Foundations of Data Science
©American Institute of Mathematical Sciences
Volume **4**, Number **1**, March **2022**
doi:10.3934/fods.2021034
pp. **37–70**

CONSTRAINED ENSEMBLE LANGEVIN MONTE CARLO

Zhiyan Ding and Qin Li*

Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53705 USA

(Communicated by Kody Law)

Abstract. The classical Langevin Monte Carlo method looks for samples from a target distribution by descending the samples along the gradient of the target distribution. The method enjoys a fast convergence rate. However, the numerical cost is sometimes high because each iteration requires the computation of a gradient. One approach to eliminate the gradient computation is to employ the concept of "ensemble." A large number of particles are evolved together so the neighboring particles provide gradient information to each other. In this article, we discuss two algorithms that integrate the ensemble feature into LMC, and the associated properties.

In particular, we find that if one directly surrogates the gradient using the ensemble approximation, the algorithm, termed Ensemble Langevin Monte Carlo, is unstable due to a high variance term. If the gradients are replaced by the ensemble approximations only in a constrained manner, to protect from the unstable points, the algorithm, termed Constrained Ensemble Langevin Monte Carlo, resembles the classical LMC up to an ensemble error but removes most of the gradient computation.

1. **Introduction.** Bayesian sampling is one of the core problems in Bayesian inference. It has a wide applications in data assimilation and inverse problems [34, 1] that arise in remote sensing and imaging [24], atmospheric science and earth science [17], petroleum engineering [28, 30] and epidemiology [25]. The goal is to find i.i.d. samples or approximately i.i.d. samples from a probability distribution that encodes the information of an unknown parameter. Throughout the paper we denote

$$p(x) \propto e^{-f(x)}, \quad x \in \mathbb{R}^d \tag{1}$$

the distribution function of the unknown parameter x, and we assume that $\nabla f(x)$ is L-smooth, meaning ∇f is Lipschitz continuous with L being its Lipschitz constant: $|\nabla f(y) - \nabla f(x)| < L|x - y|$.

There are many successful sampling algorithms [31, 2, 11, 32]. One class of classical sampling approach is the celebrated Markov chain Monte Carlo

2020 *Mathematics Subject Classification.* Primary: 62D05; Secondary: 82C31, 65C05.

Key words and phrases. Langevin Monte Carlo, ensemble methods, variance, gradient free.

Q.L. acknowledges support from Vilas Early Career award. The research of Z.D., and Q.L is supported in part by NSF via grant DMS-1750488, DMS-2023239 and Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin Madison with funding from the Wisconsin Alumni Research Foundation.

* Corresponding author: Qin Li.

(MCMC) [32, 35, 21, 12, 20]. This is a class of methods that sets the target distribution as the invariant measure of the Markov transition kernel, so after many rounds of iteration, the sample can be viewed to be drawn from the invariant measure. Since there are many ways to design the Markov chain, there are many subcategories of MCMC methods. Among them, the Langevin Monte Carlo (LMC) stands out for its simplicity, and fast convergence rate.

The key idea of LMC is to design a stochastic differential equation, whose long time equilibrium coincides with the target distribution. The samples are then drawn by following the trajectory of the (discretized) SDE. Typically the SDE converges exponentially fast, and thus the probability distribution of LMC samples, viewed as the discrete version of the SDE, also converges to the target distribution exponentially fast, up to a discretization error. The non-asymptotic convergence rate for these methods and their variations was recently made rigorous in [4, 5, 14, 13, 15, 39, 9] for log-concave probability distribution functions (or equivalently, for convex $f(x)$).

One key drawback of LMC is that it requires the frequent calculation of the gradients. For each sample, at each iteration, one needs to compute at least one full gradient. For a problem in $\mathbb{R}^d$, this is a calculation of d partial derivatives per sample per iteration, and in the case when $d \gg 1$, the cost is rather high. Therefore, in the most practical setting, one looks for substitutes of LMC that achieve "gradient-free" property so that the number of partial derivative computation is relaxed [10, 39].

Another sampling strategy that is completely parallel to the MCMC method is the ensemble type method. Unlike MCMC, or LMC in particular, ensemble methods evolve a large number of samples altogether, and these samples interplay with each other. A Fokker-Planck type PDE is formulated to drive an arbitrarily given distribution toward the target distribution, and the ensemble methods can be viewed as the particle methods applied to numerically evolve the PDE, with the ensemble distribution of the samples approximating the solution of the PDE. Two famous ensemble methods are Ensemble Kalman Inversion [23, 38] and Ensemble Kalman Sampling [18, 33, 19]. Earlier works are found in [34, 16, 29]. See also the numerical analysis and other follow up works in [7, 8, 22, 41].

The main drawbacks of ensemble methods are also obvious: The algorithms surrogate the statistical quantities with the ensemble version, introducing new computational cost and some ensemble error. Numerical analysis essentially needs to trace the propagation of such ensemble error, and is typically very involved. There is, however, one factor of ensemble methods that can potentially bring a great benefit: Since a lot of samples are evolved together on $\mathbb{R}^d$, it is easy to imagine that close neighbors of each sample can already approximately provide the gradient information. This may make gradient-free computation possible. Indeed, suppose one has a large number of particles, sampled from a certain probability distribution, in a small neighborhood of a sample x^*, then taking the average of the finite differences between these particles can give a rather good estimate to the gradient $\nabla f(x^*)$ to be used in LMC. This idea was already explored in EKS, where the authors inserted a variance term in the underlying SDE of LMC, and by combining the gradient term with the variance term, they formed a covariance that requires no gradient computation. However, such strategy holds true either if the forward map is linear, or the samples are all controllably close to each other. It is hard to justify either in real practice. Nevertheless, such exploration sets a stepping stone for designing gradient-free methods under the ensemble framework.

To summarize, the non-asymptotic convergence rate of LMC is thoroughly studied for a large class of nonlinear $f(x)$, while the validity of ensemble methods are generally lacking. On the other hand, LMC requires the computation of gradients, but the strategy of evolving a large number of samples as is done in the ensemble methods can potentially eliminate the gradient computation.

It is thus natural to ask if it is possible to bring together the two approaches for a new method that may inherit the advantages of both. To be specific, we look for an algorithm that requires as few gradient calculations as possible, while being able to sample (almost) exponentially fast in time. One attempt of breeding the two methods was taken in [41] where the authors added another layer of LMC into EnKF and designed the so-called Langevined EnKF. For linear $f(x)$ they can show the consistency, and in the nonlinear case, gradients are nevertheless needed. Therefore the advantage of removing the gradient computation using the concept of ensemble is lost. We look for the possibility of replacing gradients using the neighbor information whenever possible, and have a very different goal in this paper.

As such, we provide two sides of the answer:

- We first study the most straightforward approach. This is to sample a large number of particles altogether and in each iteration for the updates, we replace every gradient in LMC by the ensemble approximation. We term this method Ensemble LMC (EnLMC). This algorithm, despite being intuitive, will be shown to be unstable. Indeed, at the "outskirts" of $p(x)$, the accuracy of the updates very sensitively depend on the gradient, and the error induced by the surrogate can be significantly enlarged. This instability suggests that the replacement should not be enacted in these regions.
- We therefore propose an alternative, termed Constrained Ensemble LMC (CEnLMC). The constrained version of EnLMC enacts the ensemble approximation to the gradient only in the stable region, and for samples in the unstable region, we directly compute ∇f. We can show that this method provides samples that are close to LMC samples, and thus converges to the target distribution at the same rate (exponential, up to a controllable error term). Furthermore, we present how the parameters in the constraints determine the stability of the algorithm and the chance of enacting ensemble approximations.

We stress that the method CEnLMC is not completely "gradient-free" since it enacts ensemble approximation to replace the gradient computation only in the "stable" regions. However, the study conducted here presents an understanding on how to fuse the concepts of ensemble methods and LMC. While the new method provides a possibility to reduce the gradient computation, it also embraces the fast convergence that can be achieved by LMC for nonlinear f.

We also mention that there are many means for approximating the gradients. We cannot claim the optimality of the ensemble approximation used in this article. It is highly possible that one can replace the gradients in LMC using other methods that explore information from neighboring ensemble samples in a more efficient way (see Appendix B for a negative example). This line of research requires a more detailed study on multiple choices of ensemble approximation and is beyond the scope of the current paper. The current result is one of the pioneering attempts to

integrate ensemble features to LMC, and shed light on inventing algorithms that both converge fast and are gradient-free.

Lastly, we mention that in some communities (optimization for example), the algorithms that avoid or use gradients are termed zero-th order and first order methods. Similarly methods that use hessian information are of second order. The method we propose in this article can be viewed in between zero-th and first, since it eliminates a large portion of gradient calculations. Compared to zero-th order method, the advantages are obvious. All zero-th order methods converge slowly. One such example is the random walk Metropolis (RWM) that converges in $O(d^2)$ iterations [15]. On the contrary, LMC converges in $O(d)$ [5], or sometimes $O(d^{1/2})$ iterations when f is sufficiently smooth [26]. Our method matches the convergence rate as the classical LMC, but eliminates gradients, meaning it achieves the first order convergence with a zero-th order cost.

The paper is organized as follows. In Section 2, we review two main ingredients of our methods: the classical LMC, and the ensemble gradient approximation. In Section 3, we propose the two new methods and discuss the properties. More specifically, we will show the brute-force combination of LMC and the ensemble gradient approximation will lead to an unstable algorithm (EnLMC), but the constrained version (CEnLMC) recovers the target distribution with a high numerical saving. We show two numerical examples to demonstrate the saving and the accuracy in Section 4. The proof is given in Section 5.

2. **Two main ingredients.**

The main ingredients of our method are the classical Langevin Monte Carlo and an ensemble approximation to the gradient. We review them in this section.

2.1. **Langevin Monte Carlo (LMC).**

LMC is a very popular MCMC type sampling method. Under mild conditions, it provides fast convergence: after a few rounds of iterations, samples can be viewed approximately drawn from the target distribution.

The classical LMC starts with a sample, denoted as x^0, and updates the sample position according to:

$$x^{m+1} = x^m - \nabla f(x^m)h + \sqrt{2h}\xi_d^m\,, \tag{2}$$

where h is the time stepsize, and ξ_d^m is drawn i.i.d. from $\mathcal{N}(0, I_d)$, and I_d denotes the identity matrix of size $d \times d$. For a fixed small h, as $m \to \infty$, it is expected that q^m, the probability distribution of x^m, gets close to p, the target distribution.

To intuitively understand the convergence of this algorithm, we can view the updating formula as the Euler-Maruyama discretization for the following SDE:

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{3}$$

where B_t is a d-dimensional Brownian motion. The SDE characterizes the trajectory of X_t by the forcing term $\nabla f(X)\,\mathrm{d}t$ and the random walk $\mathrm{d}B_t$. While ∇f drives X_t to the minimum of f, the Brownian motion term introduces the fluctuation. Denote $q^0(x)$ the initial distribution from where X_0 is drawn, and $q(x,t)$ the probability density function of X_t, then it is a well-known result that $q(x,t)$ satisfies the following Fokker-Planck equation:

$$\partial_t q = \nabla \cdot (\nabla f q + \nabla q)\,, \quad \text{with} \quad q(x,0) = q^0\,. \tag{4}$$

It was shown in [27] that $q(x,t)$ converges to the target density function $p(x) \propto e^{-f}$ exponentially fast in time, meaning:

$$\lim_{t\to\infty} X_t \sim p(x)\,.$$

Considering that the updating formula for LMC (2) is merely a discretization of (3), then $x^m \approx X_{mh}$, and thus for large enough m, q^m, the distribution of x^m, should also be close to p. This is made rigorous recently in a number of papers [4, 5, 14, 13], most of which quantize the difference between q^m and p using the Wasserstein distance. To be more specific, it was shown in [5, 13] that for strongly-convex, gradient-Lipschitz f, to achieve ϵ accuracy in Wasserstein L_2 distance, the number of iteration needs to be $m \geq \widetilde{\mathcal{O}}(d/\epsilon^2)$. Here the notation $\widetilde{\mathcal{O}}$ hides a log factor.

We should note, however, that in each iteration of LMC, one local gradient needs to be computed, and this is equivalent to a calculation of d partial derivatives per iteration. This essentially means a cost of $\widetilde{\mathcal{O}}(d^2/\epsilon^2)$ is needed for one good sample. For a problem with high dimensionality $d \gg 1$, the cost is prohibitive. It would be desirable to combine this method with strategies that eliminate gradient computation for a gradient-free fast-converging sampling method.

2.2. **Ensemble mean gradient approximation.** Ensemble sampling methods have been gaining ground in recent years. The idea is to evolve a large number of samples altogether so that samples could provide information to each other. In particular, if two samples are close to each other, the finite difference roughly provides approximate gradient information. There are various choices of using neighbors to find approximated gradients. We look for a probability ensemble in this article. Suppose we look for an approximate gradient of f at $x^* \in \mathbb{R}^d$ using its neighbors x that are within η distance, and assume the neighbor x is drawn from an arbitrary probability density function $q(x)$, independent of x^*, then call

$$\tilde{d}_{\eta,q}(x^*) = \alpha_d \frac{\langle \nabla f(x^*), x - x^* \rangle}{|x-x^*|^2} \frac{\mathbf{1}_{|x-x^*|\leq\eta}}{q(x)}(x-x^*)\,, \tag{5}$$

where α_d is the normalization constant:

$$\alpha_d = \frac{d}{V} = \frac{d^2}{S_d \eta^d}\,, \quad \text{where} \quad V = \int_{|x-x^*|\leq\eta} 1\,\mathrm{d}x = \int_0^\eta r^{d-1} S_d dr = \frac{\eta^d S_d}{d}\,, \tag{6}$$

with S_d being the volume of unit d-sphere, we can formulate an ensemble gradient approximation:

$$\nabla f(x^*) = \mathbb{E}_q\left(\tilde{d}_{\eta,q}(x^*)\right)\,. \tag{7}$$

The formula (7) is valid merely because:

$$\begin{aligned}
\nabla f(x^*) &= \frac{d}{V}\int_{|x-x^*|\leq\eta} \frac{(x-x^*)\otimes(x-x^*)}{|x-x^*|^2}\,\mathrm{d}x\cdot\nabla f(x^*) \\
&= \alpha_d \int_{|x-x^*|\leq\eta} \frac{(x-x^*)\otimes(x-x^*)}{|x-x^*|^2}\,\mathrm{d}x\cdot\nabla f(x^*) \\
&= \alpha_d \int_{\mathbb{R}^d} \frac{\langle \nabla f(x^*), x-x^*\rangle}{|x-x^*|^2}\frac{\mathbf{1}_{|x-x^*|\leq\eta}}{q(x)}(x-x^*)q(x)\,\mathrm{d}x \\
&= \alpha_d \mathbb{E}_q\left(\frac{\langle \nabla f(x^*), x-x^*\rangle}{|x-x^*|^2}\frac{\mathbf{1}_{|x-x^*|\leq\eta}}{q(x)}(x-x^*)\right)\,.
\end{aligned}$$

