Breaking the $\sqrt{d}$ Communication Barrier in Federated Sampling with Adaptive Hamiltonian Monte Carlo

Jiajun Liang
Independent Researcher
liangjiajun0529@gmail.com

Linxuan Wang
Purdue University
wang6559@purdue.edu

Guang Lin
Purdue University
guanglin@purdue.edu

Qifan Song
Purdue University
qfsong@purdue.edu

Abstract

This work proposes the Adaptive Federated Hamiltonian Monte Carlo (AFHMC) for Bayesian federated learning (FL) tasks. The prior work on HMC applications for FL achieves $O(\sqrt{d}/\epsilon)$ communication cost under log-convex distributions, already demonstrating its advantage over the Langevin dynamic-based counterpart. Based on a refined analysis, this paper further reveals the benefit of HMC for the FL regime by incorporating a control mechanism for local node shifts. We prove that AFHMC at most requires $O(\{d/\epsilon^2\}^{1/3})$ communication cost up to a logarithmic term, which is significantly better than the existing results. Moreover, the proposed method adapts to client heterogeneity. Under low-heterogeneity settings (i.e., the target sampling precision ϵ^2 is higher than the heterogeneity level), the communication rate of AFHMC can be as good as $O(\log(d/\epsilon^2))$, resembling the communication costs for FL optimization tools. To the best of our knowledge, this provides the first logarithmic communication complexity result for federated sampling, even though it requires low heterogeneity. Numerically, we also justify AFHMC's advantage under the low-heterogeneity regime. Our code is available at `https://github.com/laaaarrywang/Adaptive-Federated-HMC`.

1 Introduction

Federated learning (FL) provides a collaborative training paradigm in which multiple clients jointly learn a global model without sharing their raw data, thereby preserving data privacy while leveraging distributed data resources (McMahan et al., 2017; Bonawitz et al., 2019; Li et al., 2020a). A typical FL algorithm involves iterative rounds of local training followed by aggregating updates via node-to-center communications. When FL is applied to real-time systems, e.g., low-earth orbit (LEO) satellite constellations (Chen et al., 2022), it is critical that the FL algorithm design aligns well with the system constraints, such as limited connection bandwidth/windows and high communication latency. Therefore, a successful FL framework must be communication-efficient and tolerate low aggregation frequency. Despite its empirical success in large-scale applications, most existing FL algorithms (e.g., Konečnỳ et al., 2015, 2016; Li et al., 2020b; Wang et al., 2020), are designed within a frequentist framework and focus on point estimation, which makes it difficult to quantify uncertainty in model predictions and parameters. Bayesian federated learning (BFL) addresses this limitation by inferring the full posterior, thus enabling principled uncertainty quantification and robust decision-making under data heterogeneity and limited local samples. However, scalable Bayesian FL inference remains challenging due to multiple factors: (i) high-dimensional parameter spaces, (ii) heterogeneous and potentially non-IID client data, and (iii) strict communication and privacy constraints.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Hamiltonian Monte Carlo (HMC, Duane et al. (1987); Neal (2012)) is a MCMC method for sampling from complex posteriors, which exploits gradient information and an auxiliary momentum variable to construct proposals that can efficiently explore high-dimensional distributions. Compared with first-order methods such as stochastic gradient Langevin dynamics (SGLD, Welling & Teh (2011); Sato & Nakagawa (2014)), HMC and its stochastic variants often achieve superior mixing and lower asymptotic variance but its second-order nature also makes theoretical analysis and distributed implementation more involved. Extending HMC to the federated setting is nontrivial, because naive combinations of stochastic HMC with FedAvg-style aggregation (McMahan et al., 2017) can easily suffer from bias accumulation, instability of global dynamics, and inefficient communication.

To address these challenges, recent work proposes the Federated Averaging stochastic Hamiltonian Monte Carlo (FA-HMC, Liang et al. (2025)) algorithm, which integrates stochastic HMC updates on local clients with periodic aggregation of both positions and momenta at a central server. Liang et al. (2025) showed that FA-HMC requires $O(\sqrt{d}/\epsilon)$ communication cost (equivalently, $O(\sqrt{d}/\epsilon)$ updates occur in the center node) to achieve ϵ Wasserstein sampling error. In contrast, under non-distributed settings, HMC achieves a tight $O(\{d/\epsilon^2\}^{1/4})$ gradient computation cost (Mangoubi & Vishnoi, 2018). It naturally leads to the question: can we further close the gap between the communication cost in FL paradigm and the computation cost in the non-FL paradigm, in terms of the dimension dependency?

This work introduces an adaptive variant, Adaptive Federated Hamiltonian Monte Carlo (AFHMC), which incorporates surrogate leapfrog steps and an additional global shift vector to better align local and global dynamics. We conduct refine analysis which explicitly characterizes the discrepancy between surrogate and exact leapfrog dynamics, and establish non-asymptotic convergence guarantees under standard smoothness and strong convexity assumptions. AFHMC can achieve improved dependence on the number of clients, communication rounds, and stochastic gradient noise level, and our experiments demonstrate competitive or superior performance compared with existing Bayesian federated baselines.

2 Related Works and Our Contributions

Federated optimization Early studies, such as FedAvg (McMahan et al., 2017), established convergence guarantees for smooth and strongly convex objectives, typically achieving sublinear rates in the presence of data heterogeneity. To address this challenge, subsequent work introduced variance reduction and control variate techniques, such as SCAFFOLD (Karimireddy et al., 2020) and its variants, which improve convergence by mitigating client drift and reducing bias induced by heterogeneous local updates. In particular, Mitra et al. (2021) provided the first result establishing linear convergence rates in terms of communication complexity using control variates, while Mishchenko et al. (2022) further advanced this line of research by proposing randomized algorithms that achieve constant communication complexity under suitable conditions.

Bayesian federated learning. Bayesian extensions of federated learning aim to characterize posterior uncertainty over model parameters or predictions while respecting client-level privacy and communication constraints. Existing approaches include variational Bayes formulations (e.g., Zhang et al., 2022; Guo et al., 2023), hierarchical Bayesian models for heterogeneous clients (e.g., Vedadi et al., 2024; Kim & Hospedales, 2025), and Bayesian variants of federated averaging where each client maintains local posterior approximations that are periodically aggregated (e.g., Al-Shedivat et al., 2021; Chen & Chao, 2021; Liu et al., 2024). These methods often rely on factorized approximations or conjugate priors to keep the communication and computation overhead manageable.Our work instead targets exact (or asymptotically exact) MCMC-based Bayesian inference in federated environments, and develops HMC-based algorithms that can operate under realistic communication budgets.

HMC and stochastic variants. HMC is an efficient sampler for complex high-dimensional Bayesian inference, with rigorous guarantees on convergence and mixing under smoothness and curvature conditions of the target distribution (Durmus et al., 2020; Mangoubi & Smith, 2017, 2021). To scale HMC to large datasets, stochastic gradient HMC (SGHMC Chen et al. (2014)) utilizes mini-batch estimates, typically introducing extra friction or correction terms to control the additional stochastic noise. Subsequent works have investigated the convergence behavior of SGHMC (Zou & Gu, 2021), variance-reduction techniques (Zou et al., 2019), low-precision sampling techniques (Wang et al., 2024) and etc. These developments, however, are primarily designed for centralized

settings, and their direct application to federated learning is hindered by client heterogeneity and limited access to the full dataset.

Adaptive and geometry-aware HMC. A line of research has explored adaptive HMC schemes that adjust hyperparameters such as step sizes, mass matrices, and integrator structures using information gathered during sampling. Examples include Hoffman & Gelman (2014); Hoffman et al. (2021); Hirt et al. (2021); Fernández-Pendás et al. (2016); Girolami & Calderhead (2011), all aiming to enhance stability and sampling efficiency in complex geometries. More recently, adaptive tuning strategies have been proposed to automatically select splitting integrators and hyperparameter ranges for (generalized) HMC, reducing reliance on heuristic tuning in practice (Akhmatskaya et al., 2026; Nagar et al., 2024). Our AFHMC algorithm is conceptually related to these works in that it uses a surrogate Hamiltonian flow informed by global information, but it is tailored to the federated setting, where the adaptation must be communication-efficient and robust to client heterogeneity.

Beyond existing results, the main contributions of this paper are summarized as follows:

1. We propose AFHMC, an adaptive HMC algorithm for FL tasks based on surrogate leapfrog updates. The new algorithm introduces a global drift control variate to address the heterogeneity among local gradient update trajectories.

2. We introduce a principled framework to analyze divergence errors induced by approximate Hamiltonian dynamics.

3. Under strong convexity, we derive non-asymptotic error bounds that quantify the interplay between the surrogate leapfrog integrator and data heterogeneity. In particular, we prove that AFHMC achieves $O(\{d/\epsilon^2\}^{1/3})$ communication cost under extreme heterogeneity, and $O(\log(d/\epsilon^2))$ communication cost under low heterogeneity. To the best of our knowledge, both rates of communication complexity have never been achieved by prior federated sampling algorithms. Not only do we obtain a better communication rate, but we also don't rely on stringent Hessian smoothness conditions that are commonly required to achieve an advanced rate (e.g., Liang et al., 2025; Mangoubi & Vishnoi, 2018; Plassier et al., 2023). Please refer to Table 1 for a comprehensive comparison against prior theoretical works. Note that if we indeed enforce an extra Hessian smoothness condition, we could technically obtain even better results.

4. We empirically demonstrate that the proposed method achieves improved computational efficiency and stability in distributed and approximate-gradient settings.

3 Motivation and Algorithm

The goal of a standard sampling problem is to generate approximate samples from the target posterior $\pi(\theta) \propto \exp(-f(\theta))$ where f is the potential function. In the standard leapfrog HMC, as illustrated in Figure 1(left), the parameter θ is evolved using a randomly sampled Gaussian momentum for K leapfrog steps (defined later in (2)), after which the momentum is resampled before starting a new sequence of leapfrog updates. In Mangoubi & Vishnoi (2018), it is shown that performing $O(\log d)$ such leapfrog trajectories, each consisting of $O(\{d/\epsilon^2\}^{1/4})$ steps, suffices to ensure convergence of the task distribution. Therefore, the total gradient computational cost of the overall procedure is dominated by the number of leapfrog steps, leading to a total complexity of order $O(d^{1/4})$ up to logarithmic factors.

Let $[N] = \{1, \dots, N\}$ index the clients. In federated settings, each client c has a local potential $f^{(c)}(\theta)$ with weight $w_c \geq 0$ and thus global potential $f(\theta) = \sum_{c=1}^{N} w_c f^{(c)}(\theta)$, where $\sum_{c=1}^{N} w_c = 1$. To achieve logarithmic communication complexity in federated sampling, the above theoretical insight motivates us to integrate communication steps into the alternating leapfrog procedure, expecting that the order of FL communication cost matches $O(\log d)$ order of leapfrog trajectories in centralized HMC. To be concrete, as illustrated in Figure 1 (right), we aim that each local node performs K leapfrog steps followed by a communication round (e.g., averaging the local parameters across nodes), before proceeding to the next leapfrog phase. This corresponds to the FA-HMC algorithm (Liang et al., 2025) with $T = 1$. Unfortunately, Liang et al. (2025) establishes a lower bound showing that, under high federated heterogeneity, the communication cost of such an HMC scheme is at least on the order of $O(\{d/\epsilon^2\}^{1/2})$. Thus, additional non-trivial FL training designs are necessary to reach a better rate.

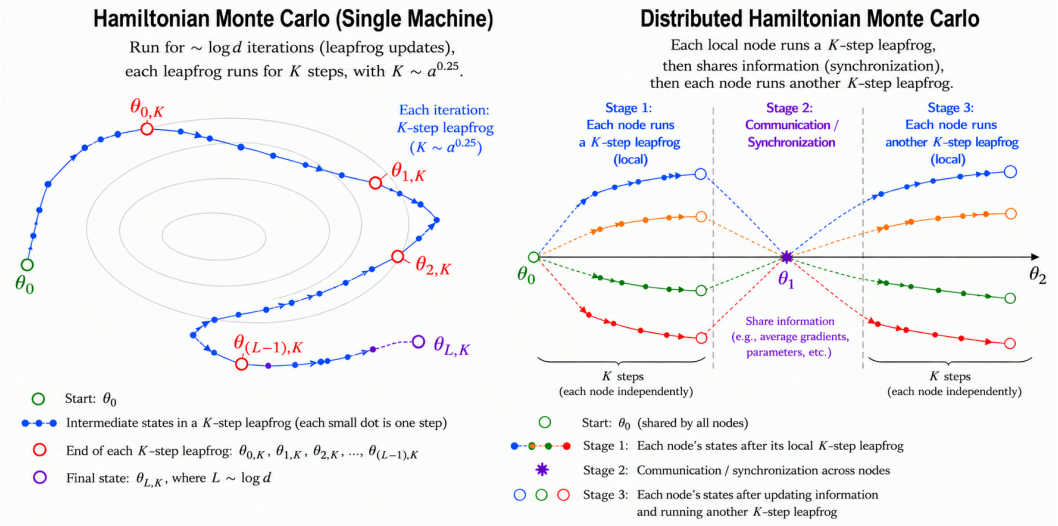

Figure 1: An illustration of the centralized HMC algorithm (left panel) and the federated HMC algorithm (right panel).

In any FL optimization tasks, a key challenge to achieve efficient communication is to control the local shift caused by heterogeneity across local nodes. Local gradient updates move the local parameter toward the location optimum, resulting in a global bias after aggregating local solutions. A breakthrough was introduced by Karimireddy et al. (2020), which introduced a control shift technique that reduces the communication cost of FedAvg (McMahan et al., 2017) from $O(d/\epsilon^2)$ to $O(\log(d/\epsilon^2))$. In the FL sampling regime, FA-LD (Deng et al., 2024), a Langevin-counterpart of FedAvg also incurs a communication cost of $O(d/\epsilon^2)$. Plassier et al. (2023) further improved the sampling algorithm by adapting the control shift technique of (Karimireddy et al., 2020).

Inspired by existing works on local shift control, we propose an adaptive HMC federated update rule by introducing a drift variate. Given an energy function F, the HMC dynamics follow the Hamiltonian equations:

$$d\theta(t) = p(t)dt, \qquad dp(t) = -\nabla_\theta F(\theta(t))dt. \tag{1}$$

The above continuous-time dynamic cannot be solved analytically. Commonly, one uses the leapfrog integrator to discretize the dynamics over a step size η by

$$\theta_{k+1} = \theta_k + \eta p_k - \frac{\eta^2}{2}\nabla F(\theta_k), \qquad p_{k+1} = p_k - \frac{\eta}{2}\nabla F(\theta_k) - \frac{\eta}{2}\nabla F(\theta_{k+1}). \tag{2}$$

Under the FL setting, update algorithm (2) is running on local nodes, and F represents a local energy function which can differ substantially from the global energy function f due to node heterogeneity. To account for this potential gap between local F and global f, we inject a bias correction term to the gradient and replace all $\nabla F(\cdot)$ terms in (2) with $\nabla F(\cdot) - \nabla F(\theta_0) + \nabla f(\theta_0)$ (see, Algorithm 1). This bias correction term $\nabla f(\theta_0) - \nabla F(\theta_0)$ addresses the discrepancy between local and global losses, and the drift variate $\nabla f(\theta_0)$ needs to be broadcast to all local nodes. This novel local update rule is named the surrogate leapfrog update.

The AFHMC algorithm consists of the following iterative steps: 1. After synchronizing parameter θ_t across all nodes, each node c shares the local $\nabla f^{(c)}(\theta_t)$ to obtain a global drift variate $\nabla f(\theta_t)$. 2. All local nodes run K surrogate leapfrog steps, with the shared global drift variate, initialized from the same position and momentum. 3. All nodes share and synchronize the local parameter θ_{t+1}, after K leapfrog steps. The algorithm can be found in Algorithm 2. In contrast to FA-HMC (Liang et al., 2025), which runs multiple local HMC steps where each HMC step consists of K leapfrog updates, the proposed AFHMC only runs one single local HMC step (i.e., K updates of surrogate leapfrog, see Figure 1). As discussed previously, this design is to ensure minimal communication cost, because every local HMC step will need a fresh new drift control variate, which requires communication.

The AFHMC algorithm can be easily implemented with stochastic gradients for acceleration purposes, similarly to Liang et al. (2025), but our analysis in this paper focuses on the use of the exact gradient.

