Enhancing Low-Precision Sampling via Stochastic Gradient Hamiltonian Monte Carlo

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Low-precision training has emerged as a promising low-cost technique to enhance the training efficiency of deep neural networks without sacrificing much accuracy. Its Bayesian counterpart can further provide uncertainty quantification and improved generalization accuracy. This paper investigates low-precision samplers via Stochastics Gradient Hamiltonian Monte Carlo (SGHMC) with low-precision and full-precision gradients accumulators for both strongly log-concave and non-log-concave distributions. Theoretically, our results show that, to achieve ϵ-error in the 2-Wasserstein distance for non-log-concave distributions, low-precision SGHMC achieves quadratic improvement ($\tilde{\mathcal{O}}\left(\epsilon^{-2}{\mu^*}^{-2}\log^2\left(\epsilon^{-1}\right)\right)$) compared to the state-of-the-art low-precision sampler, Stochastic Gradient Langevin Dynamics (SGLD) ($\tilde{\mathcal{O}}\left(\epsilon^{-4}{\lambda^*}^{-1}\log^5\left(\epsilon^{-1}\right)\right)$). Moreover, we prove that low-precision SGHMC is more robust to the quantization error compared to low-precision SGLD due to the robustness of the momentum-based update w.r.t. gradient noise. Empirically, we conduct experiments on synthetic and MNIST, CIFAR-10 & CIFAR-100 datasets which successfully validate our theoretical findings. Our study highlights the potential of low-precision SGHMC as an efficient and accurate sampling method for large-scale and resource-limited deep learning.

1 Introduction

In recent years, deep neural networks (DNNs) have achieved remarkable success, accompanied by an increase in model complexity [Simonyan and Zisserman, 2014, He et al., 2016, Vaswani et al., 2017, Radford et al., 2018, Chen et al., 2023]. Consequently, there is a growing interest in utilizing low-precision optimization techniques to address the computational and memory costs associated with these complex models [Wang et al., 2018, Banner et al., 2018, Wu et al., 2018, Lin et al., 2019, Sun et al., 2019, Wortsman et al., 2023]. As a counterpart of low-precision optimization, low-precision sampling is relatively unexplored but has shown promising preliminary results. Zhang et al. [2022] studied the effectiveness of Stochastic Gradient Langevin Dynamics (SGLD) [Welling and Teh, 2011] in the context of low-precision arithmetic, highlighting its superiority over the optimization counterpart, Stochastic Gradient Descent (SGD). This superiority stems from SGLD's inherent robustness to system noise compared with SGD.

Other than SGLD, Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [Chen et al., 2014] is another popular gradient-based sampling method, closely related to the underdamped Langevin dynamics. Recently, Cheng et al. [2018], Gao et al. [2022] have shown that the SGHMC converges to its target distribution faster than the best-known convergence rate of SGLD in the 2-Wasserstein distance under both strongly log-concave and non-log-concave assumptions. Beyond this, SGHMC is analogous to stochastic gradient methods augmented with momentum, which is shown to have

Submitted to the First Workshop on Machine Learning with New Compute Paradigms at NeurIPS (MLNPCP 2023). Do not distribute.

more robust updates w.r.t. gradient estimation noise [Liu et al., 2020]. Note that the stochastic error induced by the quantization function in the low-precision update is equivalent to an extra noise of the stochastic gradient, causing an increase in the gradient variance. Thus, we believe the SGHMC is particularly suited for low-precision arithmetic.

Our main contributions of this paper are threefold:

First, we conduct the first study of low-precision SGHMC. We adopt low-precision arithmetic (including full- and low-precision gradient accumulators and variance correction (VC) version of low-precision gradient accumulators) to SGHMC.

Second, we provide a comprehensive theoretical analysis of low-precision SGHMC for both strongly log-concave and non-log-concave target distributions. All our theoretical results are summarized in Table 3 (deferred in Appendix A), where we compare the 2-Wasserstein convergence limit and the required gradient complexity. Our analysis exhibits the superiority of HMC-based low-precision algorithms over SGLD counterpart w.r.t. convergence speed and robustness to quantization error, especially under the non-log concave distributions.

Third, we provide promising empirical results in deep learning. We show the sampling capabilities of HMC-based low-precision algorithms and the effectiveness of the VC function in both strongly log-concave and non-log-concave target distributions. We also provide evidence of the superior performance of HMC-based low-precision algorithms compared to SGLD in real-world tasks.

In summary, low-precision SGHMC emerges as a compelling alternative to standard SGHMC due to its ability to enhance speed and memory efficiency without sacrificing accuracy.

2 Preliminaries

2.1 Stochastic Gradient Hamiltonian Monte Carlo

Given a dataset D, a model with weights (i.e., model parameters) $\mathbf{x} \in \mathbb{R}^d$, and a prior $p(\mathbf{x})$, we are interested in sampling from the posterior $p(\mathbf{x}|D) \propto \exp(-U(\mathbf{x}))$, where $U(\mathbf{x})$ is some energy function. In order to sample from the target distribution, SGHMC [Chen et al., 2014] is proposed and strongly related to the underdamped Langevin dynamics. Cheng et al. [2018] proposes the following discretization of underdamped Langevin dynamics (9) with stochastic gradient:

$$
\begin{aligned}
\mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{-\gamma\eta})\nabla\tilde{U}(\mathbf{x}_k) + \xi_k^{\mathbf{v}} \\
\mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)\nabla\tilde{U}(\mathbf{x}_k) + \xi_k^{\mathbf{x}},
\end{aligned} \tag{1}
$$

where u, γ denote the hyperparameters of inverse mass and friction respectively, $\nabla\tilde{U}$ is unbiased gradient estimation of U and $\xi_k^{\mathbf{v}}$, and η is the step size. $\xi_k^{\mathbf{x}}$ are normal distributed in $\mathbb{R}^d$ satisfying that :

$$
\begin{aligned}
\mathbb{E}\xi_k^{\mathbf{v}}(\xi_k^{\mathbf{v}})^\top &= u(1-e^{-2\gamma\eta}) \cdot \mathbf{I}, \\
\mathbb{E}\xi_k^{\mathbf{x}}(\xi_k^{\mathbf{x}})^\top &= u\gamma^{-2}(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3) \cdot \mathbf{I}, \\
\mathbb{E}\xi_k^{\mathbf{x}}(\xi_k^{\mathbf{v}})^\top &= u\gamma^{-1}(1 - 2e^{-\gamma\eta} + e^{-2\gamma\eta}) \cdot \mathbf{I}.
\end{aligned} \tag{2}
$$

2.2 Low-Precision Quantization

Two popular formats to represent low-precision numbers are known as the *fixed point* (FP) and *block floating point* [Song et al., 2018] (BFP). The quantization error which is defined as the gap between two adjacent representable numbers is denoted as Δ. Furthermore, all representable numbers are truncated to an upper limit $\bar{U}$ and a lower limit $\bar{L}$.

Given the low-precision number representation, a quantization function is desired to round real-valued numbers to their low-precision counterparts. Two common quantization functions are *deterministic rounding* and *stochastic rounding*. The deterministic rounding function, denoted as Q^d, quantizes a number to its nearest representable neighbor. The stochastic rounding denoted as Q^s (refer to (10) of Appendix A), randomly quantizes a number to the two closest representable neighbors satisfying the unbiased condition, i.e. $\mathbb{E}[Q^s(\theta)] = \theta$. In what follows, we use Q_W and Q_G

to denote the stochastic rounding quantizer we used for the weights and gradients respectively, allowing different quantization errors. But for simplicity in the analysis and experiments, we use the same number of bits to represent the weights and gradients.

3 Low-Precision Stochastic Gradient Hamiltonian Monte Carlo

In this section, we investigate the convergence property of low-precision SGHMC for non-log-concave target distributions. We defer the analysis of the low-precision SGHMC under strongly log-concave target distributions, as well as the analysis of low-precision SGLD [Zhang et al., 2022] to Appendix A and B respectively. All of our theorems are based on the fixed point representation and omit the clipping effect.

In order to derive a convergence analysis for non-log-concave target distribution, we assume the energy function $U(\cdot)$ is M-smooth (Assumption 1) also satisfied the dissaptiveness assumption (Assumption 3), and the mean squared error of stochastic gradients is bounded by constant σ^2 (Assumption 4). Detailed assumptions and explanations are deferred in Appendix A. In the statement of theorems, the big-O notation $\tilde{\mathcal{O}}$ gives explicitly dependence on the quantization error Δ and concentration parameters (λ^*, μ^*) but hides multiplicative terms that depend polynomially on the other parameters (e.g., dimension d, friction γ, inverse mass u and gradients variance σ^2).

3.1 Full- and low-Precision Gradient Accumulators

Adopting the updating rule in equations 1, we propose the low-precision SGHMC with full gradient accumulator (SGHMCLP-F) as the following:

$$\begin{aligned} \mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{-\gamma\eta})Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}_k))) + \xi_k^{\mathbf{v}} \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}_k))) + \xi_k^{\mathbf{x}}, \end{aligned} \tag{3}$$

The storage and computation costs can be further reduced by the low-precision gradient accumulators, i.e., the low-precision SGHMC with low-precision gradient accumulators (SGHMCLP-L):

$$\begin{aligned} \mathbf{v}_{k+1} &= Q_W\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{v}}\right), \\ \mathbf{x}_{k+1} &= Q_W\left(\mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{x}}\right). \end{aligned} \tag{4}$$

Our analysis for the above two algorithms utilizes similar techniques in Raginsky et al. [2017].

Theorem 1 (Informal version of Theorem 5)**.** *Given the smoothness, dissaptivity and assumption for stochastic gradients, let p^* denote the target distribution of $\mathbf{x}$ and $\mathbf{v}$. Given initialization $\mathbf{x}_0 = \mathbf{v}_0 = 0$ and $\gamma^2 \le 4Mu$, for some sufficiently small ϵ and step size η, the K-th iteration of the SGHMCLP-F update (3), i.e., $\mathbf{x}_K$ and $\mathbf{v}_K$, satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) \le \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\Delta \log(1/\epsilon)}\right), \tag{5}$$

for some K satisfying

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2 \mu^{*2}} \log^2\left(\frac{1}{\epsilon}\right)\right),$$

where μ^* is a constant w.r.t. dimension d, denoting the concentration rate of the underdamped Langevin dynamics [Zou et al., 2019].

Theorem 2 (Informal version of Theorem 7)**.** *Given the smoothness, dissaptivity and assumption for stochastic gradients, let p^* denote the target distribution of $\mathbf{x}$ and $\mathbf{v}$. Given initialization $\mathbf{x}_0 = \mathbf{v}_0 = 0$ and $\gamma^2 \le 4Mu$, for some sufficiently small ϵ and step size η, the K-th iteration of the SGHMCLP-L update (4), i.e., $\mathbf{x}_K$ and $\mathbf{v}_K$, satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\max\{\sigma^2, \sigma\} \log\left(\frac{1}{\epsilon}\right)} + \frac{\log^{3/2}\left(\frac{1}{\epsilon}\right)}{\epsilon^2}\sqrt{\Delta}\right), \tag{6}$$

for some K satisfying

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2 \mu^{*2}} \log^2\left(\frac{1}{\epsilon}\right)\right).$$

Similar to the convergence result of full-precision SGHMC or SGLD [Raginsky et al., 2017, Gao et al., 2022], the above upper bound (5) of SGHMCLP-F contains a ϵ term and a $\log(\epsilon^{-1})$ term. The difference is that for the SGHMCLP-F algorithm, the quantization error Δ affects the multiplicative constant of the $\log(\epsilon^{-1})$ term. Without Δ, one can choose a small ϵ and a larger batch size (i.e., a smaller σ^2) to offset $\log\left(\epsilon^{-1}\right)$ term, such that the 2-Wasserstein distance can be sufficiently small. With the same technical tools, we conduct a similar convergence analysis of SGLDLF-P for non-log-concave target distributions (refer to Theorem 10 of Appendix B). Comparing Theorems 1 and 10, we show that SGHMCLP-F can achieve lower 2-Wasserstein (i.e. $\tilde{\mathcal{O}}\left(\epsilon + \left(\log\left(\epsilon^{-1}\right)\Delta\right)^{1/2}\right)$ versus $\tilde{\mathcal{O}}\left(\epsilon + \log\left(\epsilon^{-1}\right)\Delta^{1/2}\right)$) distance for non-log-concave target distribution within fewer iterations (i.e., $\tilde{\mathcal{O}}\left(\epsilon^{-2}\mu^{*-2}\log^2\left(\epsilon^{-1}\right)\right)$ versus $\tilde{\mathcal{O}}\left(\epsilon^{-4}\lambda^{*-1}\log^5\left(\epsilon^{-1}\right)\right)$).

We verify the advantage of SGHMCLF-P over SGLDLF-P by our simulations in section 4.

As for SGHMCLP-L, which additionally quantizes the weights after each update, a small stepsize can result in staying at the starting point. In such cases, ensuring convergence becomes challenging, and the output of the SGHMCLP-L has a worse convergence upper bound compared to Theorem 1. Empirically, we observe that the output $\mathbf{x}_K$'s distribution has an overdispersion problem (i.e. Figure 1 (a) and 5 (a)). In Theorem 11, we generalize the result of the naïve SGLDLP-L in [Zhang et al., 2022] to non-log-concave target distribution. Similarly, we observe that SGHMCLP-L needs fewer iterations than SGLDLP-L in terms of the order w.r.t. ϵ and achieves better upper bound $\tilde{\mathcal{O}}\left(\epsilon^{-2}\log^{3/2}\left(\epsilon^{-1}\right)\Delta^{1/2}\right)$ versus $\tilde{\mathcal{O}}\left(\epsilon^{-4}\log^5\left(\epsilon^{-1}\right)\Delta^{1/2}\right)$.

3.2 Variance Correction

To resolve the overdispersion caused by the low-precision gradient accumulators, Zhang et al. [2022] propose a quantization function Q^{vc} (refer to Algorithm 1 in Appendix A) that directly samples from the discrete weight space instead of quantizing a real-valued Gaussian sample. This quantization function aims to reduce the discrepancy between the ideal sampling variance (i.e., the required variance of full-precision counterpart algorithms) and the actual sampling variance in our low-precision algorithms.

In this work, we study the effect of Q^{vc} on low-precision SGHMC. Let $\mathrm{Var}_{\mathbf{v}}^{hmc} = u(1 - e^{-2\gamma\eta})$ and $\mathrm{Var}_{\mathbf{x}}^{hmc} = u\gamma^{-2}(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3)$, the VC SGHMCLP-L can be done as:

$$\begin{aligned} \mathbf{v}_{k+1} &= Q^{vc}\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{-\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_k)), \mathrm{Var}_{\mathbf{v}}^{hmc}, \Delta\right) \\ \mathbf{x}_{k+1} &= Q^{vc}\left(\mathbf{x}_k + \gamma^{-1}(1 - e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(\mathbf{x}_k)), \mathrm{Var}_{\mathbf{x}}^{hmc}, \Delta\right) \end{aligned} \tag{7}$$

Now, we are ready to present the convergence analysis of VC SGHMC-L.

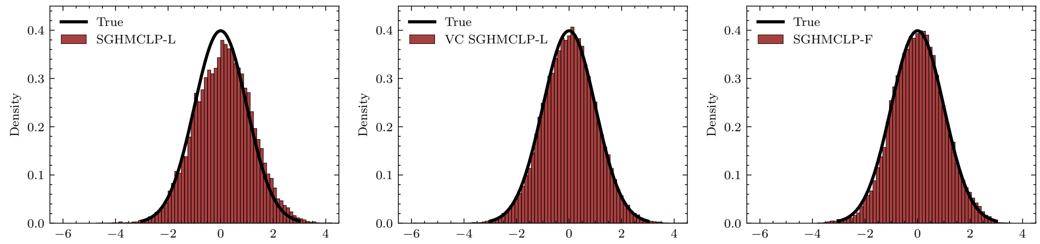
Theorem 3 (Informal version of Theorem 9). *Given the smoothness, dissaptivity and assumption for stochastic gradients, let p^* denote the target distribution of $\mathbf{x}$. Given initialization $\mathbf{x}_0 = \mathbf{v}_0 = 0$ and $\gamma^2 \le 4Mu$, for some sufficiently small ϵ and step size η, the K-th iteration of the VC SGHMCLP-L update (4), i.e., $\mathbf{x}_K$, satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\max\{\sigma^2, \sigma\}\log\left(\frac{1}{\epsilon}\right)} + \frac{\log\left(\frac{1}{\epsilon}\right)}{\epsilon}\sqrt{\Delta}\right), \tag{8}$$

for some K satisfying

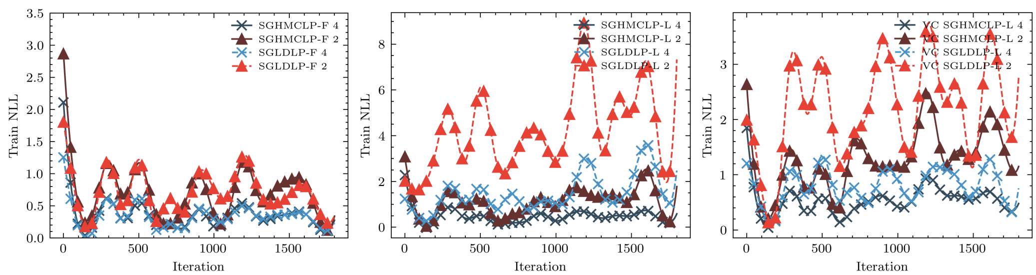
$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2\mu^{*2}}\log^2\left(\frac{1}{\epsilon}\right)\right).$$

Comparing with Theorem 2, the variance corrected quantization can improve the upper bound w.r.t. ϵ from $\tilde{\mathcal{O}}\left(\epsilon^{-2}\log^{3/2}\left(\epsilon^{-1}\right)\Delta^{1/2}\right)$ to $\tilde{\mathcal{O}}\left(\epsilon^{-1}\log\left(\epsilon^{-1}\right)\Delta^{1/2}\right)$. In Theorem 12, we generalize the result of the VC SGLDLP-L in [Zhang et al., 2022] to non-log-concave target distribution. Similarly, we observe that VC SGHMCLP-L needs fewer iterations than VC SGLDLP-L in terms of the order w.r.t. ϵ and achieves better upper bounds ($\tilde{\mathcal{O}}\left(\epsilon + \log\left(\epsilon^{-1}\right)\epsilon^{-1}\Delta^{1/2}\right)$ versus $\tilde{\mathcal{O}}\left(\epsilon + \log^3\left(\epsilon^{-1}\right)\epsilon^{-2}\Delta^{1/2}\right)$).



(a) (b) (c)

Figure 1: Low-precision SGHMC on Gaussian distribution. (a): SGHMCLP-L. (b): VC SGHMCLP-L. (c): SGHMCLP-F.



(a) (b) (c)

Figure 2: Training NLL of low-precision SGHMC and SGLD on logistic model with MNIST in terms of different numbers of fractional bits. (a): Methods with full-precision gradient accumulators. (b): Methods with low-precision gradient accumulators. (c): Variance corrected quantization.

Interestingly, the naïve SGHMCLP-L has similar dependence on the quantization error Δ with VC SGLDLP-L but saves more computation resources since the variance corrected quantization requires sampling discrete random variables. We verify our finding in Table 2.

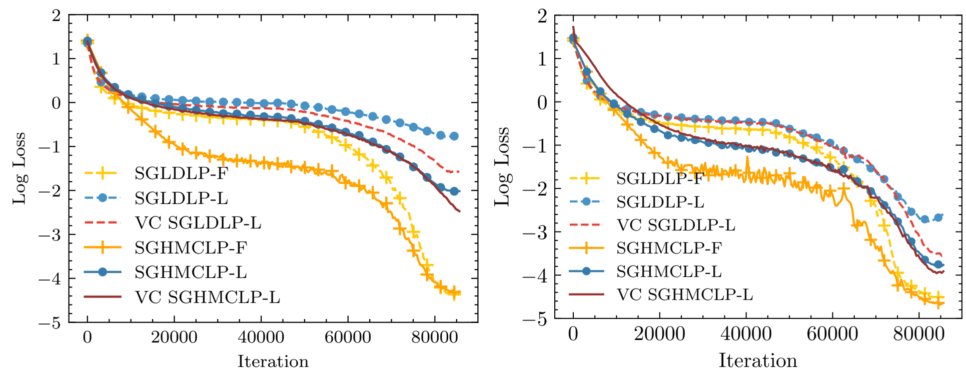
4 Experiments

We assess the performance of the proposed low-precision SGHMC algorithms through sampling a Gaussian distribution and implementing a Bayesian logistic regression to the MNIST dataset (Section 4.1), and training a Bayesian ResNet-18 on the CIFAR-10 and CIFAR-100 datasets (Section 4.2). We compare our proposed algorithms with their SGLD counterparts. Details and additional experiment results (e.g., sampling Gaussian mixture distribution and MLP training on MNIST dataset) can be found in Appendix F. In all experiments, *qtorch* [Zhang et al., 2019] is employed for Low-Precision sampling with the same quantization.

4.1 Sampling Gaussian distributions & MNIST

We use a Gaussian distribution to represent the log-concave distribution. The simulation results are shown in Figure 1. It shows that the SGHMCLP-F samples fit the true Gaussian distribution well. Regarding the naïve SGHMCLP-L, we observe an overdispersion problem and the variance corrected function solves this problem.

We further examine the sampling performance of low-precision SGHMC and SGLD on real-world data. We use logistic models to represent the class of strongly log-concave distributions. The results are in Figure 2. We use fixed point numbers with 2 integer bits and vary the number of fractional bits which corresponds to varying the quantization gap Δ. We report train negative log-likelihood (NLL) with different numbers of fractional bits in Figure 2. From the results on MNIST, we can see that when adopted to full-precision gradient accumulators low-precision SGHMC are robust to the quantization error. Even when we use only 2 fractional bits, SGHMCLP-F can still converge to a good distribution but with more iteration. As the precision error increases, both SGHMCLP-L and SGLDLP-L have a worse convergence pattern compared to SGHMCLP-F and SGLDLP-F.



(a) (b)

Figure 3: Log of training NLL of low-precision SGHMC and SGLD on ResNet-18 with CIFAR100 and constant step sizes. (a): 8-bit Fixed Point. (b): 8-bit Block Float Point.

Table 1: Test errors (%) of full-precision gradient accumulators on CIFAR with ResNet-18.

	32-bit Floating			8-bit Fixed Point			8-bit Block Floating Point		
	SGD	SGLD	SGHMC	SGD	SGLD	SGHMC	SGD	SGLD	SGHMC
CIFAR-10	4.73 ± 0.10	**4.52** ± **0.07**	4.78 ± 0.08	5.19 ± 0.09	**5.07** ± **0.04**	5.08 ± 0.08	4.75 ± 0.21	**4.58** ± **0.07**	4.93 ± 0.09
CIFAR-100	**22.34** ± **0.22**	22.40 ± 0.04	22.37 ± 0.04	23.71 ± 0.18	**23.36** ± **0.10**	23.54 ± 0.10	22.86 ± 0.14	22.70 ± 0.22	**22.39** ± **0.11**

We showed empirically that SGHMCLP-L and VC SGHMCLP-L outperform SGLDLP-L and VC SGLDLP in Figure 2, showing low-precision SGHMC is more robust to the quantization error.

4.2 CIFAR-10 & CIFAR-100

We consider computer vision tasks CIFAR10 and CIFAR100 on the ResNet-18. We use 8-bit number representation as it becomes increasingly popular and powered by new chips. We report the average test errors over 3 runs in Tables 1 and 2. We use 8-bit fixed point (FP) and block floating point (BFP) representing weights and gradients. SGHMCLP-F is comparable with SGDLP-F and the naïve SGHMCLP-L significantly outperforms naïve SGLDLP-L and SGDLP-L across datasets. Furthermore, from the result in Figure 3, we empirically show that the convergence speed of SGHMC is way better than the SGLD. Besides the variance corrected quantization function can bring some gain on the test accuracy, the performance of SGHMCLP-L is good enough and comparable with the performance of VC SGLDLP-L. By using BFP, the performance of all low-precision methods improves over fixed point, and we observe similar results as the FP.