One key idea of the ensemble gradient approximation is to realize that the term in $\tilde{d}_{\eta,q}$ can be approximated when η is small, namely:

$$\langle \nabla f(x^*), x - x^* \rangle \approx f(x) - f(x^*).$$

Replace the $\langle \nabla f, x - x^* \rangle$ term in $\tilde{d}_{\eta,q}$ by the finite difference term, and define

$$d_{\eta,q}(x^*) = \alpha_d \frac{f(x) - f(x^*)}{|x - x^*|^2} \frac{\mathbf{1}_{|x-x^*|\le\eta}}{q(x)} (x - x^*), \tag{8}$$

then the gradient $\nabla f(x^*)$ has a finite difference approximation, replacing (7):

$$\nabla f(x^*) \approx \mathbb{E}_q(d_{\eta,q}(x^*)) = \mathbb{E}_q\left(\alpha_d \frac{f(x) - f(x^*)}{|x - x^*|^2} \frac{\mathbf{1}_{|x-x^*|\le\eta}}{q(x)} (x - x^*) \right). \tag{9}$$

We can further justify the error in this approximation. Suppose ∇f is Lipschitz continuous, then

$$|f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle| \le L|x - x^*|^2 \le L\eta^2, \tag{10}$$

we have:

$$\begin{aligned}
&|\nabla f(x^*) - \mathbb{E}_q(d_{\eta,q}(x^*))| \\
\le &\mathbb{E}_q\left(|d_{\eta,q}(x^*) - \tilde{d}_{\eta,q}(x^*)| \right) \\
= &\mathbb{E}_q\left(\left| \alpha_d \frac{\langle \nabla f(x^*), x - x^* \rangle}{|x - x^*|^2} \frac{\mathbf{1}_{|x-x^*|\le\eta}}{q(x)} (x - x^*) - \alpha_d \frac{f(x) - f(x^*)}{|x - x^*|^2} \frac{\mathbf{1}_{|x-x^*|\le\eta}}{q(x)} (x - x^*) \right| \right) \\
= &\mathbb{E}_q\left(\left| \alpha_d \frac{|f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle|}{|x - x^*|^2} \frac{\mathbf{1}_{|x-x^*|\le\eta}}{q(x)} (x - x^*) \right| \right) \\
\le &\mathbb{E}_q\left(\left| \alpha_d L \frac{\mathbf{1}_{|x-x^*|\le\eta}}{q(x)} (x - x^*) \right| \right) \le L\eta d.
\end{aligned} \tag{11}$$

This formula suggests the approximation is first order in η, and the smallness of η needs to dominate the largeness in d.

Remark 1. We also stress that the derivation is valid only if the neighbors are distributed according to $q(x)$, a known distribution, and that this $q(x)$ needs to be independent of x^*.

Suppose in reality, we have N independent particles around x^*, denoted as $\{x_j\}_{j=1}^N$, sampled from $q_j(x)$ respectively, then the ensemble gradient approximation formula is further reduced to:

$$\nabla f(x^*) \approx \alpha_d \frac{1}{N} \sum_{j=1}^{N} \frac{f(x_j) - f(x^*)}{|x_j - x^*|^2} \frac{\mathbf{1}_{|x_j-x^*|\le\eta}}{q_j(x_j)} (x_j - x^*). \tag{12}$$

We note that $q_j(x)$ do not have to be the same.

3. Algorithms and properties.

We propose our new methods in this section. The strategy is to sample a large number of particles according to LMC (2), and replace the gradients in LMC using the ensemble gradient approximation (12). Then immediately the samples are no longer i.i.d. but they share the same marginal distribution.

We discuss in Section 3.1 the straightforward combination of the two. We term the method the Ensemble LMC (EnLMC). However, we will find the algorithm is rather unstable due to the gradient approximation in the unstable regions. This suggests us to enact the ensemble gradient approximation only in a constrained

manner. The new algorithm, termed the Constrained Ensemble LMC (CEnLMC), will be discussed in Section 3.2, in which we provide a number of constraints, and enact the ensemble gradient approximation only when these constraints are satisfied. The intuition of how these constraints are formulated will also be discussed. The theoretical results will also be summarized in Section 3.3.

3.1. **Ensemble LMC, a direct combination.** We now study the direct combination of LMC and the ensemble gradient approximation. Denote $\{x_i^m\}_{i=1}^N$ the N samples at the m-th step iteration, then following the LMC formula, we would like to write Ensemble LMC (EnLMC) in the form of:

$$x_i^{m+1} = x_i^m - hF_i^m + \sqrt{2h}\xi_i^m\,, \tag{13}$$

with the force $F_i^m = \frac{1}{N-1}\sum F_{ij}^m$ approximating $\nabla f(x_i^m)$. Here F_{ij}^m stands for the contribution of x_j^m towards calculating $\nabla f(x_i^m)$.

Denote $\mathcal{F}^{m-1} = \sigma\left(x_{j\le N}^{n\le m-1}\right)$ the filtration, and p_j^m the marginal distribution of x_j^m conditioned on $\mathcal{F}^{m-1}$, we can replace x^* and $q(x)$ by x_i^m and $p_j^m(x)$ respectively in (5) to define:

$$G_{ij}^m = \alpha_d \frac{\left\langle \nabla f(x_i^m), x_j^m - x_i^m\right\rangle}{|x_j^m - x_i^m|^2}\frac{x_j^m - x_i^m}{\mathrm{p}_j^m}\mathbf{1}_{|x_j^m - x_i^m|\le\eta} \tag{14}$$

where $\mathrm{p}_j^m = p_j^m(x_j^m)$ and α_d is defined in (6). Then, we still have (7) holds true, meaning, for all $j \neq i$,

$$\nabla f(x_i^m) = \mathbb{E}_{p_j^m}(G_{ij}^m) = \mathbb{E}\left(G_{ij}^m \middle| \mathcal{F}^{m-1}, x_i^m\right)\,. \tag{15}$$

Recall the definition of $d_{\eta,q}$ in (8), we define

$$F_{ij}^m = \alpha_d \frac{\delta f_{ij}^m}{|\delta x_{ij}^m|^2}\frac{\delta x_{ij}^m}{\mathrm{p}_j^m}\mathbf{1}_{|\delta x_{ij}^m|\le\eta}\,, \quad \text{with} \quad \begin{cases}\delta f_{ij}^m = f(x_j^m) - f(x_i^m)\,,\\ \delta x_{ij}^m = x_j^m - x_i^m\,,\end{cases} \tag{16}$$

and thus, citing (11), we have

$$\mathbb{E}\left(\left|G_{i,j}^m - F_{i,j}^m\right|\middle|\mathcal{F}^{m-1}, x_i^m\right) \le L\eta d\,. \tag{17}$$

Summing up contribution from all $j \neq i$, we approximate $\nabla f(x_i^m)$ by:

$$\nabla f(x_i^m) \approx F_i^m = \frac{1}{N-1}\sum_{j\neq i}^N F_{ij}^m\,. \tag{18}$$

We note that according to (13), $\mathrm{p}_j^m = p_j^m(x_j^m)$ can be explicitly calculated. Indeed to update x_j^m from x_j^{m-1}, we need x_j^{m-1}, F_j^{m-1} and a random variable ξ_j^{m-1}. Realizing that when conditioned on $\mathcal{F}^{m-1}$, both x_j^{m-1} and F_j^{m-1} are determined, and the only randomness comes from the Gaussian variable ξ_j^{m-1}, meaning x_j^m is merely a Gaussian variable as well when conditioned on $\mathcal{F}^{m-1}$:

$$x_j^m|\mathcal{F}^{m-1} \sim \mathcal{N}(x_j^{m-1} - hF_j^{m-1}\,, 2hI_d)\,,$$

or in other words:

$$p_j^m(x) = \frac{1}{(4\pi h)^{d/2}}\exp\left(-|x - \left(x_j^{m-1} - hF_j^{m-1}\right)|^2/(4h)\right)\,. \tag{19}$$

Plugging in the definition of x_j^m, we can compute p_j^m explicitly:

$$\mathrm{p}_j^m = \frac{1}{(4\pi h)^{d/2}}\exp\left(-|\xi_j^{m-1}|^2/2\right)\,. \tag{20}$$

Remark 2. This is to resonate the discussion in Remark 1. In the derivation above we used the conditional distribution, conditioned on $\mathcal{F}^{m-1}$. If one uses (12) in a brute-force manner, including all randomness, then we arrive at

$$F_{ij}^m = \alpha_d \frac{\delta f_{ij}^m}{|\delta x_{ij}^m|^2} \frac{\mathbf{1}_{|\delta x_{ij}^m| \le \eta}}{p^m(x_j^m)} \delta x_{ij}^m \,,$$

where p^m is the true distribution of x_j^m without the conditioning. However, this definition of F_{ij}^m cannot be used in the ensemble approximation: The x_i^m and x_j^m are not independent to each other and thus the ensemble $\mathbb{E}_{p^m}(F_{ij}^m)$ may not recover $\nabla f(x_i^m)$. More importantly, $p^m(x)$ is unknown in practice, making the calculation impossible.

We plug (20) into (18) and run (13) for the update. The method is termed Ensemble Langevin Monte Carlo (EnLMC), as presented in Algorithm 1.

Algorithm 1 Ensemble Langevin Monte Carlo (EnLMC)

Preparation:

1. Input: h (time stepsize); N (particle number); η (parameter); d (dimension); M (stopping index); α_d (6); $f(x)$.
2. Initial: $\{x_i^0\}_{i=1}^N$ i.i.d. sampled from an initial distribution induced by $q^0(x)$.

Run: **For** $m = 0, 1, \cdots M$

For $i = 1, 2, \cdots, N$

– Define

$$F_i^m = \frac{1}{N-1} \sum_{j \neq i}^N F_{ij}^m \,, \quad \text{with} \quad F_{ij}^m = \alpha_d \frac{\delta f_{ij}^m}{|\delta x_{ij}^m|^2} \frac{\mathbf{1}_{|\delta x_{ij}^m| < \eta}}{\mathrm{p}_j^m} \delta x_{ij}^m \,, \tag{21}$$

where δf_{ij}^m and δx_{ij}^m are defined in (16).

– Draw ξ_i^m from $\mathcal{N}(0, I_d)$;

– Update

$$\begin{cases} x_i^{m+1} = x_i^m - hF_i^m + \sqrt{2h}\xi_i^m \\ \mathrm{p}_i^{m+1} = \dfrac{1}{(4\pi h)^{d/2}} \exp\left(-|\xi_i^m|^2/2\right) \end{cases} . \tag{22}$$

end

end

Output: $\{x_i^M\}_{i=1}^N$.

The design of this algorithm follows straightforwardly from intuition: One replaces the gradient in LMC by the ensemble approximation using the neighbors' information. Since the difference between the true gradient and the ensemble approximation shrinks to zero as η, the neighboring range vanishes, one may incline to conclude that this method would converge also, as long as η is small enough.

However, this is not true. This ensemble surrogate of the gradient induces strong instability to the algorithm. Indeed, ξ_j^m is a Gaussian variable, and for every fixed ϵ, there is non-trivial probability that makes $p_j^m(x_j^m) < \epsilon$, which blows up the force term (21). We explicitly show this instability using the following example with $d = 1$ and $f(x) = x^2/2$:

Theorem 3.1. *Assume $\{x_i^m\}_{i=1}^N$ are generated from Algorithm 1, then for $d = 1$ and $f(x) = x^2/2$, we have: for any $m > 0$, $1 \le i \le N$*

$$\mathbb{E}|x_i^m|^2 = \infty\,. \tag{23}$$

This negative example suggests that directly replacing the gradient by the ensemble approximation leads to an unstable method.

We leave the proof to Section 5.1, but quickly discuss the intuition of the proof here. Indeed, to compute the variance of x^{m+1} term: $\mathbb{E}|x_i^{m+1}|^2$, it is necessary to compute the variance of the force term $\mathbb{E}\left(\left|F_{i,j}^m\right|^2\right)$. The trajectory of $\{x_i\}_{i=1}^N$ is hard to trace, but one can nevertheless compute the conditional variance, conditioned on $\mathcal{F}^{m-1}$:

$$\mathbb{E}\left(\left|F_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right) = \int \left|F_{i,j}^m\right|^2 p_j^m(x_j^m) p_i^m(x_i^m)\,\mathrm{d}x_j^m\,\mathrm{d}x_i^m\,, \tag{24}$$

where p_i^m are the conditional probability distribution given $\mathcal{F}^{m-1}$.

Noting that according to the definition of F_{ij}^m in (21), for $f(x) = |x|^2/2$, we have:

$$F_{i,j}^m = \frac{1}{\eta}\frac{(x_j^m + x_i^m)(x_j^m - x_i^m)}{2|x_j^m - x_i^m|^2}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_j^m)}(x_j^m - x_i^m) = \frac{(x_j^m + x_i^m)}{2\eta}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_j^m)}\,. \tag{25}$$

At the same time, denoting $w_i^m = x_i^{m-1} - hF_i^{m-1}$ the deterministic part of the update for x_i^m, we know that, for all i:

$$x_i^m - w_i^m = \sqrt{2h}\xi_i^{m-1} \sim N(0, 2h) \quad \Rightarrow \quad p_i^m(x_i^m) = \exp\left(-\frac{|x_i^m - w_i^m|^2}{4h}\right)\,. \tag{26}$$

Plugging (25) and (26) into (24), we have:

$$\begin{aligned}&\mathbb{E}\left(\left|F_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right)\\ =&\int_{\mathbb{R}}\int_{B_\eta(x_i^m)} \frac{\left(x_j^m + x_i^m\right)^2}{4\eta^2}\exp\left(\frac{-|x_i^m - w_i^m|^2 + |x_j^m - w_j^m|^2}{4h}\right)\mathrm{d}x_j^m\,\mathrm{d}x_i^m\,.\end{aligned} \tag{27}$$

Since the p_j^m term is in the denominator in (25), and when one takes the variance, this term gets squared. In the end this exponential term from x_j^m appears in a positive manner in (27). This already suggests the blowing up of this variance term. A more careful derivation shows:

$$\begin{aligned}&\mathbb{E}\left(\left|F_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right)\\ =&\int_{\mathbb{R}} e^{-\frac{|x_i^m - w_i^m|^2}{4h}}\int_{B_\eta(0)}\frac{(z + 2x_i^m)^2}{4\eta^2}e^{\frac{|z + x_i^m - w_j^m|^2}{4h}}\,\mathrm{d}z\,\mathrm{d}x_i^m\\ =&\int_{B_\eta(0)} e^{\frac{-|w_i^m|^2 + |z - w_j^m|^2}{4h}}\int_{\mathbb{R}}\frac{(z + 2x_i^m)^2}{4\eta^2}e^{\frac{x_i^m(z + w_i^m - w_j^m)}{2h}}\,\mathrm{d}x_i^m\,\mathrm{d}z\\ =&\infty\,.\end{aligned} \tag{28}$$

In the second equality we used the change of variables $z = x_j^m - x_i^m$. The infinity comes from the inner integral, where we are essentially looking at the second moment of an exponential function.

