ELF: Federated Langevin Algorithms with Primal, Dual and Bidirectional Compression

Avetik Karagulyan[1] **Peter Richtárik**[2]

[1]CNRS, CentraleSupélec, Université Paris-Saclay, Laboratoire des Signaux et Systèmes, France
[2]King Abdullah University of Science and Technology, Saudi Arabia

Abstract

Federated sampling algorithms have recently gained great popularity in the community of machine learning and statistics. This paper proposes a new federated sampling algorithm called Error Feedback Langevin algorithms (ELF). In particular, we analyze the combinations of EF21 and EF21-P with the federated Langevin Monte-Carlo. We propose three algorithms, P-ELF, D-ELF, and B-ELF, that use primal, dual, and bidirectional compressors. We analyze the proposed methods under Log-Sobolev inequality and provide non-asymptotic convergence guarantees. Simple experimental results support our theoretical findings.

1 INTRODUCTION

Sampling from high-dimensional distributions holds immense significance in modern statistics and machine learning. This problem is particularly relevant in Bayesian inference [Robert, 2007], where sampling from high-dimensional posterior distributions is the bottleneck. This work will focus specifically on sampling from posteriors that arise in Bayesian federated learning [Kassab and Simeone, 2022, Vono et al., 2022, Liu and Simeone, 2022].

Federated learning is a machine learning framework that assumes data is distributed across different devices/clients, with a central server coordinating them. This scenario commonly arises in mobile applications, where each device possesses data and maintains a (limited) internet connection with the server [Konečnỳ et al., 2016, McMahan et al., 2017]. Consequently, the communication complexity becomes the computational bottleneck in most cases. The objective is to train a global model by performing local updates while minimizing the information communicated.

More formally, we want to sample from a target distribution π, defined on the Euclidean space $\mathbb{R}^d$ and is absolutely continuous with respect to the Lebesgue measure. For convenience, we will use π to refer to both the target distribution and its density function, given by:

$$\pi(x) \propto \exp(-F(x)), \tag{1}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is called the potential function. In particular, when solving a Bayesian inference problem, F corresponds to the negative log-likelihood. In the federated setting, the potential function is assumed to be sum-decomposable, with each component stored on one of the clients or nodes/devices:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x),$$

where n is the number of nodes and $F_i(x)$ represents the potential function of the i-th node. Each node only has access to its respective score, the gradient $\nabla F_i(x)$.

Building upon this framework, we propose three sampling algorithms that combine Langevin Monte Carlo (LMC) with well-known federated optimization techniques called EF21 [Richtárik et al., 2021] and EF21-P [Gruntkowska et al., 2022]. The algorithms are as follows:

- D-ELF: LMC with dual compression (Section 3.1);
- P-ELF: LMC with primal compression (Section 3.2);
- B-ELF: LMC with bidirectional compression (Section 3.3).

The first algorithm, D-ELF, focuses on client-to-server (uplink) compression to reduce communication complexity. Early papers of federated learning, such as [Konečnỳ et al., 2016] assumerd that the uplink communication is more costly than server-to-client communication. However, more recent reports[1], indicate that the difference between uploading and downloading speeds is negligible [Philippenko and

[1] https://www.speedtest.net/global-index

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:1965–1989.

Dieuleveut, 2020]. As a result, downlink compression becomes equally important. The second algorithm, P-ELF, adopts the EF21 scheme for the primal space, applying compression to the server-to-client (downlink) communication [Gruntkowska et al., 2022]. This approach leverages compression in the direction opposite to the traditional uplink compression. The third algorithm, B-ELF, combines uplink and downlink compression, hence the term "bidirectional." In the frequentist setting, bidirectional federated learning has been explored by several authors Liu et al. [2020], Philippenko and Dieuleveut [2020], Gruntkowska et al. [2022]. However, this setting has not yet been extensively developed and studied for sampling problems. In this work, we analyze the first federated sampling algorithm incorporating bidirectional compression.