Algorithm 1 Surrogate leapfrog approximation $\widetilde{h}_{\mathrm{LF}}(\cdot)$

Input: energy function $F(\cdot)$, initial position θ_0, drift vector $\nabla f(\theta_0)$, momentum p_0, stepsize η, leapfrog step K
for $k = 0$ **to** $K - 1$ **do**
$\theta_{k+1} = \theta_k + \eta p_k - \frac{\eta^2}{2}(\nabla F(\theta_k) - \nabla F(\theta_0) + \nabla f(\theta_0))$
$p_{k+1} = p_k - \frac{\eta}{2}\nabla F(\theta_k) - \frac{\eta}{2}\nabla F(\theta_{k+1}) + \eta(\nabla F(\theta_0) - \nabla f(\theta_0))$
end for
Output: $\widetilde{h}_{\mathrm{LF}}(F, \theta_0, \nabla f(\theta_0), p_0, \eta, K) = \theta_K$

Algorithm 2 Adaptive federated hamiltonian monte carlo (AFHMC) algorithm

Input: initial position θ_0, stepsize η; leapfrog update step K
Each local node receives the initial parameter θ_0 from the center
for $t = 0$ **to** $t_{\max}$ **do**
$1st$ **broadcast:** The c-th local node sends $\nabla f^{(c)}(\theta_t)$ to the center, for all $c = 1, \ldots, N$;
$1st$ **communicate:** The center node samples Gaussian momentum p^t, and sends p_t and drift vector $\nabla f(\theta_t) := \sum_{c=1}^N w_c \nabla f^{(c)}(\theta_t)$ to all local nodes;
for $c = 1, 2 \ldots, N$ **do**
Initialize $\theta_{t,0}^{(c)} = \theta_t$, $p_{t,0}^{(c)} = p_t$;
update $\theta_{t,K}^{(c)} = \widetilde{h}_{\mathrm{LF}}(f^{(c)}, \theta_{t,0}^{(c)}, \nabla f(\theta_t), p_{t,0}^{(c)}, \eta, K)$;
end for
$2nd$ **broadcast:** The c-th local node sends $\theta_{t,K}^{(c)}$ to the center for all $c = 1, \ldots, N$;
$2nd$ **communicate:** The center node sends back $\theta_{t+1} := \sum_{c=1}^N w_c \theta_{t,K}^{(c)}$ to all local nodes;
end for
Output: $\{\theta_t : t = 0, \ldots, t_{\max}\}$

4 Convergence Analysis

Let $z_i^c, 1 \le i \le n_c, 1 \le c \le N$ be the local data at the c-th node, $\theta \in \mathbb{R}^d$ and $\ell(\theta; z_i^c)$ be the negative log-likelihood (or loss) function. Define $w_c = n_c / \sum_{c=1}^N n_c$, and $f^{(c)}(\theta) := \sum_{i=1}^{n_c} \ell(\theta; z_i^c)/w_c$ is the local loss function of accessible to the c-th local node. The goal is to simulate the global target distribution $\pi(\theta) \propto \exp(-f(\theta))$, where $f(\theta) = \sum_{c=1}^N w_c f^{(c)}(\theta)$, $w_c \ge 0$ with $\sum_c w_c = 1$.

This section will investigate the theoretical properties of AFHMC algorithm. Before stating our main results, we specific the model assumption as follows:

Assumption 4.1 (m-Strongly Convex)**.** For each $c = 1, 2, \ldots, N$, $f^{(c)}$ is m-strongly convex for some $\mu > 0$, i.e., $\forall x, y \in \mathbb{R}^d$ $f^{(c)}(y) \ge f^{(c)}(x) + \langle \nabla f^{(c)}(x), y - x \rangle + \frac{m}{2}\|y - x\|_2^2$.

Assumption 4.2 (M-Smoothness)**.** For each $c = 1, 2, \ldots, N$, $f^{(c)}$ is M-smooth for some $M > 0$, i.e., $\forall x, y \in \mathbb{R}^d$, we have $\|\nabla f^{(c)}(y) - \nabla f^{(c)}(x)\| \le M\|x - y\|$.

The assumptions are fairly standard in the theoretical literature of gradient-based MCMC algorithms (e.g., Dalalyan, 2017; Mangoubi & Vishnoi, 2018; Dalalyan & Karagulyan, 2019; Erdogdu & Hosseinzadeh, 2021). Unlike prior advanced results (e.g., Dalalyan & Karagulyan, 2019; Liang et al., 2025; Plassier et al., 2023; Mangoubi & Vishnoi, 2018), we don't require any second-order smoothness (e.g., Hessian) of the energy functions.

Our first result provides a non-asymptotic bound, studying how the gap between θ_t and θ_t^π contracts, where θ_t^π is a coupled stationary continuous HMC process satisfying $\theta_t^\pi \sim \pi$ for all $t \ge 0$ (refer to Appendix Section A.1 for formal definitions).

Theorem 4.3 (Convergence)**.** *Under Assumptions 4.1-4.2, there exist sufficient large constants C_0, C_1, C_2 such that $\{\theta_t\}_t$ satisfies that, for $K\eta \le \sqrt{m}/(C_0 M) \le \sqrt{m}/(2M)$,*

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^\pi\|^2 \le \left(1 - \frac{m(K\eta)^2}{16}\right)\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 + C_1 K^2 \eta^4 \frac{M^2}{m} d + C_2 (K\eta)^8 \frac{M^2}{m} H$$

where $H := \max_{i\le K}\sum_{c=1}^{N} w_c \mathbb{E}\|\frac{1}{i\eta}(\nabla f^{(c)}(\bar{\mathcal{H}}_\eta^i(\theta^\pi)) - \nabla f^{(c)}(\theta^\pi) + \nabla f(\theta^\pi) - \nabla f(\bar{\mathcal{H}}_\eta^i(\theta^\pi))\|^2$ *is a metric quantifying the federated heterogeneity.*

The $\bar{\mathcal{H}}_\eta^i(\theta)$ (formally defined in Appendix Section A.1) denotes the output by running $i\eta$-length ideal continuous HMC dynamic (i.e., (4) with F being replaced by f) initialized from θ.

Remark 4.4. H is a metric only related to the target distribution and local distribution. It characterizes the client-level gradient heterogeneity under the target distribution. H is a theoretical quantity usually unknown in practice. But in terms of its dependency on the dimension, due to the nature of the squared L_2 norm, one can prove that $H = O(d)$. The proof of this bound and examples of calculating the exact H are given in Appendix Section C.

Sketch of proof We decompose $\|\theta_{t+1} - \theta_{t+1}^\pi\|^2$ into two terms: $\|\bar{\mathcal{H}}_\eta^K(\theta_t) - \theta_{t+1}^\pi\|^2$ and $\|\theta_{t+1} - \bar{\mathcal{H}}_\eta^K(\theta_t)\|^2$. The first term is subject to the classic contraction theorem of ideal HMC dynamics. The second term can be characterized by (i) the divergence between $\theta_{t,k}^{(c)}$ and $\sum_{c=1}^N w_c\theta_{t,k}^{(c)}$, and (ii) the divergence between ideal HMC dynamics $\bar{\mathcal{H}}_\eta^i(\theta)$ and the discretized HMC dynamics $\mathcal{H}_\eta^i(\theta)$. Thanks to the local shift control mechanism, we establish a new technical bound (Lemma B.6)) to study the surrogate error and give $\mathbb{E}\|\theta_{t,k}^{(c)} - \sum_{c=1}^N w_c\theta_{t,k}\|^2 \approx O((K\eta)^6 H)$, and bound the divergence caused by discretizations by $O(\eta^4 d)$ (via Lemma B.10).

Based on the recursive contraction result in Theorem 4.3, one can easily derive the convergence rate for $\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^\pi\|^2$. Let π_t be the marginal distribution of θ_t and $\mathcal{W}_2(\mu,\nu)$ denote the Wassertein-2 distance between probability measures μ and ν. Then, by definition, $\mathcal{W}_2^2(\pi_t,\pi) = \inf_{\theta\sim\pi}\mathbb{E}\|\theta_t - \theta\|^2 \le \mathbb{E}\|\theta_t - \theta_t^\pi\|^2$, leading to

$$\mathcal{W}_2^2(\pi_t,\pi) \le \left(1 - \frac{m(K\eta)^2}{16}\right)^t \mathcal{W}_2^2(\pi_0,\pi) + C_1'\eta^2\frac{M^2 d}{m^2} + C_2' K^6\eta^6\frac{M^2 H}{m^2}.$$

This leads to our next theorem, which studies the communication cost rate for AFHMC.

Theorem 4.5. *Under Assumptions 4.1-4.2 and if* $H \gg M^4\epsilon^2/m > 0$*, then there exists some small absolute constant* C *such that if we choose*

$$K\eta = C\left(\frac{\epsilon^2 m^2}{HM^2}\right)^{1/6} \ll C\frac{\sqrt{m}}{M} \quad \textit{and} \quad \eta = C\left(\frac{\epsilon^2 m^2}{M^2 d}\right)^{1/2},$$

$\mathcal{W}_2(\pi_{t_\epsilon},\pi) \le \epsilon$ *can be attained with communication times*

$$t_\epsilon \asymp \left(\frac{HM^2}{\epsilon^2 m^5}\right)^{1/3}\log(\mathcal{W}_2^2(\pi_0,\pi)/\epsilon^2).$$

The corresponding local iteration number is

$$K \asymp \frac{M^{2/3}d^{1/2}}{\epsilon^{2/3}m^{2/3}H^{1/6}}.$$

Moreover, if $H \ll M^4\epsilon^2/m$ *(i.e., heterogeneity across local nodes is very small), it only requires*

$$t_\epsilon \asymp \log(\mathcal{W}_2^2(\pi_0,\pi)/\epsilon^2),$$

achieving a linear rate with respect to the communication cost.

For a heterogeneous situation, the communication cost t_ϵ implicitly depends on the model dimension via the quantity H. As discussed in Remark 4.4, H can scale linearly with d, leading to $t_\epsilon \asymp (d/\epsilon^2)^{1/3}\log(d/\epsilon^2)$. For a homogeneous situation where $H \approx 0$, the communication cost reduces to $O(\log(d/\epsilon^2))$. To the best of our knowledge, both rates have never been achieved by prior Bayesian FL methods in the literature. Moreover, our analysis do not require extra stringent Hessian smoothness conditions that are commonly required to achieve an advanced rate e.g., Assumption A3 in Plassier et al. (2023), Assumption 3.3 in Liang et al. (2025). We compare all these results in Table 1.

This theorem can also be interpreted from a different perspective. AFHMC can have two training phases: in the first phase (fast convergence), the sampling error linearly decays until $\mathcal{W}_2^2(\pi_t,\pi) \sim mH/M^4$; in the second phase (slow convergence), it needs $O((d/\epsilon^2)^{1/3}\log(d/\epsilon^2))$ communications to reach $\mathcal{W}_2^2(\pi_t,\pi) = \epsilon^2 = o(mH/M^4)$. Note that these two phases do require different tuning of K and η, as shown by the proof of Theorem 4.5.

Algorithm	Condition beyond Assumptions 4.1-4.2	Rate of communication	Rate of gradients (up to logarithmic terms)
Centralized HMC Mangoubi & Vishnoi (2018)	Hessian Lipschitz under infinity norm (Assum. 1 in their paper)	none	$(d/\epsilon^2)^{1/4}$
VR-FALD Plassier et al. (2023)	Hessian Lipschitz under L2 norm (Assum. HX1 in their paper)	$(d/\epsilon^2)^{1/2}$	$(d/\epsilon^2)^{1/2}$
FA-HMC Liang et al. (2025)	Hessian Lipschitz under L2 norm (Assum. 3 in their paper)	$(d/\epsilon^2)^{1/2}$	$(d/\epsilon^2)^{1/2}$
AFHMC (this paper)	none	$\log(d/\epsilon^2) \sim (d/\epsilon^2)^{1/3}$	$(d/\epsilon^2)^{1/2} \sim (d/\epsilon^2)^{2/3}$

Table 1: Comparison of Bayesian theoretical results: requires assumptions, rate of communication and rate of gradient computation to achieve ϵ Wasserstein convergence

Remark 4.6. Our main theory is derived based on using the exact gradient updates. While it is of great interest to investigate the theoretical behavior when stochastic gradient is used instead, it poses great challenges to obtain a sharp rate that surpasses the Langevin counterpart. Such a result doesn't exist so far, even for the centralized HMC algorithm. With a lack of theoretical guarantee, we still believe in the advantage of AFHMC under noisy gradient, which is empirically justified by our numerical studies.

5 Numerical Evidences

In this section, we present our numerical results, including both simulation studies and real data applications, to support our theoretical claims. We compare the empirical performance of AFHMC, FA-HMC and FA-LD, and demonstrate the superiority of the proposed AFHMC algorithm. Throughout this section, a *communication round* refers to one outer iteration of the federated algorithm. Each AFHMC round consists of two server exchanges (the gradient aggregation for the drift correction and the parameter aggregation after the surrogate trajectory; see the two broadcasts in Algorithm 2), whereas each FA-HMC or FA-LD round consists of one. This constant factor does not change the communication complexity rates. For an empirical comparison at an equal number of server exchanges, we compare AFHMC at round R with FA-HMC at round $2R$.

Simulated Bayesian logistic regression. Following Liang et al. (2025); Mangoubi & Vishnoi (2018), we consider Bayesian logistic regression with a standard Gaussian prior on $\theta \in \mathbb{R}^d$. Given feature vectors $x_i \in \mathbb{R}^d$ and binary labels $y_i \in \{-1, +1\}$, the posterior is

$$\pi(\theta) \propto \exp\left(-\frac{1}{2}\|\theta\|^2\right) \prod_{i=1}^{n} \sigma(y_i x_i^\top \theta),$$

where $\sigma(\cdot)$ denotes the sigmoid function. The $n = 1000$ observations are split evenly across $N = 20$ clients. We evaluate the samplers in two settings: (1) a high-dimensional case with $d = 1000$ and exact local gradients; (2) a lower-dimensional case with $d = 10$ with inaccurate gradients. Specifically, we add $\mathcal{N}(0, 100 I_d)$ noise to the exact gradients, simulating the stochastic mini-batching.

The algorithms (FA-LD, FA-HMC, and AFHMC) are implemented with the following tuning parameter settings. FA-LD uses $T = 10$ local Langevin steps (each Langevin step is equivalent to $K = 1$ leapfrog step). FA-HMC uses $T = 10$ local HMC steps, and each HMC step consists of K_F classic leapfrog steps. We sweep K_{F} over a grid centered around the trajectory-length rule $K_{\mathrm{F}}\eta \approx \pi/3$ (Mangoubi & Vishnoi, 2018). AFHMC uses K_{A} local surrogate leapfrog steps, and we consider two candidates of K_{A}: (i) $K_{\mathrm{A}} = K_{\mathrm{F}}$ such that AFHMC and FA-HMC match the HMC dynamic trajectory length, and (ii) $K_{\mathrm{A}} = 10K_{\mathrm{F}}$ such that AFHMC and FA-HMC match the same number of leapfrog steps between consecutive rounds of communication. For different η, the K_{A} and K_{F} values that deliver the best performance are chosen. For each $(\text{algorithm}, \eta, K)$ configuration, we run $M = 20$ independent chains with a total leapfrog budget of 2×10^6 and discard the first half of each chain as burn-in. We measure performance using the marginal error

$$\mathrm{ME} \;=\; \frac{1}{d}\sum_{i=1}^{d} \mathcal{W}_1\big(\hat{\mu}_i,\, \mu_i^{\mathrm{MHMC}}\big)\,, \tag{3}$$

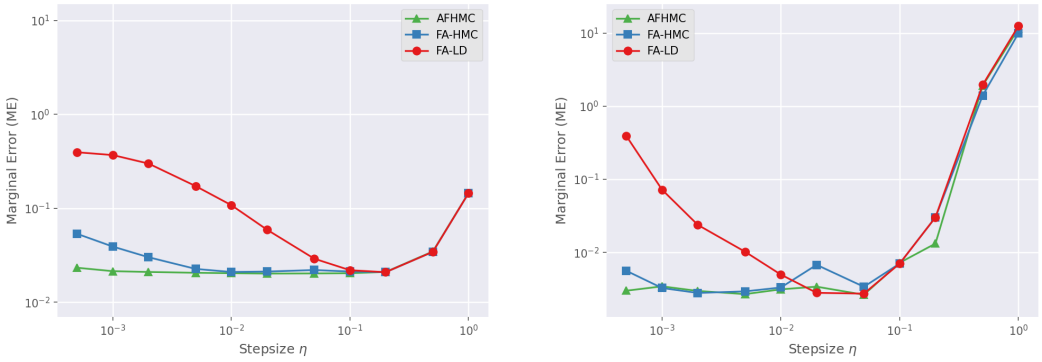

Figure 2: Simulated Bayesian logistic regression. Left: $d = 1000$ with exact gradients. Right: $d = 10$ with coordinate-wise $\mathcal{N}(0, 100)$ gradient noise. Each panel reports marginal error relative to a centralized MHMC reference chain under a fixed total leapfrog budget. AFHMC is more stable across stepsizes and improves the small-stepsize behavior of FA-HMC and FA-LD.

where $\hat{\mu}_i$ is the empirical marginal of the i-th coordinate produced by the candidate sampler, μ_i^{MHMC} is the surrogate true posterior margins, approximated by a centralized Metropolis-adjusted HMC reference chain with 10^7 steps, and $\mathcal{W}_1$ denotes the Wasserstein-1 distance. For each $(\text{algorithm}, \eta)$ pair, we report the best-performing K from the candidate grid; the full grid is provided in Appendix D.1.