5 Conclusion

We provide the first comprehensive investigation for low-precision SGHMC in both strongly log-concave and non-log-concave target distributions with several variants of low-precision training. In particular, we prove that for non-log-concave distributions, low-precision SGHMC with full-precision, low-precision, and variance-corrected gradient accumulators, all achieve an acceleration in iterations and have a better convergence upper bound w.r.t the quantization error compared to the low-precision SGLD counterpart. Moreover, we study the improvement of variance-corrected quantization applied to low-precision SGHMC under different cases. Under certain conditions, the naïve SGHMCLP-L can replace the VC SGLDLP-L to get comparable results saving more computation

Table 2: Test errors (%) of low-precision gradient accumulators on CIFAR with ResNet-18.

	8-bit Fixed Point					8-bit Block Floating Point				
	SGD	SGLD	VC SGLD	SGHMC	VC SGHMC	SGD	SGLD	VC SGLD	SGHMC	VC SGHMC
CIFAR-10	8.50 ± 0.22	7.81 ± 0.07	7.03 ± 0.23	6.63 ±0.01	**6.60** ± **0.06**	5.86 ± 0.18	5.75 ± 0.05	5.51 ± 0.01	5.38 ± 0.06	**5.15** ± **0.08**
CIFAR-100	28.42 ± 0.35	27.15 ± 0.35	26.73 ± 0.12	26.57 ± 0.10	**26.43** ± **0.19**	26.75 ± 0.11	26.11 ± 0.38	25.14 ± 0.11	25.29 ± 0.03	**24.45** ± **0.16**

resources. We conduct empirical experiments on Gaussian, Gaussian mixture distribution, logistic regression, and Bayesian deep learning tasks to justify our theoretical findings.

References

R. Banner, I. Hubara, E. Hoffer, and D. Soudry. Scalable methods for 8-bit training of neural networks. *Advances in neural information processing systems*, 31, 2018.

F. Bolley and C. Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pages 331–352, 2005.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

X. Chen, C. Liang, D. Huang, E. Real, K. Wang, Y. Liu, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

C. De Sa, M. Leszczynski, J. Zhang, A. Marzoev, C. R. Aberger, K. Olukotun, and C. Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Z. Li and C. M. De Sa. Dimension-free bounds for low-precision training. *Advances in Neural Information Processing Systems*, 32, 2019.

P.-C. Lin, M.-K. Sun, C. Kung, and T.-D. Chiueh. Floatsd: A new weight representation and associated update method for efficient convolutional neural network training. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2):267–279, 2019.

Y. Liu, Y. Gao, and W. Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Z. Song, Z. Liu, and D. Wang. Computation error analysis of block floating point arithmetic oriented convolution neural network accelerator design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

X. Sun, J. Choi, C.-Y. Chen, N. Wang, S. Venkataramani, V. V. Srinivasan, X. Cui, W. Zhang, and K. Gopalakrishnan. Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. *Advances in neural information processing systems*, 31, 2018.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

M. Wortsman, T. Dettmers, L. Zettlemoyer, A. Morcos, A. Farhadi, and L. Schmidt. Stable and low-precision training for large-scale vision-language models. *arXiv preprint arXiv:2304.13013*, 2023.

S. Wu, G. Li, F. Chen, and L. Shi. Training and inference with integers in deep neural networks. *arXiv preprint arXiv:1802.04680*, 2018.

R. Zhang, A. G. Wilson, and C. De Sa. Low-precision stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, pages 26624–26644. PMLR, 2022.

T. Zhang, Z. Lin, G. Yang, and C. De Sa. Qpytorch: A low-precision arithmetic simulation framework. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 10–13. IEEE, 2019.

D. Zou, P. Xu, and Q. Gu. Stochastic gradient hamiltonian monte carlo methods with recursive variance reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

A Additional Results for Low-precision Stochastic Gradient Hamiltonian Monte Carlo

The underdamped Langevin dynamics has a continuous-time diffusion form:

$$\begin{aligned} d\mathbf{v}_t &= -\gamma \mathbf{v}_t dt - u\nabla U(\mathbf{x}_t)dt + \sqrt{2\gamma u} d\mathbf{B}_t \\ d\mathbf{x}_t &= \mathbf{v}_t dt. \end{aligned} \tag{9}$$

And we formally define the stochastic rounding quantization function as:

$$Q^s(\theta) = \begin{cases} \Delta \left\lfloor \frac{\theta}{\Delta} \right\rfloor, & \text{w.p. } \left\lceil \frac{\theta}{\Delta} \right\rceil - \frac{\theta}{\Delta} \\ \Delta \left\lceil \frac{\theta}{\Delta} \right\rceil, & \text{w.p. } 1 - \left(\left\lceil \frac{\theta}{\Delta} \right\rceil - \frac{\theta}{\Delta} \right). \end{cases} \tag{10}$$

Before diving into the theorems, we introduce some necessary assumptions.

Assumption 1 (Smoothness)**.** *The energy function U is M-smooth, i.e., there exists a positive constant M such that*

$$\|\nabla U(\mathbf{x}) - \nabla U(\mathbf{y})\|^2 \le M^2 \|\mathbf{x} - \mathbf{y}\|^2, \quad \textit{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Assumption 2 (Strongly Log-Convex)**.** *The energy function U is m-strongly log-convex, i.e., there exists a positive constant m such that,*

$$U(\mathbf{y}) \ge U(\mathbf{x}) + \langle \nabla U(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{m_1}{2} \|\mathbf{y} - \mathbf{x}\|^2, \quad \textit{for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Assumption 3 (Dissaptiveness)**.** *There exist constants $m_2, b > 0$, such that the following holds*

$$\langle \nabla U(\mathbf{x}), \mathbf{x} \rangle \ge m_2 \|\mathbf{x}\|^2 - b, \quad \textit{for any } \mathbf{x} \in \mathbb{R}^d.$$

Assumption 4 (Bounded Variance)**.** *There exists a constant $\sigma^2 > 0$, such that the following holds*

$$\mathbb{E}\left\|\nabla \tilde{U}(\mathbf{x}) - \nabla U(\mathbf{x})\right\|^2 \le \sigma^2, \quad \textit{for any } \mathbf{x} \in \mathbb{R}^d.$$

Beyond the above assumptions, we further define $\kappa_1 = M/m_1$ and $\kappa_2 = M/m_2$ as the condition number for strongly log-concave and non-log-concave target distribution respectively, and denote the global minimum of $U(\mathbf{x})$ as $\mathbf{x}^*$. Assumption 3 is the standard assumption [Raginsky et al., 2017, Zou et al., 2019, Gao et al., 2022] in the analysis of sampling from non-log-concave distributions and is essential to guarantee the convergence of underdamped Langevin dynamics. Now we introduce the of SGHMCLP-F for strongly log-concave and non-log-concave target distribution in Theorem 4 and 5 respectively.

Theorem 4. *Suppose Assumptions 1, 2 and 4 hold and the minimum satisfies $\|\mathbf{x}^*\|^2 < \mathcal{D}^2$. Furthermore, let p^* denote the target distribution of $\mathbf{x}$ and $\mathbf{v}$. Given any sufficiently small ϵ, if we set the step size to be*

$$\eta = \min\left\{ \frac{\epsilon \kappa_1^{-1}}{\sqrt{479232/5(d/m_1 + \mathcal{D}^2)}}, \frac{\epsilon^2}{1440\kappa_1 u^2 \left[(M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right]} \right\},$$

then after K steps starting with initial points $\mathbf{x}_0 = \mathbf{v}_0 = 0$, the output $(\mathbf{x}_K, \mathbf{v}_K)$ of the SGHMCLP-F in (3) *satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) \le \tilde{\mathcal{O}}\left(\epsilon + \Delta\right),$$

for some K satisfying

$$K \le \frac{\kappa_1}{\eta} \log\left(\frac{36\left(\frac{d}{m_1} + \mathcal{D}^2\right)}{\epsilon} \right) = \tilde{\mathcal{O}}\left(\epsilon^{-2} \log\left(\epsilon^{-1}\right) \Delta^2\right).$$

Theorem 5. *Suppose Assumptions 1, 3 and 4 hold. Furthermore, let p^* denote the target distribution of $\mathbf{x}$ and $\mathbf{v}$. Given initialization $\mathbf{x}_0 = \mathbf{v}_0 = 0$ and $\gamma^2 \leq 4Mu$, for any sufficiently small ϵ, if we set the step size to be $\eta = \tilde{\mathcal{O}}\left(\frac{\mu^*\epsilon^2}{\log(1/\epsilon)}\right)$ and also satisfy*

$$\eta \leq \min\left\{\frac{\gamma}{4\left(8Mu + u\gamma + 22\gamma^2\right)}, \sqrt{\frac{4u^2}{4Mu + 3\gamma^2}}, \frac{6\gamma bu}{\left(4Mu + 3\gamma^2\right)d}, \frac{1}{8\gamma}, \frac{\gamma m_2}{12(21u + \gamma)M^2}, \frac{8(\gamma^2 + 2u)}{(20u + \gamma)\gamma}\right\},$$

then, the K-th iteration of the SGHMCLP-F update (3)*, i.e., $\mathbf{x}_K$ and $\mathbf{v}_K$, satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) \leq \tilde{\mathcal{O}}\left(\epsilon + \widetilde{A}\sqrt{\log\left(\frac{1}{\epsilon}\right)}\right),$$

for some K satisfying

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2 \mu^{*2}} \log^2\left(\frac{1}{\epsilon}\right)\right),$$

where constants are defined as: $\widetilde{A} = \max\left\{\sqrt{\Delta^2 d + \sigma^2}, \sqrt[4]{\Delta^2 d + \sigma^2}\right\}$, and μ^* is a constant w.r.t. dimension d, denoting the concentration rate of the underdamped Langevin dynamics [Zou et al., 2019].

Theorem 1 in Zhang et al. [2022] implies that for strongly log-concave target distribution, the low-precision SGLD with full-precision gradient accumulators can achieve ϵ accuracy within $\tilde{\mathcal{O}}\left(\epsilon^{-2}\log\left(\epsilon^{-1}\right)\Delta^2\right)$ iterations.

Thus, the theorem of SGHMCLP-F does not showcase any advantage over SGLDLP-F. This is not surprising, since the quantization applied to the gradients in the full-precision gradient accumulator algorithm is equivalent to adding extra noise to the stochastic gradients. As theoretically shown by Cheng et al. [2018] for strongly-log-concave target distribution, HMC doesn't exhibit any advantage over the unadjusted Langevin algorithm when stochastic gradients are used.

However, as shown in the Theorem 5, for non-log-concave distributions, the low-precision SGHMC displays faster convergence speed and a better dependence on the quantization error Δ compared to SGLD. Besides the discussion in Theorem 1, we can discuss the upper w.r.t. to Δ, due to the fact that $\log(x) \leq x^{1/e}$, one can tune the choice of ϵ and η, and achieve a $\tilde{\mathcal{O}}\left(\Delta^{e/(1+2e)}\right)$ 2-Wasserstein bound for non-log-concave target distribution. Furthermore, based on Theorem 10, after carefully choosing the stepsize η, the 2-Wasserstein distance of the SGLDLF-P algorithm can be further bounded by $\tilde{\mathcal{O}}\left(\Delta^{e/(2+2e)}\right)$ which is worse than the bound $\tilde{\mathcal{O}}\left(\Delta^{e/(1+2e)}\right)$ obtained by SGHMC. Next, we introduce the convergence analysis of SGHMCLP-L for strongly log-concave and non-log-concave target distribution in Theorem 6 and 7 respectively.

Theorem 6. *Let Assumption 1, 2 and 4 hold and the minimum satisfies $\|\mathbf{x}^*\|^2 < \mathcal{D}^2$. Furthermore, let p^* denote the target distribution of $\mathbf{v}$ and $\mathbf{x}$. Given any sufficiently small ϵ, if we set the step size η to be*

$$\eta = \min\left\{\frac{\epsilon\kappa_1^{-1}}{\sqrt{663552/5\left(\frac{d}{m_1} + \mathcal{D}^2\right)}}, \frac{\epsilon^2}{2880\kappa_1 u\left(\frac{\Delta^2 d}{4} + \sigma^2\right)}\right\},$$

then after K steps starting with initial points $\mathbf{x}_0 = \mathbf{v}_0 = 0$, the output $(\mathbf{x}_K, \mathbf{v}_K)$ of the SGHMCLP-L in (4) *satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \frac{\Delta}{\epsilon}\right), \tag{11}$$

for some K satisfying

$$K \leq \frac{\kappa_1}{\eta}\log\left(\frac{36\left(\frac{d}{m_1} + \mathcal{D}^2\right)}{\epsilon}\right) = \tilde{\mathcal{O}}\left(\epsilon^{-2}\log\left(\epsilon^{-1}\right)\Delta^2\right).$$

Compared with Theorem 2 in Zhang et al. [2022], We cannot show the advantages of low-precision SGHMC over SGLD for strongly log-concave target distribution. However, for non-log-concave target distribution, we show SGHMCLP-L can achieve lower distance in smaller iterations. Next, we present the convergence theorem of SGHMCLP-L for non-log-concave target distribution. Besides the discussion in Theorem 2, by the same argument in Theorem 1's discussion after carefully choosing the stepsize η, the 2-Wasserstein distance of SGHMCLP-L to non-log-concave target distribution can be further bounded as $\tilde{\mathcal{O}}\left(\Delta^{e/(3+6e)}\right)$, and the distance of the sample obtained by SGLDLP-L can be bounded as $\tilde{\mathcal{O}}\left(\Delta^{e/10(1+e)}\right)$. Thus the low-precision SGHMC is more robust to the quantization error than SGLD. Next, we present the convergence analysis of VC SGHMCLP-L in (8). We begin with the formal definition of the variance-corrected quantization function Q^{vc}. Instead of adding real value Gaussian noise and quantizing the weights, we can design a categorical sampler that samples from the space $\{\Delta, -\Delta, 0\}$ with the desired expectation μ and variance v as

$$\mathrm{Cat}(\mu, v) = \begin{cases} \Delta, & w.p. \frac{v+\mu^2+\mu\Delta}{2\Delta^2} \\ -\Delta, & w.p. \frac{v+\mu^2-\mu\Delta}{2\Delta^2} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Based on the sampler 12, we design the variance correction quantization function Q^{vc} in the algorithm 1.

Theorem 7. *Let Assumptions 1, 3 and 4 hold. If* $\gamma^2 \leq 4Mu$ *and we set the step size to be* $\eta = \tilde{\mathcal{O}}\left(\frac{\mu^*\epsilon^2}{\log(1/\epsilon)}\right)$*, also satisfied*

$$\eta \leq \min\left\{\frac{\gamma}{4\left(8Mu+u\gamma+22\gamma^2\right)}, \sqrt{\frac{4u^2}{4Mu+3\gamma^2}}, \frac{6\gamma bu}{\left(4Mu+3\gamma^2\right)d}, \frac{1}{8\gamma}, \frac{\gamma m_2}{12(21u+\gamma)M^2}, \frac{8(\gamma^2+2u)}{(20u+\gamma)\gamma}\right\},$$

let p^* *denote the target distribution of* $(\mathbf{x}, \mathbf{v})$ *then after* K *steps starting at the initial point* $\mathbf{x}_0 = \mathbf{v}_0 = 0$ *the output* $(\mathbf{x}_K, \mathbf{v}_K)$ *of SGHMCLP-L in 4 satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\max\left\{\sigma^2, \sigma\right\}\log\left(\frac{1}{\epsilon}\right)} + \frac{\log^{3/2}\left(\frac{1}{\epsilon}\right)}{\epsilon^2}\sqrt{\Delta}\right), \tag{13}$$

for some K *satisfying*

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2\mu^{*2}}\log^2\left(\frac{1}{\epsilon}\right)\right).$$

Theorem 8. *Let Assumption 1, 2 and 4 hold and the minimum satisfies* $\|\mathbf{x}^*\|^2 < \mathcal{D}^2$*. Furthermore, let* p^* *denote the target distribution of* $\mathbf{x}$ *and* $\mathbf{v}$*. Given any sufficiently small* ϵ*, if we set the stepsize to be*

$$\eta = \min\left\{\frac{\epsilon^2}{663552/5\left(\frac{d}{m_1}+\mathcal{D}^2\right)\kappa_1^2}, \frac{\epsilon^2}{90u^2\Delta^2 d\kappa_1 + 360u^2\sigma^2\kappa_1}\right\}$$

after K *steps starting from the initial point* $\mathbf{x}_0 = \mathbf{v}_0 = 0$ *the output* $(\mathbf{x}_K, \mathbf{v}_K)$ *of the VC SGHMCLP-L in algorithm 2 satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K, \mathbf{v}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\Delta}\right), \tag{14}$$

for some K *satisfying*

$$K \leq \frac{\kappa_1}{\eta}\log\left(\frac{36\left(\frac{d}{m_1}+\mathcal{D}^2\right)}{\epsilon}\right) = \tilde{\mathcal{O}}\left(\epsilon^{-2}\log\left(\epsilon^{-1}\right)\Delta^2\right).$$

Theorem 8 shows that the variance corrected quantization function can solve the overdispersion problem we observe for the naïve SGHMCLP-L algorithm for strongly log-concave distribution. The $\mathcal{W}_2$ distance between the sample distribution and target distribution can be arbitrarily close to $\tilde{\mathcal{O}}(\sqrt{\Delta})$. Compared to the Theorem 3 in Zhang et al. [2022], the VC SGHMCLP-L doesn't showcase its advantage over VC SGLDLP-L for strongly log-concave distribtuion, however for non-log-concave target distribution we show VC SGHMCLP-L can achieve lower 2-Wasserstein distance in smaller iterations. Next, we provide the convergence analysis of the VC SGHMCLP-L for non-log-concave distribution.

Algorithm 1 Variance-Corrected Quantization Function Q^{vc}.

input: (μ, v, Δ) $\{Q^{vc}$ returns a variable with mean μ and variance $v\}$
$v_0 \leftarrow \Delta^2/4$ $\{\Delta^2/4$ is the largest possible variance that stochastic rounding can cause$\}$
if $v > v_0$ **then** $\{$add a small Gaussian noise and sample from the discrete grid to make up the remaining variance$\}$
  $x \leftarrow \mu + \sqrt{v - v_0}\xi$, where $\xi \sim \mathcal{N}(0, I_d)$
  $r \leftarrow x - Q^d(x)$
  for all i **do**
    sample c_i from Cat$(|r_i|, v_0)$ as in (12)
  end for
  $\theta \leftarrow Q^d(x) + \text{sign}(r) \odot c$
else $\{$sample from the discrete grid to achieve the target variance$\}$
  $r \leftarrow \mu - Q^s(\mu)$
  for all i **do**
    $v_s \leftarrow \left(1 - \frac{|r_i|}{\Delta}\right) \cdot r_i^2 + \frac{|r_i|}{\Delta} \cdot (-r_i + \text{sign}(r_i)\Delta)^2$
    if $v > v_s$ **then**
      sample c_i from Cat$(0, v - v_s)$ as in (12)
      $\theta_i \leftarrow Q^s(\mu)_i + c_i$
    else
      $\theta_i \leftarrow Q^s(\mu)_i$
    end if
  end for
end if
clip θ if outside representable range
return θ

Theorem 9. *Let Assumption 1, 3 and 4 hold. If $\gamma^2 \leq 4Mu$ and we set the step size to be $\eta = \tilde{\mathcal{O}}\left(\frac{\mu^* \epsilon^2}{\log(1/\epsilon)}\right)$, also satisfied*

$$\eta \leq \min\left\{\frac{\gamma}{4\left(8Mu + u\gamma + 22\gamma^2\right)}, \sqrt{\frac{4u^2}{4Mu + 3\gamma^2}}, \frac{6\gamma bu}{\left(4Mu + 3\gamma^2\right)d}, \frac{1}{8\gamma}, \frac{\gamma m_2}{12(21u + \gamma)M^2}, \frac{8(\gamma^2 + 2u)}{(20u + \gamma)\gamma}\right\}.$$

We further assume that $\mathbb{E}\left\|Q_G(\nabla\tilde{U}(x))\right\|_2^2 \leq G^2$, let p^* be the target distribution of $\mathbf{x}$ then after K steps starting at the initial point $\mathbf{x}_0 = \mathbf{v}_0 = 0$ the output $(\mathbf{x}_K)$ of the VC SGHMCLP-L in algorithm 2 satisfies

$$\mathcal{W}_2(p(\mathbf{x}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\max\left\{\sigma^2, \sigma\right\}\log\left(\frac{1}{\epsilon}\right)} + \frac{\log\left(\frac{1}{\epsilon}\right)}{\epsilon}\sqrt{\Delta}\right), \tag{15}$$

for some K satisfying

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2 \mu^{*2}}\log^2\left(\frac{1}{\epsilon}\right)\right).$$

B Stochastic Gradient Langevin Dynamics Result

In order to sample from the target distribution, Langevin dynamics-based samplers, such as overdamped Langevin MCMC and underdamped Langevin MCMC methods, are widely used when the evaluation of $U(\mathbf{x})$ is expansive due to a large sample size. The continuous-time overdamped Langevin MCMC can be represented by the following stochastic differential equation(SDE):

$$d\mathbf{x}_t = -\nabla U(\mathbf{x}_t) + \sqrt{2}d\mathbf{B}_t, \tag{16}$$

where $\mathbf{B}_t$ represents the standard Brownian motion in $\mathbb{R}^d$. Under some mild conditions, it can be proved that the invariant distribution of (16) converges the target distribution $\exp(-U(\mathbf{x}))$. To

Table 3: Theoretical results of the achievable 2-Wasserstein distance and the required gradient complexity for both log-concave (*italic*) non-log-concave (**bold**) target distributions, where ϵ is any sufficiently small constant, Δ is the quantization error, and μ^* and λ^* denote the concentration rate of underdamped and overdamped Langevin dynamics respectively.

	Gradient Complexity	Achieved 2-Wasserstein
Full-precision gradient accumulators		
SGLD/SGHMC (Theorem 4)	$\tilde{\mathcal{O}}\left(\log\left(\epsilon^{-1}\right)\epsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\Delta\right)$
SGLD (Theorem 10)	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\lambda^{*-1}\log^5\left(\epsilon^{-1}\right)\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\log\left(\epsilon^{-1}\right)\sqrt{\Delta}\right)$
SGHMC (Theorem 5)	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\mu^{*-2}\log^2\left(\epsilon^{-1}\right)\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\sqrt{\log\left(\epsilon^{-1}\right)\Delta}\right)$
Low-precision gradient accumulators		
SGLD/SGHMC (Theorem 6)	$\tilde{\mathcal{O}}\left(\log\left(\epsilon^{-1}\right)\epsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\epsilon^{-1}\Delta\right)$
VC SGLD/VC SGHMC (Theorem 8)	$\tilde{\mathcal{O}}\left(\log\left(\epsilon^{-1}\right)\epsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\sqrt{\Delta}\right)$
SGLD (Theorem 11)	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\lambda^{*-1}\log^5\left(\epsilon^{-1}\right)\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\log^5\left(\epsilon^{-1}\right)\epsilon^{-4}\sqrt{\Delta}\right)$
VC SGLD (Theorem 12)	$\tilde{\mathcal{O}}\left(\epsilon^{-4}\lambda^{*-1}\log^3\left(\epsilon^{-1}\right)\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\log^3\left(\epsilon^{-1}\right)\epsilon^{-2}\sqrt{\Delta}\right)$
SGHMC (Theorem 7)	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\mu^{*-2}\log^2\left(\epsilon^{-1}\right)\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\log^{3/2}\left(\epsilon^{-1}\right)\epsilon^{-2}\sqrt{\Delta}\right)$
VC SGHMC (Theorem 9)	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\mu^{*-2}\log^2\left(\epsilon^{-1}\right)\right)$	$\tilde{\mathcal{O}}\left(\epsilon+\log\left(\epsilon^{-1}\right)\epsilon^{-1}\sqrt{\Delta}\right)$

reduce the computational cost of evaluating $\nabla U(\mathbf{x})$, Welling and Teh [2011] proposed the Stochastic Gradient Langevin Dynamics (SGLD) and updates the weights using stochastic gradients:

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\eta\nabla\tilde{U}(\mathbf{x}_k)+\sqrt{2\eta}\xi_{k+1}, \tag{17}$$

where η is the stepsize, the ξ_{k+1} is a standard Gaussian noise, and $\nabla\tilde{U}(\mathbf{x}_k)$ is an unbiased estimation of $\nabla U(\mathbf{x}_k)$. Despite the additional noise induced by stochastic gradient estimations, SGLD can still converge to the target distribution.