1.1 LANGEVIN SAMPLING

Langevin Monte-Carlo is one of the most common methods of sampling. It is based on discretizing a stochastic differential equation (SDE) called Langevin diffusion (LD). The latter is formulated as follows:

$$\mathrm{d}L_t = -\nabla F(L_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where B_t is the Brownian motion and F is the potential function from (1). The critical property of this SDE is that it has a solution and is ergodic under mild conditions. Moreover, the target π is its invariant distribution [Bhattacharya, 1978]. Let us now define by ρ_t the density of L_t. Then, the evolution of ρ_t is characterized by the Fokker-Planck equation corresponding to LD [Pavliotis, 2014, Risken, 1996]:

$$\frac{\partial \rho_t(x)}{\partial t} = \nabla \cdot (F(x)\rho_t(x)) + \Delta \rho_t(x).$$

Using the chain rule in the Fokker-Planck equation, one can verify that π is indeed the stationary distribution for the Langevin diffusion.

Langevin Monte-Carlo (LMC) is the Euler-Maruyama discretization of the Langevin diffusion [Parisi, 1981]. That is,

$$x_{k+1} = x_k - \gamma \nabla F(x_k) + \sqrt{2\gamma} Z_k, \tag{2}$$

where $(Z_k)_k$ is a sequence of i.i.d. standard Gaussians on $\mathbb{R}^d$ that are independent of previous iterations. If the gradient of the potential (score) function is Lipschitz continuous, and the target satisfies the Log-Sobolev inequality, then the distribution of the K-th iterate converges to π [Vempala and Wibisono, 2019]. See Section 1.3 for more context on the LMC.

1.2 EF21 AND EF21-P

The Error Feedback algorithm first appeared in a heuristic manner in the paper by Seide et al. [2014]. It was proposed as a stabilization mechanism for supervised learning using contractive compressors. Later, Alistarh et al. [2018], Stich et al. [2018] analyzed the method theoretically. Nevertheless, the initial EF has issues. Namely, it does not generalize to the distributed setting, which is crucial to federated learning, and the convergence analysis requires unrealistic assumptions, such as bounds on the gradient norm. See also Section 2 of Horváth and Richtárik [2020] for more details on the shortcomings of the Error Feedback method. The EF21 (Error Feedback 21) algorithm modifies the original EF proposed by Richtárik et al. [2021]. The method proposes Markov compressors and uses them to compress gradient differences before communicating them to the server. It solves the above issues, and in particular, it applies to the distributed setting. The method is state of the art in theory and practice amongst error feedback mechanisms [Fatkhullin et al., 2021].

Interestingly, theoretical guarantees on EF21 are rather conservative. Compared with other methods, it does not gain in terms of communication complexity. However, simple experiments show that EF21 beats all the other FL methods, hinting that the worst-case analysis is not informative in this case. We refer the reader to Section 3.1 for the exact definition and mathematical details of the EF21.

EF21-P is a primal error-feedback method largely inspired by EF21. The method is essentially the analog of EF21 on the primal space. Contrary to the dominating approach in federated learning [Konečnỳ et al., 2016, Stich et al., 2018, Mishchenko et al., 2019, Richtárik et al., 2021, Fatkhullin et al., 2021], it performs compression on iterates of the algorithm rather than their gradients. Hence, it reduces the complexity of downlink communication. In general, efficient server-to-client compression may play a key role when the model is extremely large [Dean et al., 2012, Brown et al., 2020]. Furthermore, according to Gruntkowska et al. [2022], EF21-P can also be viewed as an iteration perturbation method. These methods are used in various settings in machine learning, including generalization [Orvieto et al., 2022] and smoothing [Duchi et al., 2012]. For the complete definition of the method, see Section 3.2.