Figure 2 reports the resulting marginal error curves for the exact-gradient and noisy-gradient settings. In the exact-gradient high-dimensional setting, AFHMC maintains low marginal error over a broader range of small stepsizes: for example, at $\eta = 5 \times 10^{-4}$, AFHMC attains $\mathrm{ME} = 2.3 \times 10^{-2}$, compared with 5.3×10^{-2} for FA-HMC and 3.9×10^{-1} for FA-LD. In the noisy-gradient setting, AFHMC achieves the best overall marginal error in the sweep, $\mathrm{ME} = 2.7 \times 10^{-3}$, and remains competitive with the best tuned FA-HMC and FA-LD configurations at the other stepsizes.

Dimension scaling on heterogeneous Gaussian targets. Following the heterogeneous Gaussian experiment of Liang et al. (2025), we study how the communication cost scales with dimension on a controlled federated Gaussian benchmark. There are $N = 10$ clients. For half of the clients the local target is $\mathcal{N}(20\mathbf{1}_d, I_d)$, and for the remaining clients it is $\mathcal{N}(\mathbf{1}_d, 4I_d)$. Averaging the local quadratic potentials gives the global target $\mathcal{N}(16.2\mathbf{1}_d, 1.6I_d)$. We vary $d \in \{2, 50, 100, 150, \ldots, 500\}$ and measure convergence by the squared 2-Wasserstein distance to this global Gaussian target. Note that the empirical 2-Wasserstein distance is expensive to compute for high-dimensional data, therefore, we approximate the posterior distribution by fitting a multivariate Gaussian and compute the 2-Wasserstein distance between the fitted Gaussian distribution and the target distribution via the closed-form formula.

Suggested by the theoretical choice provided by Theorem 4.5, for AFHMC, the leapfrog count and stepsize are scaled with dimension as

$$K(d) = \max\left\{1, \left\lfloor 50(d/100)^{1/3} \right\rceil\right\}, \qquad \eta(d) = 0.006(d/100)^{-1/2},$$

so that $K(100) = 50$ and $\eta(100) = 0.006$. Thus, as d increases, AFHMC uses longer trajectories but smaller leapfrog steps. We run AFHMC until $\mathcal{W}_2^2 < 0.1$ for 20 consecutive checks and record the first communication round at which the threshold is crossed. The 2-Wassertein distance calculation is based on $M(d) = 100d(d-1)$ independent parallel chains (note that we need to estimate a d-dimensional covariance matrix).

As a matched-budget baseline, we run FA-HMC with the same $\eta(d)$ and the same number of communication rounds as AFHMC. FA-HMC uses $T = 10$ local HMC update, and the leapfrog step count within each HMC update is chosen as

$$K_{\mathrm{F}}(d) = \max\left\{1, \lfloor K(d)/10 \rceil\right\},$$

so that the total number of leapfrog updates per communication round is matched to AFHMC up to rounding. To check that the gap below is not caused by giving FA-HMC the AFHMC stepsize, we

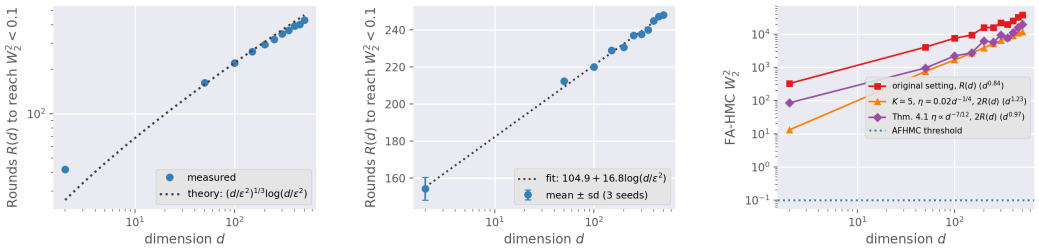

Figure 3: Dimension scaling on Gaussian targets. Left: communication rounds required by AFHMC to reach $\mathcal{W}_2^2 < 0.1$ on the heterogeneous target as dimension increases. Middle: the same quantity on the low-heterogeneity target (mean ± sd over three seeds), with the fitted logarithmic law. Right: FA-HMC $\mathcal{W}_2^2$ on the heterogeneous target under the original setting ($R(d)$ rounds) and under two FA-HMC-specific stepsize schedules at the equal server-exchange budget $2R(d)$; the horizontal line is the AFHMC stopping threshold $\mathcal{W}_2^2 = 0.1$.

additionally run FA-HMC for $2R(d)$ rounds, i.e., the same number of server exchanges as AFHMC at round $R(d)$, under two FA-HMC-specific schedules: the dimension-scaling setting of Liang et al. (2025), $K_{\mathrm{F}} = 5$, $T = 10$ and $\eta(d) = 0.02d^{-1/4}$; and the stepsize suggested by Theorem 4.1 of Liang et al. (2025), $\eta(d) = 0.006(d/100)^{-7/12}$, with $T = 10$ and $K_{\mathrm{F}}(d)$ as above (see Appendix D.2).

Figure 3 summarizes the dimension-scaling experiment. The left panel shows that the number of communication rounds required by AFHMC increases from 42 at $d = 2$ to 430 at $d = 500$. As shown by the figure, the empirical behavior closely resembles the theoretical trend illustrated by the dotted black curve. When fitting the plot by regression $C(d/\epsilon^2)^{\alpha}\log(d/\epsilon^2)$ with $\epsilon^2 = 0.1$ (excluding the outlier case $d = 2$), we have $\hat{\alpha} = 0.282$. This log-corrected scaling is close to the predicted $O((d/\epsilon^2)^{1/3}\log(d/\epsilon^2))$ communication complexity, especially on the larger-dimensional regime. The right panel reports the $\mathcal{W}_2^2$ error of FA-HMC when FA-HMC spends the same dimension-dependent budget used by AFHMC, together with the two FA-HMC-specific schedules at the equal server-exchange budget. The horizontal "adaptive threshold" line marks $\mathcal{W}_2^2 = 0.1$, the stopping accuracy used for AFHMC in the left panel. Under the matched budget, FA-HMC remains orders of magnitude above this threshold: its final $\mathcal{W}_2^2$ ranges from 3.28×10^2 at $d = 2$ to 3.81×10^4 at $d = 500$, with empirical growth speed $O(d^{0.843})$. The FA-HMC-specific schedules do not close the gap even with twice as many rounds: at $d = 500$, they reach $\mathcal{W}_2^2 = 1.19 \times 10^4$ and 1.98×10^4 after 860 rounds, whereas AFHMC crosses 0.1 at round 430, and their errors grow as $O(d^{1.23})$ and $O(d^{0.97})$, i.e., roughly linearly in d, as expected from a heterogeneity bias. This significant performance gap between AFHMC and FA-HMC indicates that the FA-HMC dynamics retain a persistent heterogeneity bias, whereas AFHMC reaches a higher level of accuracy and efficiency.

Low-heterogeneity regime. To examine the $O(\log(d/\epsilon^2))$ communication complexity predicted by Theorem 4.5 when the heterogeneity is small, we keep $N = 10$ clients but estimate every local Gaussian from 10^5 samples drawn from the same distribution $\mathcal{N}(16.2\mathbf{1}_d, 1.6I_d)$, so that the heterogeneity only comes from sampling error; the pooled samples define the target. We use $K(d) = \lfloor 50(d/100)^{1/2} \rceil$ and $\eta(d) = 0.006(d/100)^{-1/2}$, and repeat each setting with three random seeds. AFHMC reaches $\mathcal{W}_2^2 < 0.1$ for every dimension and seed (Figure 3, middle), and the first-crossing round is well described by $R(d) = 104.9 + 16.8\log(d/\epsilon^2)$ with $R^2 = 0.996$. As d grows from 50 to 500, $R(d)$ only increases from 212 to 248 (a factor of 1.17), compared with an increase from 162 to 430 (a factor of 2.65) on the heterogeneous target. This provides direct numerical evidence for the logarithmic communication complexity in the low-heterogeneity regime (details in Appendix D.2).

Logistic regression for FMNIST. We next evaluate the proposed method on a Bayesian multinomial logistic regression task using the Fashion-MNIST (FMNIST) dataset (Xiao et al., 2017), which consists of 60,000 training images and 10,000 test images from 10 clothing categories. The dataset is released by Zalando Research under the MIT license. Each image is a 28×28 grayscale image, flattened into a 784-dimensional feature vector. To model client heterogeneity, we partition the

training data across $N = 10$ clients using a label-skewed Dirichlet split with concentration parameter $\alpha = 0.1$, see Appendix D.3 for details.

We compare FA-LD, FA-HMC, and AFHMC under the same heterogeneous data split. For FA-LD we use the $K = 1$ case, while for FA-HMC and AFHMC we evaluate $K \in \{10, 50, 100\}$. Each method is run for 3600 server exchanges, i.e., 3600 rounds for FA-LD and FA-HMC and 1800 rounds for AFHMC. Between two synchronizations, each client performs 50 leapfrog steps in FA-LD and FA-HMC and K leapfrog steps in AFHMC. For each reported metric, we select the stepsize separately for every method and K value.

Figure 4 reports the test accuracy against the number of server exchanges. AFHMC with $K \in \{10, 100\}$ overtakes all FA-HMC variants after about 1000 exchanges, stays ahead thereafter, and attains the highest final accuracy (0.842, versus 0.841 for the best FA-HMC at the same number of exchanges and 0.836 at the same number of rounds). With $K = 10$, AFHMC achieves this using about ten times fewer gradient evaluations per client than FA-HMC. This result shows that the drift correction in AFHMC continues to be useful beyond the Gaussian experiments and transfers to a non-Gaussian Bayesian logistic regression posterior with real image data. Additional predictive-quality metrics, including Brier score and negative log-likelihood, are reported in Appendix D.3.

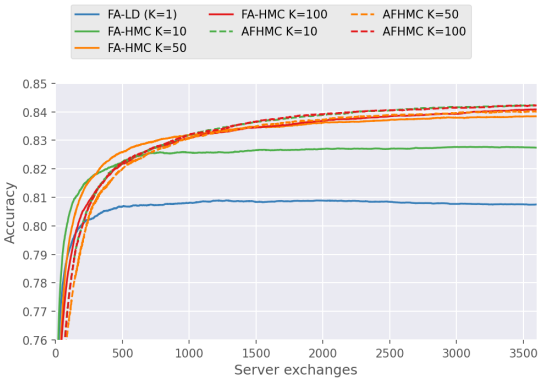

Figure 4: Test accuracy for Bayesian logistic regression on heterogeneous FMNIST against the number of server exchanges (one per round for FA-LD and FA-HMC, two per round for AFHMC).

6 Conclusions and Discussions

In this work, we propose a novel Bayesian Federated sampling algorithm based on the Hamiltonian dynamics. We incorporate a global drift control variate into the leapfrog integrator to alleviate the local drifting caused by the data heterogeneity. This surrogate leapfrog update rule leads to a significant advancement of the theoretical behavior, especially for the communication cost with respect to the system's high dimensionality. Our analysis shows that to achieve an ϵ sampling accuracy, AFHMC only requires $O((d/\epsilon^2)^{1/3})$ communication rounds. Compared with the $O((d/\epsilon^2)^{1/2})$ rate achieved by Langevin-based federated sampler (Plassier et al., 2023; Deng et al., 2024), our result reaffirms the superiority of Hamiltonian dynamics over Langevin dynamics for the Federated sampling task, promoting the communities to utilize HMC-based Federated learning methods for complicated high-dimensional tasks. The advantage claimed by our theory is justified by empirical evidence. In addition, AFHMC can achieve $O(\log(d/\epsilon^2))$ rate for communication complexity under low heterogeneity, matching the rate of communication cost for the Federated optimization realm; our low-heterogeneity Gaussian experiment in Section 5 confirms this logarithmic scaling numerically.

Limitations. The scope of our theoretical investigation is limited to the log-convex target distribution and doesn't consider the stochastic gradient. The theoretical analysis for Hamiltonian dynamics or HMC is highly non-trivial. In the literature, not limited to FL, no result achieves an advanced dimension-dependent rate for HMC under nonconvexity or stochastic gradients. Our contribution is mainly theoretical, hence our numerical experiments are on a small scale, similarly to prior theoretical works, serving for the purpose of validating our theorems.

Acknowledgments and Disclosure of Funding

Lin gratefully acknowledge support from the National Science Foundation (DMS-2533878, DMS-2053746, DMS-2134209, ECCS-2328241, CBET-2347401 and OAC-2311848), and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program under the

"Uncertainty Quantification for Multifidelity Operator Learning (MOLUcQ)" project (Project No. 81739), DE-SC0023161, the SciDAC LEADS Institute, and DOE–Fusion Energy Science, under grant number: DE-SC0024583.

References

Akhmatskaya, E., Nagar, L., Carrillo, J. A., Gavira Balmacz, L., Inouzhe, H., Parga Pazos, M., and Rodríguez Álvarez, M. X. Adaptive tuning of hamiltonian monte carlo methods. *Applied Mathematical Modelling*, 157:116892, 2026. ISSN 0307-904X. doi: https://doi.org/10.1016/j.apm.2026.116892. URL https://www.sciencedirect.com/science/article/pii/S0307904X26001538.

Al-Shedivat, M., Gillenwater, J., Xing, E., and Rostamizadeh, A. Federated learning via posterior averaging: A new perspective and practical algorithms. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=GFsU8a0sGB.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., and McMahan, B. Towards Federated Learning at Scale: System design. *Proceedings of Machine Learning and Systems*, 1:374–388, 2019.

Chen, H., Xiao, M., and Pang, Z. Satellite-based computing networks with federated learning. *IEEE Wireless Communications*, 29(1):78–84, 2022.

Chen, H.-Y. and Chao, W.-L. Fed{be}: Making bayesian model ensemble applicable to federated learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dgtpE6gKjHn.

Chen, T., Fox, E., and Guestrin, C. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*, pp. 1683–1691, 2014.

Dalalyan, A. S. Theoretical Guarantees for Approximate Sampling from Smooth and Log-concave Densities. *Journal of the Royal Statistical Society: Series B*, 79(3):651–676, 2017.

Dalalyan, A. S. and Karagulyan, A. User-friendly Guarantees for the Langevin Monte Carlo with Inaccurate Gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Deng, W., Zhang, Q., Ma, Y.-A., Song, Z., and Lin, G. On convergence of federated averaging langevin dynamics. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24, 2024.

Duane, S., Kennedy, A., Pendleton, B., and Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2): 216–222, 1987.

Durmus, A. and Moulines, E. High-dimensional Bayesian Inference via The Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Durmus, A., Éric Moulines, and Saksman, E. Irreducibility and Geometric Ergodicity of Hamiltonian Monte Carlo. *Annals of Statistics*, 48(6):3545–3564, 2020.

Erdogdu, M. A. and Hosseinzadeh, R. On the Convergence of Langevin Monte Carlo: the Interplay between Tail Growth and Smoothness. In *Proc. of Conference on Learning Theory (COLT)*, pp. 1776–1822, 2021.

Fernández-Pendás, M., Akhmatskaya, E., and Sanz-Serna, J. M. Adaptive multi-stage integrators for optimal energy conservation in molecular simulations. *Journal of Computational Physics*, 327: 434–449, 2016.

Girolami, M. and Calderhead, B. Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Guo, H., Greengard, P., Wang, H., Gelman, A., Kim, Y., and Xing, E. Federated learning as variational inference: A scalable expectation propagation approach. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dZrQR7OR11.

Hirt, M., Titsias, M., and Dellaportas, P. Entropy-based adaptive hamiltonian monte carlo. *Advances in neural information processing systems*, 34:28482–28495, 2021.

Hoffman, M., Radul, A., and Sountsov, P. An adaptive-mcmc scheme for setting trajectory lengths in hamiltonian monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pp. 3907–3915. PMLR, 2021.

Hoffman, M. D. and Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, pp. 1593–1623, 2014.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic Controlled Averaging for Federated Learning. In *International Conference on Machine Learning (ICML)*, pp. 5132–5143, 2020.

Kim, M. and Hospedales, T. Fedhb: Hierarchical bayesian federated learning. *Journal of Machine Learning Research*, 26(272):1–50, 2025.

Konečnỳ, J., McMahan, B., and Ramage, D. Federated optimization: Distributed optimization Beyond the Datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. URL `https://arxiv.org/abs/1610.05492`.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smithy, V. Federated Optimization in Heterogeneous Networks. In *Proceedings of the 3rd MLSys Conference*, 2020b.

Liang, J., Zhang, Q., Deng, W., Song, Q., and Lin, G. Bayesian federated learning with hamiltonian monte carlo: Algorithm and theory. *Journal of Computational and Graphical Statistics*, 34(2): 509–518, 2025.

