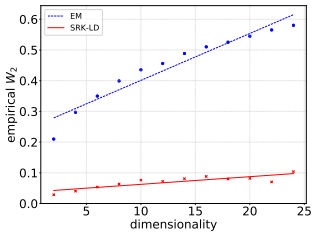
We thank all reviewers for their valuable and constructive feedback. Below are our responses.

Revs. 1,2,3 - Formatting & Citation: We will fix any typos, cite the reference (Durmus et al., 2016), clarify the definitions of $\mathrm{d}B_u^{(i)}$, $\bar{X}_t$, and "smoothness", supply a reference for eq (6), and correct the reference letter case issue.

Revs. 1,3 - Eq (13-14): We will provide a short instructive derivation for eqs (13-14) in the appendix.

Rev. 1 - Sec. 4.2: We will add more clarification on the class of non-convex potentials at the beginning of the section.

Rev. 1 - dimension dependence: Figure on the right shows the asymptotic bias vs dimensionality for a Gaussian mixture problem. We estimated W_2 with the Monte Carlo estimator detailed in App. F and fitted a line with least squares.



Rev. 1 - x-axis in Fig. 1(b): The original motivation was to showcase the evolution of the distance of SRK-LD iterates to the target distribution across iterations. The key point made is that SRK-LD produces a sequence of iterates with lower asymptotic bias than that produced by EM. However, we agree that a refined measure is required to demonstrate the algorithm's performance in terms of computational resource usage. To this end, we performed a wall time analysis, whose results were included in App. G.1.3 of our original submission. To summarize the findings, SRK-LD is 2.5-3 times as costly as EM (on a CPU) per iteration. However, the scheme can still beat EM in terms of total cost if we select a much larger step size for it than that for EM. Recall, Sec. 5.1 and Fig. 1(a) demonstrated that SRK-LD can be run with large step sizes, at which running EM results in divergence.

Rev. 1 - MSE in Fig. 1c: We simulate with EM using a small step size of 10^{-6} until we obtain an initial batch of particles roughly distributed as the target distribution. We evolve this same initial batch in three ways: *(i)* using EM with a small step size of 10^{-6}, *(ii)* using SRK-ID with a step size of 10^{-3}, and *(iii)* using EM with a step size of 10^{-3}. We treat *(i)* as the true continuous-time process and estimate the MSE of EM and SRK-ID at each iteration with the formula $\frac{1}{N}\sum_{i=1}^{N}\|X_t^{(i)} - X_{hk}^{(i)}\|$, where X_t is the continuous-time process computed by *(i)*, and X_{hk} is the discrete-time Markov chain (either *(ii)* or *(iii)*). Here, $hk = t$, h is the step size, and k is the iteration index. The superscript is used to index entries in a batch with N samples. The MSE increases since we started off from an initial batch that is roughly distributed as the target distribution, and the discrete-time Markov chains *(ii)* and *(iii)* each has their own biases. The purpose was to demonstrate that *(ii)* has a smaller bias than *(iii)*. Note the setting was detailed in Section 5.2 and App. G.2 in our original submission. We preferred to show these plots as opposed to error-vs-iteration plots since we found it hard to choose a good initialization such that neither method took very long to converge.

Rev. 2 - condition number dependence in Thm. 2: A convergence bound where the exponent of the Lipschitz constant L in the numerator matches up with the exponent of the strong convexity constant m in the denominator would give us insight into the convergence behaviour at varying worst-case curvature cases. Yet, the two exponents don't match up in our bound (such type of bounds aren't uncommon; see e.g. [17, Thm. 5]). Plugging constants C_2 and C_3 into (35) shows that the dominating term decided by the condition number is $\mathcal{O}(\kappa^{5/3})$, where $\kappa = L/m$. However, our (slightly pessimistic) bound has additional terms dependent on smoothness with the dominant term being $\mathcal{O}(1 \vee L^6)$.

Rev. 2 - uniform deviation: Local deviation orders are well-studied in the SDE literature typically using the Itô-Taylor expansion on the continuous process [38, 42]. In our case, to unroll the recursion and obtain a converging bound, it is instructive to rely on contraction of diffusion, high-order Lipschitz smoothness, and Markov chain moment bounds. We chose the more hands-on approach for tight constants, but note there are established results without explicit constants [42, Lem. 2.2.2]. We believe our approach might be the most straightforward in deriving uniform bounds.

Rev. 2 - benefit with higher-order discretization: Indeed, Metropolis-adjusted algorithms allow exact sampling with fast convergence rates. We thank the reviewer for the pointer to Chen et al., 2019; we will discuss this in the final version. On the other hand, to the best of our knowledge, there has been limited success in modifying these methods to take advantage of data subsampling, which is typically required for scaling to large datasets. By contrast, Bayesian learning has witnessed considerable success in practice with SGLD [54], a gradient-subsampled version of the (asymptotically biased) gradient Langevin dynamics algorithm. We may expect direct discretizations (w/o Metropolis-adjustment) of exponentially contracting diffusions to take advantage of the data subsampling paradigm. A systematic study on this topic warrants an independent investigation. Yet, SRK-LD can be easily adapted to this scenario using existing techniques in the literature; see the next comment for more on this.

Revs. 2,3 - stochastic oracle: Thank you for this recommendation. We assume an oracle model where *(i)* the randomness in the oracle is independent of that of the Brownian motion, *(ii)* the stochastic gradient $\hat{\nabla} f$ is unbiased, i.e. $\mathbb{E}[\hat{\nabla} f(x)] = \nabla f(x)$ for all $x \in \mathbb{R}^d$, and *(iii)* the stochastic gradient $\hat{\nabla} f$ has bounded variance at iterates of the Markov chain and the "interpolated" random variables, i.e. $\mathbb{E}[\|\nabla f(Y)\|_2^2] \leq d\sigma^2$, where Y may be $\tilde{X}_k$, $\tilde{H}_1$ or $\tilde{H}_2$. Compared to the full gradient case, the key differences in this setting are *(a)* Markov chain moment bounds become slightly "inflated", and *(b)* local deviations have extra terms dependent on the variance of the oracle. Note that only *(b)* affects the rate's dependence on ϵ. Under assumptions *(i,ii,iii)*, one may show that SRK-LD achieves the same convergence rate when the variance $d\sigma^2 = \mathcal{O}(\epsilon^2/h)$ for step size h. On the other hand, we recover the rate of $\tilde{\mathcal{O}}(d\epsilon^{-2})$ for gradient Langevin dynamics if the variance is large. The analysis and proof are mostly mechanical, following along similar lines of [12]. We will include a detailed discussion on this result.