This infinite variance of $F_{i,j}^m$, calculated in (28), suggests the variance of x_i^{m+1}, to be showed in (23), is also infinite. Proving Theorem 3.1 then amounts to carrying out the detailed derivation on how $\mathbb{E}|x_i^{m+1}|^2$ depends on $\mathbb{E}\left|F_{i,j}^m\right|^2$, and we leave this to Section 5.1.

3.2. **Constrained Ensemble LMC, a modification.** We now take a more careful look at the instability in the ensemble gradient approximation to LMC. Intuitively there are two sources of instability:

- When x_i^m is at the "outskirt" of $p(x)$, $f(x_i^m)$ is high, and $p(x_i^m) \propto \exp\{-f(x_i^m)\}$ is extremely small. This could bring high relative error, and we should avoid making any approximations in this region.
- In the formula (18), $p_j^m(x_j^m)$ is in the denominator. Considering the way the term is defined in (20), it takes an $\mathcal{O}(1)$ value with high probability when ξ_j^m is moderately small. However, there is a small chance for $|\xi_j^m|$ to take large values, which will make $p_j^m(x_j^m)$ extremely small, bringing infinite variance, as shown in (28).

To avoid these two scenarios, we essentially need to identify:

- x_i^m who are at the "outskirt" of p;
- x_j^m that is within η distance from x_i^m but has large $|\xi_j^{m-1}|$.

When these happen, the ensemble approximation is disabled and we come back to use the true gradient $\nabla f(x_i^m)$.

To identify the first scenario is relatively straightforward: We simply set a threshold, call it M_f, and will only employ ensemble gradient approximation when $f(x_i^m)$ is smaller than M_f:

$$f(x_i^m) < M_f\,.$$

To identify the second scenario is slightly more involved. We now consider

$$\begin{aligned}\sqrt{2h}|\xi_j^{m-1}| = |x_j^m - w_j^m| \leq& |x_j^m - x_i^m| + |x_i^m - w_i^m| + |w_i^m - w_j^m| \\ =& |\delta x_{ij}^m| + \sqrt{2h}|\xi_i^{m-1}| + |\delta w_{ij}^m|\end{aligned}$$

where we denote the deterministic component of the updating formula:

$$w_i^m = x_i^{m-1} - hF_i^{m-1}\,, \quad \delta w_{ij}^m = w_j^m - w_i^m\,. \tag{29}$$

A sufficient condition to have a moderate $|\xi_j^{m-1}|$ is to have all three terms on the right hand side moderate. For a fixed x_i^m, since we only consider x_j^m who are already within η distance, the first term is already bounded by η and is small. We therefore need to ensure the remaining two terms are bounded as well. To do so, we propose to enact the ensemble gradient approximation only if $|\xi_i^{m-1}|$ is at most moderately large, and for those x_i^m, we include the x_j^m contribution in the calculation of F_i^m only if $|\delta w_{ij}^m|$ is at most moderately large. This is to say, for a fixed preset constant pairs $(R_1\,, R_2)$:

- When $\sqrt{2h}|\xi_i^{m-1}| > R_1$:

$$F_i^m = \nabla f(x_i^m)\,, \tag{30}$$

- When $\sqrt{2h}|\xi_i^{m-1}| \leq R_1$:

$$F_i^m = \frac{1}{N_i^m}\sum_{j\neq i}^N F_{ij}^m\,, \quad \text{with} \quad F_{ij}^m = \alpha_d \frac{\delta f_{ij}^m}{|\delta x_{ij}^m|^2}\frac{\delta x_{ij}^m}{\mathrm{p}_j^m}\mathbf{1}_{|\delta x_{ij}^m|\leq\eta\,,|\delta w_{ij}^m|\leq R_2}\,, \tag{31}$$

where p_j^m is defined in (20) and

$$N_i^m = \sum_{j\neq i}^N \mathbf{1}_{|\delta w_{ij}^m|\leq R_2}\,, \tag{32}$$

is the number of neighbors within η distance whose corresponding $|\delta w_{ij}^m|$ is controlled.

Note that compared with (16), we add another indicator function in (31) to ensure δw_{ij}^m is controlled by R_2. Furthermore, numerically to have statistical stability, we also preset a value for N^* and require $N_i^m \geq N^*$. If $N_i^m < N^*$, we do not enact the ensemble approximation and use the true gradient $\nabla f(x_i^m)$.

Summarizing the discussion above, we have:

$$F_i^m = \begin{cases} \nabla f(x_i^m), & \sqrt{2h}|\xi_i^{m-1}| > R_1 \text{ or } f(x_i^m) > M_f \text{ or } N^* > N_i^m \\ \dfrac{1}{N_i^m}\displaystyle\sum_{j\neq i}^{N} F_{ij}^m\,, & \text{otherwise}\,. \end{cases} \tag{33}$$

Replacing the gradient term in LMC using (33), we arrive at a new algorithm. We term it Constrained Ensemble Langevin Monte Carlo (CEnLMC), as summarized in Algorithm 2.

Algorithm 2 Constrained Ensemble Langevin Monte Carlo (CEnLMC)

Preparation:

1. Input: h (time stepsize); N (particle number); η, R_1, R_2, N^*, M_f (parameters); d (dimension); M (stopping index); α_d (6); $\nabla f(x)$; $f(x)$; f^* (minimal value).
2. Initial: $\{x_i^0\}_{i=1}^N$ i.i.d. sampled from an initial distribution induced by $q^0(x)$. Set $w_i^{-1} = \infty$ for $1 \leq i \leq N$.

Run: **For** $m = 0, 1, \cdots, M$

For $i = 1, 2, \cdots, N$

– Define

$$N_i^m = \sum_{j\neq i}^{N} \mathbf{1}_{|\delta w_{ij}^m| < R_2}\,.$$

– **If** $\sqrt{2h}|\xi_i^{m-1}| > R_1$ or $f(x_i^m) > M_f$ or $N^* > N_i^m$, define

$$F_i^m = \nabla f(x_i^m)\,.$$

else define

$$F_i^m = \frac{1}{N_i^m}\sum_{j\neq i}^{N} F_{ij}^m\,, \quad \text{with} \quad F_{ij}^m = \alpha_d \frac{\delta f_{ij}^m}{|\delta x_{ij}^m|^2}\frac{\delta x_{ij}^m}{\mathrm{p}_j^m}\mathbf{1}_{|\delta x_{ij}^m|\leq\eta\,,|\delta w_{ij}^m|\leq R_2}\,. \tag{34}$$

where δf_{ij}^m, δx_{ij}^m are defined in (16), and $\delta w_{i,j}^m$ is defined in (29).

end

– Draw ξ_i^m from $\mathcal{N}(0, I_d)$.

– Update

$$\begin{cases} x_i^{m+1} = x_i^m - hF_i^m + \sqrt{2h}\xi_i^m\,, \\ \mathrm{p}_i^{m+1} = \dfrac{1}{(4\pi h)^{d/2}}\exp\left(-|\xi_i^m|^2/2\right)\,, \\ w_i^{m+1} = x_i^m - hF_i^m \end{cases} \tag{35}$$

end

end

Output: $\{x_i^M\}_{i=1}^N$.

3.3. **Properties of CEnLMC.** There are two types of properties of CEnLMC that we would like to discuss: 1. the convergence: We would like to show that the distribution of x_i^m, as $m \to \infty$ converges to the target distribution; 2. the numerical cost: We would like to show that the probability of computing the gradients is low with a proper tuning of R_1, R_2 and M_f, and thus most gradients are replaced by its cheaper ensemble version. This makes CEnLMC cheaper than the classical LMC.

These two properties are discussed in the following subsections respectively.

3.3.1. *Convergence of CEnLMC.* To show the method converges is to show that the distribution of x_i^m, as $m \to \infty$, converges to the target distribution p up to a small discretization error.

Our strategy is to show that particles computed from CEnLMC are close to the particles computed from the classical LMC if they start with the same initial data. Since it is well-known that the distribution of LMC samples converges to the target distribution, the samples found by CEnLMC then recover the target distribution as $m \to \infty$ as well.

We first introduce the particle system that solves the classical LMC (2). Define $z_i^0 = x_i^0$ for $1 \le i \le N$ and update

$$z_i^{m+1} = z_i^m - \nabla f(z_i^m)h + \sqrt{2h}\xi_i^m \,, \tag{36}$$

where ξ_i^m is the same as (35). This is the classical LMC algorithm, and all samples z_i are decoupled from each other. Our first goal is to show that x_i^m and z_i^m are approximately the same, as seen in the following theorem.

Theorem 3.2. *Assume $\{x_i^m\}_{i=1}^N$ are generated from Algorithm 2, and $\{z_i^m\}_{i=1}^N$ are generated from (36), with the parameters chosen to satisfy*

$$h \le \min\left\{\frac{1}{L}, \frac{1}{d}\right\} \,, \ \max\{\eta, 1\} \le R_2 \,, \ M_f > f^* \,,$$

where f^* is the optimal (minimum) of $f(x)$. Assume f is L-smooth, then, for $m \ge 0$, $1 \le i \le N$:

$$\mathbb{E}|x_i^m - z_i^m| \le \mathcal{O}\left(\exp(Lmh)\left(\sqrt{\frac{R_1^d(M_f - f^*)d^2}{L\eta^d N^*}}\exp\left(\frac{R_2(R_2+R_1)}{2h}\right) + \eta d\right)\right). \tag{37}$$

If we further assume f is μ-convex, then, denoting $\kappa = L/\mu$, for any $m \ge 0$, $1 \le i \le N$:

$$\mathbb{E}|x_i^m - z_i^m| \le \mathcal{O}\left(\sqrt{\frac{R_1^d\kappa(M_f - f^*)d^2}{\mu\eta^d N^*}}\exp\left(\frac{R_2(R_2+R_1)}{2h}\right) + \kappa\eta d\right). \tag{38}$$

We leave the proof to Section 5.2.

We stress the importance of this theorem. The theorem estimates the distance between the proposed samples and the classical LMC samples. With the properly tuned parameters, we can make the bound in (37)-(38) small, forcing the two sets of samples close to each other. LMC is a classical algorithm that we have rich understanding about. In particular, we have results from [5, 6, 14] that give non-asymptotic error estimate: The error, in Wasserstein distance, converges to zero, exponentially fast, up to the discretization error that depends on d, the dimension of the problem, and h, the stepsize. This means, the newly proposed algorithm

CEnLMC also converges exponentially fast, up to the discretization error and this newly induced approximation error.

We now take a closer look at this approximation error. Use the convex case as an example, we examine the two terms in (38). The second bound mainly comes from the finite difference approximation, induced in (17), and the first term traces back to ensemble error $(\mathbb{E}|\nabla f(x_i^m) - G_i^m|^2)$. After adding constraints (30)-(33), this error contributes to $1/\sqrt{N^*}$ term. This is optimal in terms of N^* according to the central limit theorem.

To make the distance small, we first need to let η be small so that the error from the finite differencing is small. Upon choosing small η, with $R_{1,2}$ fixed, we need to select a moderate $(M_f - f^*)/N^*$ to make the first term small. Since M_f is the bound we set to turn on or off the ensemble gradient approximation, we expect it to be relatively large. N^* is the minimum number of neighbors needed to enact the ensemble approximation to ensure statistical accuracy and is thus also expected to be large. To accommodate both, we set $M_f = (N^*)^\rho + f^*$ with $\rho < 1$.

We summarize this choice of parameters in the following corollary:

Corollary 1. *Under the same assumption as in Theorem 3.2 and let f be μ-convex, for any small number $\epsilon > 0$ and $0 < \rho < 1$, by setting*

$$M_f = (N^*)^\rho + f^*, \quad \eta < \frac{\epsilon}{\kappa d}, N^* = \frac{R_1^{d/(1-\rho)} \kappa^{1/(1-\rho)} d^{2/(1-\rho)}}{\mu^{1/(1-\rho)} \eta^{d/(1-\rho)} \epsilon^{2/(1-\rho)}} \exp\left(\frac{R_2(R_2+R_1)}{2(1-\rho)h}\right), \tag{39}$$

we have: for any $m \geq 0$, $1 \leq i \leq N$:

$$\mathbb{E}|x_i^m - z_i^m| \leq \mathcal{O}(\epsilon) . \tag{40}$$

This is obtained by simply setting both terms in (38) smaller than ϵ. We omit the proof.

Now we are ready to combine this result with the well-known convergence result of LMC to show the convergence of CEnLMC. The convergence is discussed in both Wasserstein distance sense, and weak sense.

Theorem 3.3. *Under the same assumption as in Theorem 3.2 and let f be μ-convex, we denote $\kappa = L/\mu$ the condition number, q_i^m the probability density of x_i^m. Assume $\int |x| q^0 \, \mathrm{d}x < \infty$, we have:*

1. *W_1 convergence: For any $m \geq 0$, $1 \leq i \leq N$,*

$$W_1(q_i^m, p) \leq \exp\left(-\frac{\mu h m}{2}\right) W_1(q^0, p) + \mathcal{O}\left(\kappa(\sqrt{hd} + \eta d) + \sqrt{\frac{R_1^d \kappa d^2 (M_f - f^*)}{\mu \eta^d N^*}} \exp\left(\frac{R_2(R_2+R_1)}{2h}\right)\right) . \tag{41}$$

2. *Weak convergence: For any Lipschitz function* $g : \mathbb{R}^d \to \mathbb{R}$ *with* $\mathbb{E}_p(g^2) < \infty$ *and* $m \geq 0$*, we have*

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} g(x_i^m) - \mathbb{E}_p(g)\right| \\
\leq &\mathcal{O}\left(\exp\left(-\frac{\mu h m}{2}\right) W_1(q^0, p)\right) \\
+&\mathcal{O}\left(\frac{1}{\sqrt{N}} + \kappa(\sqrt{hd} + \eta d) + \sqrt{\frac{R_1^d \kappa d^2 (M_f - f^*)}{\mu \eta^d N^*}} \exp\left(\frac{R_2(R_2+R_1)}{2h}\right)\right).
\end{aligned} \tag{42}$$

We leave the proof to Section 5.2. We note that in both (41) and (42), there is one exponentially decaying term, and the rest can be seen as the remainder term. Therefore we can call the convergence rate exponential, up to a controllable discretization and ensemble error. The exponentially decaying term comes from the fact that the distribution of z_i^m decays to the target distribution exponentially fast, and the remainder term mostly comes from the distance between $\{x_i^m\}$ and $\{z_i^m\}$ systems.

Remark 3. This theorem gives a clear guidance on the choice of some parameters. To have fast convergence and small error term, the parameters need to be tuned to have second term in (41) as small as possible. Assume we have enough particles ($N \to \infty$), we set this term to be smaller than ϵ, then:

$$\eta \leq \mathcal{O}\left(\frac{\epsilon}{\kappa d}\right), \; h \leq \mathcal{O}\left(\frac{\epsilon^2}{36\kappa^2 d}\right), \; N^* > \mathcal{O}\left(\frac{36 R_1^d \kappa d^2 (M_f - f^*)}{\mu \eta^d \epsilon^2} \exp\left(\frac{R_2(R_2+R_1)}{h}\right)\right).$$

We then set the first term to be smaller than ϵ as well, then the lower bound for the needed number of iteration is:

$$m > \mathcal{O}\left(\frac{\kappa^2 d}{\epsilon^2} \log\left(\frac{W_1(q^0, p)}{\epsilon}\right)\right),$$

meaning after these many iterations, $W_1(q_i^m, p) \leq 2\epsilon$, where q_i^m is the distribution of x_i^m.