Liu, L., Jiang, X., Zheng, F., Chen, H., Qi, G.-J., Huang, H., and Shao, L. A bayesian federated learning framework with online laplace approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):1–16, 2024. doi: 10.1109/TPAMI.2023.3322743.

Mangoubi, O. and Smith, A. Rapid mixing of hamiltonian monte carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Mangoubi, O. and Smith, A. Mixing of hamiltonian monte carlo on strongly log-concave distributions: Continuous dynamics. *The Annals of Applied Probability*, 31(5):2019–2045, 2021.

Mangoubi, O. and Vishnoi, N. Dimensionally tight bounds for second-order hamiltonian monte carlo. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/e07bceab69529b0f0b43625953fbf2a0-Paper.pdf`.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient Learning of Deep Networks from Decentralized Data. In *Proc. of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtarik, P. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15750–15769. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/mishchenko22b.html`.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Nagar, L., Fernández-Pendás, M., Sanz-Serna, J. M., and Akhmatskaya, E. Adaptive multi-stage integration schemes for hamiltonian monte carlo. *Journal of Computational Physics*, 502:112800, 2024.

Neal, R. M. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, volume 54, pp. 113–162. Chapman and Hall/CRC, 2012.

Plassier, V., Moulines, E., and Durmus, A. Federated averaging langevin dynamics: Toward a unified theory and new algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 5299–5356. PMLR, 2023.

Sato, I. and Nakagawa, H. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pp. 982–990. PMLR, 2014.

Vedadi, E., Dillon, J. V., Mansfield, P. A., Singhal, K., Afkanpour, A., and Morningstar, W. R. Federated variational inference: Towards improved personalization and generalization. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=6OmRkUHgs5.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Federated Learning with Matched Averaging. In *International Conference on Learning Representations (ICLR)*, 2020.

Wang, Z., Chen, Y., Song, Q., and Zhang, R. Enhancing low-precision sampling via stochastic gradient hamiltonian monte carlo. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=uSLNzzuiDJ.

Welling, M. and Teh, Y. W. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *International Conference on Machine Learning (ICML)*, 2011.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, X., Li, Y., Li, W., Guo, K., and Shao, Y. Personalized federated learning via variational bayesian inference. In *International Conference on Machine Learning*, pp. 26293–26310. PMLR, 2022.

Zongchen Chen, S. S. V. Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions. In *Approximation, Randomization, and Combinatorial Optimization*, 2019.

Zou, D. and Gu, Q. On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients. In *International Conference on Machine Learning (ICML)*, pp. 13012–13022, 2021.

Zou, D., Xu, P., and Gu, Q. Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the paper's contributions and scope. We propose a novel AFHMC algorithm. The theoretical contributions claimed in the abstract and introduction (i.e., better communication complexity w.r.t. the model dimension) are supported by our main theorems and numerical results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discussed the limitations of our work in the end of the main text.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All the theorems, formulas, and proofs in the paper are numbered and cross-referenced. All the proofs are provided in the Appendix Section B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code for reproduction is available at: `https://github.com/laaaarrywang/Adaptive-Federated-HMC`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We discuss the source and license of Fashion-MNIST data in the Numerical Evidences section. The other experiments are all using simulated data. We also include the code for reproduction at: `https://github.com/laaaarrywang/Adaptive-Federated-HMC`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the training and testing details in the Numerical Evidences section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include the standard errors of results of the Bayesian logistic regression on Fashion-MNIST in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on GPU nodes of an internal HPC cluster. We report the per-round computational cost of each method (leapfrog steps and gradient evaluations per client) in Appendix D.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics. We maintained anonymity and provided credits where due, to the best of our knowledge.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contribution of this work is mainly theoretical. There are no foreseeable societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper proposes a novel HMC sampling algorithm for the Federated Learning problem. This work doesn't produce any generative models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite the existing datasets used in our real data applications. We will respect the licenses and terms of use of these assets when releasing our code and trained checkpoints.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide the implementation code for all proposed algorithms and experiments at `https://github.com/laaaarrywang/Adaptive-Federated-HMC`, released under the MIT license and documented with README files describing how to run each experiment.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This project does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This project does not involve human subjects or human data collections.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

A Algorithm, assumption and notations

A.1 notations and definitions

The ideal HMC dynamic for the global loss f is defined as:

$$d\theta(t) = p(t)dt, \qquad dp(t) = -\nabla_\theta f(\theta(t))dt. \tag{4}$$

According to Algorithm 2, we define $\theta_{t,j}^{(c)}$ and $p_{t,j}^{(c)}$ to be the local position variable and local momentum variable at the j-th surrogate leapfrog step of the c-th local node within the t-th global iteration. In particular, $p_{t,0}^{(c)} = p_t$ is a freshly sampled Gaussian variable at iteration t shared by all local nodes; for $j > 0$, $p_{t,j}^{(c)}$ is generated following the local surrogate leapfrog. Also, we have $\theta_t := \sum_{c=1}^N w_c \theta_{t-1,K}^{(c)}$ aggregated from local nodes and $\theta_{t,0}^{(c)} := \theta_t$ broadcast from the center node.

Define $\theta_{t,k} := \sum_{c=1}^N w_c \theta_{t,j}^{(c)}$ as the corresponding (implicit) global position variable, and $p_{t,k} = \sum_{c=1}^N w_c p_{t,j}^{(c)}$ is the corresponding (implicit) global momentum variable.

Define $\mathcal{H}_\eta(\theta) := \theta + \eta p - (\eta^2/2)\nabla f(\theta)$ as the classical discrete leapfrog update from position θ associated with momentum p and stepsize η and $\bar{\mathcal{H}}_\eta(\theta) := \theta + \int_0^\eta p(t)dt$ as the classical continuous leapfrog update from position θ associated with momentum $p(t)$ and length η, following the HMC dynamic (4). Similarly, define $\mathcal{H}_\eta(p) := p - \eta/2(\nabla f(\theta) + \nabla f(\theta + \eta p - \eta^2/2\nabla f(\theta)))$ as a classical discrete leapfrog update from p associated with position θ and stepsize η and $\bar{\mathcal{H}}_\eta(p) := p - \int_0^\eta \theta(t)dt$ as a classical continuous leapfrog update from p associated with position $\theta(t)$ and length η following the HMC dynamic (4).

Based on these, we use the shorthands $\mathcal{H}_\eta^1(\theta) := \mathcal{H}_\eta(\theta)$, $\mathcal{H}_\eta^{i+1}(\theta) := \mathcal{H}_\eta(\mathcal{H}_\eta^i(\theta)) = \mathcal{H}_\eta^i(\theta) + \eta\mathcal{H}_\eta^i(p) - \eta^2/2\nabla f(\mathcal{H}_\eta^i(\theta))$, and $\mathcal{H}_\eta^1(p) := \mathcal{H}_\eta(p)$, $\mathcal{H}_\eta^{i+1}(p) := \mathcal{H}_\eta(\mathcal{H}_\eta^i(p)) = \mathcal{H}_\eta^i(p) - \eta^2/2(\nabla f(\mathcal{H}_\eta^i(\theta)) + \nabla f(\mathcal{H}_\eta^i(\theta) + \eta\mathcal{H}_\eta^i(p) - \eta^2/2\nabla f(\mathcal{H}_\eta^i(\theta))))$. We also use the shorthands $\bar{\mathcal{H}}_\eta^i(\theta) = \theta + \int_0^{i\eta} p(s)ds$ and $\bar{\mathcal{H}}_\eta^i(p) := p - \int_0^{i\eta} \theta(s)ds$.

In addition, we define a coupled continuous HMC process as follows. θ_0^π follows the target distribution $\propto \exp(-f)$. $\theta_{t,0}^\pi := \theta_t^\pi$, $\theta_{t,k+1}^\pi := \bar{\mathcal{H}}_\eta(\theta_{t,k}^\pi)$, and $\theta_{t+1}^\pi := \theta_{t,K}^\pi = \bar{\mathcal{H}}_{K\eta}(\theta_{t,0}^\pi)$. For this process, the associated initial momentum is the same as the discrete HMC process, i.e., $p_{t,0}^\pi = p_t^\pi = p_t$.

We also denote $\theta^* = \text{argmin}_\theta f(\theta)$.

B Proofs

B.1 main proof

Theorem 4.5. *Under Assumptions 4.1-4.2 and if $H \gg M^4\epsilon^2/m > 0$, then there exists some small absolute constant C such that if we choose*

$$K\eta = C\left(\frac{\epsilon^2 m^2}{HM^2}\right)^{1/6} \ll C\frac{\sqrt{m}}{M} \quad \textit{and} \quad \eta = C\left(\frac{\epsilon^2 m^2}{M^2 d}\right)^{1/2},$$

$\mathcal{W}_2(\pi_{t_\epsilon}, \pi) \le \epsilon$ can be attained with communication times

$$t_\epsilon \asymp \left(\frac{HM^2}{\epsilon^2 m^5}\right)^{1/3} \log(\mathcal{W}_2^2(\pi_0, \pi)/\epsilon^2).$$

The corresponding local iteration number is

$$K \asymp \frac{M^{2/3} d^{1/2}}{\epsilon^{2/3} m^{2/3} H^{1/6}}.$$

Moreover, if $H \ll M^4\epsilon^2/m$ (i.e., heterogeneity across local nodes is very small), it only requires

$$t_\epsilon \asymp \log(\mathcal{W}_2^2(\pi_0, \pi)/\epsilon^2),$$

achieving a linear rate with respect to the communication cost.

Proof. Let us define

$$K^2\eta^2 = C(\frac{\epsilon^2 m^2}{HM^2})^{1/3}, \qquad \eta^2 = C\frac{\epsilon^2 m^2}{M^2 d} \tag{5}$$

for some sufficiently small C, we have

$$\frac{K^2\eta^4\frac{M^2}{m}d}{mK^2\eta^2} = C\epsilon^2 \le \frac{1}{3}\epsilon^2, \qquad \frac{(K\eta)^8\frac{M^2}{m}H}{mK^2\eta^2} = C^3\epsilon^2 \le \frac{1}{3}\epsilon^2. \tag{6}$$

One can check that if

$$t \ge \frac{\log(\epsilon^2/(3\mathbb{E}\|\theta_0 - \theta_0^\pi\|^2))}{\log(1 - Cm(\frac{\epsilon^2 m^2}{HM^2})^{1/3}/16)} \asymp \frac{1}{C}(\frac{HM^2}{\epsilon^2 m^5})^{1/3}(\log(3\mathbb{E}\|\theta_0 - \theta_0^\pi\|^2/\epsilon^2)),$$

we have

$$(1 - m(K\eta)^2/16)^t\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 \le \frac{1}{3}\epsilon^2. \tag{7}$$

Combining (6)-(7) and by Theorem 4.3, we can see that, for enough small C and set (5), we can have

$$\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 \le \epsilon^2.$$

This justifies the required communication round

$$t_\epsilon \ge \frac{1}{C}(\frac{HM^2}{\epsilon^2 m^5})^{1/3}\log(3\mathbb{E}\|\theta_0 - \theta_0^\pi\|^2/\epsilon^2)$$

On the other hand, if $H \ll M^4\epsilon^2/m$ (i.e., the heterogeneity cross local nodes is small), by choosing enough small C such that with $K^2\eta^2 = Cm/M^2$ and $\eta^2 = C\epsilon^2 m^2/(M^2 d)$, we have

$$\frac{K^2\eta^4\frac{M^2}{m}d}{mK^2\eta^2} \le C^2\epsilon^2 \le \frac{1}{3}\epsilon^2, \qquad \frac{(K\eta)^8\frac{M^2}{m}H}{mK^2\eta^2} = C^3\frac{m}{M^4}H \ll \epsilon^2 \le \frac{1}{3}\epsilon^2. \tag{8}$$

One can check that if

$$t \ge \frac{\log(\epsilon^2/(3\mathbb{E}\|\theta_0 - \theta_0^\pi\|^2))}{\log(1 - Cm^2/(16M^2))},$$

we have

$$(1 - \frac{Cm^2}{M^2})^t\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 \le \frac{1}{3}\epsilon^2. \tag{9}$$

Combining (8)-(9) and by Theorem 4.3, we obtain that $\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 \le \epsilon^2$. This justified the required communication round to be

$$t_\epsilon \ge C'\log(3\mathbb{E}\|\theta_0 - \theta_0^\pi\|^2/\epsilon^2).$$

Finally, we can choose θ_0^π such that $\mathcal{W}^2(\pi_t, \pi) = \mathbb{E}\|\theta_0 - \theta_0^\pi\|^2$, which concludes the proof. □

Theorem 4.3 (Convergence)**.** *Under Assumptions 4.1-4.2, there exist sufficient large constants C_0, C_1, C_2 such that $\{\theta_t\}_t$ satisfies that, for $K\eta \le \sqrt{m}/(C_0 M) \le \sqrt{m}/(2M)$,*

$$\mathbb{E}\|\theta_{t+1} - \theta_{t+1}^\pi\|^2 \le \left(1 - \frac{m(K\eta)^2}{16}\right)\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 + C_1K^2\eta^4\frac{M^2}{m}d + C_2(K\eta)^8\frac{M^2}{m}H$$

where $H := \max_{i\le K}\sum_{c=1}^N w_c\mathbb{E}\|\frac{1}{i\eta}(\nabla f^{(c)}(\bar{\mathcal{H}}_\eta^i(\theta^\pi)) - \nabla f^{(c)}(\theta^\pi) + \nabla f(\theta^\pi) - \nabla f(\bar{\mathcal{H}}_\eta^i(\theta^\pi))\|^2$ is a metric quantifying the federated heterogeneity.

Proof. We introduce an intermediate variable $\bar{\mathcal{H}}_\eta^K(\theta_t)$ being the continuously updated version of $\theta_{t,K}$ from θ_t. For $c_1, c_2 > 0$ with $c_1 + c_2 = 1$, by Lemma B.12, we can separate the errors into two parts

$$\|\theta_{t,K} - \theta_{t,K}^\pi\|^2 \le \frac{1}{c_1} \underbrace{\|\theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_t)\|^2}_{\text{discrete divergence error}} + \frac{1}{c_2} \underbrace{\|\bar{\mathcal{H}}_\eta^K(\theta_t) - \theta_{t,K}^\pi\|^2}_{\text{ideal convergence error}} \tag{10}$$

The convergence of the ideal process has been well studied in the literature Zongchen Chen (2019). Indeed, recall the definitions that that $\bar{\mathcal{H}}_\eta^K(\theta_t)$ and $\theta_{t,K}^\pi$ share the same initial momentum p_t for each t, by Lemma B.1, it can be bounded by a contraction factor of the distance between their initial positions

$$\|\bar{\mathcal{H}}_\eta^K(\theta_t) - \theta_{t,K}^\pi\|^2 \le (1 - \frac{m(K\eta)^2}{4})\|\theta_t - \theta_t^\pi\|^2, \qquad \text{for } K\eta \le 1/(2\sqrt{M}). \tag{11}$$

We choose c_1, c_2 such that $\frac{1-m(K\eta)^2/4}{c_2} = 1 - m(K\eta)^2/8$, and it follows

$$c_2 = \frac{1 - m(K\eta)^2/4}{1 - m(K\eta)^2/8}, \qquad c_1 = \frac{m(K\eta)^2/8}{1 - m(K\eta)^2/8} \ge \frac{m(K\eta)^2}{8} \tag{12}$$

This choice ensures that, after amplification by $1/c_1$, the contraction factor in (10) is still strictly smaller than 1.

