IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Markov chain Monte Carlo for Gaussian: A linear control perspective

Bo Yuan, Jiaojiao Fan, Yuqing Wang, Molei Tao and Yongxin Chen

***Abstract*— Drawing samples from a given target probability distribution is a fundamental task in many science and engineering applications. A commonly used method for sampling is the Markov chain Monte Carlo (MCMC) which simulates a Markov chain whose stationary distribution coincides with the target one. In this work, we study the convergence and complexity of MCMC algorithms from a dynamic system point of view. We focus on the special cases with Gaussian target distributions and provide a Lyapunov perspective to them using tools from linear control theory. In particular, we systematically analyze two popular MCMC algorithms: Langevin Monte Carlo (LMC) and kinetic Langevin Monte Carlo (KLMC). By applying Lyapunov theory we derive impressive complexity bounds to these algorithms: for LMC, our result is better than all existing results, and for KLMC, ours matches the best known bound. Our analysis also highlights subtle differences between sampling and optimization that could inform the more challenging task to sample from general distributions. Overall, our findings offer valuable insights for improving MCMC algorithms.**

***Index Terms*— Linear systems, Lyapunov methods, Filtering.**

I. Introduction

The task to draw random samples from an (unnormalized) distribution $\nu \propto \exp(-f(x))$ with potential $f : \mathbb{R}^d \to \mathbb{R}$, plays a crucial role in many areas of science and engineering, including Bayesian inference, filtering/estimation, uncertainty quantification, inverse problems, etc [1], [2], [3], [4]. For instance, particle filtering algorithms recursively sample from the posterior distributions of the state after each new measurement arrives. In inference problems, in contrast to optimization approaches that give point estimates, the sampling methods have the advantage of being able to quantify the uncertainties of such estimates.

A popular paradigm for sampling is Markov chain Monte Carlo (MCMC), and chief among them are those based on the Langevin dynamics, either overdamped or underdamped [5], [6], [7]. In practice, the Langevin dynamics are (time) discretized and integrated over a given stepsize. A metric that is commonly used for quantifying the performance of sampling algorithms is the number of steps required to achieve a given level of accuracy in some statistical divergence or metric, known as mixing time [8], akin to the number of iterations in optimization to achieve certain accuracy.

This work was supported by the NSF under grant 1847802, 1942523 and 2008513.

B. Yuan, J. Fan, and Y. Chen are with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA {byuan48,jiaojiaofan,yongchen}@gatech.edu

Y. Wang and M. Tao are with the School of Math, Georgia Institute of Technology, Atlanta, GA 30332, USA {ywang3398,mtao}@gatech.edu

Both the overdamped and underdamped Langevin dynamics (continuous in time) and algorithms (discrete in time) have been extensively studied with a variety of assumptions on the target distribution ν. A standard setting is when the potential energy $f(x)$ defined on $\mathbb{R}^d$ is strongly convex and smooth (i.e. having a global Lipschitz gradient). In this case, the complexity (i.e. mixing time for reaching ϵ statistical error) of Langevin Monte Carlo (LMC) can be $\widetilde{O}(d\epsilon^{-2})$ [9], [10], [11], [12], where the $\widetilde{O}$ notation means Landau's big O additionally with constant and logarithm terms in ϵ ignored. Under the same assumption, the complexity of an underdamped/kinetic Langevin Monte Carlo (KLMC) was however shown to be $\widetilde{O}(\sqrt{d}\epsilon^{-1})$ [13], [14], which gives an order of $\widetilde{O}(\sqrt{d}\epsilon^{-1})$ improvement over the overdamped one. These existing works adopted different proof techniques to analyze the convergence rates of standard LMC or KLMC for general strongly-convex and smooth potentials. It is not clear whether these complexity bounds can be further improved, even just for the LMC or KLMc algorithms.

In this work, we make inroads toward better non-asymptotic complexity bounds for sampling by examining the sampling problems with Gaussian target distributions. Any nondegenerate Gaussian distributions satisfy the standard setting considered in the prementioned existing works: the potential is strong-convex and smooth. In particular, we make the following assumption:

Assumption 1: The target distribution is Gaussian $\nu \propto \exp(-f) = \exp(-\frac{1}{2}(\cdot - m_g)^T\Sigma_g^{-1}(\cdot - m_g))$ (namely $\nu = \mathcal{N}(m_g, \Sigma_g)$) defined on $\mathbb{R}^d$ and the potential f is α-strongly convex and β-smooth, i.e., $\alpha\mathbb{I} \preceq \Sigma_g^{-1} \preceq \beta\mathbb{I}$. The initial distribution for the MCMC algorithm is also Gaussian.

We study Langevin sampling algorithms in the Gaussian setting from a linear control perspective and present a new complexity analysis for these algorithms by leveraging tools from linear Lyapunov theory. More specifically, under Assumption 1, it is sufficient to analyze the convergence behaviors of the mean and covariance matrix separately. Since the dynamics of the mean and covariance matrix can be expressed by linear systems, we can apply the Lyapunov theory in linear control to compute the complexity bound.

Main Contributions: By comparing the results for continuous dynamics and discrete algorithms, our analysis underscores the fact that the complexity of sampling is from time discretization; the continuous-time dynamics of mean and covariance of Langevin dynamics have exactly the same convergence rate. Our technique reveals that the time-discretization of the mean dynamics does not induce bias, but that of the covariance dynamics does. We conclude

that the size of the bias relies on dimension d, resulting in dimension-dependent complexity bounds for sampling, in contrast to dimension-free complexity bounds for optimization. More quantitatively, our analysis yields a complexity bound $\tilde{O}(\kappa\sqrt{d}/\epsilon)$ for LMC, better than all the existing results [9], [10], [11], [12], [15], albeit for Gaussian cases. For KLMC, we establish complexity bound $\tilde{O}(\kappa\sqrt{d}/\epsilon)$ by Theorem 4, the same as the best existing results [13], [14].