The low-precision SGLD with full-precision gradient accumulators (SGLDLP-F) only quantizes weights before computing the gradient. The update rule can be defined as:

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\eta Q_G\left(\nabla\tilde{U}(Q_W(\mathbf{x}_k))\right)+\sqrt{2\eta}\xi_{k+1}. \tag{18}$$

Zhang et al. [2022] shows that the SGLDLP-F outperforms its counterpart low-precision SGD with full-gradient accumulators (SGDLP-F). The computation costs can be further reduced using low-precision gradient accumulators by only keeping low-precision weights. Low-precision SGLD with low-precision gradient accumulators (SGLDLP-L) can be defined as the following:

$$\mathbf{x}_{k+1}=Q_W\left(\mathbf{x}_k-\eta Q_G(\nabla\tilde{U}(\mathbf{x}_k))+\sqrt{2\eta}\xi_{k+1}\right). \tag{19}$$

Zhang et al. [2022] studied the convergence property of both SGLDLP-F and SGLDLP-L under strongly-log-concave distributions, and showed that a small stepsize deteriorates the performance of SGLDLP-L. To mitigate this problem, Zhang et al. [2022] proposed a variance-corrected quantization function.

Theorem 10. *Suppose Assumptions 1, 3 and 4 hold. Let $\widetilde{A}$ have the same definition in Theorem 5, and λ^* be the concentration number of* (16). *After K steps starting with initial point $\mathbf{x}_0=0$, if we set the stepsize to be $\eta=\tilde{\mathcal{O}}\left(\left(\frac{\epsilon}{\log(1/\epsilon)}\right)^4\right)$. The output $\mathbf{x}_K$ of SGLDLP-F in* (18) *satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K),p^*)\leq\tilde{\mathcal{O}}\left(\epsilon+\widetilde{A}\log\left(\frac{1}{\epsilon}\right)\right), \tag{20}$$

provided

$$K=\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4\lambda^*}\log^5\left(\frac{1}{\epsilon}\right)\right).$$

Algorithm 2 Variance-Corrected Low-Precision SGHMC (VC SGHMCLP-L).

given: Stepsize η, friction γ, inverse mass u, number of training iterations K, gradient quantizer Q_G, quantization gap Δ and upper bound of low-precision representation U. Let $\text{Var}_{\mathbf{v}}^{hmc} = u(1-e^{-2\gamma\eta})$ and $\text{Var}_{\mathbf{x}}^{hmc} = u\gamma^{-2}(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3)$ and $S_{\mathbf{v}} = 1$ {Initialize the scaling parameter}.
for $k = 1 : K$ **do**
  rescale $\mathbf{v}_k = \mathbf{v}_k * S_v$ {Restore the velocity before update}
  update $\mu(\mathbf{v}_{k+1}) = \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{-\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_k))$
  update $\mu(\mathbf{x}_{k+1}) = \mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(\mathbf{x}_k))$
  update $S_{\mathbf{v}} = \frac{\|\mu(\mathbf{v}_{k+1})\|_\infty}{U}$ {Update the Scaling}
  update $\mathbf{v}_{k+1} \leftarrow Q^{vc}\left(\mu(\mathbf{v}_{k+1})/S_{\mathbf{v}}, Var_{\mathbf{v}}^{hmc}/S_{\mathbf{v}}^2, \Delta\right)$
  update $\mathbf{x}_{k+1} \leftarrow Q^{vc}\left(\mu(\mathbf{x}_{k+1}), Var_{\mathbf{x}}^{hmc}, \Delta\right)$
end for
output: samples $\{x_k\}$

Theorem 10 shows that the low-precision SGLD with full-precision gradient accumulators can converge to the non-log-concave target distribution provided a small gradient variance and quantization error. Next, we present the SGLDLP-L's result.

Theorem 11. *Let Assumptions 1, 3 and 4 hold. If we set the step size to be* $\eta = \tilde{\mathcal{O}}\left(\left(\frac{\epsilon}{\log(1/\epsilon)}\right)^4\right)$, *after K steps starting at the initial point $\mathbf{x}_0 = 0$ the output $\mathbf{x}_K$ of the SGLDLP-L in* (19) *satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\max\{\sigma^2, \sigma\}}\log\left(\frac{1}{\epsilon}\right) + \frac{\log^5\left(\frac{1}{\epsilon}\right)}{\epsilon^4}\sqrt{\Delta}\right), \tag{21}$$

provided

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4\lambda^*}\log^5\left(\frac{1}{\epsilon}\right)\right).$$

The VC SGLDLP-L can be done as:

$$\mathbf{x}_{k+1} = Q^{vc}\left(\mathbf{x}_k - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k)), 2\eta, \Delta\right) \tag{22}$$

Theorem 12. *Let Assumption 1, 3 and 4 hold. If we set the stepsize to be* $\eta = \tilde{\mathcal{O}}\left(\frac{\epsilon^4}{\log^4\left(\frac{1}{\epsilon}\right)}\right)$, *after K steps from the initial point $\mathbf{x}_0 = 0$ the output $\mathbf{x}_K$ of VC SGLDLP-L in* (22) *satisfies*

$$\mathcal{W}_2(p(\mathbf{x}_K), p^*) = \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\max\{\sigma^2, \sigma\}\log\left(\frac{1}{\epsilon}\right)} + \frac{\log^3\left(\frac{1}{\epsilon}\right)}{\epsilon^2}\sqrt{\Delta}\right), \tag{23}$$

provided

$$K = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^4\lambda^*}\log^5\left(\frac{1}{\epsilon}\right)\right).$$

C Technical Detail

In this section, we disclose more details of empirical experiments. When implementing low-precision SGHMC on classification task in the CIFAR-10 and CIFAR-100 dataset, we observed that the momentum term $\mathbf{v}$ tend to gather in a small range around zero in which case the low-precision representations of $\mathbf{v}$ end up in gathering only few points, thus the momentum information is seriously lost and cause in performance degradation. In order to tackle this problem and fully utilize all the low-precision representations, we borrow the idea of rescaling from the bit centering trick and adopted to the low-precision SGHMC method. The detailed algorithm is listed in Algorithm 2.

In Algorithm 2, we introduce the bit centering trick [De Sa et al., 2018] to enhance the variance corrected quantization function. Bit centering trick is a technique to increase the accuracy low-precision training algorithm by recentering and rescaling representable bits making low-precision numbers closer to its real full-precision counterpart. We borrow the idea of rescaling to enhance the variance-corrected quantization function. Based on the discussion in previous paragraph, when the desired variance v is small the variance corrected quantization has a high chance to match the variance. By scaling up the weights, additional to increasing the accuracy of low-precision representation also increase the desired variance resulting in a lower chance of fail in variance corrected quantization.

D Proof of Main Theorems

D.1 Proof of Theorem 4

Section 3.1 introduces low-precision HMC with full-precision gradient accumulators (SGHMCLP-F) as:

$$\begin{aligned}
&\mathrm{v}\mathbf{v}_{k+1} = \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{-\gamma\eta})Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}_k))) + \xi_k^{\mathbf{v}} \\
&\mathrm{v}\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}_k))) + \xi_k^{\mathbf{x}},
\end{aligned}$$

In this section, we prove the convergence of SGHMCLP-F in terms of 2-Wasserstein distance for strongly-log-concave target distribution via coupling argument. To simplify the notation we define the quantized stochastic gradients at $\mathbf{x}$ as:

$$\tilde{g}(\mathbf{x}) := Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}))) \tag{24}$$

$$=: \nabla U(\mathbf{x}) + \xi. \tag{25}$$

Lemma 13. *For any $\mathbf{x} \in \mathbb{R}^d$, the random noise ξ of the low-precision gradients defined in* (25) *satisfies:*

$$\|\mathbb{E}\xi\|^2 \leq M^2\frac{\Delta^2 d}{4}$$

$$\mathbb{E}[\|\xi\|^2] \leq (M^2+1)\frac{\Delta^2 d}{4} + \sigma^2.$$

We follow the proof in Cheng et al. [2018]. Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. Given probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we define a *transference plan* ζ between μ and ν as a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for all sets $A \in \mathbb{R}^d$, $\zeta(A \times \mathbb{R}^d) = \mu(A)$ and $\zeta(\mathbb{R}^d \times A) = \nu(A)$. We denote $\Gamma(\mu,\nu)$ as the set of all transference plans. A pair of random variables $(\mathbf{x},\mathbf{y})$ is called a coupling if there exists a $\zeta \in \Gamma(\mu,\nu)$ such that $(\mathbf{x},\mathbf{y})$ is distributed according to ζ. (With some abuse of notation, we will also refer to ζ as the coupling.)

In order to calculate the Wasserstein distance from the proposed sample $(\mathbf{x}_K, \mathbf{v}_K)$ and the target distribution sample $(\mathbf{x}^*, \mathbf{v}^*)$, we define sample $q_k = (\mathbf{x}_k, \mathbf{x}_k + \mathbf{v}_k)$ and the target distribution sample $q^* = (\mathbf{x}^*, \mathbf{x}^* + \mathbf{v}^*)$. Let $p_k = (\mathbf{x}_k, \mathbf{v}_k)$ and $\widehat{\Phi}_\eta$ be the operator that maps from p_k to p_{k+1} i.e.

$$p_{k+1} = \widehat{\Phi}_\eta p_k.$$

The solution $(\mathbf{x}_t, \mathbf{v}_t)$ of the continuous underdamped Langevin dynamics with exact gradient satisfies the following equations:

$$\mathbf{v}_t = \mathbf{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\nabla U(\mathbf{x}_s)ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s, \tag{26}$$

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t \tilde{\mathbf{v}}_s ds.$$

Let Φ_η denote the operator that maps p_0 to the solution of continuous underdamped Langevin dynamics in (26) after time step η. Notice the solution $(\tilde{\mathbf{v}}_t, \tilde{\mathbf{x}}_t)$ of the discrete underdamped Langevin dynamics with an exact gradient can be written as

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\nabla U(\tilde{\mathbf{x}}_0)ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s, \tag{27}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + \int_0^t \tilde{\mathbf{v}}_s ds.$$

We can also define a similar operator for the discrete underdamped Langevin dynamics solution $\tilde{p}_t = (\tilde{\mathbf{x}}_t, \tilde{\mathbf{v}}_t)$, let $\widetilde{\Phi}_t$ be the operator that maps $\tilde{p}_0$ to $\tilde{p}_t$. Furthermore the SGHMCLP-F can be written as:

$$\mathbf{v}_t = \mathbf{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\tilde{g}(\mathbf{x}_0)ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s, \tag{28}$$

$$\mathbf{x}_t = \tilde{\mathbf{x}}_0 + \int_0^t \mathbf{v}_s ds.$$

Given $\tilde{g}(\mathbf{x}_0) = \nabla U(\mathbf{x}_0) + \xi_0$ and $\mathbf{x}_0 = \tilde{\mathbf{x}}_0$, we know:

$$\mathbf{v}_t = \tilde{\mathbf{v}}_t - u\left(\int_0^t e^{-\gamma(t-s)}ds\right)\xi \tag{29}$$

$$\mathbf{x}_t = \tilde{\mathbf{x}}_t - u\left(\int_0^t \left(\int_0^r e^{-\gamma(t-s)}ds\right)dr\right)\xi.$$

Lemma 14. *Let q_0 be some initial distribution and $\widetilde{\Phi}_\eta$ and Φ_η be the operator we defined above for discrete Langevin dynamics with exact full-precision gradients and low-precision gradients respectively. If the stepszie $1 > \eta > 0$, then the Wasserstein distance satisfies*

$$\mathcal{W}_2^2(\Phi_\eta q_0, q^*) \leq \left(\mathcal{W}_2(\widetilde{\Phi}_\eta q_0, q^*) + \sqrt{5}/2u\eta\sqrt{d}M\Delta\right)^2 + 5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right).$$

The lemma 14 says that if starting from the same distribution after one step of low-precision update the Wasserstein distance from the target distribution is bounded by the distance after one step of exact gradients plus $\mathcal{O}(\eta^2\Delta^2)$. Furthermore from the corollary 7 in Cheng et al. [2018] we know that for any $i \in \{1, \cdots, K\}$:

$$\mathcal{W}_2^2(\Phi_\eta q_i, q^*) \leq e^{-\eta/2\kappa_1}\mathcal{W}_2^2(q_i, q^*), \tag{30}$$

where $\kappa_1 = M/m_1$ is the condtion number. Let $\mathcal{E}_K$ denote the $26\left(d/m_1 + \mathcal{D}^2\right)$, and from the discretization error bound from Theorem 9 and Lemma 8 (sandwich inequality) in Cheng et al. [2018], we get

$$\mathcal{W}_2(\Phi_\eta q_i, \widetilde{\Phi}_\eta q_i) \leq 2\mathcal{W}_2(\Phi_\eta p_i, \widetilde{\Phi}_\eta p_i) \leq \eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}.$$

By triangle inequality:

$$\begin{aligned}\mathcal{W}_2(\widetilde{\Phi}_\eta q_i, q^*) &\leq \mathcal{W}_2(\Phi_\eta q_i, \widetilde{\Phi}_\eta q_i) + \mathcal{W}_2(\Phi_\eta q_i, q^*)\\ &\leq \eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + e^{-\eta/2\kappa_1}\mathcal{W}_2(q_i, q^*).\end{aligned}$$

Combine this with the result in Lemma 14 we have,

$$\mathcal{W}_2^2(\widehat{\Phi}_\eta q_i, q^*) \leq \left(e^{-\eta/2\kappa_1}\mathcal{W}_2(q_i, q^*) + \eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{5}/2u\eta\sqrt{d}M\Delta\right)^2 + 5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right).$$

By invoking the Lemma 7 in Dalalyan and Karagulyan [2019] we can bound the 2-Wasserstein distance by:

$$\begin{aligned}\mathcal{W}_2(q_K, q^*) \leq{}& e^{-K\eta/2\kappa_1}\mathcal{W}_2(q_0, q^*) + \frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2}}{1 - e^{-\eta/2\kappa_1}}\\ &+ \frac{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2} + \sqrt{1 - e^{-\eta/\kappa_1}}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}}.\end{aligned}$$

Finally by sandwich inequality we have:

$$\begin{aligned}\mathcal{W}_2(p_K, p^*) \leq{}& 4e^{-K\eta/2\kappa}\mathcal{W}_2(p_0, p^*) + 4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2}}{1 - e^{-\eta/2\kappa}}\\ &+ \frac{20u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2} + \sqrt{1 - e^{-\eta/\kappa}}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}}.\end{aligned}$$

Now we let the first term less than $\epsilon/3$, from the lemma 13 in [Cheng et al., 2018] we know that $\mathcal{W}_2(p_K, p^*) \leq 3\left(\frac{d}{m_1} + \mathcal{D}^2\right)$. So we can choose K as the following,

$$K \leq \frac{2\kappa_1}{\eta} \log\left(36\left(\frac{d}{m_1} + \mathcal{D}^2\right)\right).$$

Next, we choose a stepsize $\eta \leq \frac{\epsilon\kappa_1^{-1}}{\sqrt{479232/5(d/m_1+\mathcal{D}^2)}}$ to ensure the second term is controlled below $\epsilon/3 + \frac{16\kappa_1 uM\Delta\sqrt{5d}}{2}$. Since $1 - e^{-\eta/2\kappa_1} \geq \eta/4\kappa_1$ and definition of $\mathcal{E}_K$,

$$\begin{aligned}
4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2}}{1 - e^{-\eta/2\kappa}} &\leq 4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2}}{\eta/4\kappa_1} \leq 16\kappa_1\left(\eta\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{uM\Delta\sqrt{5d}}{2}\right)\\
&\leq \epsilon/3 + \frac{16\kappa_1 uM\Delta\sqrt{5d}}{2}.
\end{aligned}$$

Finally by choosing the stepsize satisfied that,

$$\eta \leq \frac{\epsilon M\Delta\sqrt{5d}}{120u\left[(M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right]},$$

the third term can be bounded as:

$$\begin{aligned}
&\frac{20u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \frac{u\eta M\Delta\sqrt{5d}}{2} + \sqrt{1 - e^{-\eta/\kappa}}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}}\\
&\leq \frac{20u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}{\frac{u\eta M\Delta\sqrt{5d}}{2}} = 40u\eta\frac{\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}{M\Delta\sqrt{5d}} \leq \epsilon/3.
\end{aligned}$$

This complete the proof.

D.2 Proof of Theorem 5

In this section we analyze the Wasserstein distance between the sample $(\mathbf{x}_k, v_K)$ in (3) and the target distribution, given the target distribution satisfies the assumption 1 and 3. We follow the proof in Raginsky et al. [2017]. To analyze the Wasserstein distance, we first calculate the distance between solutions of low-precision discrete underdamped Langevin dynamics and solutions of the ideal continuous underdamped Langevin dynamics, also the distance between solutions of the ideal continuous underdamped Langevin dynamics and the target distribution.

Again let $p_k = (\mathbf{x}_k, v_k)$ denote the low-precision sample from (3) at k-th iteration, let $\hat{p}_t = (\hat{x}_t, \hat{v}_t)$ denote the sample from the ideal continuous underdamped Langevin dynamics in 26 at time t. Then the Wasserstein distance between the p_k and the target distribution p^* can be bounded as:

$$\mathcal{W}_2(p_K, p^*) \leq \mathcal{W}_2(p_K, \hat{p}_{K\eta}) + \mathcal{W}_2(\hat{p}_{K\eta}, p^*).$$

We first bound $\mathcal{W}_2(p_K, \hat{p}_{K\eta})$ by invoking the weighted CKP inequality Bolley and Villani [2005],

$$\mathcal{W}_2^2(p_K, \hat{p}_{K\eta}) \leq \Lambda\left(\sqrt{D_{KL}(p_K||\hat{p}_{K\eta})} + \sqrt[4]{D_{KL}(p_K||\hat{p}_{K\eta})}\right),$$

where $\Lambda = 2\inf_{\theta>0}\sqrt{1/\theta\left(3/2 + log\mathbb{E}_{\hat{p}_{K\eta}}\left[exp(\theta(\|\hat{x}_{K\eta}\|^2 + \|\hat{v}_{K\eta}\|^2))\right]\right)}$. We define a Lyapunov function for every $(x, v) \in \mathbb{R}^d \times \mathbb{R}^d$

$$\mathcal{E}(\mathbf{x}, \mathbf{v}) = \|\mathbf{x}\|^2 + \|\mathbf{x} + 2\mathbf{v}/\gamma\|^2 + 8u(U(\mathbf{x}) - U(\mathbf{x}^*))/\gamma^2.$$

Note that $\|a\|^2 + \|b\|^2 \geq \|a-b\|^2/2$ and $U(x) \geq U(x^*)$, we can have:

$$\mathcal{E}(x, v) \geq \|x\|^2 + \|x + 2v/\gamma\|^2 \geq \max\{\|x\|^2, 2\|v/\gamma\|^2\}.$$

Given assumptions 2 and 3 hold and apply Lemma B.4 in Zou et al. [2019], we can get

$$\begin{aligned}\Lambda \leq & 2 \inf_{0<\theta\leq\min\{\frac{\gamma}{128u},\frac{m_2}{32}\}} \sqrt{\frac{1}{\theta}\left(\frac{3}{2}+2\theta\mathcal{E}(\mathbf{X}_0,\mathbf{V}_0)+\frac{32M\theta u(4d+2b+m_2\|\mathbf{x}^*\|^2)}{\gamma^2 m_2}\right)} \\ \leq & 2\sqrt{2\mathcal{E}(\mathbf{X}_0,\mathbf{V}_0)+\frac{32M\theta u(4d+2b+m_2\|\mathbf{x}^*\|^2)+16(12um_2+3\gamma^2)}{\gamma^2 m_2}} := \bar{\Lambda}.\end{aligned}$$

It remains to bound the divergence between the distribution p_K and $\hat{p}_{K\eta}$. We first define a continuous interpolation of the low-precision sample $(\mathbf{x}_k, \mathbf{v}_k)$,

$$d\mathbf{v}_t = -\gamma\mathbf{v}_t dt - uG_t dt + \sqrt{2\gamma u} dB_t \tag{31}$$

$$d\mathbf{x}_t = \mathbf{v}_t dt, \tag{32}$$

where $G_t = \sum_{k=0}^{K} \tilde{g}(\mathbf{x}_k)\mathbf{1}_{\mathrm{t}\in[\mathrm{k}\eta,(\mathrm{k}+1)\eta)}$. Integrating this equation from time 0 to t, we can get

$$\begin{aligned}\mathbf{v}_t &= \mathbf{v}_0 - \int_0^t \gamma\mathbf{v}_s ds - \int_0^t uG_s dt + \int_0^t \sqrt{2\gamma u} dB_s \\ \mathbf{x}_t &= \mathbf{x}_0 + \int_0^t \mathbf{v}_s ds.\end{aligned}$$

Notice that when $t = k\eta$, the solution of (31) has the same distribution with the low-precision sample $(\mathbf{x}_k, \mathbf{v}_k)$. Now by Girsanov formula we can compute the Radon-Nikodym derivative of $\hat{p}_{K\eta}$ with respect to p_K as follow:

$$\frac{d\hat{p}_{K\eta}}{dp_K} = exp\left\{\sqrt{\frac{\gamma u}{2}}\int_0^t (\nabla U(\mathbf{x}_s) - G_s) d\mathbf{B}s - \frac{\gamma u}{4}\int_0^T \|\nabla U(\mathbf{x}_s) - G_s\| ds\right\}.$$

It follows that

$$\begin{aligned}D_{KL}(p_K\|\hat{p}_{K\eta}) &= \mathbb{E}_{p_K}\left[\log\left(\frac{d\hat{p}_{K\eta}}{dp_K}\right)\right] \\ &= \frac{\gamma u}{4}\mathbb{E}\int_0^{K\eta} \|\nabla U(\mathbf{x}_s) - G_s\|^2 ds \\ &= \frac{\gamma u}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - G_s\|^2\right] ds \\ &= \frac{\gamma u}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \tilde{g}(\mathbf{x}_k)\|^2\right] ds.\end{aligned} \tag{33}$$

Furthermore, in the k-th interval, we have

$$\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \tilde{g}(\mathbf{x}_k)\|^2\right] \leq 2\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k)\|^2\right] + 2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right]. \tag{34}$$

We now bound the first term in the RHS of the (34). By the smooth Assumption1, we have

$$\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k)\|^2\right] \leq M^2\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_k\|^2\right].$$

Notice that

$$\begin{aligned}\mathbf{x}_s &= \mathbf{x}_k + \int_{k\eta}^s \mathbf{v}_r dr \\ &= \mathbf{x}_k + \int_{k\eta}^s \left(\mathbf{v}_{k\eta}e^{-\gamma(r-k\eta)} - u\left(\int_{k\eta}^r e^{-\gamma(r-z)}\tilde{g}(\mathbf{x}_k)dz\right) + \sqrt{2\gamma u}\int_{k\eta}^r e^{-\gamma(r-z)}dB_z\right) dr.\end{aligned}$$

This further implies that:

$$
\begin{aligned}
\|\mathbf{x}_s - \mathbf{x}_k\|^2 =& \left\| \int_{k\eta}^{s} \left(\mathbf{v}_{k\eta} e^{-\gamma(r-k\eta)} - u \left(\int_{k\eta}^{r} e^{-\gamma(r-z)} \tilde{g}(\mathbf{x}_k) dz \right) + \sqrt{2\gamma u} \int_{k\eta}^{r} e^{-\gamma(r-z)} dB_z \right) dr \right\|^2 \\
\le& 3 \left\| \int_{k\eta}^{s} \mathbf{v}_{k\eta} e^{\gamma(k\eta - r)} dr \right\|^2 + 3 \left\| \int_{k\eta}^{s} \int_{k\eta}^{r} u \tilde{g}(\mathbf{x}_k) e^{\gamma(z-r)} dz dr \right\|^2 + 6ru \left\| \int_{k\eta}^{s} \int_{0}^{s} e^{-\gamma(r-z)} dB_z dr \right\|^2 \\
\le& 3\eta^2 \left\| \mathbf{v}_k \right\|^2 + 3u^2\eta^4 \left\| \tilde{g}(\mathbf{x}_k) \right\|^2 + 3 \left[\frac{u}{\gamma^2} \left(2\gamma(s-k\eta) + 4e^{-\gamma(s-k\eta)} - e^{-2\gamma(s-k\eta)} - 3 \right) d \right] \\
\le& 3\eta^2 \left(\left\| \mathbf{v}_k \right\|^2 + u^2\eta^2 \left\| \tilde{g}(\mathbf{x}_k) \right\|^2 + 2du \right), \qquad (35)
\end{aligned}
$$

where we use inequality $1 - x \le e^{-x} \le 1 - x + x^2/2$ for $x > 0$ and $k\eta \le s \le (k+1)\eta$ to get the last inequality. Given this analysis we can bound the first term in the RHS of (34)

$$
\mathbb{E}\left[\left\| \nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k) \right\|^2 \right] \le 3M^2\eta^2 \left(\mathbb{E}\left\| v_k \right\|^2 + u^2\eta^2 \mathbb{E}\left\| \tilde{g}(\mathbf{x}_k) \right\|^2 + 2du \right).
$$

By lemma 13, the second term in the RHS of (34) can be bounded as:

$$
\mathbb{E}\left[\left\| \nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k) \right\|^2 \right] \le (M^2+1)\frac{\Delta^2 d}{4} + \sigma^2.
$$

We need to introduce a lemma to bound the $\sup_k \|\mathbf{x}_k\|^2$, $\sup_k \|v_k\|^2$ and $\sup_k \|\tilde{g}(\mathbf{x}_k)\|^2$.