In the following, to finish the proof, it suffice to show that the divergence error

$$\mathbb{E}\|\theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 \le \frac{m(K\eta)^2}{16}\|\theta_t - \theta_t^\pi\|^2 + C_1 K^2\eta^4\frac{M^2}{m}d + C_2(K\eta)^8\frac{M^2}{m}H,$$

or some constants $C_1, C_2 > 0$. To study the discrete divergence error, we separate the error within the leapfrog-step level. Since by definitions $\theta_{t,K} = \sum_{c=1}^N \theta_{t,K}^{(c)}$ and $\theta_t = \theta_{t,0}$ we can write

$$\theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_t) = \theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_{t,0}) = \sum_{k=1}^K \bar{\mathcal{H}}_\eta^{K-k}(\theta_{t,k}) - \bar{\mathcal{H}}_\eta^{K-k+1}(\theta_{t,k-1}).$$

By convexity of $\|\cdot\|^2$ and Lemma B.12,

$$\begin{aligned}\|\theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 =& \|\sum_{k=1}^K \bar{\mathcal{H}}_\eta^{K-k}(\theta_{t,k}) - \bar{\mathcal{H}}_\eta^{K-k+1}(\theta_{t,k-1})\|^2 \\ \le& K\sum_{k=1}^K \|\bar{\mathcal{H}}_\eta^{K-k}(\theta_{t,k}) - \bar{\mathcal{H}}_\eta^{K-k+1}(\theta_{t,k-1})\|^2\end{aligned}$$

We emphasis that $\bar{\mathcal{H}}_\eta^{K-k}(\theta_{t,k})$ starts with position $\theta_{t,k}$ and momentum $p_{t,k}$ while $\bar{\mathcal{H}}_\eta^{K-k+1}(\theta_{t,k-1})$ can be viewed as a $(K-k)\eta$-length process starting with position $\bar{\mathcal{H}}_\eta(\theta_{t,k-1})$ and momentum $\bar{\mathcal{H}}_\eta(p_{t,k-1})$. Lemma B.1 yields
$\|\bar{\mathcal{H}}_\eta^{K-k}(\theta_{t,k}) - \bar{\mathcal{H}}_\eta^{K-k+1}(\theta_{t,k-1})\|^2 \le 2\|\theta_{t,k} - \bar{\mathcal{H}}_\eta(\theta_{t,k-1})\|^2 + 2(K-k)^2\eta^2\|p_{t,k} - \bar{\mathcal{H}}_\eta(p_{t,k-1})\|^2$.
It follows that, we can bound the divergence error as follow

$$\|\theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 \le 2K\sum_{k=1}^K \Big(\underbrace{\|\theta_{t,k} - \bar{\mathcal{H}}_\eta(\theta_{t,k-1})\|^2}_{\text{error induced by inaccurate positions}} + (K-k)^2\eta^2 \underbrace{\|p_{t,k} - \bar{\mathcal{H}}_\eta(p_{t,k-1})\|^2}_{\text{error induced by inaccurate momentums}} \Big). \tag{13}$$

The divergence errors can be further separated into systematic leapfrog approximation error and surrogate-induced error. To see this, by definitions

$$\theta_{t,k} - \bar{\mathcal{H}}_\eta(\theta_{t,k-1}) = \underbrace{\mathcal{H}_\eta(\theta_{t,k-1}) - \bar{\mathcal{H}}_\eta(\theta_{t,k-1})}_{\text{systematic leapfrog gap}} - \frac{\eta^2}{2}\sum_{c=1}^N w_c(\underbrace{\nabla f^{(c)}(\theta_{t,k-1}^{(c)}) - \nabla f^{(c)}(\theta_{t,k-1})}_{\text{surrogate induced gap}})$$

$$\begin{aligned}p_{t,k} - \bar{\mathcal{H}}_\eta(p_{t,k-1}) =& \underbrace{\mathcal{H}_\eta(p_{t,k-1}) - \bar{\mathcal{H}}_\eta(p_{t,k-1})}_{\text{systematic leapfrog gap}} \\ &- \frac{\eta}{2}\sum_{c=1}^N w_c(\underbrace{\nabla f^{(c)}(\theta_{t,k-1}^{(c)}) - \nabla f^{(c)}(\theta_{t,k-1}) + \nabla f^{(c)}(\theta_{t,k}^{(c)}) - \nabla f^{(c)}(\theta_{t,k})}_{\text{surrogate induced gap}}).\end{aligned}$$

Applying Lemma B.12, we can have

$$\|\theta_{t,k} - \bar{\mathcal{H}}_\eta(\theta_{t,k-1})\|^2 \leq 2\|\mathcal{H}_\eta(\theta_{t,k-1}) - \bar{\mathcal{H}}_\eta(\theta_{t,k-1})\|^2 + 2\frac{\eta^4}{4}\sum_{c=1}^N w_c \|\nabla f^{(c)}(\theta_{t,k-1}^{(c)}) - \nabla f^{(c)}(\theta_{t,k-1})\|^2$$
$$\|p_{t,k} - \bar{\mathcal{H}}_\eta(p_{t,k-1})\|^2 \leq 2\|\mathcal{H}_\eta(p_{t,k-1}) - \bar{\mathcal{H}}_\eta(p_{t,k-1})\|^2 + 2\frac{\eta^2}{4}\sum_{c=1}^N w_c \sum_{j=k-1}^k 2\|\nabla f^{(c)}(\theta_{t,j}^{(c)}) - \nabla f^{(c)}(\theta_{t,j})\|^2.$$

Inserting these beck to (13), we have

$$\begin{aligned}
&\|\theta_{t,K} - \bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 \\
\leq& 2K\sum_{k=1}^K \Big(2\|\mathcal{H}_\eta(\theta_{t,k-1}) - \bar{\mathcal{H}}_\eta(\theta_{t,k-1})\|^2 + 2(K-k)^2\eta^2\|\mathcal{H}_\eta(p_{t,k-1}) - \bar{\mathcal{H}}_\eta(p_{t,k-1})\|^2 \\
&+ \frac{\eta^4}{2}\sum_{c=1}^N w_c(\|\nabla f^{(c)}(\theta_{t,k-1}^{(c)}) - \nabla f^{(c)}(\theta_{t,k-1})\|^2 + (K-k)^2\sum_{j=k-1}^k \|\nabla f^{(c)}(\theta_{t,j}^{(c)}) - \nabla f^{(c)}(\theta_{t,j})\|^2)\Big) \\
\leq& 5K\sum_{k=0}^{K-1}\Big(\underbrace{\|\mathcal{H}_\eta(\theta_{t,k}) - \bar{\mathcal{H}}_\eta(\theta_{t,k})\|^2 + (K-k)^2\eta^2\|\mathcal{H}_\eta(p_{t,k}) - \bar{\mathcal{H}}_\eta(p_{t,k})\|^2}_{\text{divergence error induced by leapfrog approximation}} \\
&+ (K\eta)^2\eta^2\underbrace{\sum_{c=1}^N w_c\|\nabla f^{(c)}(\theta_{t,k}^{(c)}) - \nabla f^{(c)}(\theta_{t,k})\|^2}_{\text{divergence error induced by surrogate leapfrog update}}\Big),
\end{aligned} \tag{14}$$

where we have applied $2\eta^4(1/2 + 2(K-k)^2) \leq 5K^2\eta^4$ for $k < K$ in the last inequality.
By M-smoothness assumption,

$$\sum_{c=1}^N w_c\mathbb{E}\|\nabla f^{(c)}(\theta_{t,k}^{(c)}) - \nabla f^{(c)}(\theta_{t,k})\|^2 \leq \sum_{c=1}^N w_c M^2\mathbb{E}\|\theta_{t,k}^{(c)} - \theta_{t,k}\|^2$$

and further by Lemma B.6 that $\sum_{c=1}^N w_c\mathbb{E}\|\theta_{t,k}^{(c)} - \theta_{t,k}\|^2 \leq k^3\eta^4\sum_{i=1}^{k-1}(16M^2\mathbb{E}\|\theta_{t,i}^\pi - \theta_{t,i}\|^2 + 16M^2\mathbb{E}\|\theta_t^\pi - \theta_t\|^2 + 4(k\eta)^2 H)$, we have

$$\sum_{c=1}^N w_c\mathbb{E}\|\nabla f^{(c)}(\theta_{t,k}^{(c)}) - \nabla f^{(c)}(\theta_{t,k})\|^2 \leq k^3\eta^4 M^2\sum_{i=1}^{k-1}(16M^2\mathbb{E}\|\theta_{t,i}^\pi - \theta_{t,i}\|^2 + 16M^2\mathbb{E}\|\theta_t^\pi - \theta_t\|^2 + 4(k\eta)^2 H). \tag{15}$$

Here we recall that $H := \max_{i\leq K}\sum_{c=1}^N w_c\mathbb{E}\|\frac{1}{i\eta}(\nabla f^{(c)}(\bar{\mathcal{H}}_\eta^i(\theta^\pi)) - \nabla f^{(c)}(\theta^\pi) + \nabla f(\theta^\pi) - \nabla f(\bar{\mathcal{H}}_\eta^i(\theta^\pi)))\|^2$ is a quantity measuring the federated heterogeneity.
As for the systematic leapfrog approximation errors, by Lemma B.3 and Lemma B.4, we have

$$\begin{aligned}
\|\mathcal{H}_\eta(\theta_{t,k}) - \bar{\mathcal{H}}_\eta(\theta_{t,k})\|^2 \leq& \eta^6\Big(\frac{M^2}{10}\|p_{t,k}\|^2 + \frac{M^4\eta^2}{60}\|\theta_{t,k} - \theta^*\|^2\Big) \\
\|\mathcal{H}_\eta(p_{t,k}) - \bar{\mathcal{H}}_\eta(p_{t,k})\|^2 \leq& \eta^4\Big(\frac{5M^2}{2}\|p_{t,k}\|^2 + \frac{2\eta^2M^4}{3}\|\theta_{t,k} - \theta^*\|^2\Big)
\end{aligned} \tag{16}$$

Inserting (15)-(16) back to (14) and applying Lemma B.10 on $p_{t,k}$ and $\theta_{t,k}$, we obtain

$$\begin{aligned}\mathbb{E}\|\theta_{t,K}-\bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 \leq & 5K\sum_{k=0}^{K-1}\Big((\frac{\eta^6M^2}{10}+\frac{(K\eta)^2\eta^4 5M^2}{2})\mathbb{E}\|p_{t,k}\|^2+(\frac{\eta^8M^4}{60}+\frac{(K\eta)^2 2\eta^6M^4}{3})\mathbb{E}\|\theta_{t,k}-\theta^*\|^2\\ &+K^5\eta^8M^2\sum_{i=1}^{k-1}(16M^2\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2+16M^2\mathbb{E}\|\theta_t^\pi-\theta_t\|^2+4(k\eta)^2H)\Big)\\ \leq & 5K\sum_{k=0}^{K-1}\Big((\frac{\eta^6M^2}{10}+\frac{(K\eta)^2\eta^4 5M^2}{2})(5\mathbb{E}|p_t\|^2+5M^2(k\eta)^2\mathbb{E}\|\theta_t-\theta^*\|^2)\\ &+(\frac{\eta^8M^4}{60}+\frac{(K\eta)^2 2\eta^6M^4}{3})(4(k\eta)^2\mathbb{E}\|p_t\|^2+3\mathbb{E}\|\theta_t-\theta^*\|^2)\\ &+K^5\eta^8M^2\sum_{i=1}^{k-1}(16M^2\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2+16M^2\mathbb{E}\|\theta_t^\pi-\theta_t\|^2+4(k\eta)^2H)\Big).\end{aligned}$$

Rearranging the coefficients, it can be seen that

$$\begin{aligned}\mathbb{E}\|\theta_{t,K}-\bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 \leq & \Big(\frac{5}{2}K^2\eta^6M^2+\frac{125}{2}K^4\eta^6M^2+\frac{K^4\eta^{10}M^4}{3}+\frac{40K^6\eta^{10}M^4}{3}\Big)\mathbb{E}\|p_t\|^2\\ &+\Big(\frac{5}{2}K^4\eta^8M^4+\frac{125}{2}K^6\eta^8M^4+\frac{K^2\eta^8M^4}{4}+80K^4\eta^8M^4\Big)\mathbb{E}\|\theta_t-\theta^*\|^2\\ &+80\sum_{k=0}^{K-1}K^6\eta^8M^4\sum_{i=1}^{k-1}\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2\\ &+80(K\eta)^8M^4\mathbb{E}\|\theta_t^\pi-\theta_t\|^2\\ &+20(K\eta)^{10}M^2H\\ \leq & CK^4\eta^6M^2d+CK^6\eta^8M^4\mathbb{E}\|\theta_t-\theta^*\|^2+CK^7\eta^8M^4\sum_{i=1}^{K-2}\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2\\ &+C(K\eta)^8M^4\mathbb{E}\|\theta_t^\pi-\theta_t\|^2+C(K\eta)^{10}M^2H\end{aligned}$$

for enough large C and $K\eta\leq 1/(2\sqrt{M})$.

Combining this with $\mathbb{E}\|\theta_t-\theta^*\|^2\leq 2\mathbb{E}\|\theta_t-\theta_t^\pi\|^2+2\mathbb{E}\|\theta_t^\pi-\theta^*\|^2$ and Lemma B.2 that $\mathbb{E}\|\theta_t^\pi-\theta^*\|^2\leq d/m$, we get

$$\begin{aligned}\mathbb{E}\|\theta_{t,K}-\bar{\mathcal{H}}_\eta^K(\theta_t)\|^2 \leq & C'K^4\eta^6M^2d+C'K^7\eta^8M^4\sum_{i=1}^{K-2}\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2\\ &+C'(K\eta)^8M^4\mathbb{E}\|\theta_t^\pi-\theta_t\|^2+C'(K\eta)^{10}M^2H\end{aligned}\tag{17}$$

for enough large C'.

Finally, combining it with (10)-(12), it can be seen that for an enough large C''

$$\begin{aligned}\mathbb{E}\|\theta_{t,K}-\theta_{t,K}^\pi\|^2 \leq & (1-\frac{m(K\eta)^2}{8}+\frac{C''(K\eta)^6M^4}{m})\mathbb{E}\|\theta_t-\theta_t^\pi\|^2+C''K^2\eta^4\frac{M^2}{m}d\\ &+C''K^5\eta^6\frac{M^4}{m}\sum_{i=1}^{K-2}\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2+C''(K\eta)^8\frac{M^2}{m}H.\end{aligned}\tag{18}$$

To prove our final claim of this theorem, we use induction to tackle $\sum_{i=1}^{K-2}\mathbb{E}\|\theta_{t,i}^\pi-\theta_{t,i}\|^2$. In specific, we are going to show that there exist constants C_0,C_1,C_2 (to be determined later) such that for $K\eta\leq\sqrt{m}/(C_0M)\leq 1/(2\sqrt{M})$

$$\mathbb{E}\|\theta_{t,K}-\theta_{t,K}^\pi\|^2\leq(1-\frac{m(K\eta)^2}{16})\mathbb{E}\|\theta_t-\theta_t^\pi\|^2+C_1K^2\eta^4\frac{M^2}{m}d+C_2(K\eta)^8\frac{M^2}{m}H.\tag{19}$$

Firstly, (19) holds for $K = 0, 1, 2$ by (18) for a C_0 with $16C_0^4 > C''$.

Next, assume that (19) holds for all $K \leq \widetilde{K} - 1$, we are going to show that (19) also holds for $K = \widetilde{K}$.

Applying our induction assumption on (18), we have

$$\begin{aligned}
&\mathbb{E}\|\theta_{t,\widetilde{K}} - \bar{\mathcal{H}}_\eta^{\widetilde{K}}(\theta_t)\|^2 \\
\leq&(1 - \frac{m(K\eta)^2}{8} + \frac{C''(K\eta)^6 M^4}{m})\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 + C''K^2\eta^4\frac{M^2}{m}d + C''(K\eta)^8\frac{M^2}{m}H \\
&+ C''K^5\eta^6\frac{M^4}{m}\sum_{i=1}^{K-2}\Big((1 - \frac{m(i\eta)^2}{16})\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 + C_1 i^2\eta^4\frac{M^2}{m}d + C_2(i\eta)^8\frac{M^2}{m}H\Big) \\
\leq&(1 - \frac{m(K\eta)^2}{8} + \frac{C''(K\eta)^6 M^4}{m} + C''(K\eta)^6\frac{M^4}{m})\mathbb{E}\|\theta_t - \theta_t^\pi\|^2 \\
&+ (C'' + C''K^6\eta^6\frac{M^4}{m}C_1)K^2\eta^4\frac{M^2}{m}d + (C'' + C''K^6\eta^6 C_2)(K\eta)^8\frac{M^2}{m}H
\end{aligned}$$

To finish the induction, we need to satisfy that for $K\eta \leq \sqrt{m}/(C_0 M)$,

$$\frac{m(K\eta)^2}{8} - \frac{2C''(K\eta)^6 M^4}{m} \geq \frac{m(K\eta)^2}{16},\ (C'' + C''K^6\eta^6\frac{M^4}{m}C_1) \leq C_1,\ (C'' + C''K^6\eta^6 C_2) \leq C_2.$$

which holds trivally for an enough large constant C_0.

Noting that $M \geq m$ and that $\theta_{t+1} = \theta_{t,K}$ and $\theta_{t+1}^\pi = \theta_{t,K}^\pi$, this finishes the proof.