Notation: For any complex diagonal matrix Λ, $|\Lambda|$ and $\Re(\Lambda)$ stand for the magnitude and real part of each element, respectively, and Λ^H is the Hermitian transpose of Λ. We use the weighted norm induced by a matrix P, i.e., for any vector x, $\|x\|_P^2 = x^T P x$ and for any matrix M, $\|M\|_P^2 = \|PM\|_F^2$. Denote the eigenvalues of a matrix M by $\sigma(M)$. Θ is Landau's big theta meaning asymptotical equality. We also use the standard convention of diagonal matrices: the blank space stands for zero elements.

II. Sampling via overdamped Langevin

In this section, we analyze the convergence behavior of overdamped Langevin dynamics and the LMC algorithm in the special case with Gaussian target distribution.

A. Overdamped Langevin dynamics

For a given target distribution $\nu \propto \exp(-f)$ on $\mathbb{R}^d$, the associated Langevin dynamics [9] reads

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

where W_t is a standard Wiener process. Note this is a stochastic process and X_t is a random vector for any t. Under mild assumptions, the distribution of X_t converges to its stationary distribution that coincides with the target distribution ν. Thus, one can in principle simulate the Langevin dynamics for a sufficiently long time to draw samples from ν.

When $\nu = \mathcal{N}(m_g, \Sigma_g)$, the Langevin dynamics becomes a linear stochastic differential equation (SDE) [16]. More specifically, invoking the quadratic expression of $f(x) = \frac{1}{2}(x - m_g)^T\Sigma_g^{-1}(x - m_g)$, it corresponds to the multi-dimensional version of the Ornstein–Uhlenbeck process [17]

$$\mathrm{d}X_t = -\Sigma_g^{-1}(X_t - m_g)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t. \tag{1}$$

By solving the linear stochastic differential equation (1), one has X_t also follows a Gaussian distribution as long as X_0 follows a Gaussian distribution. Thus, the evolution of the random vector X_t can be fully captured by that of its mean and covariance. Denote the mean of X_t as m_t and the covariance as Σ_t, then following standard stochastic calculus we obtain

$$\begin{aligned} \dot{m}_t &= -\Sigma_g^{-1}(m_t - m_g) && \text{(2a)} \\ \dot{\Sigma}_t &= -\Sigma_g^{-1}\Sigma_t - \Sigma_t\Sigma_g^{-1} + 2\mathbb{I}. && \text{(2b)} \end{aligned}$$

Specifically, (2a) follows by taking the expectation of (1). To get (2b), we first apply stochastic calculus to get $\mathrm{d}(X_t X_t^T)$ and then take expectation [16].

Both (2a) and (2b) are linear systems. Clearly, the equilibrium point of (2) is (m_g, Σ_g). Applying linear system theory, we can establish linear convergence of (m_t, Σ_t) to the equilibrium point, as follows.

Theorem 1 (Convergence rate of overdamped Langevin dynamics for Gaussian distributions) *Under Assumption 1, the mean m_t and covariance Σ_t of X_t evolving according to the Langevin dynamics* (1) *satisfy*

$$\begin{aligned} \|m_t - m_g\|_2^2 &\leq \exp(-2\alpha t)\|m_0 - m_g\|_2^2 && \text{(3a)} \\ \|\Sigma_t - \Sigma_g\|_F &\leq \exp(-2\alpha t)\|\Sigma_0 - \Sigma_g\|_F. && \text{(3b)} \end{aligned}$$

Proof: Denote $m_t - m_g$ and $\Sigma_t - \Sigma_g$ by δm_t and $\delta\Sigma_t$ respectively, then the linear system (2) is equivalent to

$$\begin{aligned} \dot{\delta m_t} &= -\Sigma_g^{-1}\delta m_t && \text{(4a)} \\ \dot{\delta\Sigma_t} &= -\Sigma_g^{-1}\delta\Sigma_t - \delta\Sigma_t\Sigma_g^{-1}. && \text{(4b)} \end{aligned}$$

Under Assumption 1, specifically $-\Sigma_g^{-1} \preceq -\alpha I$, we obtain

$$\|\delta m_t\|_2^2 \leq \exp(-2\alpha t)\|\delta m_0\|_2^2.$$

Vectorizing (4b) with Kronecker products yields

$$\begin{aligned} \mathrm{vec}(\dot{\delta\Sigma_t}) &= (\mathbb{I}\otimes -\Sigma_g^{-1})\mathrm{vec}(\delta\Sigma_t) + (-\Sigma_g^{-1}\otimes\mathbb{I})\mathrm{vec}(\delta\Sigma_t) \\ &= (-\Sigma_g^{-1}\oplus -\Sigma_g^{-1})\mathrm{vec}(\delta\Sigma_t). \end{aligned}$$

It is a standard result that the largest eigenvalues of $(-\Sigma_g^{-1}\oplus -\Sigma_g^{-1})$ is $-2\alpha < 0$. Since for any matrix A, $\|\mathrm{vec}(A)\|_2$ is the Frobenius norm of A, we arrive at

$$\|\delta\Sigma_t\|_F \leq \exp(-2\alpha t)\|\delta\Sigma_0\|_F. \tag{5}$$

■

B. Overdamped Langevin Monte Carlo

One popular way to discretize overdamped Langevin dynamics (1), thus turning it into a practical sampling algorithm, is the (overdamped) Langevin Monte Carlo (a.k.a. Unadjusted Langevin Algorithm). For a target distribution $\nu \propto \exp(-f)$, it runs as

$$X_{k+1} = X_k - \eta\nabla f(X_k) + \sqrt{2\eta}\xi_k,\ \xi_k \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbb{I})$$

where $\eta > 0$ is the stepsize. In the Gaussian case where $\nu = \mathcal{N}(m_g, \Sigma_g)$, it becomes

$$X_{k+1} = X_k - \eta\Sigma_g^{-1}(X_k - m_g) + \sqrt{2\eta}\xi_k,\ \xi_k \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbb{I}). \tag{6}$$

