PH} = L^{(0)}_\nu, \quad L^{(1)}_{PH} = L^{(1)}_\nu.\]

Then, with probability at least 1 − \( \delta \), we have

\[A_1 \leq (2\lambda_M L^{(0)}_\nu)^2 \sum_{k=1}^n \gamma_{k+1}^2 | h (s_k) |^2 + \log \frac{1}{\delta} ,\]

\[A_2 \leq L^{(1)}_\nu \lambda_M \left( \sigma \sum_{k=1}^n \gamma_k^2 + \left( \frac{1}{2} + \alpha_1 \sigma + \alpha_1 \frac{1}{2} \right) \sum_{k=0}^n \gamma_k^2 | h (s_k) |^2 \right) ,\]
\[
A_3 \leq L_V L^{(0)}_\nu \left( (1 + \sigma) \sum_{k=1}^{n} \gamma_k^2 + \sum_{k=1}^{n} \gamma_k^2 |h(s_{k-1})|^2 \right)
\leq L_V L^{(0)}_\nu \left( (1 + \sigma) \sum_{k=1}^{n} \gamma_k^2 + \sum_{k=0}^{n} \gamma_{k+1}^2 |h(s_k)|^2 \right),
\]

\[
A_4 \leq L^{(0)}_\nu \left( (\gamma_0 - \gamma_n) + \alpha_2 \lambda_M \sum_{k=1}^{n} \gamma_k^2 |h(s_{k-1})|^2 \right)
\leq L^{(0)}_\nu \left( \gamma_0 + \alpha_2 \lambda_M \sum_{k=0}^{n} \gamma_{k+1}^2 |h(s_k)|^2 \right),
\]

\[
A_5 \leq L^{(0)}_\nu \lambda_M \left( 2 + \sum_{k=0}^{n} \gamma_k^2 |h(s_k)|^2 \right).
\]

By reorganizing the terms,

\[
\sum_{k=0}^{n} \gamma_{k+1} (\lambda_m - C_{b_1} - C_{n_1} \gamma_{k+1}) |h(s_k)|^2 \leq V(s_0) - V^* + C_{n_2} \sum_{k=0}^{n} \gamma_{k+1}^2 + C_0 + \log \frac{1}{\delta} + C_{b_2} \sum_{k=0}^{n} \gamma_{k+1}.
\]

From this, we obtain the condition on the stepsize that \( \gamma_{k+1} \leq (\lambda_m - C_{b_1}) / C_{n_1} \) for all \( k = 0, \ldots, n \). Furthermore, constant progress can be guaranteed by further enforcing \( \gamma_{k+1} \leq \frac{1}{2} (\lambda_m - C_{b_1}) / C_{n_1} \). Then, since \( \gamma_{k+1} \leq \gamma_k \), the following inequalities hold:

\[
\sum_{k=0}^{n} \gamma_{k+1} |h(s_k)|^2 \leq \frac{2}{\lambda_m - C_{b_1}} \left( V(s_0) - V^* + C_{n_2} \sum_{k=0}^{n} \gamma_{k+1}^2 + C_0 + \log \frac{1}{\delta} + C_{b_2} \sum_{k=0}^{n} \gamma_{k+1} \right).
\]

Finally, dividing both sides by \( \sum_{k=0}^{n} \gamma_{k+1} \), we have

\[
\frac{1}{\sum_{k=0}^{n} \gamma_{k+1}} \sum_{k=0}^{n} \gamma_{k+1} |h(s_k)|^2 \leq \frac{2}{\lambda_m - C_{b_1}} \left( \frac{V(s_0) - V^* + C_{n_2} \sum_{k=0}^{n} \gamma_{k+1}^2 + C_0 + \log \frac{1}{\delta}}{\sum_{k=0}^{n} \gamma_{k+1}} + C_{b_2} \right).
\]

Since the lower bound forms a weighted average, lower bounding it with the minimum over the iterates yields the result.
### C COMPUTATIONAL RESOURCES

Table 3: Computational Resources

<table>
<thead>
<tr>
<th>Type</th>
<th>Model and Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Topology</td>
<td>1 socket with 8 physical cores</td>
</tr>
<tr>
<td>Processor</td>
<td>1 Intel i9-11900F, 2.5 GHz (maximum 5.2 GHz) per socket</td>
</tr>
<tr>
<td>Cache</td>
<td>80 KB L1, 512 KB L2, and 16 MB L3</td>
</tr>
<tr>
<td>Memory</td>
<td>64 GiB RAM</td>
</tr>
</tbody>
</table>

All experiments took about 50 hours to complete.