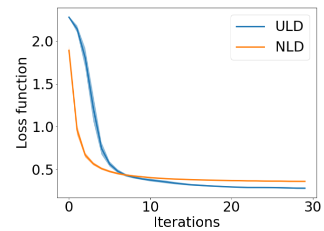
We thank the reviewers for their careful consideration and their feedback, our replies are provided below. We believe that we addressed all the raised issues, the detailed responses are given below.

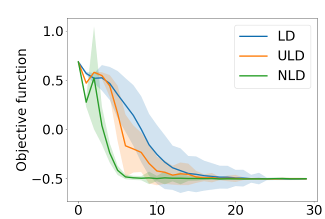
Novelty of our analysis and comparison to [TLR18]: Our analysis is significantly more complicated compared to the LD case in [TLR18] due to non-reversibility, and requires us to develop new estimates, e.g. Lemma 2, where the eigenvalue and the norm estimates require a significant amount of work because the forward iterations correspond to *non-symmetric* matrices H_γ (defined in (2.2)) and achieving the acceleration behavior requires careful estimates. The analysis here also requires us to establish novel uniform L^2 bounds for NLD in both continuous and discrete times. We have also new results and insights about the mean exit times for ULD and NLD, which is a difficult problem to study.



R.1: **(1)** We will add a conclusion section to summarize our paper. **(2)** The name NLD is indeed a bit unfortunate but the name "non-reversible" for such dynamics is standard, see e.g. "Duncan, A.B., Lelièvre, T. & Pavliotis, G.A. Variance Reduction Using Nonreversible Langevin Samplers. *J Stat Phys* 163, 457-491 (2016)". **(3)** When condition number is close to 1 or m large, ULD may not improve upon LD. **(4)** We will add discussions on dependence on other parameters in the revision. We focus on the comparison in m since it is a natural choice. For example, for convergence rate to Gibbs distribution, it is known that for m-strongly convex objectives the continuous-time overdamped Langevin diffusion has rate e^{-mt} in $\mathcal{W}_2$ independent of other parameters. **(5)** Line 220. The improvement will occur if $C_J(\tilde{\varepsilon}) = \tilde{O}(1)$. **(6)** We will add discussions comparing ULD and NLD. For example, when smallest eigenvalue m is close to the largest eigenvalue M, NLD will not improve much upon LD, but ULD will improve upon LD if m is small and ULD will be faster than NLD. The figure on the right is an example for training fully-connected neural networks on MNIST where ULD was faster when both methods were tuned. **(7)** H_γ is defined in (2.2). We will define it earlier than Line 114. We will also define H and M more appropriately. **(8)** We will correct all the typos the reviewer pointed out and add clarifications to all the bullet points.

R.2: We thank the reviewer for the insightful comments. We absolutely agree with the reviewer that smaller value of empirical risk achieved, better generalization will be in general. However, for many interesting problems such as deep learning with modern neural networks, it has been empirically found that most local minima are equivalent in the sense that they lead to similar generalization performance and that finding a global minima may sometimes lead to overfitting (see e.g. the paper "The Loss Surfaces of Multilayer Networks" by Choromanska et al.). Therefore, there is also incentive to find a local minima to achieve good generalization performance where our results would be relevant.

R.3: **(2)** We sincerely apologize for mis-citing Lemma EC.6 in [GGZ18]. Regarding Lemma 14 in [GGZ18], we would like to point out that in our current paper, our ULD uses the Cheng et al. discretization of underdamped Langevin diffusion as in [CCBJ17]; therefore it corresponds to Lemma 18 in [GGZ18] because in [GGZ18] two discretizations of underdamped Langevin diffusion are considered: Euler discretization (Lemma 14), and Cheng et al. discretization (Lemma 18). Indeed, the proof of Lemma 18 provides the bound $O(K\eta^3)$ for KL divergence; see Equation (D.20) on p.63 of [GGZ18]). We will provide a self-contained proof in the revision. **(3)** Carrying out metastability analysis without relying on discretization error is a good idea, and is worth exploring in the future. One reason we follow the current approach is to make it easier for us to compare our results with [TLR18] and show advantage and improvement when breaking the reversibility. If the improvement comes from avoiding discretization error, it might confuse the readers and undermine and main message of our paper. **(4)** In the exit-time part, we do not know any rigorous results for continuous-time Langevin beyond the double well example. Analyzing the behavior of these processes around a saddle point becomes very hard as the surface that contain the saddle point is characteristic, i.e. the drift and the normal to the surface is orthogonal which makes standard boundary layer approaches inapplicable (see Sec 5 of [BR16]) and we agree with the reviewer that exploring beyond the double well example will be a very interesting research direction to pursue which would lead to a major breakthrough in this research area. **(5)** We will cite the papers A. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities" A. Durmus and E. Moulines, "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". Thanks for bringing this up!



R.4: **(2)** Our analysis is for the empirical risk F, however it is straightforward to obtain standard generalization bounds to get results for the population risk $\overline{F}$ by an analysis similar to [TLR18]. **(3)** We agree that it would be interesting to relax the twice continuously differentiability assumption, however we note that due to the difficulty of analyzing ULD and NLD algorithms, even for such smooth functions many basic questions are open such as the mean exit time or a sharp characterization of time it takes to be in a neighorhood of a local minima. **(4-5)** In practice, the matrix J can be chosen as a random anti-symmetric matrix. For quadratic objectives, there is a formula for optimal J matrix (see the paper "Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion" by Lelievre et al, 2013). We can take the parameter $\gamma = 2\sqrt{m}$ as predicted by our theory (Lemma 2) for quadratics. In the figure, we compare ULD and NLD to LD for the double well example with random initialization over 100 runs where J is chosen randomly and $\gamma = 2\sqrt{m}$. In this simple example, we observe NLD and ULD have smaller mean exit times (from a barrier) compared to LD. We will add these discussions in the revised version.