Again, this linear dynamics is fully captured by the mean m_k and covariance Σ_k of X_k, which evolves according to

$$\begin{aligned} m_{k+1} - m_g &= (\mathbb{I} - \eta\Sigma_g^{-1})(m_k - m_g) && \text{(7a)} \\ \Sigma_{k+1} &= (\mathbb{I} - \eta\Sigma_g^{-1})\Sigma_k(\mathbb{I} - \eta\Sigma_g^{-1}) + 2\eta\mathbb{I}. && \text{(7b)} \end{aligned}$$

The convergence of m_k, Σ_k is characterized by the following result.

Theorem 2 (Convergence rate of Langevin Monte Carlo for Gaussian distributions) *Under Assumption 1, the mean m_k and covariance Σ_k of X_k evolving according to the Langevin Monte Carlo* (6) *satisfy*

$$\begin{aligned} &\|m_k - m_g\|_2^2 \leq (1 - \eta\alpha)^{2k}\|m_0 - m_g\|_2^2 \\ &\|\Sigma_k - \Sigma_g\|_F \leq (1 - \eta\alpha)^{2k}\|\Sigma_0 - \Sigma_s\|_F + \frac{\sqrt{d}\eta}{2 - \eta\beta}. \end{aligned}$$

Moreover, if $\eta = \Theta(\epsilon/\beta\sqrt{d})$, then for any $\epsilon \in [0, \sqrt{d}]$, we have $\alpha\|\Sigma_k - \Sigma_g\|_F \le \epsilon$ after

$$N = \widetilde{O}(\frac{\kappa\sqrt{d}}{\epsilon})$$

iterations, where $\kappa = \beta/\alpha$ is the condition number.

Proof: Under Assumption 1, by linear dynamics theory, the linear system (7) is globally asymptotically stable when $\eta < 1/\beta$. Moreover, since the stationary point of (7a) is $m_k = m_g$, we have

$$\|m_k - m_g\|_2^2 \le (1-\eta\alpha)^{2k}\|m_0 - m_g\|_2^2.$$

In contrast, the stationary point of (7b) is not Σ_g due to the discretization error. We denote the true equilibrium point of (7b) by Σ_s. By definition of equilibrium point, it satisfies

$$\Sigma_s = (\mathbb{I} - \eta\Sigma_g^{-1})\Sigma_s(\mathbb{I} - \eta\Sigma_g^{-1}) + 2\eta\mathbb{I}. \tag{8}$$

Combining (7b) and (8), we obtain

$$\Sigma_{k+1} - \Sigma_s = (\mathbb{I} - \eta\Sigma_g^{-1})(\Sigma_k - \Sigma_s)(\mathbb{I} - \eta\Sigma_g^{-1}).$$

It follows that

$$\mathrm{vec}(\Sigma_{k+1} - \Sigma_s) = \left((\mathbb{I} - \eta\Sigma_g^{-1}) \otimes (\mathbb{I} - \eta\Sigma_g^{-1})\right)\mathrm{vec}(\Sigma_k - \Sigma_s).$$

The operator norm of $\left((\mathbb{I} - \eta\Sigma_g^{-1}) \otimes (\mathbb{I} - \eta\Sigma_g^{-1})\right)$ is $(1-\eta\alpha)^2$, implying

$$\|\Sigma_k - \Sigma_s\|_F \le (1-\eta\alpha)^{2k}\|\Sigma_0 - \Sigma_s\|_F. \tag{9}$$

The above inequality characterizes the convergence rate of Σ_k to the stationary state of (7b). We next bound the deviation of Σ_s from Σ_g, i.e., the Frobenius norm of $\Sigma_k - \Sigma_g$. Solving the Lyapunov equation (8) yields that the explicit expression of Σ_s is

$$\Sigma_s = 2\eta(\mathbb{I} - (\mathbb{I} - \eta\Sigma_g^{-1})^2)^{-1}.$$

As Σ_g^{-1} is positive definite, we write its eigendecomposition as $\Sigma_g^{-1} = U\Lambda U^T$ with $UU^T = \mathbb{I}$ and Λ being diagonal. It follows that

$$\begin{aligned}\|\Sigma_s - \Sigma_g\|_F^2 &= \mathrm{Tr}((\Sigma_s - \Sigma_g)^2) \\ &= \frac{\eta^2}{4}\,\mathrm{Tr}\left((\mathbb{I} - \eta\Lambda/2)^{-2}\right) \\ &\le \frac{d\eta^2}{(2-\eta\beta)^2}\end{aligned} \tag{10}$$

where the second equality comes from the fact that Σ_s and Σ_g share the same eigenspace, and the last inequality is from $\alpha\mathbb{I} \preceq \Lambda \preceq \beta\mathbb{I}$ following the assumption that $\alpha\mathbb{I} \preceq \Sigma_g^{-1} \preceq \beta\mathbb{I}$. By (9), (10) and the assumption $\eta < 1/\beta$, we conclude

$$\begin{aligned}\|\Sigma_k - \Sigma_g\|_F &\le \|\Sigma_k - \Sigma_s\|_F + \|\Sigma_s - \Sigma_g\|_F \\ &\le (1-\eta\alpha)^{2k}\|\Sigma_0 - \Sigma_s\|_F + \frac{\sqrt{d}\eta}{2-\eta\beta}.\end{aligned}$$