1.3 RELATED WORK

Langevin Monte-Carlo In their seminal work, Roberts and Tweedie [1996] investigated the convergence properties of the Langevin Monte-Carlo (LMC) algorithm and found that a bias occurs when discretizing the continuous SDE. This bias leads to the stationary distribution of the generated homogeneous Markov chain differing from the target distribution π. To address this issue, Roberts and Tweedie [1996] proposed a Metropolis-Hastings adjustment step at each iteration of the LMC, resulting in the Metropolis Adjusted Langevin Algorithm (MALA) [Roberts and Rosenthal, 1998, Roberts and Stramer, 2002,

Xifara et al., 2014, Dwivedi et al., 2018]. The bias of LMC depends on the discretization step size γ, and Dalalyan [2017] proved a bound on this error. Later, several researchers studied different properties of LMC [Durmus and Moulines, 2017, Cheng et al., 2018, Cheng and Bartlett, 2018, Dalalyan and Karagulyan, 2019, Durmus and Moulines, 2019, Vempala and Wibisono, 2019].

Connecting LMC and SGD The LMC algorithm can be viewed as an instance of stochastic gradient descent (SGD) with independent Gaussian noise, as seen in (2). This similarity has been exploited in various settings for sampling problems [Raginsky et al., 2017, Chatterji et al., 2018, Wibisono, 2019, Salim et al., 2019, Karagulyan and Dalalyan, 2020]. Specifically, federated Langevin algorithms combine LMC with existing optimization mechanisms, such as LMC+FedAvg [McMahan et al., 2017, Deng et al., 2021, Plassier et al., 2022], LMC+MARINA [Gorbunov et al., 2021, Sun et al., 2022], and LMC+QSGD [Alistarh et al., 2017, Vono et al., 2022]. Our work extends this line of research by introducing error-feedback mechanisms EF21 and EF21-P to the classic LMC algorithm in the federated setting.

Relaxing strong convexity Strong convexity of the potential function plays a crucial role in the analysis of LMC. Non-convex optimization has long been a central topic in the domain, while sampling from non-strongly log-concave distributions is less studied. Previous studies on LMC convergence focused on strong convexity outside a ball [Cheng et al., 2018], penalization of the convex potential [Dalalyan et al., 2019, Karagulyan and Dalalyan, 2020], and non-convex regimes [Mangoubi and Vishnoi, 2019]. However, these results either do not cover the general non-convex case or require conditions that scale poorly with the dimension. A more efficient approach is based on isoperimetric inequalities, as they imply a rapid mixture of continuous stochastic processes [Villani, 2008]. Vempala and Wibisono [2019] proved LMC convergence under Log-Sobolev inequality, and Sun et al. [2022] extended this scheme to LMC with stochastic gradient estimators in the context of federated Langevin sampling. Our work simplifies their proof and adapts it to our setting.

Bayesian approach to FL Most FL algorithms currently focus on minimizing the training loss. However, they fail to provide reliable uncertainty quantification mechanisms, which is necessary for safety-critical applications according to some studies [Coglianese and Lehr, 2016, Fatima et al., 2017]. To address this issue, various authors [Welling and Teh, 2011, Yurochkin et al., 2019, Chen and Chao, 2021, Izmailov et al., 2021, Wilson et al., 2022, Vedadi et al., 2024] have proposed using the federated version of Bayesian inference. For example, the aim can be to calculate the regions with the highest posterior density of the predictive distribution. An important particular case is the Bayesian Neural Networks. Using Bayesian inference in neural networks can lead to better predictions, more accurate uncertainty measurements, and a systematic way of comparing different models. It can also support active learning, continual learning, and decision-making when there is uncertainty. The Bayesian deep learning community has developed several practical methods that use the Bayesian approach [Gal and Ghahramani, 2016], which have been successful in various fields, including astrophysics [Cranmer et al., 2021], diagnosing diabetic retinopathy [Filos et al., 2019], predicting click-through rates in advertising [Liu et al., 2017], and analyzing fluid dynamics [Geneva and Zabaras, 2020].