Lemma 15. *Under Assumptions 1 and 3, if we set the stepsize statisfied the following condition:*

$$
\begin{aligned}
\eta \le min \Bigg\{ & \frac{\gamma}{4\left(8Mu + u\gamma + 22\gamma^2\right)}, \sqrt{\frac{4u^2}{4Mu + 3\gamma^2}}, \frac{6\gamma bu}{\left(4Mu + 3\gamma^2\right) d}, \\
& \frac{1}{8\gamma}, \frac{\gamma m_2}{12(21u+\gamma)M^2}, \frac{8(\gamma^2+2u)}{(20u+\gamma)\gamma} \Bigg\},
\end{aligned}
$$

then for all $k \ge 0$ the $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$, $\mathbb{E}\left[\|v_k\|^2\right]$ and $\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right]$ can be bounded as

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_k\|^2 \right] &\le \overline{\mathcal{E}} + C_0 \left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2 \right) \\
\mathbb{E}\left[\|v_k\|^2 \right] &\le \gamma^2 \overline{\mathcal{E}}/2 + \gamma^2 C_0/2 \left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2 \right) \\
\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2 \right] &\le 2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2 \right) + 4M^2\overline{\mathcal{E}} + 4G^2
\end{aligned}
$$

where $\overline{\mathcal{E}}$ and C_0 are defined as:

$$
\begin{aligned}
\overline{\mathcal{E}} &= \mathbb{E}\left[\mathcal{E}(\mathbf{x}_0, \mathbf{v}_0)\right] + \frac{24(21u+\gamma)uM}{m_2\gamma^3} G^2 + \frac{96(d+b)uM}{m_2\gamma^2}, \quad G = \|\nabla U(0)\| \\
C_0 &= \frac{96u\left(\gamma^2 + 2u\right)}{m_2\gamma^4}.
\end{aligned}
$$

We now ready to bound $\mathbb{E}\left[\|\nabla U(\mathbf{x}_s - \tilde{g}(\mathbf{x}_k))\|^2\right]$ as:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \tilde{g}(\mathbf{x}_k)\|^2\right] &\leq 2\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k)\|^2\right] + 2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right] \\
&\leq 6M^2\eta^2\left(\mathbb{E}\|v_k\|^2 + u^2\eta^2\mathbb{E}\|\tilde{g}(\mathbf{x}_k)\|^2 + 2du\right) + 2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) \\
&\leq 6M^2\eta^2\left((\gamma^2/2 + 4M^2u^2\eta^2)\overline{\mathcal{E}} + (\gamma^2 C_0/2 + 2u^2\eta^2)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) + 4u^2\eta^2 G^2 + 2du\right. \\
&+ 2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) \\
&\leq 6M^2\eta^2\left[(\gamma^2/2 + 4M^2u^2\eta^2)\overline{\mathcal{E}} + 4u^2\eta^2 G^2 + 2du\right] \\
&+ \left(6M^2\eta^2(\gamma^2 C_0/2 + 2u^2\eta^2) + 2\right)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right).
\end{aligned}$$

Thus the divergence can be bounded as:

$$\begin{aligned}
D_{KL}(p_K||\hat{p}_{K\eta}) \leq \frac{3\gamma u}{2}M^2 K\eta^3\left[(\gamma^2/2 + 4M^2u^2\eta^2)\overline{\mathcal{E}} + 4u^2\eta^2 G^2 + 2du\right] \\
+ \frac{\gamma u}{4}K\eta\left(6M^2\eta^2(\gamma^2 C_0/2 + 2u^2\eta^2) + 2\right)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right).
\end{aligned}$$

By the weighted CKP inequality and given $K\eta \geq 1$,

$$\begin{aligned}
\mathcal{W}_2(p_K, \hat{p}_{K\eta}) &\leq \overline{\Lambda}\left(\sqrt{D_{KL}(p_K||\hat{p}_{K\eta})} + \sqrt[4]{D_{KL}(p_K||\hat{p}_{K\eta})}\right) \\
&\leq \overline{\Lambda}\left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\widetilde{A}\right)\sqrt{K\eta},
\end{aligned}$$

where the constants $\widetilde{C_0}$, $\widetilde{C}_1$ and $\widetilde{A}$ are defined as:

$$\begin{aligned}
\widetilde{C_0} &= \sqrt{\frac{3\gamma u}{2}M^2\left[(\gamma^2/2 + 4M^2u^2\eta^2)\overline{\mathcal{E}} + 4u^2\eta^2 G^2 + 2du\right]} + \sqrt[4]{\frac{3\gamma u}{2}M^2\left[(\gamma^2/2 + 4M^2u^2\eta^2)\overline{\mathcal{E}} + 4u^2\eta^2 G^2 + 2du\right]} \\
\widetilde{C_1} &= \sqrt{\frac{\gamma u}{4}\left(6M^2\eta^2(\gamma^2 C_0/2 + 2u^2\eta^2) + 2\right)} + \sqrt[4]{\frac{\gamma u}{4}\left(6M^2\eta^2(\gamma^2 C_0/2 + 2u^2\eta^2) + 2\right)} \\
\widetilde{A} &= \max\left\{\sqrt{\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}, \sqrt[4]{\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)}\right\}.
\end{aligned}$$

Finally by the Lemma A.2 in Zou et al. [2019], we can have

$$\mathcal{W}_2(\hat{p}_{K\eta}, p^*) \leq \Gamma_0 e^{-\mu^* K\eta},$$

where $\mu^* = e^{-\widetilde{\mathcal{O}}(d)}$ denotes the concentration rate of the underdamped Langevin dynamics and Γ_0 is a constant of order $\mathcal{O}(1/\mu^*)$. Combining this inequality with previous analysis we can prove:

$$\mathcal{W}_2(p_K, p^*) \leq \overline{\Lambda}\left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\widetilde{A}\right)\sqrt{K\eta} + \Gamma_0 e^{-\mu^* K\eta}. \tag{36}$$

In order to bound the Wasserstein distance, we need to set

$$\overline{\Lambda}\widetilde{C_0}\sqrt{K\eta^2} = \frac{\epsilon}{2} \quad \text{and} \quad \Gamma_0 e^{-\mu^* K\eta} = \frac{\epsilon}{2}. \tag{37}$$

Solving the equation (37), we can have

$$K\eta = \frac{\log\left(\frac{2\Gamma_0}{\epsilon}\right)}{\mu^*} \quad \text{and} \quad \eta = \frac{\epsilon^2}{4\overline{\Lambda}^2\widetilde{C_0}^2 K\eta}.$$

Combining these two we can have

$$\eta = \frac{\epsilon^2\mu^*}{4\overline{\Lambda}^2\widetilde{C_0}^2\log\left(\frac{2\Gamma_0}{\epsilon}\right)} \quad \text{and} \quad K = \frac{4\overline{\Lambda}^2\widetilde{C_0}^2\log^2\left(\frac{2\Gamma_0}{\epsilon}\right)}{\epsilon^2\left(\mu^*\right)^2}.$$

Plugging in (36) completes the proof.

D.3 Proof of Thoerem 10

In this section we generalize the convergence analysis of LPSGLDLP-F in Zhang et al. [2022] to non-log-concave target distribution. We prove a more general version of theorem 10 following the same proof outlines in Raginsky et al. [2017]. We further introduce an assumption about the initial distribution p_0.

Assumption 5. *The probability p_0 of the initial hypothesis $\mathbf{x}_0$ has a bounded and strictly positive density and satisfies the following:*

$$\kappa_0 := \log \int_{\mathbb{R}^d} e^{\|x\|^2} p_0(x)dx < \infty.$$

Note that the for initial distribution $\mathbf{x}_0 = 0$, the value $\kappa_0 = 0$ is bounded and the assumption is satisfied. Recall the Overdamped Langevin dynamics is

$$d\mathbf{x}_t = -\nabla U(\mathbf{x}_t)dt + \sqrt{2}dB_t. \tag{38}$$

We further define the value of the energy function and the gradient at point 0 at the following:

$$|U(0)| = G_0, \quad \|\nabla U(0)\| = G_1.$$

In order to analyze the convergence of SGLD for non-log-concave distribution, we need to introduce extra assumptions.

Then the solution of the Langevin dynamics should satisfies

$$\mathbf{x}_t = \mathbf{x}_0 - \int_0^t \nabla U(\mathbf{x}_s)ds + \sqrt{2}\int_0^t dB_s. \tag{39}$$

To analysis the LPSGLDLP-F in (18), we define a counituous interpolation of the low-precison sample as:

$$\hat{x}_t = \hat{x}_0 - \int_0^t G_s ds + \sqrt{2}\int_0^t dB_s, \tag{40}$$

where $G_s = \sum_{k=0}^{K} \tilde{g}(\hat{x}_k)\mathbf{1}_{\mathrm{s}\in[\mathrm{k}\eta,(\mathrm{k}+1)\eta)}$. The Wasserstein distance can bounded as

$$\mathcal{W}_2(p_K, p^*) \leq \mathcal{W}_2(p_K, \hat{p}_{K\eta}) + \mathcal{W}_2(\hat{p}_{K\eta}, p^*),$$

where the first term of the RHS can be bounded via the weighted CKP inequality

$$\mathcal{W}_2(p_K, \hat{p}_{K\eta}) \leq C_{\hat{p}_{K\eta}} \left[\sqrt{D_{KL}\left(p_K||\hat{p}_{K\eta}\right)} + \left(\frac{D_{KL}\left(p_K||\hat{p}_{K\eta}\right)}{2}\right)^{1/4}\right],$$

where the constant $C_{\hat{p}_{K\eta}} = 2\inf_{\lambda>0}\left(\frac{1}{\lambda}\left(\frac{3}{2} + \log\int_{\mathbb{R}^d} e^{\lambda\|\omega\|^2}\hat{P}_{K\eta}(d\omega)\right)\right)$. By Lemma 4 in Raginsky et al. [2017] and assuming $K\eta > 1$, we can wrtie:

$$\mathcal{W}_2^2(p_K, \hat{p}_{K\eta}) \leq \left(12 + 8\left(\kappa_0 + 2b + 2d\right)K\eta\right)\left(D_{KL}\left(p_K||\hat{p}_{K\eta}\right) + \sqrt{D_{KL}\left(p_K||\hat{p}_{K\eta}\right)}\right).$$

Now we bound the term $D_{KL}\left(p_K||\hat{p}_{K\eta}\right)$. The Radon-Nikodym derivative of the $\hat{P}_{K\eta}$ w.r.t p_K is the following

$$\frac{d\hat{p}_{K\eta}}{dp_K} = exp\left\{\frac{1}{2}\int_0^t (\nabla U(\mathbf{x}_s) - G_s)d\mathbf{B}s - \frac{1}{4}\int_0^T \|\nabla U(\mathbf{x}_s) - G_s\|ds\right\}.$$

Thus, we have:

$$
\begin{aligned}
D_{KL}(p_K||\hat{p}_{K\eta}) &= \mathbb{E}_{p_K}\left[\log\left(\frac{d\hat{p}_{K\eta}}{dp_K}\right)\right] \\
&= \frac{1}{4}\int_0^{K\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - G_s\|^2\right]ds \\
&= \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \tilde{g}(\mathbf{x}_k)\|^2\right]ds \\
&\leq \frac{1}{2}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k)\|^2\right] \\
&+ \frac{1}{2}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right] \\
&\leq \frac{M^2}{2}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_k\|^2\right] \\
&+ \frac{1}{2}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right]. \qquad (41)
\end{aligned}
$$

We now bound the first term in the RHS of the equation 41, from the update rule in 40 we know:

$$
\begin{aligned}
\mathbf{x}_s - \mathbf{x}_k &= -(s-k\eta)\tilde{g}(\mathbf{x}_k) + \sqrt{2}\left(B_s - B_{k\eta}\right) \\
&= -(s-k\eta)\nabla U(\mathbf{x}_k) + (s-k\eta)\left(\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\right) + \sqrt{2}\left(B_s - B_{k\eta}\right),
\end{aligned}
$$

thus,

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_k\|^2\right] &\leq 3\eta^2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right] + 3\eta^2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right] + 6\eta d \\
&\leq 3\eta^2\left(M\mathbb{E}\left[\|\mathbf{x}_k\|\right] + G\right)^2 + 3\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) + 6\eta d. \qquad (42)
\end{aligned}
$$

Similarly, we need a uniform bound of $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$.

Lemma 16. *Under assumptions 1, 3 and 4, if we set the stepsize $\eta \in \left(0, 1 \wedge \frac{m_2}{2M^2}\right)$, then for all $k \geq 0$, the $\mathbb{E}\left[\|\mathrm{v}\mathbf{x}_k\|^2\right]$ can be bounded as*

$$
\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \leq \mathcal{E} + \frac{2\left(M^2+1\right)\Delta^2 d}{4m_2},
$$

provided $\mathcal{E} = \mathbb{E}\left[\|\mathbf{x}_0\|^2\right] + \frac{M}{m_2}\left(2b + 2\eta G^2 + 2d\right)$.

Using this bound, we can further bound $\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_s\|^2\right]$ as:

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_s\|^2\right] &\leq 6\eta^2 M^2\left(\mathcal{E} + \frac{2\left(M^2+1\right)}{m_2}\frac{\Delta^2 d}{4}\right) + 6\eta^2 G^2 + 3\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) + 6\eta d \\
&\leq 6\eta^2 M^2\mathcal{E} + 6\eta^2 G^2 + 6\eta d + \left(\frac{12\eta^2 M^2\left(M^2+1\right)}{m_2} + 3(M^2+1)\right)\eta^2\frac{\Delta^2 d}{4} + 3\eta^2\sigma^2, \\
&=: \overline{\mathcal{E}}\eta + C\eta^2\frac{\Delta^2 d}{4} + 3\eta^2\sigma^2
\end{aligned}
$$

where the costant $\mathcal{E}$ and C are defined as:

$$
\begin{aligned}
\overline{\mathcal{E}} &= 6M^2\mathcal{E} + 6G^2 + 6d \\
C &= \frac{12\eta^2 M^2\left(M^2+1\right)}{m_2} + 3(M^2+1).
\end{aligned}
$$

Thus the divergence can be bounded as:

$$
\begin{aligned}
D_{KL}(p_K \| \hat{p}_{K\eta}) &\leq \frac{M^2}{2}\left(\overline{\mathcal{E}} + C\eta\frac{\Delta^2 d}{4} + 3\eta\sigma^2\right)K\eta^2 + \frac{1}{2}\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)K\eta \\
&= \frac{M^2}{2}\overline{\mathcal{E}}K\eta^2 + \left(\frac{M^2}{2}C\eta^2 + \frac{1}{2}(M^2+1)\right)\frac{\Delta^2 d}{4}K\eta + \frac{3M^2\eta^2+1}{2}\sigma^2 K\eta \\
&= \frac{M^2}{2}\overline{\mathcal{E}}K\eta^2 + \left(\frac{M^2}{2}C + \frac{1}{2}(M^2+1)\right)\frac{\Delta^2 d}{4}K\eta + \frac{3M^2+1}{2}\sigma^2 K\eta \\
&=: C_0 K\eta^2 + C_1\frac{\Delta^2 d}{4}K\eta + C_2\sigma^2 K\eta.
\end{aligned}
$$

We are ready to bound the Wasserstein distance,

$$
\begin{aligned}
\mathcal{W}_2^2(p_K, \hat{p}_{K\eta}) &\leq (12 + 8(\kappa_0 + 2b + 2d))\left((C_0 + \sqrt{C_0})\sqrt{\eta} + \left(C_1 + \sqrt{C_1}\right)A + \left(C_2 + \sqrt{C_2}\right)B\right)(K\eta)^2 \\
&=: \left(\widetilde{C_0}^2\sqrt{\eta} + \widetilde{C_1}^2 A + \widetilde{C_2}^2 B\right)(K\eta)^2,
\end{aligned}
$$

where the constants are defined as:

$$
\begin{aligned}
A &= \max\left\{\frac{\Delta^2 d}{4}, \sqrt{\frac{\Delta^2 d}{4}}\right\} \\
B &= \max\left\{\sigma^2, \sqrt{\sigma^2}\right\} \\
\widetilde{C_0}^2 &= (12 + 8(\kappa_0 + 2b + 2d))\left(C_0 + \sqrt{C_0}\right) \\
\widetilde{C_1}^2 &= (12 + 8(\kappa_0 + 2b + 2d))\left(C_1 + \sqrt{C_1}\right) \\
\widetilde{C_2}^2 &= (12 + 8(\kappa_0 + 2b + 2d))\left(C_2 + \sqrt{C_2}\right).
\end{aligned}
$$

From Proposition 9 in the paper Raginsky et al. [2017], we know that

$$
\begin{aligned}
\mathcal{W}_2(\hat{p}_{K\eta}, p^*) &\leq \sqrt{2C_{LS}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{M\kappa_0}{3} + B\sqrt{\kappa_0} + G_0 + \frac{b}{2}\log 3\right)\right)}e^{-K\eta/\beta C_{LS}} \\
&=: \widetilde{C_3}e^{-K\eta/\beta C_{LS}}
\end{aligned}
$$

Finally, we can have

$$
\mathcal{W}_2(p_K, p^*) \leq \left(\widetilde{C_0}\eta^{1/4} + \widetilde{C_1}\sqrt{A} + \widetilde{C_2}\sqrt{B}\right)K\eta + \widetilde{C_3}e^{-K\eta/\beta C_{LS}}. \tag{43}
$$

In order to bound the Wasserstein distance, we need to set

$$
\widetilde{C_0}K\eta^{5/4} = \frac{\epsilon}{2} \quad \text{and} \quad \widetilde{C_3}e^{-K\eta/\beta C_{LS}} = \frac{\epsilon}{2}. \tag{44}
$$

Solving the (37), we can have

$$
K\eta = C_{LS}\log\left(\frac{2\widetilde{C_3}}{\epsilon}\right) \quad \text{and} \quad \eta = \frac{\epsilon^4}{16\widetilde{C_0}^4 (K\eta)^4}.
$$

Combining these two we can have

$$
\eta = \frac{\epsilon^4}{16\widetilde{C_0}^4 C_{LS}^4 \log^4\left(\frac{2\widetilde{C_3}}{\epsilon}\right)} \quad \text{and} \quad K = \frac{16\widetilde{C_0}^4 C_{LS}^5 \log^5\left(\frac{2\widetilde{C_3}}{\epsilon}\right)}{\epsilon^4}.
$$

Plugging K and η into (43) completes the proof.

D.4 Proof of Theorem 6

Recall the SGHMCLP-L the update rule:

$$
\begin{aligned}
\mathbf{v}_{k+1} &= Q_W\left(\mathrm{v}\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{v}}\right)\\
\mathbf{x}_{k+1} &= Q_W\left(\mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta}-1)Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{x}}\right).
\end{aligned}
$$

If we let $\alpha_k^{\mathbf{x}}$ and $\alpha_k^{\mathbf{v}}$ denote the quantization error,

$$
\begin{aligned}
\alpha_k^{\mathbf{x}} &= Q_W\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_s)) + \xi_k^{\mathbf{v}}\right) - \left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_s)) + \xi_k^{\mathbf{v}}\right)\\
\alpha_k^{\mathbf{v}} &= Q_W\left(\mathbf{x}_s + \gamma^{-1}(1-e^{-\gamma\eta})v_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta}-1)Q_G(\nabla\tilde{U}(\mathbf{x}_s)) + \xi_k^{\mathbf{x}}\right)\\
&\quad - \left(\mathbf{x}_s + \gamma^{-1}(1-e^{-\gamma\eta})v_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta}-1)Q_G(\nabla\tilde{U}(\mathbf{x}_s)) + \xi_k^{\mathbf{x}}\right),
\end{aligned}
$$

we can rewrite the update rule as:

$$
\begin{aligned}
\mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_s)) + \xi_k^{\mathbf{v}} + \alpha_k^{\mathbf{v}}\\
\mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta}-1)Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{x}} + \alpha_k^{\mathbf{x}}.
\end{aligned} \tag{45}
$$

Similarly, we can define a continuous interpolation of (45) for $t \in (0, \eta]$.

$$
\begin{aligned}
\mathbf{v}_t &= \mathbf{v}_0 e^{-\gamma t} - u\left(\int_0^t e^{-\gamma(t-s)}\left(\nabla U(\mathbf{x}_0) + \zeta\right)ds\right) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s + \int_0^t \alpha_v(s)ds\\
\mathbf{x}_t &= \mathbf{x}_0 + \int_0^t \mathbf{v}_s ds + \int_0^t \alpha_x(s)ds,
\end{aligned} \tag{46}
$$

where the $\zeta = Q_G\left(\nabla\tilde{U}(\hat{x}_0)\right) - \nabla\tilde{U}(\hat{x}_0)$ the function $\alpha_v(s)$, $\alpha_x(s)$ are defined as:

$$
\begin{aligned}
\alpha_v(s) &= \sum_{k=0}^{\infty} \alpha_k^{\mathbf{v}}/\eta \mathbf{1}_{\mathrm{s}\in(\mathrm{k}\eta,(\mathrm{k}+1)\eta)}\\
\alpha_x(s) &= \sum_{k=0}^{\infty} \alpha_k^{\mathbf{x}}/\eta \mathbf{1}_{\mathrm{s}\in(\mathrm{k}\eta,(\mathrm{k}+1)\eta)}.
\end{aligned}
$$

If we let $\hat{p}_0 = (\hat{x}_0, \hat{v}_0)$ be the initial sample and $\hat{p}_t = (\hat{x}_t, \hat{v}_t)$ be the sample that satisfies the previous equations, we can define an operator $\hat{\Phi}_t$ that maps $\hat{p}_0$ to $\hat{p}_t$ i.e., $\hat{p}_t = \hat{\Phi}_t\hat{p}_0$. Notice that since $\hat{p}_t$ is the continuous interpolation of (4), thus $\hat{p}_{k\eta} = p_k = (\mathbf{x}_k, v_k)$. Similarly, we define $q_k = (\mathbf{x}_k, v_k + \mathbf{x}_k) =: (\mathbf{x}_k, \omega_k)$ as a tool to analyze the convergence of p_k.