Finally, by choosing $\eta = \Theta(\epsilon/\beta\sqrt{d})$, the discretization error $\frac{\sqrt{d}\eta}{2-\eta\beta}$ is upper bounded by $\epsilon/2$. Then after $\widetilde{O}(\frac{\kappa\sqrt{d}}{\epsilon})$ iterations, with the same η, the other term $(1-\eta\alpha)^{2k}\|\Sigma_0 - \Sigma_s\|_F$ is also upper bounded by $\epsilon/2$. ∎

Remark 1: We use the metric $\alpha\|\Sigma_k - \Sigma_g\|_F$ instead of $\|\Sigma_k - \Sigma_g\|_F$ because the former is invariant with respect to rescaling of the coordinate.

Compared with the bound $\widetilde{O}(d\epsilon^{-2})$ obtained in [9], [10], [11], [12] for LMC with strongly-convex and smooth potentials, the mixing time complexity shows an order of $\widetilde{O}(\sqrt{d}\epsilon^{-1})$ improvement. In [15], the complexity is $\widetilde{O}(\kappa^2\sqrt{d}\epsilon^{-1})$. Compared with these two bounds, our bound for Gaussian distributions are tighter, which may indicate the theoretically tightest mixing time complexity of sampling for strongly-convex and smooth potentials is not achieved yet.

The complexity bound with respect to the Wasserstein-2 distance W_2, a popular metric used to measure the convergence of MCMC, in the Gaussian setting can be analyzed as follows.

Proposition 1: Under Assumption 1, consider X_k evolving according to the Langevin Monte Carlo (6). If $\eta = \Theta(\epsilon/\beta\sqrt{d})$, then for any $\epsilon \in [0, \sqrt{d}]$, we have $\alpha W_2(\mathcal{N}(m_k, \Sigma_k), \mathcal{N}(m_g, \Sigma_g)) \le \epsilon$ after

$$N = \widetilde{O}(\frac{\kappa\sqrt{d}}{\epsilon})$$

iterations.

Proof: By equation (3) and (11) in [18],

$$\|\Sigma_k^{1/2} - \Sigma_g^{1/2}\|_F^2 \ge \mathrm{Tr}(\Sigma_g + \Sigma_k - 2(\Sigma_g^{1/2}\Sigma_k\Sigma_g^{1/2})^{1/2}).$$

Since [19]

$$\begin{aligned}&W_2^2(\mathcal{N}(m_k, \Sigma_k), \mathcal{N}(m_g, \Sigma_g)) \\ &= \|m_k - m_g\|_2^2 + \mathrm{Tr}(\Sigma_g + \Sigma_k - 2(\Sigma_g^{1/2}\Sigma_k\Sigma_g^{1/2})^{1/2}),\end{aligned}$$

we have

$$\|\Sigma_k^{1/2} - \Sigma_g^{1/2}\|_F^2 + \|m_k - m_g\|_2^2 \ge W_2^2(\mathcal{N}(m_k, \Sigma_k), \mathcal{N}(m_g, \Sigma_g)).$$

To bound $\|\Sigma_k^{1/2} - \Sigma_g^{1/2}\|_F^2$, we just need to bound $\|\Sigma_s^{1/2} - \Sigma_g^{1/2}\|_F^2$. This is achieved following a similar calculation as in (10), which is $O(\frac{d\eta^2\beta}{2-\eta\beta})$. The same $\eta = \Theta(\epsilon/\beta\sqrt{d})$ can then give the same complexity bound $\widetilde{O}(\frac{\kappa\sqrt{d}}{\epsilon})$ with respect to W_2. ∎

III. Sampling via underdamped Langevin

In this section, we extend the analysis in the previous section to underdamped Langevin dynamics as well as the KLMC algorithm based on it. For the ease of presentation, we do not give general explicit expressions of the convergence rate in Theorem 3 and 4. Instead, in Section III-C we consider one specific case that is widely used in existing works.

A. Underdamped Langevin dynamics

The underdamped Langevin dynamics for target distribution $\nu \propto \exp(-f)$ is

$$\mathrm{d}X_t = V_t\mathrm{d}t \tag{11a}$$

$$\mathrm{d}V_t = -\gamma V_t\mathrm{d}t - u\nabla f(X_t)\mathrm{d}t + \sqrt{2\gamma u}\mathrm{d}W_t \tag{11b}$$

where $X_t, V_t \in \mathbb{R}^d$. The invariant distribution of (11) in the phase space is $\exp(-f(x) - \|v\|^2/2u)$ [20]. Here γ and u are

both positive parameters. We further adopt the following mild assumption to solely simplify the proof.

Assumption 2: $\det(\gamma^2\mathbb{I} - 4u\Sigma_g^{-1}) \neq 0$.