Note that this gives the control of η, h and N^* but still leaves the freedom to adjust R_1, R_2 and M_f. These parameters should be determined by the percentage of gradient that we are willing to calculate. The discussion is found in Remark 4.

3.3.2. *Numerical saving of CEnLMC.* We now discuss the numerical saving of CEnLMC compared with the classical LMC.

The main reason to utilize the ensemble gradient approximation is to avoid the gradient computation. In the algorithm, the ensemble approximation is enacted only if:

$$\sqrt{2h}|\xi_i^{m-1}| \leq R_1, \quad f(x_i^m) \leq M_f, \quad N_i^m \geq N^*,$$

where the size of N_i^m depends on the number of samples who satisfy $|\delta w_{ij}^m| \leq R_2$. Therefore the probability of not using the ensemble approximation (but using ∇f) can be bounded by:

$$\begin{aligned}\mathbb{P}\left(\{F_i^m = \nabla f(x_i^m)\}\right) \leq & \mathbb{P}\left(\left\{\sqrt{2h}|\xi_i^m| > R_1\right\}\right) \\ & + \mathbb{P}\left(\{|f(x_i^m) - f^*| > (M_f - f^*)\}\right) \\ & + \mathbb{P}\left(\{N_i^m < N^*\}\right).\end{aligned} \tag{43}$$

One thus needs to choose the parameters wisely to make such a probability as small as possible so that most gradients in LMC get replaced by the ensemble approximation. More specifically, we have the following theorem:

Theorem 3.4. *Under the same assumption as in Theorem 3.2 and let f be μ-convex. If $\mathrm{KL}(q_0|p) < \infty$, then for fixed $M \geq 0$, we have:*

$$\lim_{\eta\to 0}\lim_{N\to\infty}\sup_{0\leq m\leq M, 1\leq i\leq N}\mathbb{P}\left(\left\{\sqrt{2h}|\xi_i^{m-1}| > R_1\right\}\right) \leq C_d(R_1), \tag{44}$$

$$\lim_{\eta\to 0}\lim_{N\to\infty}\sup_{0\leq m\leq M, 1\leq i\leq N}\mathbb{P}\left(\{|f(x_i^m) - f^*| > (M_f - f^*)\}\right) \leq \frac{2\kappa d}{(M_f - f^*)}, \tag{45}$$

$$\lim_{\eta\to 0}\lim_{N\to\infty}\sup_{0\leq m\leq M, 1\leq i\leq N}\mathbb{P}\left(\{N_i^m < N^*\}\right) = 0. \tag{46}$$

where

$$C_d(R_1) = \frac{S_d}{(2\pi)^{d/2}}\int_{\frac{R_1\sqrt{d}}{\sqrt{2}}}^{\infty} r^{d-1}\exp\left(-\frac{r^2}{2}\right)\mathrm{d}r$$

diminishes to 0 for large R_1 and S_d is the volume of unit d-sphere.

We leave the proof of the theorem to Section 5.3. This theorem gives the bound to (43). According to the formula of (44)-(46), a direct corollary is the following:

Corollary 2. *Under the same assumption as in Theorem 3.4, for any $\epsilon > 0$, there exists constants R^*, F^* only depend on ϵ, d such that if*

$$R_1 > R^*, \quad M_f > F^*,$$

we have

$$\lim_{\eta\to 0}\lim_{N\to\infty}\sup_{0\leq m\leq M, 1\leq i\leq N}\mathbb{P}\left(\{F_i^m = \nabla f(x_i^m)\}\right) \leq \epsilon.$$

According to the Corollary 2, when we have enough particles, we can always tune the parameters so that most gradients in LMC get replaced by the ensemble approximation.

Remark 4. This theorem gives the guideline for the parameter choice of R_1, R_2 and M_f. Suppose the percentage of the gradient we would like to compute is α, and we equally distribute it to the three terms in (43). Then in the limit of $\eta \to 0$ and $N \to \infty$, R_1 should be chosen, according to (44), so that

$$C_d(R_1) \leq \frac{\alpha}{3}.$$

Similarly, according to (45), M_f should be chosen so that

$$M_f \geq \frac{6\kappa d}{\alpha} + f^*.$$

Lastly, we need to give a bound for R_2. This can be implicitly computed from (46). While it is true that in the $N \to \infty$ limit, the probability is necessarily $< \frac{\alpha}{3}$, for every fixed N, the size of R_2 will affect the probability. Such dependence is very

delicate, and we only give a rough bound. Suppose we are in the ideal case with $h \to 0$ so that $x_i^m = w_i^m$, and suppose we have iterated many times and the particles are approximately close to i.i.d. sampled from the target distribution. Then

$$\begin{aligned}\mathbb{P}\left(\{N_i^m < N^*\}\right) &= \mathbb{P}\left(\#\left\{w_j^m \middle| |w_j^m - w_i^m| < R_2,\ j = 1, 2, \dots, N\right\} < N^* + 1 \middle| w_i^m\right) \\ &\approx \mathbb{P}\left(\#\left\{x_j^m \middle| |x_j^m - x_i^m| < R_2,\ j = 1, 2, \dots, N\right\} < N^* + 1 \middle| x_i^m\right) \\ &= \sum_{k=0}^{N^*-1} \binom{N-1}{k} p^k(R_2)(1 - p(R_2))^{N-1-k} \ll O(1)\end{aligned}$$

where $p(R_2) = \mathbb{P}_{y,z\sim p}(|y - z| < R_2)$. The first equation comes from the definition, and the second is driven by the fact that x_i^m and w_i^m are close by. Assuming $N^* < \frac{N+1}{2}, p(R_2) < \frac{1}{4}$, then

$$\begin{aligned}\mathbb{P}\left(\{N_i^m < N^*\}\right) &\approx \sum_{k=0}^{N^*-1} \binom{N-1}{k} p^k(R_2)(1 - p(R_2))^{N-1-k} \\ &\le (1 - p(R_2))^{N-1} \binom{N-1}{N^*-1} \sum_{k=0}^{N^*-1} \left(\frac{p(R_2)}{1 - p(R_2)}\right)^k \\ &\le \binom{N-1}{N^*-1} \frac{(1 - p(R_2))^N}{1 - 2p(R_2)} \\ &\le C N^{N^*} (1 - p(R_2))^N\end{aligned}$$

where C is a uniform constant and we use Stirling's approximation in the last inequality. To have this term controlled by $\frac{\alpha}{3}$, we need to choose $p(R_2)$ so that:

$$1 - \left(\frac{\alpha}{3CN^{N^*}}\right)^{1/N} \le p(R_2) \le \frac{1}{4},$$

which permits:

$$\mathbb{P}\left(\{N_i^m < N^*\}\right) \approx \sum_{k=0}^{N^*-1} \binom{N-1}{k} p^k(R_2)(1 - p(R_2))^{N-1-k} \le \frac{\alpha}{3}.$$

4. **Numerical experiment.** We show two numerical examples to demonstrate the two main themes of the paper: the samples capture the target distribution, and the number of gradient calculations is significantly reduced. In particular, for both examples, we define the percentage of the gradient calculations:

$$\mathcal{R}_m = \frac{\#\{F_i^j = \nabla f(x_i^j) | 1 \le i \le N, 1 \le j \le m\}}{mN},$$

and we will show the evolution of this percentage in iterations. To demonstrate the accuracy, we also show the samples generated from LMC [37] and MALA (Metropolis-adjusted Langevin algorithm) [36, 39].

Example 1. In this example, we set $d = 2$, and the target distribution $p(x) \propto \exp(-|x_1|^2/2 - |x_2|^2/8)$. Suppose the initial distribution is:

$$q^0(x) \propto \exp\left(-\frac{(x_1 - 1)^2}{2} - \frac{(x_2 - 1)^2}{2}\right) + \exp\left(-\frac{(x_1 + 1)^2}{2} - \frac{(x_2 + 1)^2}{2}\right).$$

In the experiment, we choose $R_1 = \frac{3\sqrt{5}}{10}$, $h = \eta = 0.1$, $R_2 = 1.5$, $M_f = 20$, and $N^* = 10^3$. In Figure 1-2, we plot the samples generated by CEnLMC, LMC, and

MALA at different iterations, using $N = 10^4$. Since the example is logconcave in nature, the samples converge fairly quickly. Furthermore, we plot the ratio $\mathcal{R}_m$ at different iteration, using $N = 2 \times 10^3, 6 \times 10^3, 10^4$, in Figure 3. While in the case of $N = 2 \times 10^3$, most particles need to have its gradient computed in every iteration, the ratio drops significantly for the larger N, and as iteration m increases, the percentage of gradient calculation continues to decrease. This saving verifies the prediction from Section 3.3.2.

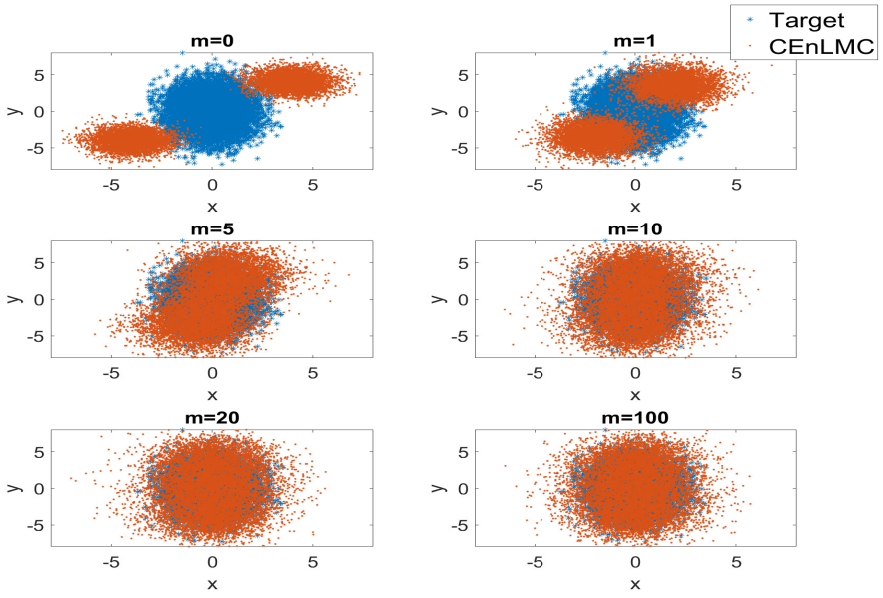

FIGURE 1. Example 1: Evolution of samples using CEnLMC. $N = 10^4$.

Example 2. In this example, we test the algorithms on a target distribution that is not logconcave. Set the target to be

$$p(x) \propto \exp\left(-\frac{(x_1-4)^2}{2} - \frac{x_2^2}{2}\right) + \exp\left(-\frac{(x_1+4)^2}{2} - \frac{x_2^2}{2}\right),$$

and the initial to be $q^0(x) \propto \exp(-|x_1|^2/2 - |x_2|^2/2)$. In the experiment, we choose $R_1 = \frac{3\sqrt{5}}{10}$, $h = \eta = 0.1$, $R_2 = 1.5$, $M_f = 20$, and $N^* = 10^3$. In Figure 4-5, we plot the samples generated by CEnLMC, LMC, and MALA at different iterations, using $N = 10^4$. Since the example is not logconcave anymore, the convergence rate of the samples is slower. We also plot the ratio $\mathcal{R}_m$ at different iteration, using $N = 2 \times 10^3, 6 \times 10^3$ and 10^4 respectively, in Figure 6. While in the case of $N = 2 \times 10^3$, most particles need to have its gradient computed in every iteration, the ratio drops significantly for the larger N.

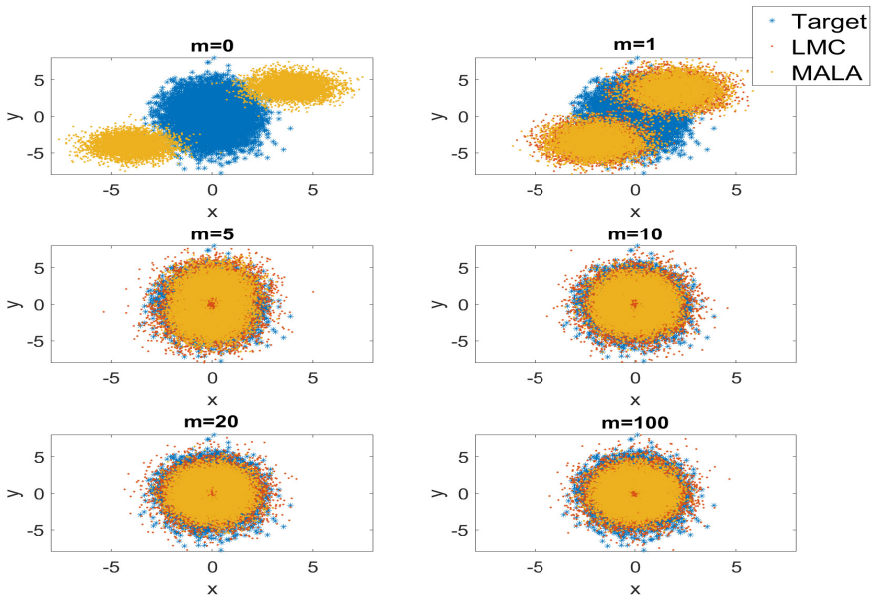

FIGURE 2. Example 1: Evolution of samples using LMC and MALA. $N = 10^4$.

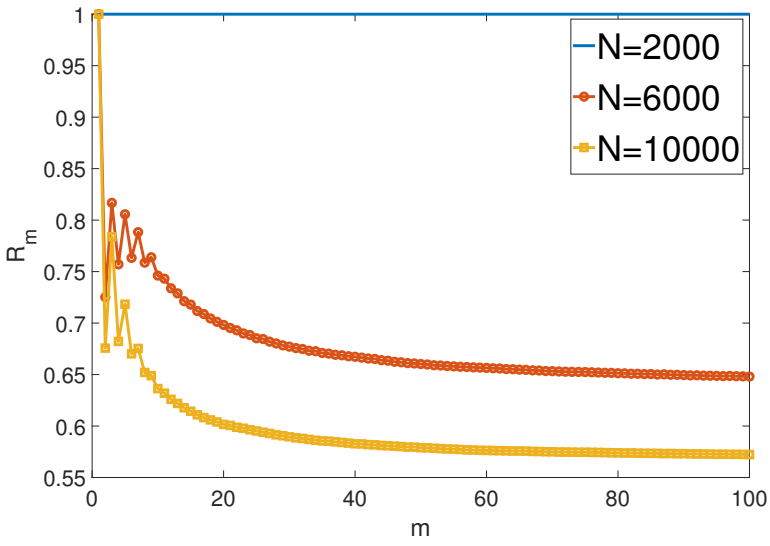

FIGURE 3. Example 1: Evolution of $\mathcal{R}_m$ when $N = 2\times10^3, 6\times10^3$ or 10^4.