We are now ready to compute the Wasserstein distance between $\hat{\Phi}_\eta q_0$ and q^*. Let Γ_1 be all of the couplings between $\widetilde{\Phi}_\eta q_0$ and q^*, and Γ_2 be all of the couplings between $\widehat{\Phi}_\eta q_0$ and q^*. Let r_1 be the optimal coupling between $\widetilde{\Phi}_\eta q_0$ and q^*. By taking the difference between (46) and (27),

$$
\begin{bmatrix} x \\ \omega \end{bmatrix} = \begin{bmatrix} \widetilde{x} \\ \widetilde{\omega} \end{bmatrix} + u \begin{bmatrix} \left(\int_0^\eta \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\zeta + \int_0^\eta \alpha_x(s)ds \\ \left(\int_0^\eta \left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\eta e^{-\gamma(s-\eta)}ds\right)\zeta + \int_0^\eta \alpha_x(s) + \alpha_v(s)ds \end{bmatrix}.
$$

Let us now analyze the Wasserstein distance between $\hat{\Phi}_\eta q_0$ and q^*,

$$
\begin{aligned}
&\mathcal{W}_2^2\left(\hat{\Phi}_\eta q_0, q^*\right)\\
&\leq \mathbb{E}_{r_1}\left\|\begin{bmatrix}\widetilde{x}\\ \widetilde{\omega}\end{bmatrix} + u\begin{bmatrix}\left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\zeta + \int_0^\eta \alpha_x(s)ds\\ \left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\eta e^{-\gamma(s-\eta)}ds\right)\zeta + \int_0^\eta\left(\alpha_x(s)+\alpha_v(s)\right)ds\end{bmatrix} - \begin{bmatrix}x^*\\ \omega^*\end{bmatrix}\right\|^2\\
&\leq \mathbb{E}_{r_1}\left\|\begin{bmatrix}\widetilde{x}\\ \widetilde{\omega}\end{bmatrix} - \begin{bmatrix}x^*\\ \omega^*\end{bmatrix}\right\|^2 + u^2\mathbb{E}\left\|\begin{bmatrix}\left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\zeta + \int_0^\eta \alpha_x(s)ds\\ \left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\eta e^{-\gamma(s-\eta)}ds\right)\zeta + \int_0^\eta\left(\alpha_x(s)+\alpha_v(s)\right)ds\end{bmatrix}\right\|^2\\
&\leq \mathcal{W}_2^2\left(\widetilde{\Phi}_\eta q_0, q^*\right) + 4u^2\left(\left(\int_0^\delta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)^2 + \left(\int_0^\delta e^{-\gamma(s-\delta)}ds\right)^2\right)\left(\frac{\Delta^2 d}{4}+\sigma^2\right)\\
&+ u^2\mathbb{E}\left[\left\|\int_0^\eta\left(\alpha_x(s)\right)ds\right\|^2\right] + u^2\mathbb{E}\left[\left\|\int_0^\eta\left(\alpha_x(s)+\alpha_v(s)\right)ds\right\|^2\right]\\
&\leq \mathcal{W}_2^2\left(\widetilde{\Phi}_\eta q_0, q^*\right) + 4u^2\left(\frac{\eta^4}{4}+\eta^2\right)\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + u^2\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] + u^2\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}+\alpha_k^{\mathbf{v}}\|^2\right]\\
&\leq \mathcal{W}_2^2\left(\widetilde{\Phi}_\eta q_0, q^*\right) + 5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 2u^2\left(\mathbb{E}\left\|\alpha_k^{\mathbf{x}}\right\|^2 + \mathbb{E}\left\|\alpha_k^{\mathbf{v}}\right\|^2\right)\\
&\leq \mathcal{W}_2^2\left(\widetilde{\Phi}_\eta q_0, q^*\right) + 5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 2u^2\left(A+B\right),
\end{aligned}
$$

where the constant A, B are the uniform bounds of $\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|\right]$ and $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|\right]$ respectively. Furthermore from the corollary 7 in Cheng et al. [2018] we know that for any $i \in \{1, \cdots, K\}$:

$$
\mathcal{W}_2^2(\Phi_\eta q_i, q^*) \leq e^{-\eta/2\kappa_1}\mathcal{W}_2^2(q_i, q^*), \tag{47}
$$

where $\kappa_1 = M/m_1$ is the condtion number. From the discretization error bound from theorem 9 and lemma 8(sandwich inequality) in Cheng et al. [2018], we get

$$
\mathcal{W}_2(\Phi_\eta q_i, \widetilde{\Phi}_\eta q_i) \leq 2\mathcal{W}_2(\Phi_\eta p_i, \widetilde{\Phi}_\eta p_i) \leq \eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}.
$$

By triangle inequality:

$$
\begin{aligned}
\mathcal{W}_2(\widetilde{\Phi}_\eta q_i, q^*) &\leq \mathcal{W}_2(\Phi_\eta q_i, \widetilde{\Phi}_\eta q_i) + \mathcal{W}_2(\Phi_\eta q_i, q^*)\\
&\leq \eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + e^{-\eta/2\kappa_1}\mathcal{W}_2(q_i, q^*),
\end{aligned}
$$

further implies the following inequality:

$$
\mathcal{W}_2^2\left(\hat{\Phi}_\eta q_i, q^*\right) \leq \left(e^{-\eta/2\kappa_1}\mathcal{W}_2\left(q_i, q^*\right) + \eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}\right)^2 + 5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 2u^2\left(A+B\right).
$$

By invoking the Lemma 7 in Dalalyan and Karagulyan [2019] we can bound the Wasserstein distance by:

$$
\begin{aligned}
\mathcal{W}_2(q_K, q^*) \leq\; & e^{-K\eta/2\kappa_1}\mathcal{W}_2(q_0, q^*) + \frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{1-e^{-\eta/2\kappa_1}}\\
&+ \frac{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 2u^2\left(A+B\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{1-e^{-\eta/2\kappa_1}}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 2u^2\left(A+B\right)}}.
\end{aligned}
$$

Finally by sandwich inequality we have:

$$
\begin{aligned}
\mathcal{W}_2(p_K, p^*) \leq\; & 4e^{-K\eta/2\kappa_1}\mathcal{W}_2(q_0, q^*) + \frac{4\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{1-e^{-\eta/2\kappa_1}}\\
&+ \frac{20u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 8u^2\left(A+B\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{1-e^{-\eta/2\kappa_1}}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right) + 2u^2\left(A+B\right)}}.
\end{aligned} \tag{48}
$$

And in this case, we know that $\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|\right]$ and $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|\right]$ can be bouned by $\frac{\Delta^2 d}{4}$. Finally, we can have:

$$\begin{aligned}\mathcal{W}_2(p_K, p^*) \leq 4e^{-K\eta/2\kappa_1}\mathcal{W}_2(q_0, q^*) + \frac{4\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{1-e^{-\eta/2\kappa_1}} \\ + \frac{20u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+4u^2\Delta^2 d}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}+\sqrt{1-e^{-\eta/2\kappa_1}}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+u^2\Delta^2 d}}.\end{aligned}$$

Now we let the first term less than $\epsilon/3$, from the lemma 13 in [Cheng et al., 2018] we know that $\mathcal{W}_2(q_0, q^*) \leq 3\left(\frac{d}{m_1}+\mathcal{D}^2\right)$. So we can choose K as the following,

$$K \leq \frac{2\kappa_1}{\eta}\log\left(36\left(\frac{d}{m_1}+\mathcal{D}^2\right)\right).$$

Next, we choose a stepsize $\eta \leq \frac{\epsilon\kappa_1^{-1}}{\sqrt{479232/5(d/m_1+\mathcal{D}^2)}}$ to ensure the second term is controlled below $\epsilon/3$. Since $1-e^{-\eta/2\kappa_1} \geq \eta/4\kappa_1$ and definition of $\mathcal{E}_K$,

$$4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{1-e^{-\eta/2\kappa}} \leq 4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{\eta/4\kappa_1} \leq 16\kappa_1\left(\eta\sqrt{\frac{8\mathcal{E}_K}{5}}\right) \leq \epsilon/3.$$

Finally by choosing the stepsize satisfied that,

$$\eta \leq \frac{\epsilon^2}{2880\kappa_1 u\left(\frac{\Delta^2 d}{4}+\sigma^2\right)},$$

the third term can be bounded as:

$$\begin{aligned}&\frac{20u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)+4u^2\Delta^2 d}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}+\sqrt{1-e^{-\eta/2\kappa_1}}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)}} \\ &\leq \frac{20u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)+4u^2\Delta^2 d}{\sqrt{1-e^{-\eta/2\kappa_1}}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)}} \leq \frac{20u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)+4u^2\Delta^2 d}{\sqrt{\eta/4\kappa_1}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)}} \\ &\leq 4\sqrt{20\kappa_1 u^2\eta\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)} + \frac{8u^2\Delta^2 d\sqrt{\kappa_1}}{\eta^{3/2}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)}} \\ &\leq \epsilon/3 + \frac{8u^2\Delta^2 d\sqrt{\kappa_1}}{\eta^{3/2}\sqrt{5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4}+\sigma^2\right)}}.\end{aligned}$$

This complete the proof.

D.5 Proof of Theorem 7

In this section, we analyze the convergence of SGHMCLP-L when the target distribution is non-log-concave. Recall the continuous interpolation of the SGHMCLP-L,

$$\begin{aligned}\mathbf{v}_t &= \mathbf{v}_0 - \int_0^t \gamma\mathbf{v}_s ds - u\int_0^t G_s ds + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s + \int_0^t \alpha_v(s)ds \\ \mathbf{x}_t &= \mathbf{x}_0 + \int_0^t \mathbf{v}_s ds + \int_0^t \alpha_x(s)ds,\end{aligned}$$

where $G_s = \sum_{k=0}^{\infty} Q_G\left(\nabla U(x_k')\right)\mathbf{1}_{\mathrm{s}\in(\mathrm{k}\eta,(\mathrm{k}+1)\eta)}$.And we define an intermediate process by let $\mathbf{v}_t' = \mathbf{v}_t + \alpha_x(t)$:

$$
\begin{aligned}
v_t' &= v_0' - \int_0^t \gamma\left(v_s' - \alpha_x(s)\right)ds - u\int_0^t G_s ds + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s + \int_0^t \left(\alpha_v(s) + \frac{1}{t}\alpha_x(t)\right)ds \\
x_t' &= x_0' + \int_0^t v_s' ds. \qquad (49)
\end{aligned}
$$

By integrating the underdamped Langevin dynamic (9), we can have:

$$
\begin{aligned}
\mathbf{v}_t &= \mathbf{v}_0 - \int_0^t \gamma\left(\mathbf{v}_s - \alpha_x(s)\right)ds - u\int_0^t \nabla U(\mathbf{x}_s)ds + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)}dB_s \\
\mathbf{x}_t &= \mathbf{x}_0 + \int_0^t \mathbf{v}_s ds. \qquad (50)
\end{aligned}
$$

Notice that the process x_t' has the same distribution with $\mathbf{x}_t$, thus in the following analysis we study the convergence of the intermediate process $p_k' = (x_{k\eta}', v_{k\eta}')$. By taking the difference of equation (49) with (50) and the Girsanov formula, we can derive the Radon-Nikodym derivative of $\hat{P}_{K\eta}$ w.r.t p_K':

$$
\begin{aligned}
\frac{d\hat{p}_{K\eta}}{dp_K'} = exp\Bigg\{&\sqrt{\frac{u}{2\gamma}}\int_0^T (\gamma\alpha_x(s) + \alpha_v(s) + \frac{1}{T}\alpha_x(T) + \nabla U(\mathbf{x}_s) - G_s)d\mathbf{B}s \\
&-\frac{u}{4\gamma}\int_0^T \|\gamma\alpha_x(s) + \alpha_v(s) + \frac{1}{T}\alpha_x(T) + \nabla U(\mathbf{x}_s) - G_s\|^2 ds\Bigg\}.
\end{aligned}
$$

Thus the divergence can be bouned as:

$$
\begin{aligned}
&D_{KL}(p_K||\hat{p}_{K\eta}) = \mathbb{E}_{p_K}\left[\log\left(\frac{d\hat{p}_{K\eta}}{dp_K}\right)\right] \\
&= \frac{u}{4\gamma}\int_0^T \mathbb{E}\left\|\gamma\alpha_x(s) + \alpha_v(s) + \frac{1}{T}\alpha_x(T) + \nabla U(\mathbf{x}_s) - G_s\right\|^2 ds \\
&= \frac{u}{4\gamma T}\mathbb{E}\left[\|\alpha_x(T)\|^2\right] + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\gamma\alpha_v(s) + \alpha_x(s) + \nabla U(\mathbf{x}_s) - G_s\|^2\right]ds \\
&\leq \frac{u}{4\gamma T\eta^2}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\gamma\alpha_v(s)\|^2\right]ds + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\alpha_x(s)\|^2\right]ds \\
&+ \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - G_s\|^2\right]ds \\
&\leq \frac{u}{4\gamma T\eta^2}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\gamma\alpha_k^{\mathbf{v}}/\eta\|^2\right]ds + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}/\eta\|^2\right]ds \\
&+ \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - Q_G(\nabla U(\mathbf{x}_k))\|^2\right]ds \\
&\leq \frac{u}{4\gamma T\eta^2}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\gamma\alpha_k^{\mathbf{v}}/\eta\|^2\right]ds + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}/\eta\|^2\right]ds \qquad (51) \\
&+ \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k)\|^2\right]ds + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - Q_G(\nabla U(\mathbf{x}_k))\|^2\right]ds.
\end{aligned}
$$

By assumption 1, we know that:

$$\mathbb{E}\left[\|\nabla U(\mathbf{x}_s)-\nabla U(\mathbf{x}_k)\|^2\right]\le M^2\mathbb{E}\left[\|\mathbf{x}_s-\mathbf{x}_k\|^2\right].$$

From the same analysis in (35), we can derive:

$$\mathbb{E}\left[\|\nabla U(\mathbf{x}_s)-\nabla U(\mathbf{x}_k)\|^2\right]\le 3M^2\eta^2\left(\mathbb{E}\left[\|\mathbf{v}_k'\|^2\right]+u^2\eta^2\mathbb{E}\left[\|Q_G(\nabla U(\mathbf{x}_k))\|^2\right]+2du\right).$$

Now we need to derive a uniform bound of $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$ and $\mathbb{E}\left[\|\mathbf{v}_k'\|^2\right]$.

Lemma 17. *Let Assumptions 3 and 1 hold. If we set the step size to the following condition*

$$\eta\le\min\left\{\frac{\gamma}{4\left(8Mu+u\gamma+22\gamma^2\right)},\sqrt{\frac{4u^2}{4Mu+3\gamma^2}},\frac{6\gamma bu}{\left(4Mu+3\gamma^2\right)d},\frac{\gamma m_2}{6\left(22u+\gamma\right)M^2}\right\},$$

then for all $k>0$ $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$ and $\mathbb{E}\left[\|v_k\|^2\right]$ can be bouned as follow:

$$\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]\le\mathcal{E}+C\Delta^2 d,\quad \mathbb{E}\left[\|v_k'\|^2\right]\le\gamma^2\mathcal{E}/2+\gamma^2C\Delta^2d/2,$$

where constants $\mathcal{E}$ and C are defined as:

$$\begin{aligned}\mathcal{E}&=\mathbb{E}\left[\mathcal{E}(\mathbf{x}_0,\mathbf{v}_0)\right]+\frac{54\left(4u+\gamma^2\right)u}{m_2\gamma^4}\sigma^2+\frac{12(22u+\gamma)uM^3}{m_2\gamma^3}G^2+\frac{96\left(d+b\right)uM}{m_2\gamma^2}\\ C&=\frac{27\left(4u+\gamma^2\right)u}{2m_2\gamma^4}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\|\nabla U(\mathbf{x}_s)-\nabla U(\mathbf{x}_k)\|^2\right]&\le 3M^2\eta^2\left(\mathbb{E}\left[\|v_k\|^2\right]+u^2\eta^2\left(\frac{\Delta^2d}{4}+\sigma^2+2M^2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]+2G^2\right)+2du\right)\\ &\le 3M^2\eta^2\left(\gamma^2\mathcal{E}/2+\gamma^2C\Delta^2d/2+u^2\eta^2\left(\frac{\Delta^2d}{4}+\sigma^2+2M^2\mathcal{E}+2M^2C\Delta^2d+2G^2\right)+2du\right)\\ &\le 3M^2\eta^2\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E}+\left(\gamma^2+2u^2M^2\right)C\Delta^2d+u^2\sigma^2+2u^2G^2+2du\right).\end{aligned}$$

Now we can go back to the divergence of p_K and $\hat{p}_{K\eta}$,

$$\begin{aligned}&D_{KL}(p_K||\hat{p}_{K\eta})\\ &\le\frac{u}{4\gamma T\eta^2}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right]+\frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\gamma\alpha_k^{\mathbf{v}}/\eta\|^2\right]ds+\frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}/\eta\|^2\right]ds\\ &+\frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E}+\left(\gamma^2+2u^2M^2\right)C\Delta^2d+u^2\sigma^2+2u^2G^2+2du\right)+\frac{u}{4\gamma}K\eta\left(\frac{\Delta^2d}{4}+\sigma^2\right)\\ &\le\frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E}+\left(\gamma^2+2u^2M^2\right)C\Delta^2d+u^2\sigma^2+2u^2G^2+2du\right)+\frac{u}{4\gamma}K\eta\left(\frac{\Delta^2d}{4}+\sigma^2\right)\\ &+\frac{u\Delta^2d}{16\gamma T\eta^2}+\frac{uK\Delta^2d}{8\gamma\eta}\\ &\le\frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E}+u^2\sigma^2+2u^2G^2+2du\right)+\frac{u}{4\gamma}K\eta\sigma^2\\ &+\left(\frac{u}{4\gamma}3M^2K\eta^3C\left(\gamma^2+2u^2M^2\right)+\frac{uK\eta}{16\gamma}+\frac{u}{16\gamma T\eta^2}+\frac{uK}{8\gamma\eta}\right)\Delta^2d\\ &=: C_0K\eta^3+C_1K\eta\sigma^2+C_2K\Delta^2,\end{aligned}$$

where the constants C_0, C_1 and C_2 are defined as:

$$\begin{aligned}C_0&=\frac{u}{4\gamma}3M^2\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E}+u^2\sigma^2+2u^2G^2+2du\right)\\ C_1&=\frac{u}{4\gamma}\\ C_2&=\left(\frac{u}{4\gamma}3M^2\eta^3C\left(\gamma^2+2u^2M^2\right)+\frac{u}{16\gamma}+\frac{u}{16\gamma T^2\eta}+\frac{u}{8\gamma\eta}\right)d.\end{aligned}$$

By the weighted CKP inequality and given $K\eta \geq 1$,

$$\begin{aligned}\mathcal{W}_2(p_K, \hat{p}_{K\eta}) &\leq \overline{\Lambda}\left(\sqrt{D_{KL}(p_K||\hat{p}_{K\eta})} + \sqrt[4]{D_{KL}(p_K||\hat{p}_{K\eta})}\right) \\ &\leq \left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\widetilde{A}\right)\sqrt{K\eta} + \widetilde{C_2}\sqrt{K\Delta},\end{aligned} \tag{52}$$

where the constants are defined as:

$$\begin{aligned}\widetilde{C_0} &= \left(\sqrt{C_0} + \sqrt[4]{C_0}\right) \\ \widetilde{C_1} &= \left(\sqrt{C_1} + \sqrt[4]{C_1}\right) \\ \widetilde{C_2} &= \left(\sqrt{C_2} + \sqrt[4]{C_2}\right) \\ \widetilde{A} &= \max\left\{\sigma, \sqrt{\sigma}\right\}.\end{aligned}$$

From the same analysis in (36), we can have:

$$\mathcal{W}_2(p_K, p^*) \leq \overline{\Lambda}\left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\widetilde{A}\right)\sqrt{K\eta} + \widetilde{C_2}\sqrt{K\eta} + \Gamma_0 e^{-\mu^* K\eta}. \tag{53}$$

In order to bound the Wasserstein distance, we need to set

$$\overline{\Lambda}\widetilde{C_0}\sqrt{K\eta^2} = \frac{\epsilon}{2} \quad \text{and} \quad \Gamma_0 e^{-\mu^* K\eta} = \frac{\epsilon}{2}. \tag{54}$$

Solving the equation (54), we can have

$$K\eta = \frac{\log\left(\frac{2\Gamma_0}{\epsilon}\right)}{\mu^*} \quad \text{and} \quad \eta = \frac{\epsilon^2}{4\overline{\Lambda}^2\widetilde{C_0}^2 K\eta}.$$

Combining these two we can have

$$\eta = \frac{\epsilon^2\mu^*}{4\overline{\Lambda}^2\widetilde{C_0}^2\log\left(\frac{2\Gamma_0}{\epsilon}\right)} \quad \text{and} \quad K = \frac{4\overline{\Lambda}^2\widetilde{C_0}^2\log^2\left(\frac{2\Gamma_0}{\epsilon}\right)}{\epsilon^2\left(\mu^*\right)^2}.$$

Plugging in (53) completes the proof.

D.6 Proof o Theorem 11

In this section we generalize the convergence analysis of SGLDLP-L in Zhang et al. [2022] to non-log-concave target distribution. Following the same proof outlines in Raginsky et al. [2017]. Recall the LPSGLDLP-L update rule 19 is the following,

$$\begin{aligned}\mathbf{x}_{k+1} &= Q_W(\mathbf{x}_k - \eta\nabla\tilde{U}(\mathbf{x}_k) + \sqrt{2\eta}\xi_{k+1}) \\ &=: \mathbf{x}_k - \eta\nabla\tilde{U}(\mathbf{x}_k) + \sqrt{2\eta}\xi_{k+1} + \alpha_k,\end{aligned}$$

where α_k is define as:

$$\alpha_k = Q_W(\mathbf{x}_k - \eta\nabla\tilde{U}(\mathbf{x}_k) + \sqrt{2\eta}\xi_{k+1}) - \mathbf{x}_k - \eta\nabla\tilde{U}(\mathbf{x}_k) + \sqrt{2\eta}\xi_{k+1}.$$

Thus, we can define a continuous interpolation of the SGLDLP-L as:

$$\mathbf{x}_t = \mathbf{x}_0 - \int_0^t G_s ds + \sqrt{2}\int_0^t dB(s) + \int_0^t \alpha(s)ds,$$

where $G_s = \sum_{k=0}^{\infty} Q_G(\nabla\tilde{U}(\mathbf{x}_k))\mathbf{1}_{\mathrm{s}\in(\mathrm{k}\eta,(\mathrm{k}+1)\eta)}$ and $\alpha(s) = \sum_{k=0}^{\infty} \alpha_k/\eta\mathbf{1}_{\mathrm{s}\in(\mathrm{k}\eta,(\mathrm{k}+1)\eta)}$. By taking the difference of the interpolation with the discrete estimation of Langevin process in equation 39, we can derive the Radon-Nikodym derivative of the $\hat{p}_{K\eta}$ w.r.t p_K as:

$$\frac{d\hat{p}_{K\eta}}{dp_K} = exp\left\{\frac{1}{2}\int_0^t(\nabla U(\mathbf{x}_s) - G_s - \alpha(s))d\mathbf{B}s - \frac{1}{4}\int_0^T\|\nabla U(\mathbf{x}_s) - G_s - \alpha(s)\|^2 ds\right\}.$$

Thus, the divergence can be computed as:

$$
\begin{aligned}
D_{KL}(p_K||\hat{p}_{K\eta}) =& \frac{1}{4}\int_0^{K\eta} \mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - G_s - \alpha(s)\|^2\right] ds \\
=& \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\left\|\nabla U(\mathbf{x}_s) - Q_G(\nabla\tilde{U}(\mathbf{x}_k)) - \alpha_k/\eta\right\|^2\right] ds \\
=& \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\left\|\nabla U(\mathbf{x}_s) - Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] ds + \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\alpha_k/\eta\|^2\right] ds \\
=& \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\nabla U(\mathbf{x}_s) - \nabla U(\mathbf{x}_k)\|^2\right] ds + \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\left\|\nabla U(\mathbf{x}_k) - Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] ds \\
&+ \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\alpha_k/\eta\|^2\right] ds \\
\leq& \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_k\|^2\right] ds + \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\left\|\nabla U(\mathbf{x}_k) - Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] ds \\
&+ \frac{1}{4}\sum_{k=0}^{K-1}\int_{k\eta}^{(k+1)\eta} \mathbb{E}\left[\|\alpha_k/\eta\|^2\right] ds. \qquad (55)
\end{aligned}
$$

From the same analysis in (35), we know that

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_k\|^2\right] &\leq 3\eta^2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right] + 3\eta^2\mathbb{E}\left[\left\|\nabla U(\mathbf{x}_k) - Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] + 6\eta d \\
&\leq 3\eta^2\left(M\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + G\right)^2 + 3\eta^2\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + 6\eta d.
\end{aligned}
$$

Again, we need to derive a uniform bound of $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$,

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_{k+1}\|^2\right] =& \mathbb{E}\left[\left\|\mathbf{x}_k - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] + 2\mathbb{E}\left[\|\xi_{k+1}\|^2\right] + \mathbb{E}\left[\|\alpha_k\|^2\right] \\
=& \mathbb{E}\left[\left\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k) + \eta\nabla U(\mathbf{x}_k) - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] + 2\eta d + \mathbb{E}\left[\|\alpha_k\|^2\right] \\
=& \mathbb{E}\left[\left\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k) + \eta\nabla U(\mathbf{x}_k) - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] + \mathbb{E}\left[\|\alpha_k\|^2\right] + 2\eta d \\
=& \mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k)\|^2\right] + \eta^2\mathbb{E}\left[\left\|\nabla U(\mathbf{x}_k) - Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] + \mathbb{E}\left[\|\alpha_k\|^2\right] + 2\eta d.
\end{aligned}
$$