Under Assumption 1, let Z_t be $\begin{pmatrix} X_t - m_g \\ V_t \end{pmatrix}$, then (11) reduces to

$$\mathrm{d}Z_t = \begin{pmatrix} 0 & \mathbb{I} \\ -u\Sigma_g^{-1} & -\gamma\mathbb{I} \end{pmatrix} Z_t \mathrm{d}t + \begin{pmatrix} 0 \\ \sqrt{2\gamma u}\mathbb{I} \end{pmatrix} \mathrm{d}W_t.$$

Again, Z_t is Gaussian as long as Z_0 is Gaussian. Denote $\begin{pmatrix} 0 & \mathbb{I} \\ -u\Sigma_g^{-1} & -\gamma\mathbb{I} \end{pmatrix}$ and $\begin{pmatrix} 0 \\ \sqrt{2\gamma u}\mathbb{I} \end{pmatrix}$ by A and B, respectively. Let δm_t and Σ_t be the mean and covariance matrix of Z_t, respectively, and $\delta\Sigma_t$ be $\Sigma_t - \widetilde{\Sigma}_g$ where

$$\widetilde{\Sigma}_g = \begin{pmatrix} \Sigma_g & \\ & u\mathbb{I} \end{pmatrix} \tag{12}$$

is the covariance of the stationary distribution of Z_t, then

$$\delta\dot{m}_t = A\delta m_t \tag{13a}$$
$$\delta\dot{\Sigma}_t = A\delta\Sigma_t + \delta\Sigma_t A^T. \tag{13b}$$

This is a linear system and its convergence property is determined by the eigenvalues of A. Standard computation for the eigenvalues of blocked matrices gives

$$\sigma(A) = \{\frac{-\gamma \pm \sqrt{\gamma^2 - 4u\lambda}}{2}, \lambda \in \sigma(\Sigma_g^{-1})\}. \tag{14}$$

Clearly, the linear system (13) is stable as γ, u, λ are all positive. Moreover, (14) implies that each eigenvalue of Σ_g^{-1} corresponds to two eigenvalues of A. Thus, we define two diagonal matrices Λ_+ and Λ_- as follows. Let λ_i represent the i-th eigenvalue of Σ_g^{-1}, then the i-th elements of Λ_+ and Λ_- are $\frac{-\gamma+\sqrt{\gamma^2-4u\lambda_i}}{2}$ and $\frac{-\gamma-\sqrt{\gamma^2-4u\lambda_i}}{2}$, respectively. Furthermore, let the eigendecomposition of Σ_g^{-1} be $U\Lambda U^T$ with $UU^T = \mathbb{I}$. By Assumption 2, one can express the eigenvectors of A by Λ_+, Λ_- and U, which follows that the eigendecomposition of A is

$$A = V \begin{pmatrix} \Lambda_+ & \\ & \Lambda_- \end{pmatrix} V^{-1} \tag{15}$$

where

$$V = \begin{pmatrix} U & U \\ U\Lambda_+ & U\Lambda_- \end{pmatrix}. \tag{16}$$

It is worth mentioning that the decomposition (16) does not hold when $\gamma^2 - 4u\lambda = 0$ for some λ, namely, when Assumption 2 does not hold. In that case, we can consider the Jordan decomposition instead of the eigendecomposition.

By linear control theory, the convergence rate of (13) is characterized by

$$r_c = -\max \Re(\sigma(A)) \tag{17}$$

and the associated Lyapunov inequality is

$$A^T P + PA \preceq -2r_c P.$$

It turns out one such choice is $P = V^{-H}V^{-1}$. Indeed,

$$A^T P + PA = 2V^{-H} \begin{pmatrix} \Re(\Lambda_+) & \\ & \Re(\Lambda_-) \end{pmatrix} V^{-1} \preceq -2r_c P. \tag{18}$$

By Lyapunov theory, it follows that

$$\|\delta m_t\|_P^2 \leq \exp(-2r_c t)\|\delta m_0\|_P^2.$$

Similarly, for the dynamics of the covariance matrix (13b), we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\|\delta\Sigma_t\|_P^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Tr}(\delta\Sigma_t P \delta\Sigma_t P) \\ &= \operatorname{Tr}(\delta\dot{\Sigma}_t P\delta\Sigma_t P + \delta\Sigma_t P\delta\dot{\Sigma}_t P) \\ &= 2\operatorname{Tr}((A^T P + PA)\delta\Sigma_t P\delta\Sigma_t) \\ &\leq -4r_c \operatorname{Tr}(P\delta\Sigma_t P\delta\Sigma_t)) \\ &= -4r_c\|\delta\Sigma_t\|_P^2, \end{aligned}$$

which follows that

$$\|\delta\Sigma_t\|_P \leq \exp(-2r_c t)\|\delta\Sigma_0\|_P.$$

Thus, we have established the following convergence results.

Theorem 3 (Convergence rate of underdamped Langevin dynamics for Gaussian distributions) *Under Assumption 1 and 2, for (X_t, V_t) evolving according to the Langevin dynamics* (11)*, one has*

$$\|m_t - m_g\|_P^2 \leq \exp(-2r_c t)\|m_0 - m_g\|_P^2 \tag{19a}$$
$$\|\Sigma_t - \widetilde{\Sigma}_g\|_P \leq \exp(-2r_c t)\|\Sigma_0 - \widetilde{\Sigma}_g\|_P. \tag{19b}$$

where m_t and Σ_t are the mean and covariance matrix of (X_t, V_t), respectively, and $\widetilde{\Sigma}_g$ is as in (12)*. Here $\|\cdot\|_P$ is the weighted norm induced by $P = V^{-H}V^{-1}$ with V given by* (16)*, and the convergence rate $r_c > 0$ is defined in* (17)*.*

B. Kinetic Langevin Monte Carlo

In [5], the implementation of underdamped Langevin dynamics is obtained by one discretization of underdamped Langevin dynamics (11) with step size η, which uses

$$\begin{pmatrix} X_{k+1} - m_g \\ V_{k+1} \end{pmatrix} = A_d \begin{pmatrix} X_k - m_g \\ V_k \end{pmatrix} + \xi_k,\ \xi_k \overset{\text{iid}}{\sim} \mathcal{N}(0, Q) \tag{20}$$

where

$$A_d = \begin{pmatrix} \mathbb{I} - \frac{u}{\gamma}\left(\eta - \frac{1}{\gamma}(1 - e^{-\gamma\eta})\right)\Sigma_g^{-1} & \frac{1}{\gamma}(1 - e^{-\gamma\eta})\mathbb{I} \\ -\frac{u}{\gamma}(1 - e^{-\gamma\eta})\Sigma_g^{-1} & e^{-\gamma\eta}\mathbb{I} \end{pmatrix}.$$