5. Proof of theoretical results.

5.1. **Proof of Theorem 3.1.** In this section, we prove Theorem 3.1. According to algorithm 1, we have

$$x_i^m = x_i^{m-1} - hF_i^{m-1} + \sqrt{2h}\xi_i^{m-1}$$

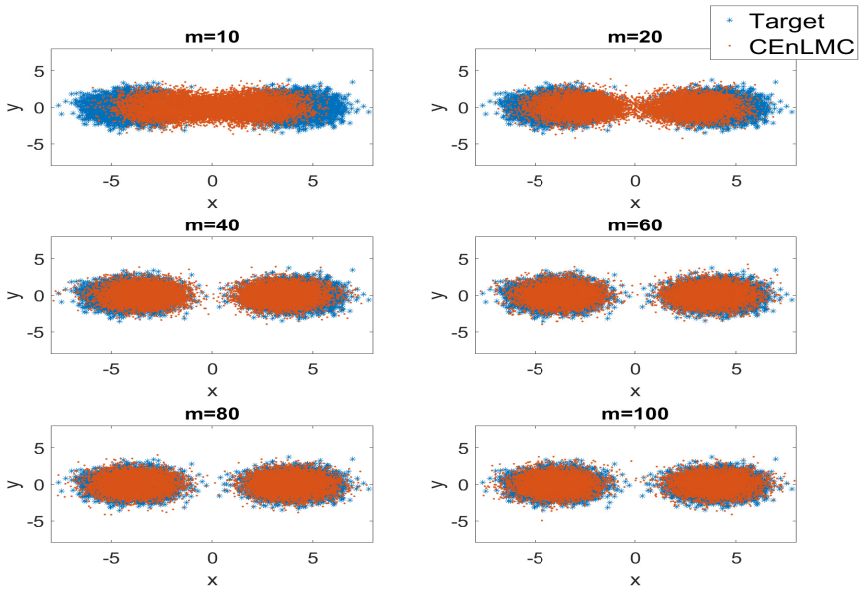

FIGURE 4. Example 2: Evolution of samples using CEnLMC when $N = 10^4$

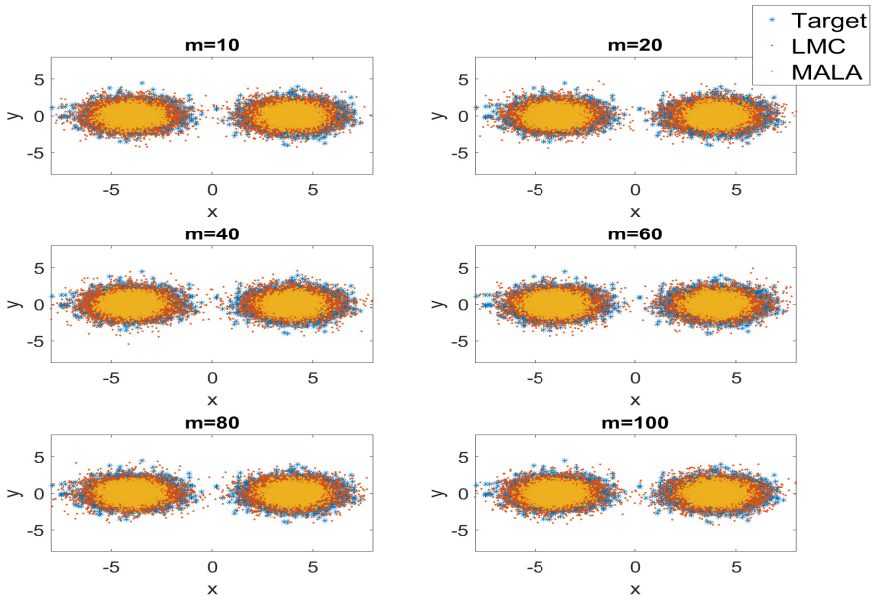

FIGURE 5. Example 2: Evolution of samples using LMC and MALA when $N = 10^4$

and $\{\xi_i^{m-1}\}_{i=1}^N$ are i.i.d. independent. Under filtration $\mathcal{F}^{m-1}$, then the conditional distribution of $\{x_i^m\}_{i=1}^N$ is independent.

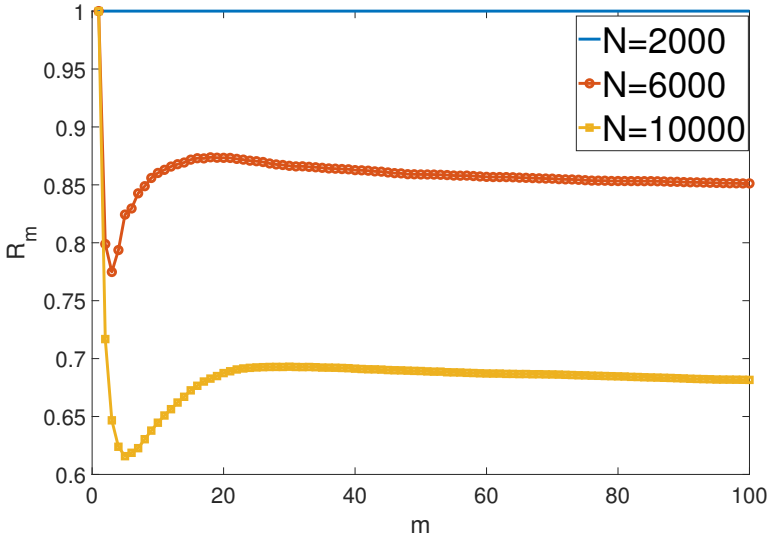

FIGURE 6. Example 2: Evolution of $\mathcal{R}_m$ with m when $N = 2 \times 10^3, 6 \times 10^3, 10^4$

To prove the theorem, we need the following proposition:

Proposition 1. *Assume $\{x_i^m\}_{i=1}^N$ are generated from Algorithm 1 with F^m defined as (21), then for $f(x) = x^2/2$, we have: for any $m > 0$, $1 \le i \le N$*

$$\mathbb{E}\left(|E_i^m|^2\right) = \mathbb{E}\,|F_i^m - \nabla f(x_i^m)|^2 = \infty\,. \tag{47}$$

Proof of Proposition 1. Since $f(x) = |x|^2/2$, we can obtain, according to (25):

$$\begin{aligned} F_{i,j}^m =& \frac{1}{\eta}\frac{(x_j^m + x_i^m)(x_j^m - x_i^m)}{2|x_j^m - x_i^m|^2}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_j^m)}(x_j^m - x_i^m) \\ =& \frac{x_j^m - x_i^m}{2\eta}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_j^m)} + \frac{x_i^m}{\eta}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_j^m)}\,. \end{aligned}$$

The two terms carry different information:

- The conditional expectation of first term equals zero:

$$\begin{aligned} &\mathbb{E}\left(\frac{x_j^m - x_i^m}{2\eta}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_j^m)}\middle|\mathcal{F}^{m-1}\right) \\ =&\frac{1}{2\eta}\int\int_{|x_j^m - x_i^m|<\eta}(x_j^m - x_i^m)p_i^m(x_i^m)\,\mathrm{d}x_j^m\,\mathrm{d}x_i^m = 0\,. \end{aligned}$$

- The second term is consistent with $\nabla f(x_i^m) = x_i^m$, meaning:

$$\mathbb{E}\left(\frac{x_i^m}{\eta}\frac{\mathbf{1}_{|\delta x_{ij}^m|<\eta}}{p_j^m(x_i^m)}\middle|\mathcal{F}^{m-1}, x_i^m\right) = x_i^m\int_{|x_j^m - x_i^m|<\eta}\frac{1}{\eta}\,\mathrm{d}x_j^m = x_i^m\,,$$

  where we use x_j^m and x_i^m is conditional independent in the first equality.

These imply, for all $j \neq i$:

$$\mathbb{E}\left(F_{i,j}^m - x_i^m\middle|\mathcal{F}^{m-1}\right) = 0\,. \tag{48}$$

Furthermore, since the conditional distribution of $x_{j_1}^m, x_{j_2}^m, x_i^m$ are independent, for $j_1 \neq j_2$, $i \neq j_1$, and $i \neq j_2$:

$$\begin{aligned}
&\mathbb{E}\left((F_{i,j_1}^m - x_i^m)(F_{i,j_2}^m - x_i^m)\middle|\mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(\mathbb{E}\left((F_{i,j_1}^m - x_i^m)(F_{i,j_2}^m - x_i^m)\middle|\mathcal{F}^{m-1}, x_i^m\right)\middle|\mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(\mathbb{E}\left(F_{i,j_1}^m - x_i^m\middle|\mathcal{F}^{m-1}, x_i^m\right)\mathbb{E}\left(F_{i,j_2}^m - x_i^m\middle|\mathcal{F}^{m-1}, x_i^m\right)\middle|\mathcal{F}^{m-1}\right)\\
=&0
\end{aligned} \tag{49}$$

Plug (48) and (49) into $\mathbb{E}\left(|E_i^m|^2\middle|\mathcal{F}^{m-1}\right) = \mathbb{E}\left(|F_i^m - \nabla f(x_i^m)|^2\middle|\mathcal{F}^{m-1}\right)$, we have

$$\begin{aligned}
\mathbb{E}\left(|E_i^m|^2\middle|\mathcal{F}^{m-1}\right) =&\mathbb{E}\left(|F_i^m - \nabla f(x_i^m)|^2\middle|\mathcal{F}^{m-1}\right)\\
=&\frac{1}{(N-1)^2}\sum_{j\neq i}^{N}\mathbb{E}\left(\left|F_{i,j}^m - x_i^m\right|^2\middle|\mathcal{F}^{m-1}\right)\\
=&\frac{1}{(N-1)^2}\sum_{j\neq i}^{N}\mathbb{E}\left(\left|F_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right) - \frac{1}{N-1}\mathbb{E}\left(|x_i^m|^2\middle|\mathcal{F}^{m-1}\right),
\end{aligned} \tag{50}$$

where we use (49) in the second equality. Noting that in (28) we already showed:

$$\mathbb{E}\left(\left|F_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right) = \infty,$$

and that the second term in (50) is finite:

$$\mathbb{E}\left(|x_i^m|^2\middle|\mathcal{F}^{m-1}\right) = \left|x_i^{m-1} - hF_i^{m-1}\right|^2 + 2h < \infty,$$

we obtain:

$$\mathbb{E}\left(|F_i^m - \nabla f(x_i^m)|^2\middle|\mathcal{F}^{m-1}\right) = \infty,$$

which proves (47), concluding this proposition. □

Now, we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. For each $m \geq 0$ and $1 \leq i \leq N$, we consider

$$x_i^{m+1} = x_i^m - h\nabla f(x_i^m) + \sqrt{2h}\xi_i^m + hE_i^m,$$

where $E_i^m = \nabla f(x_i^m) - F_i^m$ denote the differentiation from the classical LMC formula. Using x_j^m and x_i^m are conditional independent for $i \neq j$, we obtain

$$\begin{aligned}
&\mathbb{E}\left(E_i^m(x_i^m - h\nabla f(x_i^m) + \sqrt{2h}\xi_i^m)\middle|\mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(E_i^m(x_i^m - h\nabla f(x_i^m))\middle|\mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(\mathbb{E}\left(E_i^m(x_i^m - h\nabla f(x_i^m))\middle|\mathcal{F}^{m-1}, x_i^m\right)\middle|\mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(\left(\frac{1}{N-1}\sum_{j\neq i}^{N}\mathbb{E}\left(x_i^m - F_{i,j}^m\middle|\mathcal{F}^{m-1}, x_i^m\right)\right)(x_i^m - h\nabla f(x_i^m))\middle|\mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(0\,(x_i^m - h\nabla f(x_i^m))\middle|\mathcal{F}^{m-1}\right) = 0,
\end{aligned} \tag{51}$$

where we use $\mathbb{E}\left(\xi_i^m\middle|\mathcal{F}^{m-1}\right) = \mathbb{E}\left(\xi_i^m\right) = \vec{0}$ in the first equality and (48) in the second last equality.

Therefore, we have

$$\begin{aligned}&\mathbb{E}\left(|x_i^{m+1}|^2\middle|\mathcal{F}^{m-1}\right)\\=&\mathbb{E}\left(\left|x_i^m-h\nabla f(x_i^m)+\sqrt{2h}\xi_i^m\right|^2\middle|\mathcal{F}^{m-1}\right)+\mathbb{E}\left(|E_i^m|^2\middle|\mathcal{F}^{m-1}\right)\\\geq&\mathbb{E}\left(|E_i^m|^2\middle|\mathcal{F}^{m-1}\right),\end{aligned}$$

where we use (51) in the first equality. Finally, using the previous proposition, we have

$$\mathbb{E}\left(\mathbb{E}\left(|x_i^{m+1}|^2\middle|\mathcal{F}^{m-1}\right)\right)\geq\mathbb{E}\left(\mathbb{E}\left(|E_i^m|^2\middle|\mathcal{F}^{m-1}\right)\right)=\infty,$$

which proves (23). □

5.2. **Analysis of CEnLMC.** We now analyze Algorithm 2, the Constraint Ensemble LMC. The strategy is to compare the evolution of x_i^m with z_i^m, the solution to the classical LMC (36), before utilizing the convergence of z_i^m to find the convergence of x_i^m.

Theorem 3.2 discusses the closeness of x_i^m and z_i^m, while Theorem 3.3 discusses the convergence of x_i^m. The following two subsections are dedicated to these two theorems respectively.

5.2.1. *Proof of Theorem 3.2.* To show the smallness of $x_i^m-z_i^m$, we first rewrite the updating formula for x_i^m, (35), into

$$x_i^{m+1}=x_i^m-\nabla f(x_i^m)h+E_i^m h+\sqrt{2h}\xi_i^m, \tag{52}$$

where

$$E_i^m=\nabla f(x_i^m)-F_i^m. \tag{53}$$

Comparing the updating formula of z_i^m in equation (36), it is easy to see that the key lies in bounding the term E_i^m. This is shown in the following lemma.

Lemma 5.1. *Under the same conditions of Theorem 3.2, we have: for any $m\geq 0$, $1\leq i\leq N$*

$$\mathbb{E}\,|E_i^m|\lesssim\sqrt{\frac{R_1^dL(M_f-f^*)d^2}{\eta^dN^*}}\exp\left(\frac{R_2(R_2+R_1)}{2h}\right)+L\eta d. \tag{54}$$

Theorem 3.2 is a direct consequence from this lemma.