By plugging in the inequality we derived before:

$$
\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k)\|^2\right] \leq \left(1 - 2\eta m_2 + 2\eta^2 M^2\right)\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2\eta b + 2\eta^2 G^2.
$$

we can have:

$$
\mathbb{E}\left[\|\mathbf{x}_{k+1}\|^2\right] \leq \left(1 - 2\eta m_2 + 2\eta^2 M^2\right)\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2\eta b + 2\eta^2 G^2 + \frac{\eta^2\Delta^2 d}{4} + \eta^2\sigma^2 + \mathbb{E}\left[\|\alpha_k\|^2\right] + 2\eta d.
$$

Thus for any $\eta \in (0, 1 \wedge \frac{m_2}{2M^2})$ and $1 - 2\eta m_2 + 2\eta^2 M^2 > 0$, we can bound $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$ for any $k > 0$ as:

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \leq & \mathbb{E}\left[\|\mathbf{x}_0\|^2\right] + \frac{1}{2\left(m_2 - \eta M^2\right)}\left(2b + 2G^2 + \frac{\Delta^2 d}{4} + \sigma^2 + 2d\right) + \frac{\mathbb{E}\left[\|\alpha_k\|^2\right]}{2\eta\left(m_2 - \eta M^2\right)} \\
\leq & \mathbb{E}\left[\|\mathbf{x}_0\|^2\right] + \frac{1}{m_2}\left(2b + 2G^2 + \frac{\Delta^2 d}{4} + \sigma^2 + 2d\right) + \frac{\mathbb{E}\left[\|\alpha_k\|^2\right]}{\eta m_2} \\
\leq & \mathcal{E} + \frac{\Delta^2 d}{4m_2} + \frac{\mathbb{E}\left[\|\alpha_k\|^2\right]}{\eta m_2},
\end{aligned}
$$

where the constant $\mathcal{E}$ is defined as:

$$
\mathcal{E} = \mathbb{E}\left[\|\mathbf{x}_0\|^2\right] + \frac{1}{m_2}\left(2b + 2G^2 + \sigma^2 + 2d\right).
$$

Thus, we can have,

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_s - \mathbf{x}_k\|^2\right] \leq & 6\eta^2\left(\mathcal{E} + \frac{\Delta^2 d}{4m_2} + \frac{\mathbb{E}\left[\|\alpha_k\|^2\right]}{\eta m_2}\right) + 6\eta^2 G^2 + 3\eta^2\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + 6\eta d \\
\leq & \overline{\mathcal{E}}\eta + 3\eta^2\sigma^2 + \frac{6 + 3m_2}{4m_2}\eta^2\Delta^2 d + \frac{6\eta\mathbb{E}\left[\|\alpha_k\|^2\right]}{m_2}.
\end{aligned}
$$

Plugging this into the equation (55), we can have,

$$
\begin{aligned}
D_{KL}(p_K||\hat{p}_{K\eta}) \leq & \frac{M\overline{\mathcal{E}}}{4}K\eta^2 + \frac{3M\sigma^2 K\eta^3}{4} + \frac{\left(6 + 3m_2\right)M\Delta^2 d}{16m_2}K\eta^3 + \frac{6M\mathbb{E}\left[\|\alpha_k\|^2\right]K\eta^2}{4m_2} + \frac{1}{4}\left(\frac{\Delta^2 d}{4} + \sigma^2\right)K\eta + \frac{K\mathbb{E}\left[\right.}{} \\
\leq & \frac{M\overline{\mathcal{E}}}{4}K\eta^2 + \frac{3M + 1}{4}\sigma^2 K\eta + \frac{\left(\left(6 + 3m_2\right)M + m_2\right)d}{16m_2}\Delta^2 K\eta + \left(\frac{6M\eta}{4m_2} + \frac{1}{4\eta}\right)K\mathbb{E}\left[\|\alpha_k\|^2\right].
\end{aligned}
$$

By the fact that $\mathbb{E}\left[\|\alpha_k\|^2\right] \leq \frac{\Delta^2 d}{4}$, we can further bound the divergence as:

$$
\begin{aligned}
D_{KL}(p_K||\hat{p}_{K\eta}) \leq & \frac{M\overline{\mathcal{E}}}{4}K\eta^2 + \frac{3M + 1}{4}\sigma^2 K\eta + \left(\frac{\left(\left(12 + 3m_2\right)M + m_2\right)d}{16m_2} + \frac{d}{16\eta}\right)\Delta^2 K \\
=: & C_0 K\eta^2 + C_1\sigma^2 K\eta + C_2\Delta^2 K,
\end{aligned}
$$

where the constants are defined as:

$$
\begin{aligned}
C_0 &= \frac{M\overline{\mathcal{E}}}{4} \\
C_1 &= \frac{3M + 1}{4} \\
C_2 &= \left(\frac{\left(\left(12 + 3m_2\right)M + m_2\right)d}{16m_2} + \frac{d}{16\eta}\right).
\end{aligned}
$$

We are ready to bound the Wasserstein distance,

$$
\begin{aligned}
\mathcal{W}_2^2(p_K, \hat{p}_{K\eta}) \leq & \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left[\left(C_0 + \sqrt{C_0} + \left(C_1 + \sqrt{C_1}\right)A\right)(K\eta)^2 + \left(C_2 + \sqrt{C_2}\right)\Delta K^2\eta\right] \\
=: & \left(\widetilde{C_0}^2\sqrt{\eta} + \widetilde{C_1}^2 A\right)(K\eta)^2 + \widetilde{C_2}^2\Delta K^2\eta,
\end{aligned}
$$

where the constants are defined as:

$$
\begin{aligned}
A &= \max\left\{\sigma^2, \sqrt{\sigma^2}\right\} \\
\widetilde{C_0}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_0 + \sqrt{C_0}\right) \\
\widetilde{C_1}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_1 + \sqrt{C_1}\right) \\
\widetilde{C_2}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_2 + \sqrt{C_2}\right).
\end{aligned}
$$

From Proposition 9 in the paper Raginsky et al. [2017], we know that

$$\begin{aligned}\mathcal{W}_2(\hat{p}_{K\eta}, p^*) &\leq \sqrt{2C_{LS}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{M\kappa_0}{3} + B\sqrt{\kappa_0} + G_0 + \frac{b}{2}\log 3\right)\right)}e^{-K\eta/\beta C_{LS}} \\ &=: \widetilde{C_3}e^{-K\eta/\beta C_{LS}}\end{aligned}$$

Finally, we can have

$$\mathcal{W}_2(p_K, p^*) \leq \left(\widetilde{C_0}\eta^{1/4} + \widetilde{C_1}\sqrt{A}\right)K\eta + \widetilde{C_2}\sqrt{\Delta}\sqrt{K^2\eta} + \widetilde{C_3}e^{-K\eta/\beta C_{LS}}. \tag{56}$$

In order to bound the 2-Wasserstein distance, we need to set

$$\widetilde{C_0}K\eta^{5/4} \leq \frac{\epsilon}{2} \quad \text{and} \quad \widetilde{C_3}e^{-K\eta/\beta C_{LS}} = \frac{\epsilon}{2}. \tag{57}$$

Solving the (57), we can have

$$K\eta = C_{LS}\log\left(\frac{2\widetilde{C_3}}{\epsilon}\right) \quad \text{and} \quad \eta \leq \frac{\epsilon^4}{16\widetilde{C_0}^4(K\eta)^4}.$$

Combining these two we can have

$$\eta \leq \frac{\epsilon^4}{16\widetilde{C_0}^4 C_{LS}^4\log^4\left(\frac{2\widetilde{C_3}}{\epsilon}\right)} \quad \text{and} \quad K \geq \frac{16\widetilde{C_0}^4 C_{LS}^5\log^5\left(\frac{2\widetilde{C_3}}{\epsilon}\right)}{\epsilon^4}.$$

Plugging K and η into (56) completes the proof.

D.7 Proof of Theorem 8

In this section, we analyze the convergence of VC SGHMCLP-L, recall the VC SGHMCLP-L update rule is the following,

$$\begin{aligned}\mathbf{v}_{k+1} &= Q^{vc}\left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right)Q_G\left(\nabla\tilde{U}(\mathbf{x}_k)\right), Var_v, \Delta\right) \\ \mathbf{x}_{k+1} &= Q^{vc}\left(\mathbf{x}_k + \gamma^{-1}\left(1 - e^{-\gamma\eta}\right)v_k + u\gamma^{-2}\left(\gamma\eta + e^{-\gamma\eta} - 1\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k)), Var_x, \Delta\right).\end{aligned} \tag{58}$$

If we let $\alpha_k^{\mathbf{x}}$ and $\alpha_k^{\mathbf{v}}$ denote the quantization error,

$$\begin{aligned}\alpha_k^{\mathbf{v}} =& Q^{vc}\left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right)Q_G\left(\nabla\tilde{U}(\mathbf{x}_k)\right), Var_v, \Delta\right) - \left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{v}}\right) \\ \alpha_k^{\mathbf{x}} =& Q^{vc}\left(\mathbf{x}_k + \gamma^{-1}\left(1 - e^{-\gamma\eta}\right)v_k + u\gamma^{-2}\left(\gamma\eta + e^{-\gamma\eta} - 1\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k)), Var_x, \Delta\right) \\ &- \left(\mathbf{x}_k + \gamma^{-1}(1 - e^{-\gamma\eta})v_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{x}}\right),\end{aligned}$$

we can rewrite the update rule as:

$$\begin{aligned}\mathbf{v}_{k+1} &= \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{\gamma\eta})Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{v}} + \alpha_k^{\mathbf{v}} \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \gamma^{-1}(1 - e^{-\gamma\eta})v_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{x}} + \alpha_k^{\mathbf{x}}.\end{aligned}$$

Next, we first derive a uniform bound of $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right]$. In this section and the following section, we further assume the norm of quantized stochastic gradients are bounded.

Assumption 6. *For any $x \in \mathbb{R}^d$, there exists a constant $\mathcal{G}$ and the quantized stochastic gradients at x satisfies the following*

$$\mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(x))\right\|^2\right] \leq \mathcal{G}^2.$$

By the definition of the variance corrected quantization function Q^{vc}, when $Var_v > \rho_0 = \frac{\Delta^2}{4}$, if we let ψ_k denote $v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G\left(\nabla\tilde{U}(\mathbf{x}_k)\right)$,

$$
\begin{aligned}
&\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2 \middle| \psi_k\right] \\
=&\mathbb{E}\left[\left\|\left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) + \sqrt{Var_v}\xi_k \right.\right. \\
&\left.\left. - Q^d\left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \sqrt{Var_v - \rho_0}\xi_k\right) - \text{sign}(r)c\right\|^2 \middle| \psi_k\right]
\end{aligned}
$$

Let

$$
\begin{aligned}
b = Q^d&\left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \sqrt{Var_v - \rho_0}\xi_k\right) \\
&- \left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \sqrt{Var_v - \rho_0}\xi_k\right),
\end{aligned}
$$

then

$$
\begin{aligned}
&\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2 \middle| \psi_k\right] \\
=&\mathbb{E}\left[\left\|\left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) + \sqrt{Var_v}\xi_k \right.\right. \\
&\left.\left. - \left(v_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \sqrt{Var_v - \rho_0}\xi_k\right) - b - \text{sign}(r)c\right\|^2 \middle| \psi_k\right] \\
=&\mathbb{E}\left[\left\|\sqrt{Var_v}\xi_k - \sqrt{Var_v - \rho_0}\xi_k - b - \text{sign}(r)c\right\|^2 \middle| \psi_k\right] \\
\leq&\mathbb{E}\left[\left\|\sqrt{Var_v}\xi_k - \sqrt{Var_v - \rho_0}\xi_k\right\|^2\right] + \mathbb{E}\left[\|b + \text{sign}(r)c\|^2 \middle| \psi_k\right] \\
\leq&2Var_v d - \rho_0 d + \rho_0 d \\
\leq&4\gamma u d\eta.
\end{aligned} \tag{59}
$$

When $Var_v < \frac{\Delta_W^2}{4}$,

$$
\begin{aligned}
&\mathbb{E}[\|\alpha_k^{\mathbf{v}}\|^2] \\
&= \mathbb{E}\left[\left\|\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) - \mathbf{v}_{k+1} + \sqrt{Var_v}\xi_k\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) - \mathbf{v}_{k+1}\right\|^2\right] + \mathbb{E}\left[\left\|\sqrt{Var_v}\xi_k\right\|^2\right] \\
&\leq \max\left(2\mathbb{E}\left[\left\|\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) - Q^s\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right)\right\|^2\right], 2Var\right.
\end{aligned} \tag{60}
$$

Using the bound equation (6) in Li and De Sa [2019] gives us,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) - Q^s\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1 - e^{-\gamma\eta}\right) Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right)\right\|^2\right] \\
&\leq \Delta\left(1 - e^{-\gamma\eta}\right)\mathbb{E}\left[\left\|v_k - u\gamma^{-1} Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|_1\right] \\
&\leq \Delta\left(1 - e^{-\gamma\eta}\right)\sqrt{d}\left(\mathbb{E}\left[\|v_k\|\right] + \mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|\right]\right).
\end{aligned}
$$

Now we need to derive a uniform bound of $\mathbb{E}\left[\|v_k\|\right]$, by the update rule, we know that,

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_{k+1}\|^2\right] &= \mathbb{E}\left[\left\|\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1 - e^{\gamma\eta}) Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \xi_k^{\mathbf{v}} + \alpha_k^{\mathbf{v}}\right\|^2\right] \\
&\leq (1 + \gamma\eta/2)(1 - \gamma\eta/2)^2 \mathbb{E}\left[\|v_k\|^2\right] + \left(\frac{2}{\gamma\eta} + 1\right) u^2\eta^2 \mathbb{E}\left[\left\|Q_G(\nabla\tilde{U})\right\|^2\right] + 2\gamma u d\eta + \mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right] \\
&\leq (1 - \gamma\eta/2)\mathbb{E}\left[\|v_k\|^2\right] + 3u^2\eta/\gamma\mathcal{G}^2 + 2\gamma u d\eta + \mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right].
\end{aligned}
$$

When $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right] \leq 2Var_v d < 4\gamma u d\eta$, the inequality can be further written as:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_{k+1}\|^2\right] &\leq (1-\gamma\eta/2)\,\mathbb{E}\left[\|v_k\|^2\right] + 3u^2\eta/\gamma\mathcal{G}^2 + 6\gamma u d\eta \\
&\leq \mathbb{E}\left[\|\mathbf{v}_0\|^2\right] + \frac{6u^2\eta\mathcal{G}^2}{\gamma^2\eta} + \frac{12\gamma u d\eta}{\gamma\eta} \\
&\leq \mathbb{E}\left[\|\mathbf{v}_0\|^2\right] + \frac{6u^2\eta\mathcal{G}^2}{\gamma^2} + 12ud.
\end{aligned}$$

If $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right] \leq 2\mathbb{E}\left[\left\|\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1-e^{-\gamma\eta}\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) - Q^s\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1-e^{-\gamma\eta}\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right)\right\|^2\right]$, the ineuqality can be wirtten as:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_{k+1}\|^2\right] &\leq (1-\gamma\eta/2)\,\mathbb{E}\left[\|v_k\|^2\right] + 3u^2\eta/\gamma\mathcal{G}^2 + 2\gamma u d\eta + 2\Delta\left(1-e^{-\gamma\eta}\right)\sqrt{d}\left(\mathbb{E}\left[\|v_k\|\right] + \mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|\right]\right) \\
&\leq (1-\gamma\eta/2)\,\mathbb{E}\left[\|v_k\|^2\right] + 3u^2\eta/\gamma\mathcal{G}^2 + 2\gamma u d\eta + 2\Delta\gamma\eta\sqrt{d}\left(\sqrt{\mathbb{E}\left[\|v_k\|^2\right]} + \mathcal{G}\right) \\
&\leq \left(\sqrt{1-\gamma\eta/2}\sqrt{\mathbb{E}\left[\|v_k\|^2\right]} + \frac{\Delta\gamma\eta\sqrt{d}}{\sqrt{1-\gamma\eta/2}}\right)^2 + 3u^2\eta/\gamma\mathcal{G}^2 + 2\gamma u d\eta + 2\Delta\gamma\eta\sqrt{d}\mathcal{G}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\|v_k\|\right] \leq &\sqrt{\mathbb{E}\left[\|\mathbf{v}_0\|^2\right]} + \frac{\Delta\gamma\eta\sqrt{d}}{\left(1-\sqrt{1-\gamma\eta/2}\right)\sqrt{1-\gamma\eta/2}} + \frac{3u^2\eta/\gamma\mathcal{G}^2 + 2\gamma u d\eta + 2\Delta\gamma\eta\sqrt{d}\mathcal{G}}{\frac{\Delta\gamma\eta\sqrt{d}}{\sqrt{1-\gamma\eta/2}} + \sqrt{\gamma\eta/2\left(3u^2\eta/\gamma\mathcal{G}^2 + 2\gamma u d\eta + 2\Delta\gamma\eta\sqrt{d}\mathcal{G}\right)}} \\
\leq &\sqrt{\mathbb{E}\left[\|\mathbf{v}_0\|^2\right]} + \frac{\Delta\gamma\eta\sqrt{d}}{1-\gamma\eta/2} + \sqrt{6u^2/\gamma^2\mathcal{G}^2 + 4ud + 4\Delta\sqrt{d}\mathcal{G}} \\
\leq &\sqrt{\mathbb{E}\left[\|\mathbf{v}_0\|^2\right]} + \Delta\sqrt{d} + \sqrt{6u^2/\gamma^2\mathcal{G}^2 + 4ud + 4\Delta\sqrt{d}\mathcal{G}}.
\end{aligned}$$

Finally, we can have:

$$\begin{aligned}
\mathbb{E}\left[\|v_k\|\right] \leq \max\Bigg\{&\sqrt{\mathbb{E}\left[\|\mathbf{v}_0\|^2\right]} + \Delta\sqrt{d} + \sqrt{6u^2/\gamma^2\mathcal{G}^2 + 4ud + 4\Delta\sqrt{d}\mathcal{G}}, \\
&\sqrt{\mathbb{E}\left[\|\mathbf{v}_0\|^2\right]} + \sqrt{\frac{6u^2\eta\mathcal{G}^2}{\gamma^2}} + \sqrt{12ud}\Bigg\} =: A'.
\end{aligned}$$

Thus, we can have,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1-e^{-\gamma\eta}\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right) - Q^s\left(\mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}\left(1-e^{-\gamma\eta}\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right)\right\|^2\right] \\
&\leq \Delta\gamma\eta\sqrt{d}\left(A' + \mathcal{G}\right),
\end{aligned}$$

and we can bound the $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right]$ as,

$$\begin{aligned}
\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right] &\leq \max\left\{\Delta\gamma\eta\sqrt{d}\left(A'+\mathcal{G}\right), 4\gamma u d\eta\right\} \\
&= \gamma\eta\max\left\{\Delta\sqrt{d}\left(A'+\mathcal{G}\right), 4ud\right\} \\
&=: \gamma\eta A. \qquad (61)
\end{aligned}$$

Now we bound the $\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right]$. When $Var_x \geq \rho_0$, as the same analysis in (59) we can show,

$$\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] \leq 2Var_x d \leq 4ud\eta^2.$$

If $Var_x < \rho_0$, and let $\mu_x = \mathbf{x}_k + \gamma^{-1}\left(1-e^{-\gamma\eta}\right)v_k + u\gamma^{-2}\left(\gamma\eta + e^{-\gamma\eta} - 1\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k))$, by the same analysis in (60) we can have:

$$\begin{aligned}
&\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] \\
&\leq \max\left\{2\mathbb{E}\left[\|\mu_x - Q^s\left(\mu_x\right)\|^2\right], 2Var_x d\right\}.
\end{aligned}$$

Again using the bound equation (6) in Li and De Sa [2019] gives us,

$$\begin{aligned}
\mathbb{E}\left[\|\mu_x - Q^s(\mu_x)\|^2\right] &\leq \Delta\mathbb{E}\left[\left\|\gamma^{-1}\left(1-e^{-\gamma\eta}\right)v_k + u\gamma^{-2}\left(\gamma\eta + e^{-\gamma\eta} - 1\right)Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|_1\right] \\
&\leq \Delta\eta\mathbb{E}\left[\|v_k\|_1\right] + \frac{u\eta^2}{2}\mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|_1\right] \\
&\leq \Delta\eta\sqrt{d}\mathbb{E}\left[\|v_k\|\right] + \frac{u\eta^2}{2}\sqrt{d}\mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|\right] \\
&\leq \Delta\eta\sqrt{d}A' + \frac{u\eta^2}{2}\sqrt{d}\mathcal{G}.
\end{aligned}$$

Thus, we can have,

$$\begin{aligned}
\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] &\leq \max\left\{2\Delta\eta\sqrt{d}A' + u\eta^2\sqrt{d}\mathcal{G}, 4ud\eta^2\right\} \\
&\leq \eta\max\left\{2\Delta\sqrt{d}A' + u\eta\sqrt{d}\mathcal{G}, 4ud\eta\right\} \\
&=: \eta B.
\end{aligned} \tag{62}$$

Then follow the same analysis of (48), we can show

$$\begin{aligned}
\mathcal{W}_2(p_K, p^*) &\leq 4e^{-K\eta/2\kappa_1}\mathcal{W}_2(q_0, q^*) + \frac{4\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{1-e^{-\eta/2\kappa_1}} \\
&\quad + \frac{20u^2\eta^2\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + 8u^2\eta\left(\gamma A + B\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}} + \sqrt{1-e^{-\eta/\kappa_1}}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + 2u^2\eta\left(\gamma A + B\right)}}.
\end{aligned}$$

Now we let the first term less than $\epsilon/3$, from the Lemma 13 in [Cheng et al., 2018] we know that $\mathcal{W}_2(q_0, q^*) \leq 3\left(\frac{d}{m_1} + \mathcal{D}^2\right)$. So we can choose K as the following,

$$K \leq \frac{2\kappa_1}{\eta}\log\left(36\left(\frac{d}{m_1} + \mathcal{D}^2\right)\right).$$

Next, we choose a stepsize $\eta \leq \frac{\epsilon\kappa_1^{-1}}{\sqrt{479232/5(d/m_1+\mathcal{D}^2)}}$ to ensure the second term is controlled below $\epsilon/3$. Since $1 - e^{-\eta/2\kappa_1} \geq \eta/4\kappa_1$ and definition of $\mathcal{E}_K$,

$$4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{1-e^{-\eta/2\kappa_1}} \leq 4\frac{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}}{\eta/4\kappa_1} \leq 16\kappa_1\left(\eta\sqrt{\frac{8\mathcal{E}_K}{5}}\right) \leq \epsilon/3.$$

Finally by choosing the stepsize satisfied that,

$$\eta \leq \frac{\epsilon^2}{2880\kappa_1 u\left(\frac{\Delta^2 d}{4} + \sigma^2\right)},$$

the third term can be bounded as:

$$\begin{aligned}
&\frac{20u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+8u^2\eta\left(\gamma A+B\right)}{\eta^2\sqrt{\frac{8\mathcal{E}_K}{5}}+\sqrt{1-e^{-\eta/\kappa_1}}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+2u^2\eta\left(\gamma A+B\right)}}\\
&\leq\frac{20u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+8u^2\eta\left(\gamma A+B\right)}{\sqrt{1-e^{-\eta/\kappa_1}}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+2u^2\eta\left(\gamma A+B\right)}}\leq\frac{20u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+8u^2\eta\left(\gamma A+B\right)}{\sqrt{\eta/4\kappa_1}\sqrt{5u^2\eta^2\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+2u^2\eta\left(\gamma A+B\right)}}\\
&\leq 4\sqrt{20u^2\kappa_1\eta\left(\frac{\Delta^2 d}{4}+\sigma^2\right)+8\kappa_1 u^2\left(\gamma A+B\right)}\\
&\leq \epsilon/3+8\sqrt{2\kappa_1 u^2\left(\gamma A+B\right)}.
\end{aligned}$$

This complete the proof.