Here Q is a 2-by-2 block matrix with

$$\begin{aligned} Q_{11} &= \frac{2u}{\gamma}\left(\eta - \frac{3}{2\gamma} + \frac{2}{\gamma}\exp(-\gamma\eta) - \frac{1}{2\gamma}\exp(-2\gamma\eta)\right)\mathbb{I} \\ Q_{12} &= \frac{u}{\gamma}\left(1 + \exp(-2\gamma\eta) - 2\exp(-\gamma\eta)\right)\mathbb{I} \\ Q_{21} &= \frac{u}{\gamma}\left(1 + \exp(-2\gamma\eta) - 2\exp(-\gamma\eta)\right)\mathbb{I} \\ Q_{22} &= u(1 - \exp(-2\gamma\eta))\mathbb{I} \end{aligned}$$

where each block is a d-by-d matrix. To see more clearly the effects of time discretization, in what follows, we consider the first-order approximation of A_d and Q,

$$\hat{A}_d = \begin{pmatrix} \mathbb{I} & \eta\mathbb{I} \\ -u\eta\Sigma_g^{-1} & \mathbb{I} - \gamma\eta \end{pmatrix} \quad (21a)$$

$$\hat{Q} = \begin{pmatrix} 0 & 0 \\ 0 & 2u\gamma\eta\mathbb{I} \end{pmatrix}. \quad (21b)$$

Hence, (20) reduces to

$$\begin{pmatrix} \hat{X}_{k+1} - m_g \\ \hat{V}_{k+1} \end{pmatrix} = \hat{A}_d \begin{pmatrix} \hat{X}_k - m_g \\ \hat{V}_k \end{pmatrix} + \hat{\xi}_k \quad (22)$$

with $\hat{\xi}_k \sim \mathcal{N}(0, \hat{Q})$. It is worth mentioning that (22) is also one way to discretize (11). The following result characterizes the convergence of the mean m_k and covariance Σ_k of (22).

Theorem 4 (Convergence rate of kinetic Langevin Monte Carlo for Gaussian distributions) *Under Assumption 1 and 2, for $(\hat{X}_k, \hat{V}_k)$ evolving according to the kinetic Langevin Monte Carlo* (22)*, one has*

$$\|m_k - m_g\|_P^2 \le r_d^{2k} \|m_0 - m_g\|_P^2$$
$$\|\Sigma_k - \widetilde{\Sigma}_g\|_P \le r_d^{2k} \|\Sigma_0 - \Sigma_s\|_P + \|\Sigma_s - \widetilde{\Sigma}_g\|_P$$

where m_k and Σ_k are the mean and covariance matrix of $(\hat{X}_k, \hat{V}_k)$, respectively, and $\widetilde{\Sigma}_g$ is as in (12)*. Here $\|\cdot\|_P$ is the weighted norm induced by $P = V^{-H}V^{-1}$ with V given by* (16)*, and the convergence rate $r_d < 1$ is as in* (25)*.*

Proof: Denote the mean and covariance matrix of $\begin{pmatrix} \hat{X}_k - m_g \\ \hat{V}_k \end{pmatrix}$ by δm_k and Σ_k, respectively, then

$$\delta m_{k+1} = \hat{A}_d \delta m_k \quad (23a)$$

$$\Sigma_{k+1} = \hat{A}_d \Sigma_k \hat{A}_d^T + \hat{Q}. \quad (23b)$$

We next compute the eigendecomposition of $\hat{A}_d$. Notice that

$$\hat{A}_d = \mathbb{I} + \eta \begin{pmatrix} 0 & \mathbb{I} \\ -u\Sigma_g^{-1} & -\gamma\mathbb{I} \end{pmatrix}.$$

Hence, by the decomposition in (15) and (16), we have

$$\hat{A}_d = V \begin{pmatrix} 1 + \eta\Lambda_+ & \\ & 1 + \eta\Lambda_- \end{pmatrix} V^{-1} \quad (24)$$

where V, Λ_+ and Λ_- coincide with the ones used in the decomposition (16). Adopting the same weighted norm induced by $P = V^{-H}V^{-1}$, one has the Lyapunov inequality for discrete systems as

$$\begin{aligned} \hat{A}_d^T P \hat{A}_d &= \hat{A}_d^H P \hat{A}_d \\ &= V^{-H} \begin{pmatrix} |1 + \eta\Lambda_+|^2 & \\ & |1 + \eta\Lambda_-|^2 \end{pmatrix} V^{-1} \quad (25) \\ &\preceq r_d^2 P. \end{aligned}$$

Here r_d^2 is the largest value of the elements in $|1 + \eta\Lambda_+|^2$ and $|1 + \eta\Lambda_-|^2$. It follows that

$$\|\delta m_k\|_P^2 \le r_d^{2k} \|\delta m_0\|_P^2. \quad (26)$$

Since the stationary point of (23) is not $\widetilde{\Sigma}_g$, we denote the true one as Σ_s which solves

$$\Sigma_s = \hat{A}\Sigma_s\hat{A}^T + \hat{Q}. \quad (27)$$

Then Σ_k converges to Σ_s based on the following identity

$$\Sigma_{k+1} - \Sigma_s = \hat{A}_d(\Sigma_k - \Sigma_s)\hat{A}_d^T.$$

By (25), it implies that

$$\|\Sigma_{k+1} - \Sigma_s\|_P^2 = \|\hat{A}_d^H P \hat{A}_d(\Sigma_k - \Sigma_s)\|_F^2 \le r_d^4 \|\Sigma_k - \Sigma_s\|_P^2.$$