Cao et al. [2023] gives a broad overview on Bayesian federated learning, which is the Bayesian approach to federated learning, that targets issues such as data heterogeneity and client variability.

Federated sampling algorithms All the competitor papers study federated sampling *without compressing the iterate information*, unlike our algorithms D-ELF and B-ELF. See Sections 3.2 and 3.3 for formal definitions.

A standard reference of federated Langevin sampling is the QLSD algorithm by Vono et al. [2022]. However, they require strong log-concavity of the target distribution. Our analysis, instead, relies on the log-Sobolev inequality, which is a strictly more general assumption.

Another notable method is the federated averaging Langevin dynamics (FALD) Deng et al. [2021]. Federated averaging uses local methods as an alternative to compression to reduce communication complexity. As in the case of QLSD, the analysis is performed only for log-concave targets.

The paper Liang et al. [2024] studies federated averaging with Hamiltonian Monte-Carlo. The iteration of the HMC algorithm requires solving a differential equation, and thus is more computationally expensive when compared to first-order Langevin Monte-Carlo based methods. Moreover, the convergence analysis in the paper assumes a significantly stronger regularity condition, specifically, second-order smoothness, which combined with the stronger oracle of HMC might lead to faster convergence.

1.4 STRUCTURE OF THE PAPER

This paper is organized as follows. Section 2 describes the mathematical framework of the problem, the notation, definitions, and assumptions. In Sections 3.1 and 3.2, we present respectively the downlink and uplink compressed Langevin algorithms. That is D-ELF and P-ELF. In Section 3.3, we introduce our bidirectional federated Langevin algorithm: B-ELF. The main convergence results are pre-

sented in Section 4. The analysis of all three methods is influenced by [Vempala and Wibisono, 2019] and [Sun et al., 2022]. We simplify and adapt their proofs to our method; see Appendix B.1 and Appendix B.3. Section 5 provides simple experiments, comparing the proposed algorithm with LMC in the federated setting. We conclude the main part of the paper with Section 6.

2 PROBLEM SETUP

We denote by $\mathbb{R}^d$ the d-dimensional Euclidean space endowed with its usual scalar product and ℓ_2-norm defined by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$. The gradient of the function H and its Hessian evaluated at the point $x \in \mathbb{R}^d$ is denoted by $\nabla H(x)$ and $\nabla^2 H(x)$, respectively. As mentioned, we will repeatedly use the same notation for probability distributions and their corresponding densities. For the asymptotic complexity of the algorithms, we will use the $\mathcal{O}$ and $\tilde{\mathcal{O}}$ notations. We say that $f(t) = \mathcal{O}(g(t))$ when $t \to +\infty$, if $f(t) \leq Mg(t)$, for some $M > 0$ and when t is large enough. Similarly, $f(t) = \tilde{\mathcal{O}}(g(t))$, if $f(t)\log(t) = \mathcal{O}(g(t))$. For two measures μ and ν, we use $\nu \ll \mu$ to denote that ν is absolutely continuous with respect to μ.

2.1 MATHEMATICAL FRAMEWORK

The vast majority of optimization and sampling literature relies on the L-smoothness assumption.

Assumption 1 (L-smoothness)**.** The potential function is L-smooth. That is, for every $x, y \in \mathbb{R}^d$

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L \|x - y\|^2}{2}.$$

EF21 and EF21-P rely on contractive compressors to reduce the communication complexity.

Definition 2.1 (Contractive compressor)**.** A stochastic mapping $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}^d$ is a contractive compression operator with a coefficient $\alpha \in (0, 1]$ if for any $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|\mathcal{Q}(x) - x\|^2\right] \leq (1 - \alpha)\|x\|^2.$$

We denote it shortly as $\mathcal{Q} \in \mathbb{B}(\alpha)$.