□

B.2 ideal process

Lemma B.1 (Lemma D.3 in Liang et al. (2025)). *Under Assumption 4.2 consider two HMC systems $\theta(t)$ and $\bar{\theta}(t)$ starting with the same initial positions and different initial momentum $p(0)$ and $\widetilde{p}(0)$ respectively. We have*

$$\|\theta(t) - \bar{\theta}(t)\|_2 \leq \frac{1}{\sqrt{M}}\sinh(\sqrt{M}t)\|p(0) - \widetilde{p}(0)\|_2,$$

Further if we have Assumption 4.1, consider two HMC systems $\theta(t)$ and $\widetilde{\theta}(t)$ starting with initial positions $\theta(0)$ and $\widetilde{\theta}(0)$, and initial momentum $p(0)$ and $\widetilde{p}(0)$ respectively, then for $0 \leq t \leq \frac{1}{2\sqrt{M}}$, we have

$$\|\theta(t) - \widetilde{\theta}(t)\|_2 \leq (1 - \frac{m}{4}t^2)\|\theta(0) - \widetilde{\theta}(0)\|_2 + t\|p(0) - \widetilde{p}(0)\|_2.$$

Lemma B.2 (Proposition 1 in Durmus & Moulines (2019)). *Let $x^* \in \mathbb{R}^d$ be the global minimizer of a loss function $f(\cdot)$ satisfying Assumption 4.1 and x^π be a random variable following distribution $\propto e^{-f(x)}$, then*

$$\mathbb{E}\|x^* - x^\pi\|_2^2 \leq \frac{d}{m}.$$

B.3 discrete error control

Lemma B.3 (Approximation error on position of classical HMC). *Suppose p is the associated initial momentum of position q. Under Assumption 4.2,*

$$\|\mathcal{H}_\eta(q) - \bar{\mathcal{H}}_\eta(q)\|^2 \leq \eta^6\Big(\frac{M^2}{10}\|p\|^2 + \frac{M^4\eta^2}{60}\|q - \theta^*\|^2\Big).$$

Proof. By definitions,

$$\mathcal{H}_\eta(q) - \bar{\mathcal{H}}_\eta(q) = q + \eta p - \frac{\eta^2}{2}\nabla f(q) - (q + \int_0^\eta p(s)ds), \qquad \text{with} \qquad p(s) = p - \int_0^s \nabla f(q(u))du$$

Inserting $p(s) = p - \int_0^s \nabla f(q(u))du$ and $\frac{\eta^2}{2}\nabla f(q) = \int_0^\eta \int_0^s \nabla f(q)duds$, we get

$$\mathcal{H}_\eta(q) - \bar{\mathcal{H}}_\eta(q) = \int_0^\eta \int_0^s \nabla f(q(u)) - \nabla f(q)duds$$

By Jensen inequality (i.e., Lemma B.12 with $\lambda \equiv 1$) and M-smoothness assumption of ∇f,

$$\|\mathcal{H}_\eta(q) - \bar{\mathcal{H}}_\eta(q)\|^2 \le \frac{\eta^2}{2} \int_0^\eta \int_0^s M^2\|q(u) - q\|^2 duds. \tag{20}$$

In the following, we are going to apply Lemma B.1 to bound $\|q(u) - q\|^2$. First, by Cauchy–Schwarz inequality and Lemma B.12

$$\begin{aligned}\|q(u) - q\|^2 =& \|up - \int_0^u \int_0^s \nabla f(q(t))dtds\|^2 \le 2\|up\|^2 + 2\|\int_0^u \int_0^s \nabla f(q(t))dtds\|^2 \\ \le& 2\|up\|^2 + u^2 \int_0^u \int_0^s \|\nabla f(q(t))\|^2 dtds\end{aligned}$$

Inserting result from Lemma B.5 that $\|\nabla f(q(t))\|^2 \le 2M^2(\|q - \theta^*\|^2 + t^2\|p\|^2)$[1] into the last inequality, we have

$$\|q(u) - q\|^2 \le (2u^2 + \frac{M^2u^6}{6})\|p\|^2 + M^2u^4\|q - \theta^*\|^2$$

Since $u \le \eta \le 1/(2\sqrt{M})$, we also have

$$\|q(u) - q\|^2 \le \frac{24}{10}u^2\|p\|^2 + M^2u^4\|q - \theta^*\|^2$$

and inserting back to (20) yields the claim.

□

Lemma B.4 (Approximation error on momentum of classical HMC)**.** *Suppose q is the associated initial position of momentum p. Under Assumption 4.2 and $\eta \le 1/(2\sqrt{M})$,*

$$\|\mathcal{H}_\eta(p) - \bar{\mathcal{H}}_\eta(p)\|^2 \le \frac{5}{2}\eta^4M^2\|p\|^2 + \frac{2}{3}\eta^6M^4\|q - \theta^*\|^2$$

Proof. By definitions,

$$\mathcal{H}_\eta(p) - \bar{\mathcal{H}}_\eta(p) = p - \frac{\eta}{2}(\nabla f(q) + \nabla f(\mathcal{H}_\eta(q))) - (p - \int_0^\eta \nabla f(q(s))ds)$$

Inserting $\nabla f(q(s)) = \nabla f(q) + \int_0^s \nabla^2 f(q(u))p(u)du$ into above, we get

$$\mathcal{H}_\eta(p) - \bar{\mathcal{H}}_\eta(p) = \frac{\eta}{2}(\nabla f(q) - \nabla f(\mathcal{H}_\eta(q))) + \int_0^\eta \int_0^s \nabla^2 f(q(u))p(u)duds$$

Applying Lemma B.12, we can get

$$\begin{aligned}\|\mathcal{H}_\eta(p) - \bar{\mathcal{H}}_\eta(p)\|^2 \le& 2\|\frac{\eta}{2}(\nabla f(q) - \nabla f(\mathcal{H}_\eta(q)))\|^2 + 2\|\int_0^\eta \int_0^s \nabla^2 f(q(u))p(u)duds\|^2 \\ \le& 2\|\frac{\eta}{2}(\nabla f(q) - \nabla f(\mathcal{H}_\eta(q)))\|^2 + 2\frac{\eta^2}{2}\int_0^\eta \int_0^s \|\nabla^2 f(q(u))p(u)\|^2 duds \\ =& \frac{\eta^2}{2}\|\nabla f(q) - \nabla f(\mathcal{H}_\eta(q))\|^2 + \eta^2 \int_0^\eta \int_0^s \|\nabla^2 f(q(u))p(u)\|^2 duds.\end{aligned}$$

Since f satisfies M-smoothness assumption, we further have

[1] One needs to be careful that p could be the updated momentum of discrete HMC and may not have normality, losing a lot of nicely properties. Nevertheless, Lemma B.5 holds for non-normally distributed initial momentum.

$$\|\mathcal{H}_\eta(p) - \bar{\mathcal{H}}_\eta(p)\|^2 \le \frac{\eta^2 M^2}{2}\|q - \mathcal{H}_\eta(q)\|^2 + \eta^2 M^2 \int_0^\eta \int_0^s \|p(u)\|^2 du ds \tag{21}$$

To bound the RHS, setting $F = f$ in Lemma B.8 gives

$$\|q - \mathcal{H}_\eta(q)\|^2 \le 2\eta^2\|p\|^2 + \eta^4 M^2 \|q - \theta^*\|^2$$

and at the same time,

$$\begin{aligned}
\|p(u)\|^2 =& \|p - \int_0^u \nabla f(q(s)) ds\|^2 \le 2\|p\|^2 + 2\|\int_0^u \nabla f(q(s)) ds\|^2 \\
\le& 2\|p\|^2 + 2u\int_0^u \|\nabla f(q(s))\|^2 ds \le 2\|p\|^2 + 2u\int_0^u 2M^2(\|q-\theta^*\|^2 + s^2\|p\|^2) ds \\
=& (2 + \frac{u^4 M^2}{3})\|p\|^2 + 4u^2 M^2 \|q-\theta^*\|^2 \le\ 3\|p\|^2 + 4u^2 M^2 \|q-\theta^*\|^2,
\end{aligned}$$

where the first inequality is by Cauthy-Schwarz inequality, the second inequality is by Lemma B.12, the third inequality is by Lemma B.5, and the last inequality is by $u \le \eta \le 1/2\sqrt{M}$.

Inserting these back to (21) yields

$$\|\mathcal{H}_\eta(p) - \bar{\mathcal{H}}_\eta(p)\|^2 \le \frac{5}{2}\eta^4 M^2 \|p\|^2 + \frac{2}{3}\eta^6 M^4 \|q-\theta^*\|^2.$$

This finishes the proof. □

B.4 Lemmas

Lemma B.5. *Suppose p is the associated initial momentum of position q. Under Assumption 4.2*

$$\|\nabla f(q)\|^2 \le M^2\|q-\theta^*\|^2, \qquad \|\nabla f(q(t))\|^2 \le 2M^2(\|q-\theta^*\|^2 + t^2\|p\|^2)$$

Proof. Noting that $\|\nabla f(q(t))\|^2 = \|\nabla f(q(t)) - \nabla f(\theta^*)\|^2 \le M^2\|q(t) - \theta^*\|^2$, and by Lemma B.1

$$\|q(t) - \theta^*\| \le \|q - \theta^*\| + t\|p - 0\|.$$

Noting that f satisfies M-smoothness assumption

$$\|\nabla f(q(t))\|^2 \le M^2\|q(t) - \theta^*\|^2 \le M^2(\|q-\theta^*\| + t\|p\|)^2.$$

We finish the proof by applying Cauchy–Schwarz inequality on the RHS. □

Lemma B.6. *For any $q, p, q^\pi \in \mathbb{R}^d$, let $q_k^{(c)} = \widetilde{h}_{\mathrm{LF}}(f^{(c)}, q, \nabla f(q), p_0, \eta, k)$, $q_k = \sum_{c=1}^N w_c q_k^{(c)}$ and $q_k^\pi = \bar{\mathcal{H}}_\eta^k(q^\pi)$. Under the smoothness Assumption 4.2, we have*

$$\sum_{c=1}^N w_c \mathbb{E}\|q_k^{(c)} - q_k\|^2 \le 2k^3\eta^4 \sum_{i=1}^{k-1}(16M^2\mathbb{E}\|q_i^\pi - q_i\|^2 + 16M^2\mathbb{E}\|q^\pi - q\|^2 + 4(k\eta)^2 H)$$

where $H := \max_{i\le K}\sum_{c=1}^N w_c \mathbb{E}\|\frac{1}{i\eta}(\nabla f^{(c)}(q_i^\pi) - \nabla f^{(c)}(q^\pi) + \nabla f(q^\pi) - \nabla f(q_i^\pi))\|^2$ is a quantity measuring the federated heterogeneity.

Proof. To prove this lemma, we use mathematical induction to remove the discrete heterogeneous term $\sum_{c=1}^N w_c\mathbb{E}\|q_k^{(c)} - q_k\|^2$. In specific, we are going to use induction to prove that there are constants C_1, C_2 (to be determined later) such that

$$\sum_{c=1}^N w_c\mathbb{E}\|q_j^{(c)} - q_j\|^2 \le (j-1)^3\eta^4 \sum_{k=1}^{j-1}(C_1 M^2\mathbb{E}\|q_k^\pi - q_k\|^2 + C_1 M^2\mathbb{E}\|q^\pi - q\|^2 + C_2(k\eta)^2 H). \tag{22}$$

First, for $j = 1$, noting that $q_1^{(c)} = q + p - \frac{\eta}{2}(\nabla f^{(c)}(q) - \nabla f^{(c)}(q) + \nabla f(q))$ is constant for all c, thus

$$\sum_{c=1}^N w_c\mathbb{E}\|q_j^{(c)} - q_j\|^2 = 0, \qquad \text{when } j = 1.$$

Now, assume that (22) holds for $j \le k$. Next, we check it also holds for $k+1$ based on the assumption. By Lemma B.7, we have

$$\begin{aligned}
&\sum_{c=1}^{N} w_c \mathbb{E}\|q_{k+1}^{(c)} - q_{k+1}\|^2 \\
\le& 2(k+1)^3\eta^4 \sum_{i=1}^{k}\Big(4M^2 \sum_{c=1}^{N} w_c \mathbb{E}\|q_i^{(c)} - q_i\|^2 + 4M^2\mathbb{E}\|q_i^{\pi} - q_i\|^2 + 4M^2\mathbb{E}\|q^{\pi} - q\|^2 + (i\eta)^2 H\Big) \\
\le& 2(k+1)^3\eta^4 \sum_{i=1}^{k}\Big(4M^2 i^3\eta^4 \sum_{j=1}^{i-1}(C_1 M^2\mathbb{E}\|q_j^{\pi} - q_j\|^2 + C_1 M^2\mathbb{E}\|q^{\pi} - q\|^2 + C_2(j\eta)^2 H)\Big) \\
&+ 2(k+1)^3\eta^4 \sum_{i=1}^{k}\Big(4M^2\mathbb{E}\|q_i^{\pi} - q_i\|^2 + 4M^2\mathbb{E}\|q^{\pi} - q\|^2 + (i\eta)^2 H\Big)
\end{aligned}$$

where the second inequality is by our induction assumption for $i \le k$.

Noting that for any sequence $\{a_j\}$,

$$\sum_{i=1}^{k} i^3 \sum_{j=1}^{i-1} a_j = \sum_{j=1}^{k-1}\sum_{i=j+1}^{k} i^3 a_j \le \sum_{j=1}^{k-1} k^4 a_j$$

Therefore, we further have

$$\begin{aligned}
\sum_{c=1}^{N} w_c \mathbb{E}\|q_{k+1}^{(c)} - q_{k+1}\|^2 \le& (k+1)^3\eta^4 \sum_{j=1}^{k}\Big((8M^4 C_1 k^4\eta^4 + 8M^2)\mathbb{E}\|q_j^{\pi} - q_j\|^2 \\
&+ (8M^4 C_1 k^4\eta^4 + 8M^2)\mathbb{E}\|q^{\pi} - q\|^2 + (8M^2 C_2 k^4\eta^4 + 2)H\Big)
\end{aligned}$$

By choosing C_1, C_2 such that

$$8M^4 C_1 k^4\eta^4 + 8M^2 \le C_1 M^2, \qquad 8M^2 C_2 k^4\eta^4 + 2 \le C_2,$$

we can finish the induction. Recall that $K\eta$ is set to be $\le 1/(2\sqrt{M})$. One choice is that $C_1 = 16, C_2 = 4$. This finishes the proof. □

Lemma B.7. *Under the same notations and conditions of that of Lemma B.6,*

$$\sum_{c=1}^{N} w_c \mathbb{E}\|q_j^{(c)} - q_j\|^2 \le 2j^3\eta^4 \sum_{k=1}^{j-1}\Big(4M^2 \sum_{c=1}^{N} w_c \mathbb{E}\|q_k^{(c)} - q_k\|^2 + 4M^2\mathbb{E}\|q_k^{\pi} - q_k\|^2 + 4M^2\mathbb{E}\|q^{\pi} - q\|^2 + (k\eta)^2 H\Big)$$

Proof. By definitions,

$$\begin{aligned}
q_j^{(c)} - q_j =& q + j\eta p_0 - \frac{(j\eta)^2}{2}\nabla f(q) - \eta^2 \sum_{k=1}^{j-1}\sum_{i=1}^{k}(\nabla f^{(c)}(q_i^{(c)}) - \nabla f^{(c)}(q)) \\
&- \sum_{c=1}^{N} w_c\Big(q + j\eta p_0 - \frac{(j\eta)^2}{2}\nabla f(q) - \eta^2 \sum_{k=1}^{j-1}\sum_{i=1}^{k}(\nabla f^{(c)}(q_i^{(c)}) - \nabla f^{(c)}(q))\Big) \\
=& \eta^2 \sum_{k=1}^{j-1}\sum_{i=1}^{k}\Big(\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q)\Big).
\end{aligned}$$

By Lemma B.12, we have

$$\begin{aligned}
\|q_j^{(c)} - q_j\|^2 \le& \frac{j(j-1)}{2}\sum_{k=1}^{j-1}\sum_{i=1}^{k}\|\eta^2 \sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q)\|^2 \\
=& \frac{j(j-1)}{2}\sum_{i=1}^{j-1}\sum_{k=i}^{j-1}\eta^4\|\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q)\|^2
\end{aligned}$$

It follows that

$$\|q_j^{(c)} - q_j\|^2 \leq \frac{j^3\eta^4}{2}\sum_{i=1}^{j-1}\|\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q)\|^2. \quad (23)$$

The error $\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q))$ can be split into several part as follow

$$\begin{aligned}
&\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q) \\
=&\underbrace{\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q_i)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q_i)}_{\text{discrete heterogeneity}} \\
&+\underbrace{\nabla f^{(c)}(q_i^{\pi}) - \nabla f^{(c)}(q_i) + \nabla f(q_i) - \nabla f(q_i^{\pi})}_{\text{convergence error}} \\
&+\underbrace{\nabla f^{(c)}(q) - \nabla f^{(c)}(q^{\pi}) + \nabla f(q^{\pi}) - \nabla f(q)}_{\text{convergence error}} \\
&+\underbrace{\nabla f^{(c)}(q^{\pi}) - \nabla f^{(c)}(q_i^{\pi}) + \nabla f(q_i^{\pi}) - \nabla f(q^{\pi})}_{\text{oracle heterogeneity}}
\end{aligned}$$

Again, by Lemma B.12,

$$\|\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q)) - \nabla f^{(c)}(q_i^{(c)}) + \nabla f^{(c)}(q)\|^2 \leq 4(A^{(c)} + B^{(c)} + C^{(c)} + D^{(c)}) \quad (24)$$

where

$$\begin{aligned}
A^{(c)} :&\|\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q_i)) - \nabla f^{(\widetilde{c})}(q_i^{(c)}) + \nabla f^{(c)}(q_i)\|^2, \\
B^{(c)} :&\|\nabla f^{(c)}(q_i^{\pi}) - \nabla f^{(c)}(q_i) + \nabla f(q_i) - \nabla f(q^{\pi})\|^2 \\
C^{(c)} :&\|\nabla f^{(c)}(q) - \nabla f^{(c)}(q^{\pi}) + \nabla f(q^{\pi}) - \nabla f(q)\|^2, \\
D^{(c)} :&\|\nabla f^{(c)}(q^{\pi}) - \nabla f^{(c)}(q_i^{\pi}) + \nabla f(q_i^{\pi}) - \nabla f(q^{\pi})\|^2.
\end{aligned}$$