Hence,

$$\|\Sigma_k - \Sigma_s\|_P \le r_d^{2k} \|\Sigma_0 - \Sigma_s\|_P.$$

The conclusion follows the triangle inequality of the weighted norm $\|\cdot\|_P$. ∎

The non-asymptotic bound of $\Sigma_k - \widetilde{\Sigma}_g$ in terms of standard Frobenius norm can then be achieved by noticing that

$$\begin{aligned} \|\Sigma_k - \widetilde{\Sigma}_g\|_F &\le \|P^{-1}\|_O \|\Sigma_k - \widetilde{\Sigma}_g\|_P \\ &\le \|P^{-1}\|_O (r_d^{2k} \|\Sigma_0 - \Sigma_s\|_P + \|\Sigma_s - \widetilde{\Sigma}_g\|_P) \\ &\le C(P)(r_d^{2k} \|\Sigma_0 - \Sigma_s\|_F + \|\Sigma_s - \widetilde{\Sigma}_g\|_F) \end{aligned} \quad (28)$$

where $C(P) := \|P\|_O \|P^{-1}\|_O$ is the condition number of P induced by the operator norm.

We next consider the bound of $\|\Sigma_s - \widetilde{\Sigma}_g\|_P$. With (27), one has

$$\Sigma_s - \widetilde{\Sigma}_g = \hat{A}_d(\Sigma_s - \widetilde{\Sigma}_g)\hat{A}_d^T + K, \quad (29)$$

where

$$K = \hat{Q} - \widetilde{\Sigma}_g + \hat{A}_d \widetilde{\Sigma}_g \hat{A}_d^T = u\eta^2 \begin{pmatrix} \mathbb{I} & -\gamma\mathbb{I} \\ -\gamma\mathbb{I} & u\Sigma_g^{-1} + \gamma^2\mathbb{I} \end{pmatrix}. \quad (30)$$

Let $D := V^{-1}(\Sigma_s - \widetilde{\Sigma}_g)V^{-H}$. Plugging the decomposition of $\hat{A}_d$ into (29) yields that

$$D = \begin{pmatrix} 1 + \eta\Lambda_+ & \\ & 1 + \eta\Lambda_- \end{pmatrix} D \begin{pmatrix} 1 + \eta\Lambda_+ & \\ & 1 + \eta\Lambda_- \end{pmatrix}^H + V^{-1}KV^{-H}.$$

Plugging into the expressions of V and K in (16) and (30) yields that $V^{-1}KV^{-H}$ is a 2-by-2 blocked matrix, where each block $\in \mathbb{R}^{d\times d}$ is diagonal, namely,

$$V^{-1}KV^{-H} = \begin{pmatrix} E_{11} & E_{12} \\ E_{12}^H & E_{22} \end{pmatrix}, \quad (31)$$

with

$$\begin{aligned} E_{11} &= \phi\left(|\Lambda_-|^2 + u\Lambda + \gamma^2 + 2\gamma\Re(\Lambda_-)\right) \\ E_{22} &= \phi\left(|\Lambda_+|^2 + u\Lambda + \gamma^2 + 2\gamma\Re(\Lambda_+)\right) \\ E_{12} &= -\phi\left(u\Lambda + \gamma^2 + r(\Lambda_+^H + \Lambda_-) + \Lambda_+^H\Lambda_-\right), \end{aligned}$$

where $\phi = u\eta^2|\Lambda_+ - \Lambda_-|^{-2}$. It follows that

$$D = \begin{pmatrix} \frac{E_{11}}{1-|1+\eta\Lambda_+|^2} & \frac{E_{12}}{1-(1+\eta\Lambda_+)(1+\eta\Lambda_-)^H} \\ \frac{E_{12}^H}{1-(1+\eta\Lambda_+)^H(1+\eta\Lambda_-)} & \frac{E_{22}}{1-|1+\eta\Lambda_-|^2} \end{pmatrix}.$$

Lastly, with the definition of D and the decomposition of $\hat{A}_d$ in (24).

$$\begin{aligned}\|\Sigma_s - \widetilde{\Sigma}_g\|_P^2 &= \|P(\Sigma_s - \widetilde{\Sigma}_g)\|_F^2 = \|D\|_F^2 \\ &= \left\|\frac{E_{11}}{1-|1+\eta\Lambda_+|^2}\right\|_F^2 + \left\|\frac{E_{22}}{1-|1+\eta\Lambda_-|^2}\right\|_F^2 \\ &\quad + 2\left\|\frac{E_{12}}{1-(1+\eta\Lambda_+)(1+\eta\Lambda_-)^H}\right\|_F^2\end{aligned} \tag{32}$$

where the last equation is from the definition of Λ_+ and Λ_-. Further analysis on the bound of $\|\Sigma_s - \widetilde{\Sigma}_g\|_P^2$ depends on the sign of $\gamma^2 - 4u\lambda$ for each $\lambda \in \sigma(\Sigma_g^{-1})$.