Proof of Theorem 3.2. For each $m\geq 0$, $1\leq i\leq N$, we subtract (52) and (36) to obtain

$$\mathbb{E}\left|x_i^{m+1}-z_i^{m+1}\right|=\mathbb{E}\left|(x_i^m-z_i^m)-h(\nabla f(x_i^m)-\nabla f(z_i^m))\right|+h\mathbb{E}|E_i^m|. \tag{55}$$

Noting that ∇f is L-Lipschitz continuous,

$$|\nabla f(x_i^m)-\nabla f(z_i^m))|\leq Lh\,|x_i^m-z_i^m|,$$

then

$$|(x_i^m-z_i^m)-h(\nabla f(x_i^m)-\nabla f(z_i^m))|\leq(1+Lh)\,|x_i^m-z_i^m|.$$

We take the expectation, and utilize Lemma 5.1:

$$\begin{aligned}\mathbb{E}\left|x_i^{m+1}-z_i^{m+1}\right|\leq&(1+Lh)\mathbb{E}\,|x_i^m-z_i^m|\\&+h\left(\sqrt{\frac{R_1^dL(M_f-f^*)d^2}{\eta^dN^*}}\exp\left(\frac{R_2(R_2+R_1)}{2h}\right)+L\eta d\right).\end{aligned}$$

Use this formula iteratively, we have:

$$\begin{aligned}\mathbb{E}\,|x_i^m - z_i^m| \leq & (1+Lh)^m \mathbb{E}\,|x_0^m - z_0^m| \\ & + (1+Lh)^m \left(\sqrt{\frac{R_1^d (M_f - f^*) d^2}{L\eta^d N^*}} \exp\left(\frac{R_2(R_2+R_1)}{2h} \right) + \eta d \right) .\end{aligned}$$

Noting $x_0^m = z_0^m$, the first term is eliminated, and we conclude (37). When f is μ-convex,

$$\nabla f(x_i^m) - \nabla f(z_i^m) \geq \mu (x_i^m - z_i^m) ,$$

then for h small enough:

$$|(x_i^m - z_i^m) - h(\nabla f(x_i^m) - \nabla f(z_i^m))| \leq (1 - \mu h)\, |x_i^m - z_i^m| .$$

Running the same argument as above, and relaxing $(1-\mu h)^m \leq 1$, we conclude (38). □

We now prove Lemma 5.1

Proof of Lemma 5.1. We first define:

$$G_i^m = \begin{cases} \nabla f(x_i^m), & \sqrt{2h}|\xi_i^{m-1}| > R_1 \text{ or } f(x_i^m) > M_f \text{ or } N^* > N_i^m \\ \dfrac{1}{N_i^m} \displaystyle\sum_{j \neq i}^N G_{ij}^m , & \text{otherwise} . \end{cases} \tag{56}$$

where

$$G_{ij}^m = \alpha_d \frac{\langle \nabla f(x_i^m), \delta x_{ij}^m \rangle}{|\delta x_{ij}^m|^2} \frac{\mathbf{1}_{|\delta x_{ij}^m| \leq \eta\,, |\delta w_{ij}^m| \leq R_2}}{\mathrm{p}_j^m} \delta x_{ij}^m$$

is the counterpart of F_{ij}^m that eliminates the discretization error. Then

$$|E_i^m| = |\nabla f(x_i^m) - F_i^m| \leq |\nabla f(x_i^m) - G_i^m| + |G_i^m - F_i^m| .$$

Clearly the term $|\nabla f(x_i^m) - G_i^m|$ is the ensemble error and the term $|G_i^m - F_i^m|$ takes care of the discretization error.

To control $|G_i^m - F_i^m|$, we define

$$\mathbf{1}_{\Omega_i} = \mathbf{1}_{|N_i^m| \geq N^*} \mathbf{1}_{f(x_i^m) \leq M_f} \mathbf{1}_{\sqrt{2h}|\xi_i^{m-1}| \leq R_1} ,$$

then

$$\begin{aligned} \mathbb{E}\left(|G_i^m - F_i^m| \middle| \mathcal{F}^{m-1} \right) &= \mathbb{E}\left(\mathbf{1}_{\Omega_i} |G_i^m - F_i^m| \middle| \mathcal{F}^{m-1} \right) \\ &\leq \mathbb{E}\left(\frac{\mathbf{1}_{\Omega_i}}{N_i^m} \sum_{j \neq i}^N \left| G_{i,j}^m - F_{i,j}^m \right| \middle| \mathcal{F}^{m-1} \right) \\ &= \frac{1}{N_i^m} \sum_{j \neq i}^N \mathbb{E}\left(\mathbf{1}_{\Omega_i} \left| G_{i,j}^m - F_{i,j}^m \right| \middle| \mathcal{F}^{m-1} \right) \\ &\leq \max_{1 \leq j \leq N} \mathbb{E}\left(\left| G_{i,j}^m - F_{i,j}^m \right| \middle| \mathcal{F}^{m-1} \right) . \end{aligned} \tag{57}$$

Plugging (17) into (57), we obtain

$$\mathbb{E}\left(|G_i^m - F_i^m| \right) = \mathbb{E}\left(\mathbb{E}\left(|G_i^m - F_i^m| \middle| \mathcal{F}^{m-1} \right) \right) \leq L\eta d . \tag{58}$$

To control $|G_i^m - \nabla f(x_i^m)|$. We note

$$\mathbb{E}\left(|G_i^m - \nabla f(x_i^m)|^2\right) = \mathbb{E}\left(\mathbb{E}\left(\mathbf{1}_{\Omega_i}\left|G_i^m - \nabla f(x_i^m)\right|^2 \middle| \mathcal{F}^{m-1}\right)\right). \tag{59}$$

Define

$$\mathcal{E}_{i,j}^m = G_{i,j}^m - \nabla f(x_i^m)\mathbf{1}_{|\delta w_{ij}^m|<R_2},$$

then

$$\begin{aligned}
&\mathbb{E}\left(|G_i^m - \nabla f(x_i^m)|^2\right)\\
=&\mathbb{E}\left(\mathbb{E}\left(\mathbf{1}_{\Omega_i}\left|\frac{1}{N_i^m}\sum_{j\neq i}\left[G_{ij}^m - \nabla f(x_i^m)\mathbf{1}_{|\delta w_{ij}^m|<R_2}\right]\right|^2 \middle| \mathcal{F}^{m-1}\right)\right)\\
\leq&\mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}_{\Omega_i}}{(N_i^m)^2}\left|\sum_{j\neq i} G_{ij}^m - \nabla f(x_i^m)\mathbf{1}_{|\delta w_{ij}^m|<R_2}\right|^2 \middle| \mathcal{F}^{m-1}\right)\right)\\
=&\mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}_{\Omega_i}}{(N_i^m)^2}\left|\sum_{j\neq i}\mathcal{E}_{i,j}^m\right|^2 \middle| \mathcal{F}^{m-1}\right)\right)\\
\leq&\frac{1}{N^*}\mathbb{E}\left(\left\{\max_j \mathbb{E}\left(\mathbf{1}_{\Omega_i}\left|\mathcal{E}_{i,j}^m\right|^2 \middle| \mathcal{F}^{m-1}\right) + \sum_{j_1\neq j_2}^{N}\mathbb{E}\left(\mathbf{1}_{\Omega_i}\left\langle\mathcal{E}_{i,j_1}^m, \mathcal{E}_{i,j_2}^m\right\rangle \middle| \mathcal{F}^{m-1}\right)\right\}\right)\\
=&\frac{1}{N^*}\mathbb{E}\left(\max_j \mathbb{E}\left(\mathbf{1}_{\Omega_i}\left|\mathcal{E}_{i,j}^m\right|^2 \middle| \mathcal{F}^{m-1}\right)\right),
\end{aligned} \tag{60}$$

where we use $N_i^m = \sum_{j\neq i}^N \mathbf{1}_{|\delta w_{ij}^m|<R_2}$ in the first equality.

In the last equation, we note that

$$\mathbb{E}\left(\mathbf{1}_{\Omega_i}\mathcal{E}_{i,j}^m \middle| \mathcal{F}^{m-1}, x_i^m\right) = \mathbf{1}_{\Omega_i}\mathbb{E}\left(\mathcal{E}_{i,j}^m \middle| \mathcal{F}^{m-1}, x_i^m\right) = 0, \tag{61}$$

with the conditional independence, and thus

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_{\Omega_i}\left\langle\mathcal{E}_{i,j_1}^m, \mathcal{E}_{i,j_2}^m\right\rangle \middle| \mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(\mathbb{E}\left(\mathbf{1}_{\Omega_i}\left\langle\mathcal{E}_{i,j_1}^m, \mathcal{E}_{i,j_2}^m\right\rangle \middle| \mathcal{F}^{m-1}, x_i^m\right) \middle| \mathcal{F}^{m-1}\right)\\
=&\mathbb{E}\left(\left\langle\mathbb{E}\left(\mathbf{1}_{\Omega_i}\mathcal{E}_{i,j_1}^m \middle| \mathcal{F}^{m-1}, x_i^m\right), \mathbb{E}\left(\mathbf{1}_{\Omega_i}\mathcal{E}_{i,j_2}^m \middle| \mathcal{F}^{m-1}, x_i^m\right)\right\rangle \middle| \mathcal{F}^{m-1}\right)\\
=&0
\end{aligned}$$

To further control (60) we simply use the direct calculation: for any $j \neq i$

$$
\begin{aligned}
&\mathbb{E}\left(\mathbf{1}_{\Omega_i}\left|\mathcal{E}_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right) \leq \mathbb{E}\left(\mathbf{1}_{\Omega_i}\left|G_{i,j}^m\right|^2\middle|\mathcal{F}^{m-1}\right)\\
\leq & \alpha_d^2 \mathbf{1}_{|\delta w_{ij}^m|<R_2} \int_{B(w_i^m,R_1)} \int_{B(x_i^m,\eta)} \frac{|\nabla f(x_i^m)|^2}{p_j^m(x_j^m)} p_i^m(x_i^m)\,\mathrm{d}x_j^m\,\mathrm{d}x_i^m\\
\overset{\text{(I)}}{\lesssim} & C' \int_{B(w_i^m,R_1)} \int_{B(x_i^m,\eta)} \mathbf{1}_{|\delta w_{ij}^m|<R_2} \exp\left(\frac{|x_j^m - w_j^m|^2}{4h} - \frac{|x_i^m - w_i^m|^2}{4h}\right)\mathrm{d}x_j^m\,\mathrm{d}x_i^m\\
\overset{\text{(II)}}{\lesssim} & C' \int_{B(w_i^m,R_1)} \int_{B(0,\eta)} \mathbf{1}_{|\delta w_{ij}^m|<R_2} \exp\left(\frac{|y+z - w_j^m|^2}{4h} - \frac{|y - w_i^m|^2}{4h}\right)\mathrm{d}z\,\mathrm{d}y\\
\overset{\text{(III)}}{\lesssim} & C' \exp\left(\frac{\eta^2 + 2(\eta R_1 + \eta R_2 + R_2 R_1) + R_2^2}{4h}\right) \int_{B(w_i^m,R_1)} \int_{B(0,\eta)} \mathrm{d}z\,\mathrm{d}y\\
= & \frac{C' R_1^d d^2}{\eta^d \alpha_d^2} \exp\left(\frac{\eta^2 + 2(\eta R_1 + \eta R_2 + R_2 R_1) + R_2^2}{4h}\right),
\end{aligned}
\tag{62}
$$

where $C' = L(M_f - f^*)\alpha_d^2$. Here in (I) we used $\frac{1}{2L}|\nabla f(x_i^m)|^2 \leq f(x_i^m) - f^* < (M_f - f^*)$, in (II) we used change of variables $y = x_i^m, z = x_j^m - x_i^m$. In (III), we used:

$$
\begin{aligned}
&\exp\left(\frac{|y+z - w_j^m|^2}{4h} - \frac{|y - w_i^m|^2}{4h}\right)\\
=&\exp\left(\frac{|y - w_i^m + z + w_i^m - w_j^m|^2}{4h} - \frac{|y - w_i^m|^2}{4h}\right)\\
=&\exp\left(\frac{|z + w_i^m - w_j^m|^2}{4h} + \frac{\langle y - w_i^m, z + w_i^m - w_j^m\rangle}{2h}\right)\\
\lesssim&\exp\left(\frac{|z|^2}{4h} + \frac{|w_i^m - w_j^m|^2}{4h} + \frac{|z||w_i^m - w_j^m|}{2h} + \frac{|y - w_i^m|\left(|z| + |w_i^m - w_j^m|\right)}{2h}\right).
\end{aligned}
$$

Plug (62) into (60), we have

$$
\mathbb{E}\left(|G_i^m - \nabla f(x_i^m)|^2\right) \lesssim \frac{R_1^d d^2 L(M_f - f^*)}{N^* \eta^d} \exp\left(\frac{\eta^2 + 2(\eta R_1 + \eta R_2 + R_2 R_1) + R_2^2}{4h}\right). \tag{63}
$$

Using $\eta < R_2$ and Hölder inequality we have

$$
\begin{aligned}
\mathbb{E}\left(|G_i^m - \nabla f(x_i^m)|\right) &= \left(\mathbb{E}\left(|G_i^m - \nabla f(x_i^m)|^2\right)\right)^{1/2}\\
&\lesssim \sqrt{\frac{R_1^d d^2 L(M_f - f^*)}{N^* \eta^d}} \exp\left(\frac{R_2(R_2 + R_1)}{2h}\right).
\end{aligned}
$$

Combine it with (58) we prove (54). □

5.2.2. *Proof of Theorem 3.3.* The validity of Theorem 3.3 is built upon the fact that x_i^m system and z_i^m system are close, shown above, and that the z_i^m system follows LMC, which converges to the target distribution.

It is a classical result to show that the LMC solution converges. To do so, one constructs another particle system that is drawn from the target distribution. Let

y_0 be a random vector drawn from target distribution induced by p, and set

$$y_i(t) = y_i^0 - \int_0^t \nabla f(y_i(s))\,\mathrm{d}s + \sqrt{2}\int_0^t \mathrm{d}B_i(s)\,, \tag{64}$$

where we construct Brownian motion that satisfies:

$$B_i(h(m+1)) - B_i(hm) = \sqrt{h}\xi_i^m\,. \tag{65}$$

Then $y_i(t)$ is drawn from the distribution induced by p as well. On the discrete level, let $y_i^m = y_i(hm)$, then:

$$y_i^{m+1} = y_i^m - \int_{mh}^{(m+1)h} \nabla f(y_i(s))\,\mathrm{d}s + \sqrt{2h}\xi_i^m\,. \tag{66}$$

Since $y_i^m \sim p(x)$, then we have

$$W_1(q_i^m, p) \le \mathbb{E}|x_i^m - y_i^m|\,,$$

where $\mathbb{E}$ takes all randomness into account. Choose the initial data y_0 so that $W_1(q^0,p) = \mathbb{E}|x_i^0 - y_i^0|$. Then the problem boils down to showing that x_i^m is close to y_i^m. Since we already know that x_i^m and z_i^m are close, we now need to show the closeness between z and y. This classical result regarding the convergence of LMC was shown in [3, 5], and we cite it here for the completeness of the paper (with notations adjusted to our setting).