Consider $A^{(c)}$ first, observe that

$$\begin{aligned}
&\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\nabla f^{(\widetilde{c})}(q_i^{(\widetilde{c})}) - \nabla f^{(\widetilde{c})}(q_i)) - \nabla f^{(\widetilde{c})}(q_i^{(c)}) + \nabla f^{(c)}(q_i) \\
=&\sum_{\widetilde{c}=1}^{N} \frac{w_{\widetilde{c}} - \mathbb{I}(\widetilde{c}=c)}{2(1-w_c)} \cdot 2(1-w_c)\Big(\nabla f^{(\widetilde{c})}(q_k^{(c)}) - \nabla f^{(\widetilde{c})}(q_k)\Big)
\end{aligned}$$

and that $\sum_{\widetilde{c}=1}^{N} \frac{|w_{\widetilde{c}} - \mathbb{I}(\widetilde{c}=c)|}{2(1-w_c)} = 1$, by Lemma B.12

$$\begin{aligned}
\sum_{c=1}^{N} w_c A^{(c)} \leq& \sum_{c=1}^{N} w_c \sum_{\widetilde{c}=1}^{N} |w_c - \mathbb{I}(\widetilde{c}=c)|2(1-w_c)\|\nabla f^{(\widetilde{c})}(q_k^{(c)}) - \nabla f^{(\widetilde{c})}(q_k)\|^2 \\
=&2\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}(\underbrace{\sum_{c\neq\widetilde{c}}^{N} w_c(1-w_c)}_{:\leq\sum_{c=1}^{N} w_c=1} + (1-w_{\widetilde{c}})^2)\|\nabla f^{(\widetilde{c})}(q_k^{(c)}) - \nabla f^{(\widetilde{c})}(q_k)\|^2 \\
\leq& 4\sum_{\widetilde{c}=1}^{N} w_{\widetilde{c}}\|\nabla f^{(\widetilde{c})}(q_k^{(c)}) - \nabla f^{(\widetilde{c})}(q_k)\|^2
\end{aligned}$$

Further, by smoothness assumption,

$$\sum_{c=1}^{N} w_c A^{(c)} \leq 4 \sum_{c=1}^{N} w_c M^2 \|q_k^{(c)} - q_k\|^2 \tag{25}$$

Similarly, for $B^{(c)}$ and $C^{(c)}$, we have

$$\begin{aligned} \sum_{c=1}^{N} w_c B^{(c)} \leq & 4 \sum_{c=1}^{N} w_c M^2 \|q_k^{\pi} - q_k\|^2 = 4M^2 \|q_k^{\pi} - q_k\|^2 \\ \sum_{c=1}^{N} w_c C^{(c)} \leq & 4 \sum_{c=1}^{N} w_c M^2 \|q^{\pi} - q\|^2 = 4M^2 \|q^{\pi} - q\|^2 \end{aligned} \tag{26}$$

For $D^{(c)}$, by definitions, we directly have

$$\sum_{c=1}^{N} w_c \mathbb{E} D^{(c)} \leq (k\eta)^2 H \tag{27}$$

Combining (23)-(27), we have

$$\sum_{c=1}^{N} w_c \mathbb{E} \|q_j^{(c)} - q_j\|^2 \leq 2j^3 \eta^4 \sum_{k=1}^{j-1} \Big(4M^2 \sum_{c=1}^{N} w_c \mathbb{E} \|q_k^{(c)} - q_k\|^2 + 4M^2 \mathbb{E} \|q_k^{\pi} - q_k\|^2 + 4M^2 \mathbb{E} \|q^{\pi} - q\|^2 + (k\eta)^2 H \Big)$$

This finishes the proof. □

Lemma B.8 (0-th order Approximation Bound)**.** *Let $q_0, p_0 \in \mathbb{R}^d$ be two vectors. For any $F(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ satisfying smoothness Assumption 4.2, then $q_K = \widetilde{h}_{\mathrm{LF}}(F, q_0, \nabla f(q_0), p_0, \eta, k)$ satisfies*

$$\|q_k - q_0\|^2 \leq 2(k\eta)^2 \|p_0\|^2 + (k\eta)^4 M^2 \|q_0 - \theta^*\|^2 \tag{28}$$

$$\|\nabla F(q_k) - \nabla F(q_0)\|^2 \leq 2M^2 (k\eta)^2 \|p_0\|^2 + (k\eta)^4 M^4 \|q_0 - \theta^*\|^2 \tag{29}$$

Note that we actually prove 30, which is an interesting result in the sense that the k-th leapfrog step can be bounded by the shift vector $\nabla f(q_0)$ rather than $\nabla F(q_0)$.

Proof. We are going to prove

$$\|q_k - q_0\|^2 \leq 2(k\eta)^2 \|p_0\|^2 + (k\eta)^4 \|\nabla f(q_0)\|^2. \tag{30}$$

We prove claim (30) by induction. Then the claim (28) and (29) follows directly from (30) and M-smoothness Assumption 4.2 applied with $\nabla f(\theta^*) = 0$.

First, according to the definitions and Algorithm 1, we can write the full expression of q_k as follow

$$q_k = q_0 + k\eta p_0 - \frac{(k\eta)^2}{2} \nabla f(q_0) - \eta^2 \sum_{j=1}^{k-1} (k-j)(\nabla F(q_j) - \nabla F(q_0))$$

which is proved by Lemma B.9.

For claim (30) with $k = 1$, by Cauchy–Schwarz inequality, we have that

$$\|q_1 - q_0\|^2 = \|\eta p_0 - \frac{\eta^2}{2} \nabla f(q_0)\|^2 \leq 2\|\eta p_0\|^2 + 2\|\frac{\eta^2}{2} \nabla f(q_0)\|^2$$

Next, we assume that the claim (30) holds for $j = 1, 2, \ldots, k-1$. By direct calculations, for any $c_1, c_2, c_3 > 0$ such that $\sum_i c_i = 1$ we have

$$
\begin{aligned}
&\|q_k - q_0\|^2 \\
\leq &\frac{1}{c_1}(k\eta)^2\|p_0\|^2 + \frac{(k\eta)^4}{4c_2}\|\nabla f(q_0)\|^2 + \sum_{j=1}^{k-1}\frac{(k-j)\eta^4k^3}{6jc_3}\|\nabla F(q_j) - \nabla F(q_0)\|^2 \\
\leq &\frac{1}{c_1}(k\eta)^2\|p_0\|^2 + \frac{(k\eta)^4}{4c_2}\|\nabla f(q_0)\|^2 + \sum_{j=1}^{k-1}\frac{(k-j)\eta^4k^3M^2}{6jc_3}\|q_j - q_0\|^2 \\
\leq &\frac{1}{c_1}(k\eta)^2\|p_0\|^2 + \frac{(k\eta)^4}{4c_2}\|\nabla f(q_0)\|^2 + \sum_{j=1}^{k-1}\frac{(k-j)\eta^4k^3M^2}{6jc_3}\cdot(2(j\eta)^2\|p_0\|^2 + (j\eta)^4\|\nabla f(q_0)\|^2) \\
\leq &(\frac{1}{c_1} + \frac{M^2}{6c_3}\frac{2(k\eta)^4}{6})(k\eta)^2\|p_0\|^2 + (\frac{1}{4c_2} + \frac{M^2}{6c_3}\frac{(k\eta)^4}{20})(k\eta)^4\|\nabla f(q_0)\|^2
\end{aligned}
$$

where the first inequality is due to Lemma B.12 and noting that $c_1 + c_2 + \frac{\sum_{j=1}^{k-1}(k-j)j}{(k^3-k)/6}c_3 = 1$; the second inequality is due to smoothness Assumption 4.2; the third inequality holds by the induction assumption; the fourth inequality is by $\sum_{j=1}^{k-1}(k-j)j = \frac{k^3-k}{6} \leq \frac{k^3}{6}$ and $\sum_{j=1}^{k-1}(k-j)j^3 = \frac{k^5}{20} - \frac{k^3}{12} + \frac{1}{20}k \leq \frac{k^5}{20}$. The completion of the induction follows by step size assumption $K\eta \leq 1/\sqrt{M}$ and setting $(c_1, c_2, c_3) = (\frac{3}{5}, \frac{1}{3}, \frac{1}{15})$ in the last inequality. Therefore, we conclude the proof of (30). □

Lemma B.9. *For $q_k = \widetilde{h}_{\mathrm{LF}}(F, q_0, \nabla f(q_0), p_0, \eta, k)$ and $p_k = p_{k-1} - \eta(\nabla F(q_{k-1}) + \nabla F(q_k))/2 + \eta(\nabla F(q_0) - \nabla f(q_0))$, we can rewrite as*

$$q_k = q_0 + k\eta p_0 - \frac{(k\eta)^2}{2}\nabla f(q_0) - \eta^2\sum_{i=1}^{k-1}(k-i)(\nabla F(q_j) - \nabla F(q_0)) \tag{31}$$

$$p_k = p_0 - k\eta\nabla f(q_0) - \eta\sum_{j=1}^{k-1}(\nabla F(q_j) - \nabla F(q_0)) - \frac{\eta}{2}(\nabla F(q_k) - \nabla F(q_0)). \tag{32}$$

Proof. For simplicity of notations, we use $\widetilde{f}(\cdot) := F(\cdot) - F(q_0) + f(q_0)$. According to Algorithm 1,

$$q_k = q_{k-1} + \eta p_{k-1} - \frac{\eta^2}{2}\nabla\widetilde{f}(q_{k-1}), \qquad p_k = p_{k-1} - \eta(\nabla\widetilde{f}(q_{k-1}) + \nabla\widetilde{f}(q_k))/2 \tag{33}$$

Summing from $k = 1, \ldots, k$ on both side of the second equation of (33), we can get

$$p_k = p_0 - \frac{\eta}{2}\nabla\widetilde{f}(q_0) - \eta\sum_{j=1}^{k-1}\nabla\widetilde{f}(q_j) - \frac{\eta}{2}\nabla\widetilde{f}(q_k).$$

This proves (32). Inserting this into the first equality of (33), we can get

$$q_k = q_{k-1} + p_0 - \frac{\eta^2}{2}\nabla\widetilde{f}(q_0) - \eta^2\sum_{j=1}^{k-1}\nabla\widetilde{f}(q_j)$$

Summing from $k = 1, \ldots, k$ on both side yields

$$q_k = q_0 + k\eta p_0 - \frac{(k\eta)^2}{2}\nabla f(q_0) - \eta^2\sum_{i=1}^{k-1}(k-i)(\nabla F(q_j) - \nabla F(q_0))$$

which proves (31). □

Lemma B.10. *Under the same notations of Theorem 4.3 and Assumption 4.2, for $K\eta \le 1/(2\sqrt{M})$, we have*

$$\|p_{t,k}\|^2 \le 5\|p_t\|^2 + 5M^2(k\eta)^2\|\theta_t - \theta^*\|^2$$
$$\|\theta_{t,k} - \theta^*\|^2 \le 4(k\eta)^2\|p_t\|^2 + 3\|\theta_t - \theta^*\|^2.$$

Proof. Recall that $p_{t,k} := \sum_{c=1}^N w_c p_{t,k}^{(c)}$, and note that by Lemma B.11 that $\|p_{t,k}^{(c)}\|^2 \le 5\|p_t\|^2 + 5M^2(k\eta)^2\|\theta_t - \theta^*\|^2$, we have

$$\|p_{t,k}\|^2 \le \sum_{c=1}^N w_c\|p_{t,k}^{(c)}\|^2 \le 5\|p_t\|^2 + 5M^2(k\eta)^2\|\theta_t - \theta^*\|^2.$$

This proves the first claim.

For the second claim, first, by convexity of $\|\cdot\|^2$ and Lemma B.12

$$\|\theta_{t,k} - \theta^*\|^2 \le 2\|\theta_{t,k} - \theta_t\|^2 + 2\|\theta_t - \theta^*\|^2 \le \sum_{c=1}^N w_c 2\|\theta_{t,k}^{(c)} - \theta_t\|^2 + 2\|\theta_t - \theta^*\|^2.$$

At the same time, for $\|\theta_{t,k}^{(c)} - \theta_t\|^2$, Lemma B.8 states that $\|\theta_{t,k}^{(c)} - \theta_t\|^2 \le 2(k\eta)^2\|p_t\|^2 + (k\eta)^4M^2\|\theta_t - \theta^*\|^2$. And so

$$\|\theta_{t,k} - \theta^*\|^2 \le 4(k\eta)^2\|p_t\|^2 + (2 + 2(k\eta)^4M^2)\|\theta_t - \theta^*\|^2 \le 4(k\eta)^2\|p_t\|^2 + 3\|\theta_t - \theta^*\|^2$$

where the last inequality is due to $k\eta \le 1/(2\sqrt{M})$. This finishes the proof. □

Lemma B.11. *Suppose $q_k = \widetilde{h}_{\mathrm{LF}}(F, q_0, \nabla f(q_0), p_0, \eta, k)$ and $p_k = p_{k-1} - \eta(\nabla F(q_{k-1}) + \nabla F(q_k))/2 + \eta(\nabla F(q_0) - \nabla f(q_0))$. Under Assumption 4.2, for $K\eta \le 1/(2\sqrt{M})$, we have*

$$\|p_j\|^2 \le 5\|p_0\|^2 + 5M^2(j\eta)^2\|q_0 - \theta^*\|^2.$$

Proof. By definitions and Algorithm 1, we have

$$p_k = p_0 - k\eta\nabla f(q_0) - \eta\sum_{j=1}^{k-1}(\nabla F(q_j) - \nabla F(q_0)) - \frac{\eta}{2}(\nabla F(q_k) - \nabla F(q_0))$$

By Cauchy–Schwarz inequality and M-smoothness assumption,

$$\begin{aligned}
\|p_k\|^2 \le& 4\|p_0\|^2 + 4\|k\eta\nabla f(q_0)\|^2 + 4\|\eta\sum_{j=1}^{k-1}(\nabla F(q_j) - \nabla F(q_0))\|^2 + 4\|\frac{\eta}{2}(\nabla F(q_k) - \nabla F(q_0))\|^2\\
\le& 4\|p_0\|^2 + 4\|k\eta\nabla f(q_0)\|^2 + 4(k-1)\sum_{j=1}^{k-1}\eta^2\|\nabla F(q_j) - \nabla F(q_0)\|^2 + 4\|\frac{\eta}{2}(\nabla F(q_k) - \nabla F(q_0))\|^2\\
\le& 4\|p_0\|^2 + 4(k\eta)^2\|\nabla f(q_0)\|^2 + 4(k-1)\sum_{j=1}^{k-1}\eta^2M^2\|q_j - q_0\|^2 + \eta^2M^2\|q_k - q_0\|^2\\
\le& 4\|p_0\|^2 + 4(k\eta)^2\|\nabla f(q_0)\|^2 + 4(k\eta)^2M^2\max_{j\le k}\|q_j - q_0\|^2
\end{aligned}$$

where the first two inequalities are by Lemma B.12 and the third inequality is by M-smoothness assumption.

Note that Lemma B.8 gives that $\|q_j - q_0\|^2 \le 2(j\eta)^2\|p_0\|^2 + (j\eta)^4M^2\|q_0 - \theta^*\|^2$. Inserting this into the above inequality yields

$$\|p_k\|^2 \le 4(1 + 2(k\eta)^4M^2)\|p_0\|^2 + 4(k\eta)^2(1 + M^2(k\eta)^4)M^2\|q_0 - \theta^*\|^2.$$

The second claim follows by noting that $k\eta \le 1/(2\sqrt{M})$. □

Lemma B.12 (Jensen's inequality). *For any vectors $\{x_i\}_{i=1}^n$ and positive constants $\{\lambda_i\}_{i=1}^n$ with $\sum \lambda_i = 1$, due to the convexity of $\|\cdot\|^2$, we have*

$$\|\sum_i \lambda_i x_i\|^2 \le \sum_i \lambda_i \|x_i\|^2;$$
$$\|\sum_i x_i\|^2 \le \sum_i \lambda_i \|x_i/\lambda_i\|^2 = \sum_i \frac{1}{\lambda_i}\|x_i\|^2.$$

For any vector-value function $x(t)$ and non-negative integrable function $\lambda(t)$,

$$\|\int \lambda(t)x(t)dt\|^2 \le \int \lambda(t)dt \int \lambda(t)\|x(t)\|^2 dt.$$
$$\|\int \lambda(t)x(t)dt\|^2 \le \int \lambda(t)\|x(t)\|^2 dt, \quad \textit{if } \int \lambda(t)dt = 1.$$

C Study on heterogeneity metric H

C.1 Bound on H

Recall that $H := \max_{i\le K} \sum_{c=1}^N w_c \mathbb{E}\|\frac{1}{i\eta}(\nabla f^{(c)}(\bar{\mathcal{H}}_\eta^i(\theta^\pi)) - \nabla f^{(c)}(\theta^\pi) + \nabla f(\theta^\pi) - \nabla f(\bar{\mathcal{H}}_\eta^i(\theta^\pi)))\|^2$ By the M-Lipschitz condition,

$$H \le 4 \max_{1\le i\le K} \sum_{c=1}^N w_c \mathbb{E}\left\|\frac{M}{i\eta}\left(\bar{\mathcal{H}}_\eta^i(\theta^\pi) - \theta^\pi\right)\right\|^2.$$

Moreover,

$$\bar{\mathcal{H}}_\eta^i(\theta^\pi) - \theta^\pi = \int_0^{i\eta} p(t)\,dt.$$

where $p(t)$ has a stationary marginal multivariate normal distribution. Overall, we can obtain that

$$H \le CM^2 d$$

for some absolute constant C.