C. Example

In this subsection, we present one implementation of KLMC. We assume $\gamma = 2, u = 1/(2\beta)$. This set of parameters is obtained from [5]. In this scenario, the sign of each $\gamma^2 - 4u\lambda$ is nonnegative. Hence, by definition, the convergence rate of the continuous dynamics r_c in Theorem 3 is $1 - \sqrt{1 - \frac{\alpha}{2\beta}} \approx \frac{1}{4\kappa}$. Moreover, assuming $\eta < \frac{1}{1+\sqrt{1-\frac{1}{2\kappa}}}$, the convergence rate of the discrete implementation r_d in Theorem 4 is $1 + \eta(-1 + \sqrt{1 - \frac{1}{2\kappa}}) \approx 1 - \frac{\eta}{4\kappa}$. The bound of each block of D in (32) can be computed as follows.

$$\left\|\frac{E_{11}}{1-|1+\eta\Lambda_+|^2}\right\|_F = \left\|\frac{\eta^2}{2\beta}\frac{1}{4-\frac{2}{\beta}\Lambda}\frac{\frac{\Lambda}{2\beta}+(\Lambda_+)^2}{(\eta\Lambda_+)(2+\eta\Lambda_+)}\right\|_F. \tag{33}$$

Notice that each element in $\frac{1}{|(4-\frac{2}{\beta}\Lambda)(2+\eta\Lambda_+)|}$ is bounded. Hence,

$$\left\|\frac{E_{11}}{1-|1+\eta\Lambda_+|^2}\right\|_F = O(\frac{\sqrt{d}\eta}{\alpha}).$$

With the same procedure, one can show

$$\left\|\frac{E_{22}}{1-|1+\eta\Lambda_-|^2}\right\|_P = O(\frac{\sqrt{d}\eta}{\alpha})$$

and

$$\left\|\frac{E_{12}}{1-(1+\eta\Lambda_+)(1+\eta\Lambda_-)^H}\right\|_F = O(\frac{\sqrt{d}\eta}{\alpha}).$$

It follows that for (32), we have

$$\|\Sigma_s - \widetilde{\Sigma}_g\|_P = O(\frac{\sqrt{d}\eta}{\alpha}). \tag{34}$$

Hence, in particular, if we take $\eta = \Theta(\frac{\epsilon}{\sqrt{d}})$, then for any $\epsilon \in [0,1]$, we obtain $\alpha\|\Sigma_k - \widetilde{\Sigma}_g\|_P \leq \epsilon$ after

$$N = \widetilde{O}(\frac{\kappa\sqrt{d}}{\epsilon}) \tag{35}$$

iterations. Moreover, one can show

$$P^{-1} = \begin{pmatrix} 2\mathbb{I} & -2\mathbb{I} \\ -2\mathbb{I} & 4\mathbb{I} - \Lambda/\beta \end{pmatrix}.$$

Hence, the condition number of P is a bounded constant that is independent of κ and d, and by (28), we have the same mixed time complexity for the standard Frobenius norm. Our result match the best existing bound for KLMC [14].

IV. Conclusion

In this work, we present a linear control perspective to certain MCMC sampling algorithms for Gaussian target distributions. We focus on two classical algorithms: LMC and KLMC, one based on the overdamped Langevin dynamics and one based on the underdamped Langevin dynamics. Our results are better than the existing bounds in the Gaussian setting. More importantly, our analysis may shed light on complexity analysis for Langevin-based algorithms for general distributions. In the future, we plan to further investigate the KLMC algorithms with different choices of parameters γ, u as well as other algorithms such as HMC.

References

[1] A. Golightly and D. J. Wilkinson, "Bayesian sequential inference for nonlinear multivariate diffusions," *Statistics and Computing*, vol. 16, pp. 323–338, 2006.

[2] A. M. Stuart, "Inverse problems: a Bayesian perspective," *Acta numerica*, vol. 19, pp. 451–559, 2010.

[3] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[4] S. Ghosal and A. Van der Vaart, *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017, vol. 44.

[5] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, "Underdamped Langevin MCMC: A non-asymptotic analysis," in *Conference on learning theory*. PMLR, 2018, pp. 300–323.

[6] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh, "Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5504–5548, 2016.

[7] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 681–688.

[8] S. Chewi, *Log-Concave Sampling*, 2023.

[9] A. S. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 651–676, 2017.

[10] A. S. Dalalyan, "Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent," *Conference on Learning Theory*, pp. 678–689, 2017.

[11] X. Cheng and P. L. Bartlett, "Convergence of langevin mcmc in kl-divergence," *PMLR 83*, no. 83, pp. 186–211, 2018.

[12] A. Durmus, S. Majewski, and B. Miasojedow, "Analysis of langevin monte carlo via convex optimization." *Journal of Machine Learning Research*, vol. 20, pp. 73–1, 2019.

[13] A. S. Dalalyan and L. Riou-Durand, "On sampling from a log-concave density using kinetic langevin diffusions," *arXiv preprint arXiv:1807.09382*, 2018.

[14] M. Zhang, S. Chewi, M. B. Li, K. Balasubramanian, and M. A. Erdogdu, "Improved discretization analysis for underdamped langevin monte carlo," *arXiv preprint arXiv:2302.08049*, 2023.

[15] R. Li, H. Zha, and M. Tao, "Sqrt(d) dimension dependence of langevin monte carlo," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=5-2mX9_U5i

[16] A. Brissaud and U. Frisch, "Solving linear stochastic differential equations," *Journal of Mathematical Physics*, vol. 15, no. 5, pp. 524–534, 1974.

[17] C. W. Gardiner *et al.*, *Handbook of stochastic methods*. springer Berlin, 1985, vol. 3.

[18] L. Ning, X. Jiang, and T. Georgiou, "On the geometry of covariance matrices," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 787–790, 2013.

[19] C. R. Givens and R. M. Shortt, "A class of wasserstein metrics for probability distributions." *Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.

[20] G. A. Pavliotis, *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*. Springer, 2014, vol. 60.