D.8 Proof of Theorem 9

Similarily, from the analysis in (61), we know that

$$\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right]\leq\gamma\eta\mathcal{A}, \tag{63}$$

where $A=\max\left\{\Delta\sqrt{d}\left(A'+\mathcal{G}\right),4ud\right\}$. By the analysis in (59), we know that if $\mathrm{Var}_{\mathbf{x}}^{hmc}\geq\frac{\Delta^2}{4}$, we can have

$$\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right]\leq 4ud\eta^2 \tag{64}$$

by (62), if $\mathrm{Var}_{\mathbf{x}}^{hmc}<\frac{\Delta^2}{4}$,

$$\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right]\leq\eta B, \tag{65}$$

where $B=\max\left\{2\Delta\sqrt{d}A'+u\eta\sqrt{d}\mathcal{G},4ud\eta\right\}$. Thus, we can define the following:

$$\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right]=\eta\mathcal{B}, \tag{66}$$

where $\mathcal{B}$ is defined as:

$$\mathcal{B}=\begin{cases}4ud\eta, & \text{if } \mathrm{Var}_{\mathbf{x}}^{hmc}\geq\frac{\Delta^2}{4}\\ B, & \text{else.}\end{cases}$$

Combining the bound of $\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right]$, $\mathbb{E}\left[\|\alpha_k^{\mathbf{v}}\|^2\right]$ with (51), we can show,

$$
\begin{aligned}
& D_{KL}(p_K||\hat{p}_{K\eta}) \\
& \leq \frac{u}{4\gamma T\eta^2}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}\|^2\right] + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\gamma\alpha_k^{\mathbf{v}}/\eta\|^2\right]ds + \frac{u}{4\gamma}\sum_{k=0}^{K}\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left[\|\alpha_k^{\mathbf{x}}/\eta\|^2\right]ds \\
& + \frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E} + \left(\gamma^2+2u^2M^2\right)C\Delta^2 d + u^2\sigma^2 + 2u^2G^2 + 2du\right) + \frac{u}{4\gamma}K\eta\left(\frac{\Delta^2 d}{4}+\sigma^2\right) \\
& \leq \frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E} + \left(\gamma^2+2u^2M^2\right)C\Delta^2 d + u^2\sigma^2 + 2u^2G^2 + 2du\right) + \frac{u}{4\gamma}K\eta\left(\frac{\Delta^2 d}{4}+\sigma^2\right) \\
& + \frac{u\mathcal{B}}{4\gamma T} + \frac{uK\mathcal{A}}{4} + \frac{uK\mathcal{B}}{4\gamma} \\
& \leq \frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E} + \left(\gamma^2+2u^2M^2\right)C\Delta^2 d + u^2\sigma^2 + 2u^2G^2 + 2du\right) + \frac{u}{4\gamma}K\eta\left(\frac{\Delta^2 d}{4}+\sigma^2\right) \\
& + \frac{uK\mathcal{A}}{4} + \frac{uK\mathcal{B}}{2\gamma} \\
& \leq \frac{u}{4\gamma}3M^2K\eta^3\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E} + u^2\sigma^2 + 2u^2G^2 + 2du\right) + \frac{u}{4\gamma}K\eta\sigma^2 + \frac{u}{16\gamma}K\eta\Delta^2 d + \frac{uK\mathcal{A}}{4} + \frac{uK\mathcal{B}}{2\gamma} \\
& =: C_0K\eta^3 + C_1K\eta\sigma^2 + C_2K\eta\Delta^2 + C_3K\mathcal{A} + C_4K\mathcal{B},
\end{aligned}
$$

where the constants are defined as

$$
\begin{aligned}
C_0 &= \frac{u}{4\gamma}3M^2\left(\left(\gamma^2+2u^2M^2\right)\mathcal{E} + u^2\sigma^2 + 2u^2G^2 + 2du\right) \\
C_1 &= \frac{u}{4\gamma} \\
C_2 &= \frac{u}{16\gamma}d \\
C_3 &= \frac{u}{4} \\
C_4 &= \frac{u}{2\gamma}.
\end{aligned}
$$

By the weighted CKP inequality and given $K\eta \geq 1$,

$$
\begin{aligned}
\mathcal{W}_2(p_K, \hat{p}_{K\eta}) &\leq \overline{\Lambda}\left(\sqrt{D_{KL}(p_K||\hat{p}_{K\eta})} + \sqrt[4]{D_{KL}(p_K||\hat{p}_{K\eta})}\right) \\
&\leq \left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\widetilde{A} + \widetilde{C_2}\sqrt{\Delta}\right)\sqrt{K\eta} + \widetilde{C_3}\sqrt{K\mathcal{A}} + \widetilde{C_4}\sqrt{K\mathcal{B}},
\end{aligned}
$$

where the constants are defined as:

$$
\begin{aligned}
\widetilde{C_0} &= \overline{\Lambda}\left(\sqrt{C_0} + \sqrt[4]{C_0}\right) \\
\widetilde{C_1} &= \overline{\Lambda}\left(\sqrt{C_1} + \sqrt[4]{C_1}\right) \\
\widetilde{C_2} &= \overline{\Lambda}\left(\sqrt{C_2} + \sqrt[4]{C_2}\right) \\
\widetilde{C_3} &= \overline{\Lambda}\left(\sqrt{C_3} + \sqrt[4]{C_3}\right) \\
\widetilde{C_4} &= \overline{\Lambda}\left(\sqrt{C_4} + \sqrt[4]{C_4}\right) \\
\widetilde{A}^2 &= \overline{\Lambda}\max\left\{\sigma^2, \sqrt{\sigma^2}\right\}.
\end{aligned}
$$

From the same analysis of (36), we can have:

$$
\mathcal{W}_2(p_K, p^*) \leq \left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\widetilde{A}\right)\sqrt{K\eta} + \widetilde{C_2}\sqrt{K\eta}\Delta + \widetilde{C_3}\sqrt{K\mathcal{A}} + \widetilde{C_4}\sqrt{K\mathcal{B}} + \Gamma_0 e^{-\mu^* K\eta}. \quad (67)
$$

In order to bound the Wasserstein distance, we need to set

$$\overline{\Lambda}\widetilde{C_0}\sqrt{K\eta^2} = \frac{\epsilon}{2} \quad \text{and} \quad \Gamma_0 e^{-\mu^* K\eta} = \frac{\epsilon}{2}. \tag{68}$$

Solving the equation (68), we can have

$$K\eta = \frac{\log\left(\frac{2\Gamma_0}{\epsilon}\right)}{\mu^*} \quad \text{and} \quad \eta = \frac{\epsilon^2}{4\overline{\Lambda}^2\widetilde{C_0}^2 K\eta}.$$

Combining these two we can have

$$\eta = \frac{\epsilon^2\mu^*}{4\overline{\Lambda}^2\widetilde{C_0}^2\log\left(\frac{2\Gamma_0}{\epsilon}\right)} \quad \text{and} \quad K = \frac{4\overline{\Lambda}^2\widetilde{C_0}^2\log^2\left(\frac{2\Gamma_0}{\epsilon}\right)}{\epsilon^2\left(\mu^*\right)^2}.$$

Plugging in (67) completes the proof.

D.9 Proof of Theorem 12

Recall that the update of VC SGLDLP-L is

$$\begin{aligned}\mathbf{x}_{k+1} &= Q^{vc}\left(\mathbf{x}_k - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k)), 2\eta, \Delta\right)\\ &= \mathbf{x}_k - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \sqrt{2\eta}\xi_k + \alpha_k,\end{aligned}$$

where α_k is defined as

$$\alpha_k = Q^{vc}\left(\mathbf{x}_k - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k)), 2\eta, \Delta\right) - \mathbf{x}_k - \eta Q_G(\nabla\tilde{U}(\mathbf{x}_k)) + \sqrt{2\eta}\xi_k.$$

From analysis in Zhang et al. [2022], we know that

$$\begin{aligned}\mathbb{E}\left[\|\alpha_k\|^2\right] &\leq \max\left(2\Delta\eta G, 5\eta d\right)\\ &=: \eta A.\end{aligned}$$

Combining the analysis in section D.6, we can show,

$$\begin{aligned}D_{KL}(p_K||\hat{p}_{K\eta}) \leq &\frac{M\overline{\mathcal{E}}}{4}K\eta^2 + \frac{3M+1}{4}\sigma^2 K\eta + \frac{\left(\left(6+3m_2\right)M+m_2\right)d}{16m_2}\Delta^2 K\eta + \left(\frac{6M\eta}{4m_2} + \frac{1}{4\eta}\right)K\mathbb{E}\left[\|\alpha_k\|^2\right]\\ \leq &\frac{M\overline{\mathcal{E}}}{4}K\eta^2 + \frac{3M+1}{4}\sigma^2 K\eta + \frac{\left(\left(6+3m_2\right)M+m_2\right)d}{16m_2}\Delta^2 K\eta + \left(\frac{6M\eta}{4m_2} + \frac{1}{4\eta}\right)K\eta A\\ \leq &\frac{M\overline{\mathcal{E}}}{4}K\eta^2 + \frac{3M+1}{4}\sigma^2 K\eta + \frac{\left(\left(6+3m_2\right)M+m_2\right)d}{16m_2}\Delta^2 K\eta + \frac{6M+m_2}{m_2}KA\\ =:&C_0K\eta^2 + C_1K\eta\sigma^2 + C_2K\eta\Delta^2 + C_3KA,\end{aligned}$$

where the constant C_0, C_1, C_2 and C_3 are defined as:

$$\begin{aligned}C_0 &= \frac{M\overline{\mathcal{E}}}{4}\\ C_1 &= \frac{3M+1}{4}\\ C_2 &= \frac{\left(\left(6+3m_2\right)M+m_2\right)d}{16m_2}\\ C_3 &= \frac{6M+m_2}{m_2}\end{aligned}$$

We are ready to bound the Wasserstein distance,

$$\begin{aligned}\mathcal{W}_2^2(p_K, \hat{p}_{K\eta}) \leq &\left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\Big[\left(\left(C_0 + \sqrt{C_0}\right)\eta + \left(C_1 + \sqrt{C_1}\right)\widetilde{A}\right)(K\eta)^2 + \left(C_2 + \sqrt{C_2}\right)\Delta(K\eta)^2\\ &+ \left(C_3 + \sqrt{C_3}\right)\mathcal{A}K^2\eta\Big]\\ =: &\left(\widetilde{C_0}^2\eta + \widetilde{C_1}^2\widetilde{A} + \widetilde{C_2}^2\Delta\right)(K\eta)^2 + \widetilde{C_3}^2\mathcal{A}K^2\eta,\end{aligned}$$

where the constants are defined as:

$$\begin{aligned}
\widetilde{A} &= \max\left\{\sigma^2, \sqrt{\sigma^2}\right\} \\
\mathcal{A} &= \max\left\{A, \sqrt{A}\right\} \\
\widetilde{C_0}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_0 + \sqrt{C_0}\right) \\
\widetilde{C_1}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_1 + \sqrt{C_1}\right) \\
\widetilde{C_2}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_2 + \sqrt{C_2}\right) \\
\widetilde{C_3}^2 &= \left(12 + 8\left(\kappa_0 + 2b + 2d\right)\right)\left(C_3 + \sqrt{C_3}\right).
\end{aligned}$$

From Proposition 9 in the paper Raginsky et al. [2017], we know that

$$\begin{aligned}
\mathcal{W}_2(\hat{p}_{K\eta}, p^*) \leq& \sqrt{2C_{LS}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{M\kappa_0}{3} + B\sqrt{\kappa_0} + G_0 + \frac{b}{2}\log 3\right)\right)}e^{-K\eta/\beta C_{LS}} \\
=:& \widetilde{C_4}e^{-K\eta/\beta C_{LS}}
\end{aligned}$$

Finally, we can have

$$\mathcal{W}_2(p_K, p^*) \leq \left(\widetilde{C_0}\sqrt{\eta} + \widetilde{C_1}\sqrt{A} + \widetilde{C_2}\sqrt{\Delta}\right)K\eta + \widetilde{C_3}\sqrt{\mathcal{A}}\sqrt{K^2\eta} + \widetilde{C_4}e^{-K\eta/\beta C_{LS}}. \tag{69}$$

In order to bound the 2-Wasserstein distance, we need to set

$$\widetilde{C_0}K\eta^{5/4} = \frac{\epsilon}{2} \quad \text{and} \quad \widetilde{C_3}e^{-K\eta/\beta C_{LS}} = \frac{\epsilon}{2}. \tag{70}$$

Solving the (70), we can have

$$K\eta = C_{LS}\log\left(\frac{2\widetilde{C_3}}{\epsilon}\right) \quad \text{and} \quad \eta = \frac{\epsilon^4}{16\widetilde{C_0}^4\left(K\eta\right)^4}.$$

Combining these two we can have

$$\eta = \frac{\epsilon^4}{16\widetilde{C_0}^4 C_{LS}^4\log^4\left(\frac{2\widetilde{C_3}}{\epsilon}\right)} \quad \text{and} \quad K = \frac{16\widetilde{C_0}^4 C_{LS}^5\log^5\left(\frac{2\widetilde{C_3}}{\epsilon}\right)}{\epsilon^4}.$$

Plugging K and η into (69) completes the proof.

E Techinical Proofs

E.1 Proof of Lemma 13

Proof. By the definition of ξ in (25)

$$\begin{aligned}
\|\mathbb{E}\xi\|^2 &= \|\mathbb{E}\tilde{g}(\mathbf{x}) - \mathbb{E}\nabla U(\mathbf{x})\|^2 \\
&= \|\mathbb{E}\nabla U(Q_w(\mathbf{x})) - \mathbb{E}\nabla U(\mathbf{x})\|^2 \\
&\leq \mathbb{E}\left[\|\nabla U(Q_w(\mathbf{x})) - \nabla U(\mathbf{x})\|^2\right] \\
&\leq M^2\mathbb{E}\left[\|Q_w(\mathbf{x}) - \nabla U(\mathbf{x})\|^2\right] \\
&\leq M\frac{\Delta^2 d}{4}.
\end{aligned}$$

We also know that from the definition that

$$
\begin{aligned}
&\mathbb{E}\left\|\xi\right\|^2 = \mathbb{E}\left\|\tilde{g}(\mathbf{x}) - \nabla U(\mathbf{x})\right\|^2 \\
&= \mathbb{E}\left\|Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}))) - \nabla\tilde{U}(Q_W(\mathbf{x})) + \nabla\tilde{U}(Q_W(\mathbf{x})) - \nabla U(Q_W(\mathbf{x})) + \nabla U(Q_W(\mathbf{x})) - \nabla U(\mathbf{x})\right\|^2 \\
&= \mathbb{E}\left\|Q_G(\nabla\tilde{U}(Q_W(\mathbf{x}))) - \nabla\tilde{U}(Q_W(\mathbf{x}))\right\|^2 + \mathbb{E}\left\|\nabla\tilde{U}(Q_W(\mathbf{x})) - \nabla U(Q_W(\mathbf{x}))\right\|^2 + \mathbb{E}\left\|\nabla U(Q_W(\mathbf{x})) - \nabla U(\mathbf{x})\right\|^2 \\
&\le \frac{\Delta^2 d}{4} + \sigma^2 + M^2\mathbb{E}\left\|Q_W(\mathbf{x}) - \mathbf{x}\right\|^2 \\
&\le (M^2+1)\frac{\Delta^2 d}{4} + \sigma^2,
\end{aligned}
$$

where in the first inequality, we apply Assumptions 1 and 4.

□

E.2 Proof of Lemma 14

Proof. Let Γ_1 be the set of all couplings between $\widetilde{\Phi}_\eta q_0$ and q^* and Γ_2 be the set of all couplings between $\widehat{\Phi}_\eta q_0$ adn q^*. Let r_1 be the optimal coupling between $\widetilde{\Phi}_\eta q_0$ and q^*, i.e.

$$\mathbb{E}_{(\theta,\phi)\sim r_1}[\|\theta - \phi\|^2] = \mathcal{W}_2^2(\widetilde{\Phi}_\eta q_0, q^*).$$

Let $\left(\begin{bmatrix}\tilde{x}\\ \tilde{\omega}\end{bmatrix}, \begin{bmatrix}x^*\\ \omega^*\end{bmatrix}\right) \sim r_1$. We define the random variable $\begin{bmatrix}x\\ \omega\end{bmatrix}$ as

$$\begin{bmatrix}x\\ \omega\end{bmatrix} = \begin{bmatrix}\tilde{x}\\ \tilde{\omega}\end{bmatrix} + u\begin{bmatrix}\left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi \\ \left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\eta e^{-\gamma(s-\eta)}ds\right)\xi\end{bmatrix}.$$

By equation (29), $\left(\begin{bmatrix}x\\ \omega\end{bmatrix}, \begin{bmatrix}x^*\\ \omega^*\end{bmatrix}\right)$ define a valid coupling between $\Phi_\eta q_0$ and q^*. Now we can analyze the Wasserstein distance between $\Phi_\eta q_0$ and q^*.

$$
\begin{aligned}
\mathcal{W}_2^2(\widehat{\Phi}_\eta q_0, q^*) &\le \mathbb{E}_{r_1}\left[\left\|\begin{bmatrix}\tilde{x}\\ \tilde{\omega}\end{bmatrix} + u\begin{bmatrix}\left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\xi \\ \left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\delta e^{-\gamma(s-\eta)}ds\right)\xi\end{bmatrix} - \begin{bmatrix}x^*\\ \omega^*\end{bmatrix}\right\|^2\right] \qquad (71) \\
&\le \mathbb{E}_{r_1}\left[\left\|\begin{bmatrix}\tilde{x}-x^*\\ \tilde{\omega}-\omega^*\end{bmatrix} + u\begin{bmatrix}\left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)\mathbb{E}\xi \\ \left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\delta e^{-\gamma(s-\eta)}ds\right)\mathbb{E}\xi\end{bmatrix}\right\|^2\right] \\
&+ \mathbb{E}_{r_1}\left[\left\|u\begin{bmatrix}\left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr\right)(\xi-\mathbb{E}\xi) \\ \left(\int_0^\eta\left(\int_0^r e^{-\gamma(s-r)}ds\right)dr + \int_0^\eta e^{-\gamma(s-\eta)}ds\right)(\xi-\mathbb{E}\xi)\end{bmatrix}\right\|^2\right] \\
&\le \left(\mathcal{W}_2(\widetilde{\Phi}_\eta q_0, q^*) + 2u\sqrt{\eta^4/4+\eta^2}\left\|\mathbb{E}\xi\right\|\right)^2 + 4u^2(\eta^4/4+\eta^2)\mathbb{E}_{r_1}\left[\left\|\xi - \mathbb{E}\xi\right\|^2\right] \\
&\le \left(\mathcal{W}_2(\widetilde{\Phi}_\eta q_0, q^*) + \sqrt{5}/2u\eta\sqrt{d}M\Delta\right)^2 + 5u^2\eta^2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right).
\end{aligned}
$$

□

E.3 Proof of Lemma 15

Proof. In order to get the upper bound of $\|\mathbf{x}_k\|$ and $\|\mathbf{v}_k\|$, we bound the Lyapunov function $\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)$. By the smooth Assumption 1, we know

$$U(\mathbf{x}_{k+1}) - U(x^*) \le U(\mathbf{x}_k) + \langle\nabla U(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + M^2/2\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 - U(x^*).$$

Recall the definition of the Lyapunov function

$$\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1}) = \left\|\mathbf{x}_{k+1}\right\|^2 + \left\|\mathbf{x}_{k+1} + 2\mathbf{v}_{k+1}/\gamma\right\|^2 + 8u\left(U(\mathbf{x}_{k+1}) - U(x^*)\right)/\gamma^2.$$

For the first two terms we have

$$\begin{aligned}\|\mathbf{x}_{k+1}\|^2 &= \|\mathbf{x}_k\|^2 + 2\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ \|\mathbf{x}_{k+1} + 2\mathbf{v}_{k+1}/\gamma\|^2 &= \|\mathbf{x}_k + 2\mathbf{v}_k/\gamma\|^2 + 2\langle \mathbf{x}_k + 2\mathbf{v}_k/\gamma, \mathbf{x}_{k+1} - \mathbf{x}_k + 2(\mathbf{v}_{k+1} - \mathbf{v}_k)/\gamma\rangle \\ &\quad + \|\mathbf{x}_{k+1} - \mathbf{x}_k + 2(\mathbf{v}_{k+1} - \mathbf{v}_k)/\gamma\|^2 .\end{aligned}$$

This implies the following:

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \le \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] &+ 4\mathbb{E}\left[\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle\right] + \frac{4}{\gamma}\mathbb{E}\left[\langle \mathbf{x}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle\right] + \frac{4}{\gamma}\mathbb{E}\left(\langle \mathbf{v}_k, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle\right) \\ &+ \frac{8}{\gamma^2}\mathbb{E}\left[\langle \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle\right] + \frac{8u}{\gamma^2}\mathbb{E}\left[\langle \nabla U(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + M/2\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2\right] \\ &+ \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\right] + \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}_k + 2(\mathbf{v}_{k+1} - \mathbf{v}_k)/\gamma\|^2\right] .\end{aligned} \tag{72}$$

By the update rule in (3), we know that

$$\begin{aligned}\mathbb{E}\left[\langle \mathbf{x}_k, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle\right] &= \frac{1 - e^{-\gamma\eta}}{\gamma}\mathbb{E}\left[\langle \mathbf{x}_k, \mathbf{v}_k\rangle\right] + \frac{u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^2}\mathbb{E}\left[\langle \mathbf{x}_k, \tilde{g}(\mathbf{x}_k)\rangle\right], \\ \mathbb{E}\left[\langle \mathbf{x}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle\right] &= -(1 - e^{-\gamma\eta})\mathbb{E}\left[\langle \mathbf{x}_k, \mathbf{v}_k\rangle\right] - \frac{u(1 - e^{-\gamma\eta})}{\gamma}\mathbb{E}\left[\langle \mathbf{x}_k, \tilde{g}(\mathbf{x}_k)\rangle\right], \\ \mathbb{E}\left[\langle \mathbf{v}_k, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle\right] &= \frac{1 - e^{-\gamma\eta}}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^2}\mathbb{E}\left[\langle \mathbf{v}_k, \tilde{g}(\mathbf{x}_k)\rangle\right], \\ \mathbb{E}\left[\langle \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle\right] &= -(1 - e^{-\gamma\eta})\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] - \frac{u(1 - e^{-\gamma\eta})}{\gamma}\mathbb{E}\left[\langle \mathbf{v}_k, \tilde{g}(\mathbf{x}_k)\rangle\right] .\end{aligned}$$

Plug into the (72) yields:

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \le \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] &- \frac{4u(2 - \gamma\eta - 2e^{-\gamma\eta})}{\gamma^2}\mathbb{E}\left[\langle \mathbf{x}_k, \tilde{g}(\mathbf{x}_k)\rangle\right] - \frac{4(1 - e^{-\gamma\eta})}{\gamma^2}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \\ &+ \frac{4u(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^3}\mathbb{E}\left[\langle \mathbf{v}_k, \tilde{g}(\mathbf{x}_k)\rangle\right] + \frac{8u(1 - e^{-\gamma\eta})}{\gamma^3}\mathbb{E}\left[\langle \mathbf{v}_k, \nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\rangle\right] \\ &+ \frac{8u^2(\gamma\eta + e^{-\gamma\eta} - 1)}{\gamma^4}\mathbb{E}\left[\langle \nabla U(\mathbf{x}_k), \tilde{g}(\mathbf{x}_k)\rangle\right] + \left(\frac{4Mu}{\gamma^2} + 3\right)\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\right] \\ &+ \frac{8}{\gamma^2}\mathbb{E}\left[\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2\right] .\end{aligned} \tag{73}$$

By Assumption 3, we know that $\langle \mathbf{x}_k, \nabla U(\mathbf{x}_k)\rangle \ge m_2\left\|\mathbf{x}_k\right\|^2 - b$. We then assume $\eta \le 1/(8\gamma)$ and use the inequality $-x \le e^{-x} - 1 \le x^2/2 - x$ for any $x \ge 0$, it follows that

$$\begin{aligned}&- \frac{4u(2 - \gamma\eta - 2e^{-\gamma\eta})}{\gamma^2}\mathbb{E}\left[\langle \mathbf{x}_k, \tilde{g}(\mathbf{x}_k)\rangle\right] \\ &= -\frac{4u(2 - \gamma\eta - 2e^{-\gamma\eta})}{\gamma^2}\left(\mathbb{E}\left[\langle \mathbf{x}_k, \nabla U(\mathbf{x}_k)\rangle\right] + \mathbb{E}\left[\langle \mathbf{x}_k, \tilde{g}(\mathbf{x}_k) - \nabla U(\mathbf{x}_k)\rangle\right]\right) \\ &\le -\frac{4u(2 - \gamma\eta - 2e^{-\gamma\eta})}{\gamma^2}\left(m_2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - b\right) + \frac{4u(2 - \gamma\eta - 2e^{-\gamma\eta})}{\gamma^2}\left(\frac{1}{8}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k) - \nabla U(\mathbf{x}_k)\|^2\right]\right) \\ &\le -\frac{3m_2u\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + \frac{4u\eta b}{\gamma} + \frac{8u\eta}{\gamma}\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k) - \nabla U(\mathbf{x}_k)\|^2\right],\end{aligned}$$

where the first inequality is because of the Young's inequaltiy and Assumption 1 and the last inequality is based on the inequality that $\gamma\eta - (\gamma\eta)^2 \le 2 - \gamma\eta - 2e^{-\gamma\eta} \le \gamma\eta$. Again by Young's inequality and the update rule in (3) we have:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\right] &\le 2\eta^2\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + u^2\eta^4/2\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right] + \mathbb{E}\left[\|\xi_k^x\|^2\right] \\ \mathbb{E}\left[\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2\right] &\le 2\gamma^2\eta^2\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + 2u^2\eta^2\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right] + \mathbb{E}\left[\|\xi_k^v\|^2\right] .\end{aligned}$$