C.2 Solutions of H on Gaussian examples

Give normal clients with same covariance structure (e.g., $N(1, I)$ and $N(-1, I)$, and the global is $N(0, I)$), then

$$\nabla f^{(c)}(\theta_2) - \nabla f^{(c)}(\theta_1) + \nabla f(\theta_1) - \nabla f(\theta_2) \equiv 0,$$

meaning that $H \equiv 0$ always.

For non-equal covariance case, e.g., $N(0, 3I/4)$ and $N(0, 3I/2)$, and the global is $N(0, I)$, the HMC trajectory for the global has a closed form, i.e., $\theta_t = \theta_0 \cos(t) + p_0 \sin(t)$ elementwisely. The calculation of H eventually reduces to

$$d \times \sup_t \mathbb{E}_{\theta,p\sim N(0,I^2)} \left|\frac{\theta(1-\cos(t)) - p\sin(t)}{t}\right|^2 \Big/ 9 = \frac{2d}{9} \times \sup_t \frac{1-\cos(t)}{t^2} = \frac{d}{9}.$$

D Experimental Details

D.1 Additional Details for Figure 2

Table 2 reports the stepsize and leapfrog grids used for the simulated Bayesian logistic regression experiment in Figure 2. The same grid is used for both panels. FA-LD uses $T = 10$ and $K = 1$ for every stepsize. FA-HMC uses $T = 10$ and the K values listed in the table. AFHMC uses the candidate leapfrog counts $K_{\mathrm{A}} \in \mathcal{K}_{\mathrm{HMC}}(\eta) \cup 10\mathcal{K}_{\mathrm{HMC}}(\eta)$. For each $(\mathrm{algorithm}, \eta)$ pair, Figure 2 reports the best-performing K from this grid.

Table 2: Candidate grid for the simulated Bayesian logistic regression experiment. Here $\mathcal{K}_{\mathrm{HMC}}(\eta)$ denotes the FA-HMC leapfrog grid at stepsize η; the AFHMC grid is $\mathcal{K}_{\mathrm{HMC}}(\eta) \cup 10\mathcal{K}_{\mathrm{HMC}}(\eta)$.

Stepsize η	FA-LD (T, K)	FA-HMC $(T, \mathcal{K}_{\mathrm{HMC}})$	AFHMC $\mathcal{K}_{\mathrm{A}}$
5×10^{-4}	$(10, \{1\})$	$(10, \{500, 2000\})$	$\{500, 2000, 5000, 20000\}$
10^{-3}	$(10, \{1\})$	$(10, \{300, 1000\})$	$\{300, 1000, 3000, 10000\}$
2×10^{-3}	$(10, \{1\})$	$(10, \{150, 500\})$	$\{150, 500, 1500, 5000\}$
5×10^{-3}	$(10, \{1\})$	$(10, \{50, 200\})$	$\{50, 200, 500, 2000\}$
10^{-2}	$(10, \{1\})$	$(10, \{30, 100\})$	$\{30, 100, 300, 1000\}$
2×10^{-2}	$(10, \{1\})$	$(10, \{10, 50\})$	$\{10, 50, 100, 500\}$
5×10^{-2}	$(10, \{1\})$	$(10, \{2, 5, 10, 20\})$	$\{2, 5, 10, 20, 50, 100, 200\}$
10^{-1}	$(10, \{1\})$	$(10, \{1, 2, 5, 10\})$	$\{1, 2, 5, 10, 20, 50, 100\}$
2×10^{-1}	$(10, \{1\})$	$(10, \{1, 2, 5\})$	$\{1, 2, 5, 10, 20, 50\}$
5×10^{-1}	$(10, \{1\})$	$(10, \{1, 2, 5\})$	$\{1, 2, 5, 10, 20, 50\}$
1	$(10, \{1\})$	$(10, \{1, 2\})$	$\{1, 2, 10, 20\}$

Table 3: First communication round at which AFHMC reaches $\mathcal{W}_2^2 < 0.1$ on the low-heterogeneity Gaussian target, with $K(d) = \lfloor 50(d/100)^{1/2} \rceil$ and $\eta(d) = 0.006(d/100)^{-1/2}$.

d	$K(d)$	$\eta(d)$	Rounds (seeds 0, 1, 2)	Mean $\pm$ sd
2	7	0.04243	161, 153, 149	154.33 ± 6.11
50	35	0.008485	213, 212, 212	212.33 ± 0.58
100	50	0.006	220, 220, 220	220.00 ± 0.00
150	61	0.004899	229, 229, 229	229.00 ± 0.00
200	71	0.004243	231, 231, 230	230.67 ± 0.58
250	79	0.003795	237, 237, 237	237.00 ± 0.00
300	87	0.003464	238, 238, 237	237.67 ± 0.58
350	94	0.003207	240, 240, 240	240.00 ± 0.00
400	100	0.003	245, 245, 245	245.00 ± 0.00
450	106	0.002828	247, 248, 247	247.33 ± 0.58
500	112	0.002683	248, 248, 248	248.00 ± 0.00

D.2 Additional Details for the Gaussian Dimension-Scaling Experiments

Low-heterogeneity target. For each dimension d and each of the $N = 10$ clients, we draw 10^5 samples from $\mathcal{N}(16.2\mathbf{1}_d, 1.6I_d)$ and set the local target of client c to $\mathcal{N}(\hat{\mu}_c, \hat{\sigma}_c^2 I_d)$, where $\hat{\mu}_c$ is the sample mean and $\hat{\sigma}_c^2$ is the coordinate-averaged sample variance. The global target is fitted in the same way to the samples pooled over all clients, and $\mathcal{W}_2^2$ is computed as in the heterogeneous experiment. The heterogeneity therefore only comes from sampling error: across all dimensions and seeds, the root-mean-square deviation of the client means from the global mean is between 3.6×10^{-3} and 4.5×10^{-3} per coordinate, compared with 11.1 in the heterogeneous experiment, where the client means are 20 and 1 and the global mean is 16.2. AFHMC runs one trajectory of $K(d)$ surrogate leapfrog steps per round and uses the stopping rule of Section 5 ($\mathcal{W}_2^2 < 0.1$ for 20 consecutive checks, at most 1000 rounds). The random seed controls both the client samples and the HMC noise. Table 3 reports the first-crossing rounds for every seed.

FA-HMC baselines at an equal number of server exchanges. Theorem 4.1 of Liang et al. (2025) shows that, under strong convexity, Hessian smoothness and bounded stochastic-gradient variance, FA-HMC attains $\mathcal{W}_2 \le \epsilon$ when

$$\eta^2 = C \min\left\{ \frac{1}{K^2 L}, \frac{\epsilon}{K^2 \sqrt{d}\, T}, \frac{\epsilon^2}{K^2 d T^2 (1-\rho) N}, \frac{\epsilon^2}{K d \sum_{c=1}^N w_c^2 \sigma_g^2} \right\}.$$

In our Gaussian experiment all clients share the momentum, so $\rho = 1$, and the gradients are exact, so $\sigma_g^2 = 0$; the last two terms therefore impose no constraint. The remaining dimension-dependent term gives $\eta \propto (K\sqrt{T} d^{1/4})^{-1}$. With $T = 10$ and $K_{\mathrm{F}}(d) = \max\{1, \lfloor K(d)/10 \rceil\} \propto d^{1/3}$, this yields $\eta(d) \propto d^{-7/12}$, which we anchor at $d = 100$ as $\eta(d) = 0.006(d/100)^{-7/12}$. The second schedule is the dimension-scaling setting of Liang et al. (2025), $K_{\mathrm{F}} = 5$, $T = 10$ and $\eta(d) = 0.02 d^{-1/4}$.

Table 4: FA-HMC $\mathcal{W}_2^2$ on the heterogeneous Gaussian target. $R(d)$ is the AFHMC first-crossing round for $\mathcal{W}_2^2 < 0.1$. The original setting uses the AFHMC stepsize for $R(d)$ rounds; the two FA-HMC-specific schedules are run for $2R(d)$ rounds. All settings use $T = 10$; the original and Theorem-4.1 settings use $K_{\mathrm{F}}(d)$, and the second setting uses $K_{\mathrm{F}} = 5$.

				Original ($R(d)$)		K_{F}=5 ($2R(d)$)		Thm. 4.1 ($2R(d)$)	
d	$R(d)$	$2R(d)$	$K_{\mathrm{F}}(d)$	η	$\mathcal{W}_2^2$	η	$\mathcal{W}_2^2$	η	$\mathcal{W}_2^2$
2	42	84	1	0.04243	327.9	0.01682	13.0	0.05878	85.7
50	162	324	4	0.008485	4092.3	0.007521	752.1	0.00899	960.2
100	220	440	5	0.006	7621.3	0.006325	1682.8	0.006	2213.9
150	263	526	6	0.004899	9523.9	0.005715	2694.0	0.004736	2774.3
200	294	588	6	0.004243	15968.2	0.005318	3911.7	0.004005	6298.0
250	320	640	7	0.003795	16013.5	0.00503	5236.5	0.003516	5827.7
300	347	694	7	0.003464	22009.6	0.004806	6444.4	0.003161	9424.0
350	368	736	8	0.003207	20227.2	0.004624	7867.1	0.002889	7877.5
400	393	786	8	0.003	25525.1	0.004472	9014.2	0.002673	11122.8
450	403	806	8	0.002828	32540.4	0.004342	11000.1	0.002495	15877.5
500	430	860	8	0.002683	38053.9	0.004229	11872.9	0.002346	19755.7

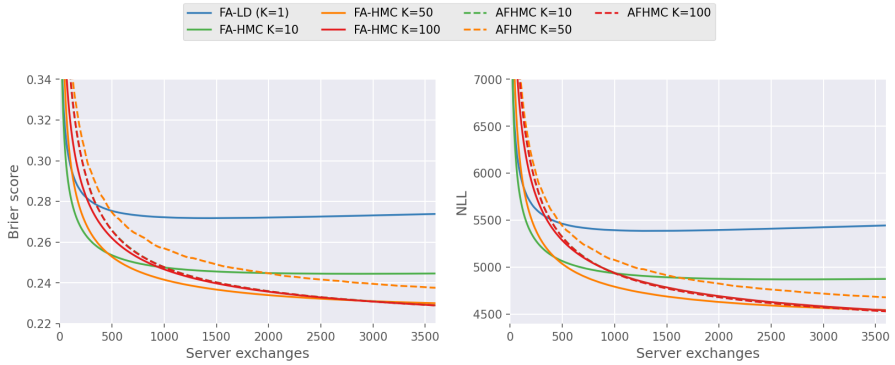

Figure 5: Additional predictive-quality metrics for Bayesian logistic regression on heterogeneous FMNIST against the number of server exchanges (one per round for FA-LD and FA-HMC, two per round for AFHMC). Left: Brier score. Right: negative log-likelihood. The data are partitioned across $N = 10$ clients using a label-skewed Dirichlet split with concentration parameter $\alpha = 0.1$.

Both schedules are run for $2R(d)$ rounds, where $R(d)$ is the AFHMC first-crossing round, so that FA-HMC uses the same number of server exchanges as AFHMC. Table 4 reports the resulting errors together with the original matched-budget setting. These schedules are theory-guided choices rather than an exhaustive search over (η, K, T), and each setting is run once.

D.3 Additional Details for MNIST Logistic Regression

To create heterogeneous federated learning scenario, the FMNIST training set is partitioned across $N = 10$ clients using a label-skewed Dirichlet split. Specifically, for each class c, we sample a client-proportion vector

$$p_c \sim \mathrm{Dirichlet}(\alpha \mathbf{1}_N), \qquad \alpha = 0.1,$$

and distribute the examples of class c across clients according to p_c. This produces heterogeneous clients with different label distributions.

Each configuration is repeated over five random seeds. The reported values in Table 5 are the final mean and standard error (SE) across seeds at 3600 server exchanges, i.e., at round 3600 for FA-LD and FA-HMC and at round 1800 for AFHMC. Since the stepsize is tuned separately for each metric, the selected stepsize may differ between Accuracy, Brier score, and NLL for the same method and K. Table 6 reports the accuracy at intermediate budgets. At 3600 server exchanges, AFHMC with $K \in \{10, 100\}$ attains the highest accuracy and the lowest NLL, and ties with FA-HMC ($K = 100$) in Brier score; FA-HMC with $K = 50$ has lower Brier score and NLL at intermediate budgets.

Table 5: Final performance on heterogeneous FMNIST at 3600 server exchanges. Entries are mean $\pm$ standard error over five random seeds.

Metric	Method	K	Rounds	η	Mean $\pm$ SE
Accuracy	FA-LD	1	3600	1×10^{-4}	0.8075 ± 0.0031
	FA-HMC	10	3600	2×10^{-5}	0.8274 ± 0.0036
	FA-HMC	50	3600	5×10^{-6}	0.8384 ± 0.0007
	FA-HMC	100	3600	2×10^{-6}	0.8408 ± 0.0003
	AFHMC	10	1800	2×10^{-5}	0.8422 ± 0.0005
	AFHMC	50	1800	5×10^{-6}	0.8402 ± 0.0005
	AFHMC	100	1800	2×10^{-6}	0.8422 ± 0.0004
Brier	FA-LD	1	3600	1×10^{-4}	0.2739 ± 0.0048
	FA-HMC	10	3600	2×10^{-5}	0.2447 ± 0.0047
	FA-HMC	50	3600	5×10^{-6}	0.2300 ± 0.0015
	FA-HMC	100	3600	2×10^{-6}	0.2290 ± 0.0003
	AFHMC	10	1800	2×10^{-5}	0.2290 ± 0.0004
	AFHMC	50	1800	5×10^{-6}	0.2376 ± 0.0008
	AFHMC	100	1800	2×10^{-6}	0.2290 ± 0.0004
NLL	FA-LD	1	3600	1×10^{-4}	5445.18 ± 136.26
	FA-HMC	10	3600	2×10^{-5}	4876.18 ± 122.35
	FA-HMC	50	3600	5×10^{-6}	4544.43 ± 34.72
	FA-HMC	100	3600	2×10^{-6}	4541.74 ± 4.99
	AFHMC	10	1800	2×10^{-5}	4532.89 ± 7.66
	AFHMC	50	1800	5×10^{-6}	4680.34 ± 15.89
	AFHMC	100	1800	2×10^{-6}	4531.58 ± 7.61

Table 6: Test accuracy on heterogeneous FMNIST at an equal number of server exchanges (AFHMC at round $x/2$, FA-LD and FA-HMC at round x). Entries are means over five random seeds; stepsizes as in Table 5.

Method	K	η	$x = 2000$	2400	2800	3200	3600
FA-LD	1	1×10^{-4}	0.8089	0.8085	0.8080	0.8076	0.8075
FA-HMC	10	2×10^{-5}	0.8270	0.8272	0.8273	0.8277	0.8274
FA-HMC	50	5×10^{-6}	0.8362	0.8369	0.8376	0.8381	0.8384
FA-HMC	100	2×10^{-6}	0.8369	0.8381	0.8391	0.8401	0.8408
AFHMC	10	2×10^{-5}	0.8389	0.8403	0.8413	0.8419	0.8422
AFHMC	50	5×10^{-6}	0.8373	0.8386	0.8391	0.8399	0.8402
AFHMC	100	2×10^{-6}	0.8392	0.8404	0.8410	0.8416	0.8422

Computational cost. Between two synchronizations, each client performs 50 leapfrog steps in FA-LD and FA-HMC, i.e., 100 mini-batch gradient evaluations per round, and K leapfrog steps plus one gradient evaluation for the drift correction in AFHMC, i.e., $2K + 1$ evaluations per round. At 3600 server exchanges, AFHMC with $K = 10$ therefore uses $1800 \times 21 = 37{,}800$ gradient evaluations per client, compared with $3600 \times 100 = 360{,}000$ for FA-HMC.