Proposition 2 (Closeness of z and y)**.** *Assume conditions of Theorem 3.2, and let f be L-smooth and μ convex with $\kappa = L/\mu$, we have: for any $m \ge 0$, $1 \le i \le N$*

$$\mathbb{E}|z_i^m - y_i^m| \le \exp\left(-\frac{\mu h m}{2}\right) W_1(q^0,p) + \mathcal{O}\left(\kappa\sqrt{hd}\right)\,. \tag{67}$$

We leave the proof to Appendix A. We should emphasize that this result is essentially the same as the one in [5, 14, 6]. The only difference is that we use L_1 norm for bounding $z_i^m - y_i^m$ for the consistency with the result in Theorem 3.2.

Now, we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. Combining Theorem 3.2 and Proposition 2 by adding (38) and (67) through the triangle inequality, we obtain

$$\begin{aligned}\mathbb{E}|x_i^m - y_i^m| \le& \mathbb{E}|x_i^m - z_i^m| + \mathbb{E}|z_i^m - y_i^m| \\ =& \exp\left(-\frac{\mu h m}{2}\right) W_1(q^0,p) \\ &+ \mathcal{O}\left(\kappa(\sqrt{hd} + \eta d) + \sqrt{\frac{R_1^d \kappa d^2 (M_f - f^*)}{\mu \eta^d N^*}} \exp\left(\frac{R_2(R_2+R_1)}{2h}\right)\right)\,.\end{aligned} \tag{68}$$

Since $W_1(q_i^m, p) \le \mathbb{E}|x_i^m - y_i^m|$, we prove (41). To prove (42), we use

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N g(x_i^m) - \mathbb{E}_p(g)\right| \le \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left|g(x_i^m) - g(y_i^m)\right| + \mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N g(y_i^m) - \mathbb{E}_p(g)\right|\,. \tag{69}$$

Using the Lipschitz continuity, the first term is easily controlled.

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left|g(x_i^m) - g(y_i^m)\right| \le \mathcal{O}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}|x_i^m - y_i^m|\right)\,. \tag{70}$$

Here the $\mathcal{O}$ notation includes the Lipschitz constant of g. The second term of (69) is a standard central limit theorem:

$$\begin{aligned}\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} g(y_i^m) - \mathbb{E}_p(g)\right| &\le \left(\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} g(y_i^m) - \mathbb{E}_p(g)\right)^2\right)^{1/2} \\ \le \left(\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left(g(y_i^m) - \mathbb{E}_p(g)\right)^2\right)^{1/2} &\le \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).\end{aligned} \tag{71}$$

Combining (68), (70) and (71) into (69), we prove the weak convergence (42). □

5.3. **Proof of Theorem 3.4.** We prove Theorem 3.4 in this section. First, we give another iteration lemma:

Lemma 5.2. *Under conditions of Theorem 3.2, let $m \ge 0$, and $\epsilon_m > 0$. Then, there exists a constant N' that is independent of η, ϵ_m such that if*

$$N > N', \quad \mathbb{E}|x_i^m - z_i^m| \le \epsilon_m, \quad \forall 1 \le i \le N$$

we have

$$\mathbb{E}|x_i^{m+1} - z_i^{m+1}| \le \epsilon_m + \frac{B(\epsilon_m)}{\eta^{d/2}} + C\eta, \quad \mathbb{P}\left(N_i^m \le N^*\right) \le 1 - B(\epsilon_m), \quad \forall 1 \le i \le N. \tag{72}$$

where C is a constant and $B : \mathbb{R} \to \mathbb{R}^+$ is a continuous function that satisfies

$$\lim_{\epsilon_m \to 0} B(\epsilon_m) = 0.$$

Remark 5. We note that in Lemma 5.2, the constants N', C and function B depend on other parameters such as $h, d, R_2, R_1, M_f, N^*, \mu, L$.

Proof of Lemma 5.2. Without loss of generality, we only consider $|x_1^m - z_1^m|$ and N_1^m. Similar to the argument in Lemma 5.1,

$$\mathbb{E}\,|E_1^m| \lesssim \sqrt{\frac{R_1^d L(M_f - f^*)d^2}{\eta^d}} \exp\left(\frac{R_2(R_2+R_1)}{2h}\right)\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right) + L\eta d.$$

According to the proof of Theorem 3.2, we obtain

$$\begin{aligned}\mathbb{E}\left|x_1^{m+1} - z_1^{m+1}\right| \le& (1-\mu h)\mathbb{E}\,|x_1^m - z_1^m| \\ &+ h\left(\sqrt{\frac{R_1^d L(M_f - f^*)d^2}{\eta^d}} \exp\left(\frac{R_2(R_2+R_1)}{2h}\right)\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right) + L\eta d\right) \\ \le& \epsilon_m + \frac{C}{\eta^{d/2}}\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right) + C\eta,\end{aligned} \tag{73}$$

where C is a constant that is independent of η and ϵ_m. Thus, it suffices to bound $\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right)$. Define

$$\widetilde{w}_i^m = z_i^m - h\nabla f(z_i^m), \quad \widetilde{N}_1^m = \sum_{j>i}^{N_z} \mathbf{1}_{|\delta\widetilde{w}_{ij}^m| < R_2/4},$$

where $N_z < N$ is a positive integer. According to [40], the KL divergence between the distribution of z_i^m and target distribution is finite for all m. This implies the distribution of z_i^m has a density. Thus, for any $M > 0$, we have

$$\lim_{N_z \to \infty} \mathbb{P}\left(\widetilde{N}_i^m > M\right) = 1\,. \tag{74}$$

Now, we start bounding $\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right)$. Since $\mathbb{E}|x_i^m - z_i^m| \le \epsilon_m$,

$$\mathbb{P}\left(|x_i^m - z_i^m| > \frac{R_2}{4}\right) \le \frac{4\epsilon_m}{R_2}\,, \quad \forall 1 \le i \le N\,.$$

which implies

$$\mathbb{P}\left(\cap_{i=1}^{N_z}\left\{|x_i^m - z_i^m| \le \frac{R_2}{4}\right\}\right) \ge 1 - \frac{4\epsilon_m N_z}{R_2}\,. \tag{75}$$

According to the definition of N_i^m (32), using (75), we obtain that for any $M < N_z$

$$\mathbb{P}\left(N_i^m > M\right) \ge \mathbb{P}\left(\widetilde{N}_i^m > M\right) - \frac{4\epsilon_m N_z}{R_2}\,. \tag{76}$$

From this,

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right) \le \frac{1}{\sqrt{M}}\left(\mathbb{P}\left(\widetilde{N}_i^m > M\right) - \frac{4\epsilon_m N_z}{R_2}\right) \\ + \frac{1}{\sqrt{N^*}}\left[1 - \left(\mathbb{P}\left(\widetilde{N}_i^m > M\right) - \frac{4\epsilon_m N_z}{R_2}\right)\right]\,.\end{aligned} \tag{77}$$

Define the right-side of (77) as $F(M, N_z, \epsilon_m)$. Since M, N_z can be arbitrarily chosen, we have

$$\mathbb{E}\left(\frac{1}{\sqrt{N_1^m}}\right) \le \inf_{M, N_z} F(M, N_z, \epsilon_m)$$

Plugging this into (73),

$$\mathbb{E}\left|x_1^{m+1} - z_1^{m+1}\right| \le \epsilon_m + \frac{C}{\eta^{d/2}} \inf_{M, N_z} F(M, N_z, \epsilon_m) + C\eta\,.$$

Noticing that

$$\lim_{M \to \infty} \lim_{N_z \to \infty} \lim_{\epsilon_m \to 0} F(M, N_z, \epsilon_m) = 0\,, \tag{78}$$

we obtain the first inequality of (72). Next, for any $M > N^*$, because

$$\mathbb{P}\left(N_i^m > N^*\right) \ge \mathbb{P}\left(N_i^m > M\right) \ge \mathbb{P}\left(\widetilde{N}_i^m > M\right) - \frac{4\epsilon_m N_z}{R_2} \ge 1 - \sqrt{N^*} F(M, N_z, \epsilon_m)\,,$$

(78) also implies the second inequality of (72). □

Now, we are ready to prove the theorem:

Proof of Theorem 3.4. Noticing that when $m = 0$,

$$\mathbb{E}|x_i^0 - z_i^0| = 0\,.$$

Using Lemma 5.2 (72), for any $\epsilon > 0$, we have

$$\lim_{\eta \to 0} \lim_{N \to \infty} \mathbb{E}|x_i^1 - z_i^1| < \epsilon, \quad \lim_{\eta \to 0} \lim_{N \to \infty} \mathbb{P}\left(\left\{N_i^1 < N^*\right\}\right) > 1 - \epsilon\,.$$

Repeating this process with Lemma 5.2, we obtain

$$\lim_{\eta \to 0} \lim_{N \to \infty} \sup_{0 \le m \le M, 1 \le i \le N} \mathbb{E}|x_i^m - z_i^m| = 0\,. \tag{79}$$

Next, to prove (44), we notice that for $m \geq 0$ and $1 \leq i \leq N$

$$x_i^m - w_i^m = \sqrt{2h}\xi_i^{m-1},$$

which implies

$$\begin{aligned}
&\mathbb{P}\left(\{|x_i^m - w_i^m| > R_1\}\right) = \mathbb{P}\left(\left\{|\xi_i^{m-1}| > \frac{R_1}{\sqrt{2h}}\right\}\right) \\
=&\int_{|x|>\frac{R_1}{\sqrt{2h}}} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{|x|^2}{2}\right) \mathrm{d}x = \frac{S_d}{(2\pi)^{d/2}} \int_{\frac{R_1}{\sqrt{2h}}}^{\infty} r^{d-1} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r \\
\leq&\frac{S_d}{(2\pi)^{d/2}} \int_{\frac{R_1\sqrt{d}}{\sqrt{2}}}^{\infty} r^{d-1} \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r\,,
\end{aligned}$$

where the last inequality comes from $h < \frac{1}{d}$.

Then, to prove (45), we first use $f(x_i^m) - f^* \leq \frac{1}{2\mu}|\nabla f(x_i^m)|^2$ to obtain

$$\begin{aligned}
&\mathbb{P}\left(\{f(x_i^m) - f^* > (M_f - f^*)\}\right) \\
=&\mathbb{P}\left(\{f(x_i^m) - f^* > (M_f - f^*)\}\right) \leq \mathbb{P}\left(\left\{|\nabla f(x_i^m)|^2 > 2\mu(M_f - f^*)\right\}\right) \\
\leq&\mathbb{P}\left(\left\{|\nabla f(y_i^m)|^2 + |\nabla f(x_i^m) - \nabla f(y_i^m)|^2 > \mu(M_f - f^*)\right\}\right) \\
\leq&\mathbb{P}\left(\left\{|\nabla f(x_i^m) - \nabla f(y_i^m)|^2 > \frac{\mu(M_f - f^*)}{2}\right\}\right) + \mathbb{P}\left(\left\{|\nabla f(y_i^m)|^2 > \frac{\mu(M_f - f^*)}{2}\right\}\right),
\end{aligned} \tag{80}$$

where y_i^m is defined in (64)-(66) and we use $2|a-b|^2 + 2|b|^2 \geq |a|^2$ in the second inequality.

The second term of (80) is easy to bound:

$$\mathbb{P}\left(\left\{|\nabla f(y_i^m)|^2 > \mu\sqrt{N^*}/2\right\}\right) \leq \frac{2}{\mu(M_f - f^*)}\mathbb{E}\left(|\nabla f(y_i^m)|^2\right) \leq \frac{2\kappa d}{(M_f - f^*)}, \tag{81}$$

where we use $\mathbb{E}_p|\nabla f(y)|^2 \leq Ld$ according to Lemma 3 in [5].

The first term can be bounded by

$$\begin{aligned}
&\mathbb{P}\left(\left\{|\nabla f(x_i^m) - \nabla f(y_i^m)|^2 > \frac{\mu(M_f - f^*)}{2}\right\}\right) \leq \mathbb{P}\left(\left\{|x_i^m - y_i^m|^2 > \frac{\mu(M_f - f^*)}{2L^2}\right\}\right) \\
\leq&\mathbb{P}\left(\left\{|x_i^m - y_i^m| > \frac{(\mu(M_f - f^*))^{1/2}}{\sqrt{2}L}\right\}\right) \leq \sqrt{\frac{2\kappa L}{(M_f - f^*)}}\mathbb{E}(|x_i^m - y_i^m|)
\end{aligned} \tag{82}$$

where we use $|\nabla f(x_i^m) - \nabla f(y_i^m)| \leq L|x_i^m - y_i^m|$ in the first inequality. Plugging (81) and (82) into right-side of (80), we prove (45) by (79).

Finally, (46) is a direct result of (79) and the second inequality in Lemma 5.2 (72). □

6. **Conclusion.** In this article, we look for the ensemble modification to the classical Langevin Monte Carlo method. This is to look for an ensemble of samples who can be viewed as i.i.d. samples drawn from a target distribution. As a modification to LMC, the gradient information is obtained by taking the average of the neighboring function evaluations, as a mean to achieve the gradient-free property. In this process of surrogation, we found two sides of the theory:

- By directly surrogating the gradient using the ensemble approximation, we develop Ensemble Langevin Monte Carlo, see Algorithm 1. We show that this method is unstable due to a potentially small denominator that induces high

variances. We provide a counterexample to explicitly show this instability in Theorem 3.1. The discovery is discussed in Section 3.1.

- We then change the strategy and enact the ensemble approximation to the gradient only in a constrained manner, to eliminate the unstable points. The algorithm is termed Constrained Ensemble Langevin Monte Carlo, see Algorithm 2. We show that, with a proper tuning, the surrogation takes place often enough to bring the reasonable numerical saving, while the induced error is still low enough for us to maintain the fast convergence rate, up to a controllable discretization and ensemble error. These properties are summarized in Section 3.3.

Such combination of ensemble method and LMC shed light on inventing gradient-free algorithms that produce i.i.d. samples almost exponentially fast. Numerical experiments are collected to demonstrate such accuracy and numerical savings.

We should note, however, the ensemble approximation to the gradient used in this article may not be the optimal one. There are potentially other means to extract properties from ensembles that permit accurate and efficient gradient evaluations. What are the optimal way to achieve gradient-free property is an ultimate task that we hope to resolve in the future.

Appendix A. **Proof of Proposition 2.** In this section, we prove Proposition 2. For convenience, we ignore i and define

$$\Delta^m = z^m - y^m \,.$$

Then it suffices to prove the smallness of $\mathbb{E}|\Delta^m|$.