It is easy to verify the fact that $\mathbb{E}\left[\|\xi_k^v\|^2\right] \leq 2\gamma u d\eta$ and $\mathbb{E}\left[\|\xi_k^x\|^2\right] \leq 2ud\eta^2$. Thus,

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \\
&\leq \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um\eta^2}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{3(1-e^{-\gamma\eta}) - \eta^2(8Mu + u\gamma + 22\gamma^2)}{\gamma^2}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \\
&+ \frac{36u^2\eta^2 + 2\gamma u\eta^2 + \left(4Mu + 3\gamma^2\right)\eta^4}{2\gamma^2}\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right] + \frac{2u^2\eta^2}{\gamma^2}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right] \\
&+ \frac{8u\eta(\gamma^2 + 2u)}{\gamma^3}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right] + \frac{(8Mu + 6\gamma^2)ud\eta^2 + 4(4d+b)u\gamma\eta}{\eta^2}.
\end{aligned}
$$

If we set

$$
\eta \leq \min\left\{\frac{\gamma}{4\left(8Mu + u\gamma + 22\gamma^2\right)}, \sqrt{\frac{4u^2}{4Mu + 3\gamma^2}}, \frac{6\gamma bu}{\left(4Mu + 3\gamma^2\right)d}\right\},
$$

we can obtain the following,

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \leq \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um_2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{(20u+\gamma)u\eta^2}{\gamma^2}\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right] \\
+ \frac{2u^2\eta^2}{\gamma^2}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right] + \frac{8u\eta\left(\gamma^2 + 2u\right)}{\gamma^3}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \tilde{g}(\mathbf{x}_k)\|^2\right] + \frac{16(d+b)u\eta}{\gamma}.
\end{aligned}
\tag{74}
$$

Furthermore we can bound $\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right]$ by the following analysis:

$$
\begin{aligned}
\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right] &\leq 2\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k) - \nabla U(\mathbf{x}_k)\|^2\right] + 2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right] \\
&\leq 2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) + 4M^2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 4G^2,
\end{aligned}
\tag{75}
$$

where G^2 is the bound of the gradient at 0, i.e. $\|\nabla U(0)\|^2 \leq G^2$. Thus we can have:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \leq \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um_2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{(21u+\gamma)4M^2u\eta^2}{\gamma^2}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \\
+ \left(\frac{2(20u+\gamma)u\eta^2}{\gamma^2} + \frac{8u\eta\left(\gamma^2+2u\right)}{\gamma^3}\right)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) \\
+ \frac{(21u+\gamma)4u\eta^2}{\gamma^2}G^2 + \frac{16(d+b)u\eta}{\gamma}.
\end{aligned}
$$

If we set the stepsize

$$
\eta \leq \min\left\{\frac{\gamma m_2}{12(21u+\gamma)M^2}, \frac{8(\gamma^2+2u)}{(20u+\gamma)\gamma}\right\},
$$

then we have:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \leq \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{8um_2\eta}{3\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \\
+ \left(\frac{16u\eta\left(\gamma^2+2u\right)}{\gamma^3}\right)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) \\
+ \frac{(21u+\gamma)4u\eta^2}{\gamma^2}G^2 + \frac{16(d+b)u\eta}{\gamma}.
\end{aligned}
$$

Furthermore by Young's inequality and Assumption 1, we can bound the Lyapunov function by the following:

$$
\mathcal{E}(x, v) \leq 5/2\|x\|^2 + \frac{12}{\gamma^2} + \frac{2uM}{\gamma^2}\left(3\|x\|^2 + 6\|x^*\|^2\right).
$$

Then if $\gamma^2 \leq 4Mu$, we have

$$\mathcal{E}(x,v) \leq \frac{16uM}{\gamma^2}\|x\|^2 + \frac{12}{\gamma^2}\|v\|^2 + \frac{12uM}{\gamma^2}\|x^*\|^2. \tag{76}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \leq &\left(1 - \frac{\gamma m_2 \eta}{6M}\right)\mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] + \left(\frac{16u\eta\left(\gamma^2 + 2u\right)}{\gamma^3}\right)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)\\ &+ \frac{(21u+\gamma)4u\eta^2}{\gamma^2}G^2 + \frac{16(d+b)u\eta}{\gamma}.\end{aligned}$$

Finally we show that

$$\begin{aligned}\sup_{k\geq 0}\mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] \leq &\,\mathbb{E}\left[\mathcal{E}(x_0, v_0)\right] + \frac{6M}{\gamma m_2 \eta}\left(\frac{16u\eta\left(\gamma^2+2u\right)}{\gamma^3}\right)\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)\\ &+ \frac{6M}{\gamma m_2\eta}\frac{(21u+\gamma)4u\eta^2}{\gamma^2}G^2 + \frac{6M}{\gamma m_2 \eta}\frac{16(d+b)u\eta}{\gamma}\\ \leq &\,\mathbb{E}\left[\mathcal{E}(x_0, v_0)\right] + \frac{96u\left(\gamma^2+2u\right)}{m_2\gamma^4}\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) + \frac{24(21u+\gamma)uM}{m_2\gamma^3}G^2 + \frac{96(d+b)uM}{m_2\gamma^2}\\ \leq &\,\overline{\mathcal{E}} + C_0\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right),\end{aligned} \tag{77}$$

where $\overline{\mathcal{E}} = \mathbb{E}\left[\mathcal{E}(x_0, v_0)\right] + \frac{24(21u+\gamma)uM}{m_2\gamma^3}G^2 + \frac{96(d+b)uM}{m_2\gamma^2}$ and $C_0 = \frac{96u\left(\gamma^2+2u\right)}{m_2\gamma^4}$. Moreover by the definition of Laypunov function, we know $\mathcal{E}(x,v) \geq \max\{\|x\|^2, 2\|v/\gamma\|^2\}$. This further implies that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] &\leq \overline{\mathcal{E}} + C_0\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right)\\ \mathbb{E}\left[\|\mathbf{v}_k\|^2\right] &\leq \gamma^2\overline{\mathcal{E}}/2 + \gamma^2 C_0/2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right).\end{aligned}$$

Combining with equation (75) we can bound $\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right]$ as:

$$\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k)\|^2\right] \leq 2\left((M^2+1)\frac{\Delta^2 d}{4} + \sigma^2\right) + 4M^2\overline{\mathcal{E}} + 4G^2. \tag{78}$$

□

E.4 Proof of Lemma 16

Proof. By the update rule in (18), we have:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{k+1}\|^2\right] =&\mathbb{E}\left[\|\mathbf{x}_k - \eta\tilde{g}(\mathbf{x}_k)\|^2\right] + \sqrt{8\eta}\mathbb{E}\left[\langle \mathbf{x}_k - \eta\tilde{g}(\mathbf{x}_k), \xi_{k+1}\rangle\right] + 2\eta\mathbb{E}\left[\|\xi_{k+1}\|^2\right]\\ =&\mathbb{E}\left[\|\mathbf{x}_k - \eta\tilde{g}(\mathbf{x}_k)\|^2\right] + 2\eta d\\ =&\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k) - \eta\left(\tilde{g}(\mathbf{x}_k) - \nabla U(Q_W(\mathbf{x}_k))\right) - \eta\left(\nabla U(Q_W(\mathbf{x}_k)) - \nabla U(\mathbf{x}_k)\right)\|^2\right] + 2\eta d\\ =&\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k) - \eta\left(\nabla U(Q_W(\mathbf{x}_k)) - \nabla U(\mathbf{x}_k)\right)\|^2\right] + \eta^2\mathbb{E}\left[\|\tilde{g}(\mathbf{x}_k) - \nabla U(Q_W(\mathbf{x}_k))\|^2\right] + 2\eta d\\ =&\left(\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k)\|\right] + \eta\mathbb{E}\left[\|\nabla U(Q_W(\mathbf{x}_k)) - \nabla U(\mathbf{x}_k)\|\right]\right)^2 + \eta^2\frac{\Delta^2 d}{4} + 2\eta d.\end{aligned}$$

We know the fact that:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k)\|^2\right] &= \mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - 2\eta\mathbb{E}\left[\langle \mathbf{x}_k, \nabla U(\mathbf{x}_k)\rangle\right] + \eta^2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right]\\ &= \mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2\eta\left(b - m_2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]\right) + 2\eta^2\left(M^2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + G^2\right)\\ &= \left(1 - 2\eta m_2 + 2\eta^2 M^2\right)\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2\eta b + 2\eta^2 G^2.\end{aligned}$$

For any $\eta \in \left(0, 1 \wedge \frac{m_2}{2M^2}\right)$, if $0 < 1 - 2\eta m_2 + 2\eta^2 M^2 < 1$ and set $c = \frac{\eta m_2 - \eta^2 M^2}{1 - 2\eta m + 2\eta^2 M^2}$, then we have:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_{k+1}\|^2\right] &\leq (1+c)\,\mathbb{E}\left[\|\mathbf{x}_k - \eta\nabla U(\mathbf{x}_k)\|^2\right] + \left(1+\frac{1}{c}\right)\eta^2\mathbb{E}\left[\|\nabla U(Q_W(\mathbf{x}_k)) - \nabla U(\mathbf{x}_k)\|^2\right] + \eta^2\frac{\Delta^2 d}{4} + 2\eta d \\ &\leq \left(1 - \eta m_2 + \eta^2 M^2\right)\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + \frac{1-\eta m_2 + \eta^2 M}{\eta m_2 - \eta^2 M}\frac{M^2\eta^2\Delta^2 d}{4} + \frac{1-\eta m_2+\eta^2 M}{1-2\eta m_2 + 2\eta^2 M^2}\left(2\eta b + 2\eta^2 G^2\right) \\ &\quad + \eta^2\frac{\Delta^2 d}{4} + 2\eta d.\end{aligned}$$

For any $k > 0$ we can bound the recursive equations as:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] &\leq \mathbb{E}\left[\|x_0\|^2\right] + \frac{1-\eta m_2+\eta^2 M^2}{\eta^2(m_2-\eta M^2)^2}\frac{M^2\eta^2\Delta^2 d}{4} + \frac{1-\eta m_2+\eta^2 M^2}{\eta(1-2\eta m_2+2\eta^2 M^2)(m_2-\eta M^2)}\left(2\eta b + 2\eta^2 G^2\right) \\ &\quad + \frac{1}{\eta(m_2-\eta M)}\left(\eta^2\frac{\Delta^2 d}{4} + 2\eta d\right) \\ &= \mathbb{E}\left[\|x_0\|^2\right] + \frac{1-\eta m_2+\eta^2 M^2}{\left(m_2-\eta M^2\right)^2}\frac{M^2\Delta^2 d}{4} + \frac{1-\eta m_2+\eta^2 M^2}{\left(1-2\eta m_2+2\eta^2 M^2\right)\left(m_2-\eta M^2\right)}\left(2b + 2\eta G^2\right) \\ &\quad + \frac{1}{m_2-\eta M^2}\left(\eta\frac{\Delta^2 d}{4} + 2d\right) \\ &\leq \mathbb{E}\left[\|x_0\|^2\right] + \frac{2M^2}{m_2}\frac{\Delta^2 d}{4} + \frac{2}{m_2}\left(2b+2\eta G^2\right) + \frac{2}{m_2}\left(\eta\frac{\Delta^2 d}{4} + 2d\right).\end{aligned}$$

Now if we let $\mathcal{E} = \mathbb{E}\left[\|x_0\|^2\right] + \frac{M}{m_2}\left(2b + 2\eta G^2 + 2d\right)$, then we can write:

$$\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \leq \mathcal{E} + \frac{2\left(M^2+1\right)}{m_2}\frac{\Delta^2 d}{4}.$$

□

E.5 Proof of Lemma 17

Proof. From the same analysis in (74), if we set

$$\eta \leq \min\left\{\frac{\gamma}{4\left(8Mu + u\gamma + 22\gamma^2\right)}, \sqrt{\frac{4u^2}{4Mu + 3\gamma^2}}, \frac{6\gamma b u}{\left(4Mu + 3\gamma^2\right)d}\right\},$$

we can obtain the following,

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] &\leq \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um_2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{(20u+\gamma)u\eta^2}{\gamma^2}\mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] \\ &\quad + \frac{2u^2\eta^2}{\gamma^2}\mathbb{E}\left[\|\nabla U(\mathbf{x}_k)\|^2\right] + \frac{8u\eta\left(\gamma^2+2u\right)}{\gamma^3}\mathbb{E}\left[\left\|\nabla U(\mathbf{x}_k) - Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right] + \frac{16(d+b)u\eta}{\gamma}.\end{aligned} \tag{79}$$

By assumption 1, we can bound $\mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k))\right\|^2\right]$ by the following,

$$\begin{aligned}\mathbb{E}\left[\|Q_G(\nabla U(\mathbf{x}_k))\|^2\right] &= \mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k)) - \nabla U(\mathbf{x}_k) + \nabla U(\mathbf{x}_k) - \nabla U(0) + \nabla U(0)\right\|^2\right] \\ &\leq \mathbb{E}\left[\left\|Q_G(\nabla\tilde{U}(\mathbf{x}_k)) - \nabla U(\mathbf{x}_k)\right\|^2\right] + 2\mathbb{E}\left[\|\nabla U(\mathbf{x}_k) - \nabla U(0)\|^2\right] + 2\mathbb{E}\left[\|\nabla U(0)\|^2\right] \\ &\leq \left(\frac{\Delta^2 d}{4} + \sigma^2\right) + 2M^2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2G^2.\end{aligned}$$

Plugging this bound into equation 79, we can have:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}\left(\mathbf{x}_{k+1}, \mathbf{v}_{k+1}\right)\right] \leq\ & \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um_2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{2(20u+\gamma)u\eta^2M^2}{\gamma^2}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \\
& + \frac{(20u+\gamma)u\eta^2}{\gamma^2}\left(\frac{\Delta^2 d}{4} + \sigma^2 + 2G^2\right) + \frac{2u^2\eta^2}{\gamma^2}\left(2M^2\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 2G^2\right) \\
& + \frac{8u\eta\left(\gamma^2+2u\right)}{\gamma^3}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + \frac{16\left(d+b\right)u\eta}{\gamma} \\
\leq\ & \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um_2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \\
& + \frac{(20u+\gamma)\gamma u\eta^2 + 8\left(\gamma^2+2u\right)u\eta}{\gamma^3}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + \frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}G^2 + \frac{16\left(d+b\right)u\eta}{\gamma} \\
\leq\ & \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{3um_2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] + \frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \\
& + \frac{\left(36u+9\gamma^2\right)u\eta}{\gamma^3}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + \frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}G^2 + \frac{16\left(d+b\right)u\eta}{\gamma}.
\end{aligned}$$

If we set the step size $\eta \leq \frac{\gamma m_2}{6(22u+\gamma)M^2}$, we can have:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}\left(\mathbf{x}_{k+1}, \mathbf{v}_{k+1}\right)\right] \leq\ & \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] - \frac{8um_2\eta}{3\gamma}\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] - \frac{2\eta}{\gamma}\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \\
& + \frac{\left(36u+9\gamma^2\right)u\eta}{\gamma^3}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + \frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}G^2 + \frac{16\left(d+b\right)u\eta}{\gamma}.
\end{aligned}$$

Again from the same analysis in (76), if $\gamma^2 \leq 4Mu$, we have

$$\mathcal{E}(x,v) \leq \frac{16uM}{\gamma^2}\|x\|^2 + \frac{12}{\gamma^2}\|v\|^2 + \frac{12uM}{\gamma^2}\|x^*\|^2.$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\mathbf{x}_{k+1}, \mathbf{v}_{k+1})\right] \leq\ & \left(1 - \frac{\gamma m_2\eta}{6M}\right)\mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] + \frac{\left(36u+9\gamma^2\right)u\eta}{\gamma^3}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) \\
& + \frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}G^2 + \frac{16\left(d+b\right)u\eta}{\gamma}.
\end{aligned}$$

Finally, we show that for any $k > 0$,

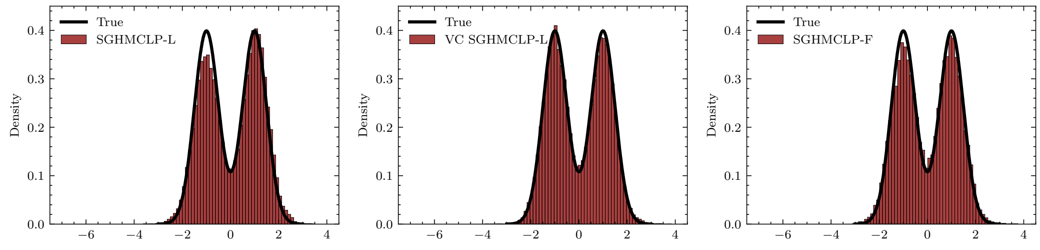
$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right] \leq\ & \mathbb{E}\left[\mathcal{E}(x_0, v_0)\right] + \frac{6M}{\gamma m_2\eta}\frac{\left(36u+9\gamma^2\right)u\eta}{\gamma^3}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) \\
& + \frac{6M}{\gamma m_2\eta}\frac{2(22u+\gamma)u\eta^2M^2}{\gamma^2}G^2 + \frac{6M}{\gamma m_2\eta}\frac{16\left(d+b\right)u\eta}{\gamma} \\
\leq\ & \mathbb{E}\left[\mathcal{E}(x_0, v_0)\right] + \frac{54\left(4u+\gamma^2\right)u}{m_2\gamma^4}\left(\frac{\Delta^2 d}{4} + \sigma^2\right) + \frac{12(22u+\gamma)uM^3}{m_2\gamma^3}G^2 + \frac{96\left(d+b\right)uM}{m_2\gamma^2} \\
=:\ & \mathcal{E} + C\Delta^2 d.
\end{aligned}$$

Finally by the fact that $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] \leq \mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right]$ and $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \leq \gamma^2\mathbb{E}\left[\mathcal{E}(\mathbf{x}_k, \mathbf{v}_k)\right]/2$ we can get our claim in Lemma 17.

□

F Additional Experiment Results

In this section, we provide additional experiment results.



(a) (b) (c)

Figure 5: Low-precision SGHMC with stepsize equal to 0.01 on a Gaussian mixture distribution. (a): SGHMCLP-L. (b): VC SGHMCLP-L. (c): SGHMCLP-F.

F.1 Sampling from Gaussian mixture Distribution

We first demonstrate the performance of Low-precision SGHMC for fitting a strongly log-concave distribution. In this case, we use the standard Gaussian distribution as the representative of the strongly log-concave distribution. The simulation result is shown in Figure 1. As in the Figure 1 and 5 displayed, the sample obtained from naïve SGHMCLP-L has a larger variance than the target distribution. This verifies the results we prove in Theorem 6 and 7. This is because in addition to the Gaussian noise the naïve quantizer in order to be unbiased introduces an extra noise which increases the variance of the sample. The variance corrected quantizer solves this problem by quantizing the mean of each sample and letting the variance of the quantizer equal to the variance $\mathrm{Var}_{\mathbf{x}}^{hmc}$ defined by the Hamiltonian dynamics 9. The variance-corrected SGHMC with low-precision gradient accumulators (VC SGHMCLP-L) doesn't suffer from the larger variance problem as the variance corrected quantization matches the variance defined in (2).

We also study in which case the variance corrected quantization function is advantageous over the naïve stochastic quantization function. We test the 2-Wasserstein distance of VC SGHMCLP-L and SGHMCLP-L over different variances. The result is shown in Figure 4. We found that when the variance $\mathrm{Var}_{\mathbf{x}}^{hmc}$ is close to the largest quantization variance $\Delta^2/4$, the variance corrected quantization function shows the largest advantage over the naïve quantization. When the variance $\mathrm{Var}_{\mathbf{x}}^{hmc}$ is less than $\Delta^2/4$ the correction has a chance to fail and when it is 100 times the quantization variance, the advantage of variance corrected quantization shows less advantage. One possible reason is the quantization noise eliminated by variance corrected quantization function is not critical compared with the intrinsic variance needed.

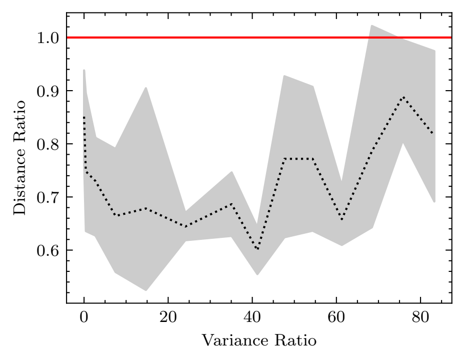


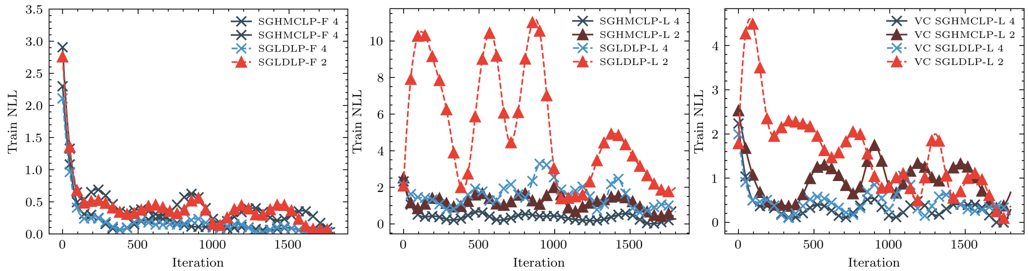
Figure 4: Wasserstein Distance Ratio of VC SGHMCLP-L & SGHMCLP-L (Smaller is better). The dashed line is the 2-Wasserstein distance to the target distribution ratio between the sample obtained by VC SGHMCLP-L and SGHMCLP-L.

F.2 Multi-layer perception

We present the low-precision SGHMC with MLP on the MNIST dataset in Figure 6. We observe similar results as the low-precision SGHMC with the logistic model.

F.3 CIFAR-10 & CIFAR-100

In this section, we present some additional results for experiments on computer vision tasks in CIFAR datasets.



(a) (b) (c)

Figure 6: Training NLL of low-precision SGHMC and SGLD on MLP with MNIST in terms of different numbers of fractional bits. (a): Methods with Full-Precision Gradients Accumulators. (b): Methods with Low-Precision Gradients Accumulators. (c): Variance corrected quantization. The low-precision SGHMC adopted with full-precision gradient accumulators achieves comparable results with SGLD. However, when adopted with low-precision gradient accumulators and variance-corrected quantization SGHMC shows more robustness to quantization error especially when the number of representable bits is low.

Table 4: Test errors (%) of Low-precision gradient accumulators on CIFAR with ResNet-18.

	CIFAR-10	CIFAR-100
32-bit Float		
SGD	4.73 ± 0.10	**22.34 ± 0.22**
SGLD	**4.52 ± 0.07**	22.40 ± 0.04
SGHMC	4.78 ± 0.08	22.37 ± 0.04
8-bit Fixed Point		
SGD	8.50 ± 0.22	28.42 ± 0.35
SGLD	7.81 ± 0.07	27.15 ± 0.35
VC SGLD	7.03 ±0.23	26.73 ±0.12
SGHMC	6.63 ± 0.10	26.57 ± 0.10
VC SGHMC	**6.60 ± 0.06**	**26.43 ± 0.19**
8-bit Block Float Point		
SGD	5.86 ±0.18	26.75 ±0.11
SGLD	5.75 ±0.05	26.11±0.38
VC SGLD	5.51 ±0.01	25.14 ±0.11
SGHMC	5.38 ±0.06	25.29 ±0.03
VC SGHMC	**5.15 ±0.08**	**24.45 ±0.16**