Proof of Proposition 2. We first divide Δ^{m+1} into several parts:

$$\begin{aligned}
\Delta^{m+1} =& \Delta^m + (y^{m+1} - y^m) - (z^{m+1} - z^m) \\
=& \Delta^m + \left(- \int_{mh}^{(m+1)h} \nabla f(y(s)) \, \mathrm{d}s + \sqrt{2h}\xi_m \right) \\
& - \left(- \int_{mh}^{(m+1)h} \nabla f(z^m) \, \mathrm{d}s + \sqrt{2h}\xi_m \right) \\
=& \Delta^m - \left(\int_{mh}^{(m+1)h} (\nabla f(y(s)) - \nabla f(z^m)) \, \mathrm{d}s \right) \qquad (83) \\
=& \Delta^m - \left(\int_{mh}^{(m+1)h} (\nabla f(y(s)) - \nabla f(y^m) + \nabla f(y^m) - \nabla f(z^m)) \, \mathrm{d}s \right) \\
=& \Delta^m - h \left(\nabla f(y^m) - \nabla f(z^m) \right) - \int_{mh}^{(m+1)h} (\nabla f(y(s)) - \nabla f(y^m)) \, \mathrm{d}s \\
=& \Delta^m - hU^m - V^m
\end{aligned}$$

where

$$\begin{aligned}
U^m &= \nabla f(y^m) - \nabla f(z^m) \,, \\
V^m &= \int_{mh}^{(m+1)h} (\nabla f(y(s)) - \nabla f(y^m)) \, \mathrm{d}s \,.
\end{aligned}$$

Now the first two terms of (83) can be bounded by

$$|\Delta^m - hU^m| \le (1 - \mu h) \, |\Delta^m| \,, \qquad (84)$$

where we use f is μ-convex.

Next, for the second term on the right-hand side of (83), we first bound L^2-norm:

$$\begin{aligned}
\mathbb{E}\left(|V^m|^2\right) &\overset{\text{(I)}}{\leq} h\int_{mh}^{(m+1)h}\mathbb{E}\left(|\nabla f(y(s))-\nabla f(y^m)|^2\right)\mathrm{d}s\\
&\overset{\text{(II)}}{\leq} hL^2\int_{mh}^{(m+1)h}\mathbb{E}\left(|y(s)-y^m|^2\right)\mathrm{d}s\\
&= hL^2\int_{mh}^{(m+1)h}\mathbb{E}\left(\left|\int^s -mh\nabla f(y(t))\,\mathrm{d}t+\sqrt{2}(B(s)-B(nh))\right|^2\right)\mathrm{d}s\\
&\overset{\text{(III)}}{\leq} 2h^2L^2\int_{mh}^{(m+1)h}\int^s -mh\mathbb{E}\left(|\nabla f(y(t))|^2\right)\mathrm{d}t\,\mathrm{d}s\\
&\quad + 4h^2L^2\int_{mh}^{(m+1)h}\mathbb{E}|\xi^m|^2\,\mathrm{d}s\\
&\overset{\text{(IV)}}{=} h^4L^2\mathbb{E}\left(|\nabla f(y^m)|^2\right)+4h^3L^2d\\
&\overset{\text{(V)}}{=} h^4L^2\mathbb{E}_p|\nabla f|^2+4h^3L^2\overset{\text{(VI)}}{\leq} h^4L^3d+4h^3L^2d\,,
\end{aligned}\tag{85}$$

where (II) comes from L-Lipschitz condition, (I) and (III) come from the use of Young's inequality and Jensen's inequality when we move the $|\cdot|^2$ from outside to inside of the integral, and (IV) and (V) hold true because $y(t)\sim p$ for all t. In (VI) we use $\mathbb{E}_p|\nabla f|^2\leq Ld$ using [5, Lemma 3].

Using Hölder's inequality and $h\leq\frac{1}{L}$, (85) implies

$$\mathbb{E}\left(|V^m|\right)\leq\left(\mathbb{E}\left(|V^m|^2\right)\right)^{1/2}\leq 5h^{3/2}Ld^{1/2}\,.$$

Plugging this and (84) into (83), we obtain

$$\mathbb{E}\left(\left|\Delta^{m+1}\right|\right)\leq\mathbb{E}\left(|\Delta^m-hU^m|\right)+\mathbb{E}\left(|V^m|\right)\leq(1-\mu h)\mathbb{E}\left(|\Delta^m|\right)+5h^{3/2}Ld^{1/2}\,.$$

Using this iteratively and $\mathbb{E}|\Delta^0|=\mathbb{E}|z^0-y^0|=W_1(q^0,p)$, we prove (67). □

Appendix B. **Other choices of ensemble gradient approximation.** The ensemble gradient approximation we present in Section 2.2 is of probability type, namely, we take the ensemble average of finite difference around x^*. There are other ways to find gradient approximations as well, and probably the most straightforward method is to solve a linear algebra problem formulated by the closest d neighbors.

More specifically, let $\eta>0$ and $x^*\in\mathbb{R}^d$. Assume that there are d points $\{x_i\}_{i=1}^d$ in the ball $B_\eta(x^*)$, then we have

$$\Delta_x\cdot\nabla f(x^*)=\Delta_f+o(\eta)\,,$$

where

$$\Delta_x=\begin{bmatrix}(x_1-x^*)^\top\\(x_2-x^*)^\top\\\cdots\\(x_d-x^*)^\top\end{bmatrix},\quad \Delta_f=\begin{bmatrix}f(x_1)-f(x^*)\\f(x_2)-f(x^*)\\\cdots\\f(x_d)-f(x^*)\end{bmatrix}.\tag{86}$$

If Δ_x is full rank, then by solving the equation $\Delta_x\cdot z=\Delta_f$, we obtain an approximation of the gradient

$$z\approx\nabla f(x^*)\,.$$

A natural question to ask is, how likely is it to find d neighbors in a small neighborhood of a given sample? To quantify such probability, we use the following lemma:

Lemma B.1. *Suppose $|p(x)| \le M < \infty$ and $\{x_i\}_{i=1}^N$ are i.i.d. drawn from p with $N > 0$. Let $N = c/\eta^d$, where c is a positive constant. Then we have*

$$\limsup_{\eta\to 0} \mathbb{P}\left(\#\left\{x_i \middle| |x_i - x_1| < \eta,\ i = 1, 2, \dots, N\right\} \ge d+1\right) \le 1 - \exp\left(-cM\right) .$$

This lemma can be viewed as a negative result: even with N exponentially big on d, there is still a nontrivial chance for a sample to not have enough neighbors around for the gradient computation.

Proof of Lemma B.1. Fixed $x_1 \in \mathbb{R}^d$,

$$\mathbb{P}\left(|x_2 - x_1| < \eta | x_1\right) = \int_{|z|<\eta} p(x_1 + z)\, \mathrm{d}z \le \eta^d M .$$

Denote $p = \mathbb{P}\left(|x_2 - x_1| < \eta | x_1\right)$. Because $\{x_i\}_{i=1}^N$ are independent, we have

$$\begin{aligned}
&\mathbb{P}\left(\#\left\{x_i \middle| |x_i - x_1| < \eta,\ i = 1, 2, \dots, N\right\} < d+1 | x_1\right) \\
=&\sum_{k=1}^{d} \mathbb{P}\left(\#\left\{x_i \middle| |x_i - x_1| < \eta,\ i = 1, 2, \dots, N\right\} = k | x_1\right) \\
=&\sum_{k=0}^{d-1} \binom{N-1}{k} p^k (1-p)^{N-1-k} \ge (1-p)^{N-1} .
\end{aligned}$$

Since $c = N\eta^d$,

$$\begin{aligned}
&\limsup_{\eta\to 0} \mathbb{P}\left(\#\left\{x_i \middle| |x_i - x_1| < \eta,\ i = 1, 2, \dots, N\right\} < d+1 | x_1\right) \\
\ge &\limsup_{\eta\to 0} (1-p)^{N-1} \ge \limsup_{\eta\to 0} \left(1 - \eta^d M\right)^{\frac{c}{\eta^d} - 1} \\
\ge &\exp\left(-cM\right) .
\end{aligned}$$

This implies

$$\limsup_{\eta\to 0} \mathbb{P}\left(\#\left\{x_i \middle| |x_i - x_1| < \eta,\ i = 1, 2, \dots, N\right\} < d+1\right) \ge \exp\left(-cM\right) .$$

which concludes the proof. □

REFERENCES

[1] C. Andrieu, N. de Freitas, A. Doucet and M. I. Jordan, An introduction to MCMC for machine learning, *Machine Learning*, **50** (2003), 5–43.

[2] A. Beskos, A. Jasra, K. Law, R. Tempone and Y. Zhou, Multilevel sequential Monte Carlo samplers, *Stochastic Process. Appl.*, **127** (2017), 1417–1440.

[3] N. S. Chatterji, N. Flammarion, Y.-A. Ma, P. L. Bartlett and M. I. Jordan, On the theory of variance reduction for stochastic gradient Monte Carlo, *Proceedings of the 35th international Conference on Machine Learning*, **80** (2018), 764–773. Available from: `http://proceedings.mlr.press/v80/chatterji18a/chatterji18a.pdf`.

[4] A. S. Dalalyan, Theoretical guarantees for approximate sampling from smooth and log-concave densities, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **79** (2017), 651–676.

[5] A. S. Dalalyan and A. Karagulyan, User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient, *Stochastic Process. Appl.*, **129** (2019), 5278–5311.

[6] A. S. Dalalyan and L. Riou-Durand, On sampling from a log-concave density using kinetic Langevin diffusions, *Bernoulli*, **26** (2020), 1956–1988.

[7] Z. Ding and Q. Li, Ensemble Kalman inversion: Mean-field limit and convergence analysis, *Stat. Comput.*, **31** (2021), 21pp.
[8] Z. Ding and Q. Li, Ensemble Kalman sampler: Mean-field limit and convergence analysis, *SIAM J. Math. Anal.*, **53** (2021), 1546–1578.
[9] Z. Ding and Q. Li, Langevin Monte Carlo: Random coordinate descent and variance reduction, *J. Mach. Learn. Res.*, **22** (2021), 51pp.
[10] Z. Ding and Q. Li, Variance reduction for random coordinate descent-Langevin Monte Carlo, *Proceedings of the 34th Conference on Neural Information Processing Systems*, **33** (2020), 3748–3760. Available from: https://proceedings.neurips.cc/paper/2020/file/272e11700558e27be60f7489d2d782e7-Paper.pdf.
[11] A. Doucet, N. de Freitas and N. Gordon, An introduction to sequential Monte Carlo Methods, in *Sequential Monte Carlo Methods in Practice*, Stat. Eng. Inf. Sci., Springer, New York, 2001, 3–14.
[12] S. Duane, A. D. Kennedy, B. J. Pendleton and D. Roweth, Hybrid Monte Carlo, *Phys. Lett. B*, **195** (1987), 216–222.
[13] A. Durmus, S. Majewski and B. Miasojedow, Analysis of Langevin Monte Carlo via convex optimization, *J. Mach. Learn. Res.*, **20** (2019), 46pp.
[14] A. Durmus and É. Moulines, Non-asymptotic convergence analysis for the unadjusted Langevin algorithm, *Ann. Appl. Probab.*, **27** (2017), 1551–1587.
[15] R. Dwivedi, Y. Chen, M. J. Wainwright and B. Yu, Log-concave sampling: Metropolis-Hastings algorithms are fast, *J. Mach. Learn. Res.*, **20** (2019), 42pp.
[16] G. Evensen, *Data Assimilation. The Ensemble Kalman Filter*, Springer-Verlag, Berlin, 2009.
[17] P. Fabian, Atmospheric sampling, *Adv. Space Res.*, **1** (1981), 17–27.
[18] A. Garbuno-Inigo, F. Hoffmann, W. Li and A. M. Stuart, Interacting Langevin diffusions: Gradient structure and Ensemble Kalman sampler, *SIAM J. Appl. Dyn. Syst.*, **19** (2020), 412–441.
[19] A. Garbuno-Inigo, N. Nüsken and S. Reich, Affine invariant interacting Langevin dynamics for Bayesian inference, *SIAM J. Appl. Dyn. Syst.*, **19** (2020), 1633–1658.
[20] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.*, **6** (1984), 721–741.
[21] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57** (1970), 97–109.
[22] M. Herty and G. Visconti, Continuous limits for constrained ensemble Kalman filter, *Inverse Problems*, **36** (2020), 28pp.
[23] M. A. Iglesias, K. J. H. Law and A. M. Stuart, Ensemble Kalman methods for inverse problems, *Inverse Problems*, **29** (2013), 20pp.
[24] Q. Li and K. Newton, Diffusion equation-assisted Markov chain Monte Carlo methods for the inverse radiative transfer equation, *Entropy*, **21** (2019), 25pp.
[25] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang and J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), *Science*, **368** (2020), 489–493.
[26] R. Li, H. Zha and M. Tao, Sqrt(d) dimension dependence of Langevin Monte Carlo, preprint, 2021, arXiv:2109.03839.
[27] P. A. Markowich and C. Villani, On the trend to equilibrium for the Fokker-Planck equation: An interplay between physics and functional analysis. VI Workshop on Partial Differential Equations, Part II (Rio de Janeiro, 1999), *Mat. Contemp.*, **19** (2000), 1–29.
[28] J. Martin, L. C. Wilcox, C. Burstedde and O. Ghattas, A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion, *SIAM J. Sci. Comput.*, **34** (2012), A1460–A1487.
[29] B. Leimkuhler, C. Matthews and J. Weare, Ensemble preconditioning for Markov chain Monte Carlo simulation, *Stat. Comput.*, **28** (2018), 277–290.
[30] N. R. Nagarajan, M. M. Honarpour and K. Sampath, Reservoir-fluid sampling and characterization–Key to efficient reservoir management, *J. Petroleum Technology*, **59** (2007).
[31] R. M. Neal, Annealed importance sampling, *Stat. Comput.*, **11** (2001), 125–139.
[32] R. M. Neal, Probabilistic inference using Markov chain Monte Carlo methods, Technical Report CRG-TR-93-1. Dept. of Computer Science, University of Toronto, 1993.
[33] N. Nüsken and S. Reich, Note on interacting Langevin diffusions: Gradient structure and ensemble Kalman Sampler by Garbuno-Inigo, Hoffmann, Li and Stuart, preprint, arXiv:1908.10890.

[34] S. Reich, A dynamical systems framework for intermittent data assimilation, *BIT*, **51** (2011), 235–249.
[35] G. O. Roberts and J. S. Rosenthal, General state space Markov chains and MCMC algorithms, *Probab. Surv.*, **1** (2004), 20–71.
[36] G. O. Roberts and O. Stramer, Langevin diffusions and Metropolis-Hastings algorithms. International Workshop in Applied Probability (Caracas, 2002), *Methodol. Comput. Appl. Probab.*, **4** (2002), 337–357.
[37] G. O. Roberts and R. L. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations, *Bernoulli*, **2** (1996), 341–363.
[38] C. Schillings and A. M. Stuart, Analysis of the ensemble Kalman filter for inverse problems, *SIAM J. Numer. Anal*, **55** (2017), 1264–1290.
[39] X. T. Tong, M. Morzfeld and Y. M. Marzouk, MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure, *SIAM J. Sci. Comput.*, **42** (2020), A1765–A1788.
[40] S. S. Vempala and A. Wibisono, Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices, *Proceedings of the 33rd Conference on Neural Information Processing Systems*, **32** (2019). Available from: https://proceedings.neurips.cc/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf.
[41] P. Zhang, Q. Song and F. Liang, A Langevinized ensemble Kalman filter for large-scale static and dynamic learning, preprint, 2021, arXiv:2105.05363.

Received September 2021; revised October 2021; early access December 2021.

E-mail address: zding49@math.wisc.edu
E-mail address: qinli@math.wisc.edu