Here, we notice that we do not require unbiasedness. In many federated learning algorithms, unbiased compressors with bounded variance are used (see e.g. [Konečnỳ et al., 2016, Alistarh et al., 2017, Mishchenko et al., 2019, Gorbunov et al., 2021]). Unbiased compressors are defined as (possibly stochastic) mappings such that $\mathbb{E}[\mathcal{Q}(x)] = x$ and $\mathbb{E}[\|\mathcal{Q}(x) - x\|^2] \leq \omega \|x\|^2$. Then, simple computation shows that $\frac{1}{\omega+1}\mathcal{Q}$ is a $\frac{1}{\omega+1}$-contractive compressor. However, the class of contractive compressors is strictly larger. Indeed. Let us look at the Top-τ compressor [Alistarh et al., 2017]. This compressor returns only the τ coordinates with the largest absolute values of the input vector. For example, if $x = (-4, 3, 10, -1, 2)^\top$, then we have $\mathcal{Q}_{\text{Top-2}}(x) = (-4, 0, 10, 0, 0)^\top$. It is obvious that Top-τ cannot be represented with unbiased compressors, as it is deterministic. This, concludes the argument.

Our analysis relies on the interpretation of sampling as an optimization problem over the space of measures. In order to reformulate our problem, let us first recall the definition of the Kullback-Leibler divergence.

Definition 2.2 (Kullback-Leibler divergence)**.** The Kullback-Leibler divergence between two probability measures ν and π is defined as

$$H_\pi(\nu) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\nu(x)}{\pi(x)}\right) \nu(x)\mathrm{d}x, & \text{if } \nu \ll \pi; \\ +\infty, & \text{otherwise.} \end{cases}$$

We aim to construct approximate samples from π with ε accuracy. That is to sample from some other distribution ν such that $H_\pi(\nu) < \varepsilon$. Alternatively, it means that we want to minimize the functional:

$$\min_{\nu \in \mathcal{P}(\mathbb{R}^d)} H_\pi(\nu).$$

Indeed, the minimum of this functional is equal to zero and is attained only when $\nu = \pi$. Recall now the classical problem of optimization, that is minimizing a $H : \mathbb{R}^d \to \mathbb{R}$. Polyak [1963] and Łojasiewicz [1963] independently proposed an inequality, which is weaker than strong convexity, but it nevertheless implies linear convergence of the gradient descent. It is known under the joint name of Polyak-Łojasiewicz inequality:

$$H(x) - \min_x H(x) \leq \frac{1}{\mu}\|\nabla H(x)\|^2,$$

assuming the objective has a minimum. See [Karimi et al., 2016, Khaled and Richtárik, 2020] for more details on the P inequality, as well as its comparison with other similar conditions for non-convex optimization. In the problem of sampling, the objective functional is defined on the space of measures $\mathcal{P}(\mathbb{R}^d)$. One can define the usual notions of differentiability and convexity on this space using the Wasserstein distance [Ambrosio et al., 2008]. Then, the Langevin Monte-Carlo algorithm becomes a first order minimization method for the KL divergence [Wibisono, 2018]. Furthermore, Fisher information takes the role of the square norm of the gradient.

Definition 2.3 (Fisher information)**.** The Fisher information of probability measures ν and π is denoted by $J_\pi(\nu)$ and it is defined as below:

$$J_\pi(\nu) := \begin{cases} \int_{\mathbb{R}^d} \left\|\nabla \log\left(\frac{\nu}{\pi}\right)\right\|^2 \nu(x)\mathrm{d}x, & \text{if } \nu \ll \pi; \\ +\infty, & \text{otherwise.} \end{cases}$$

Algorithm 1 D-ELF

Input: Initialization $x_0 \sim \rho_0$, $g_0^i = \nabla F_i(x_0)$ $g_0 = \nabla F(x_0)$, step-size h, iterations K
for $k = 0, 1, 2, \ldots, K-1$ **do**
  The server:
  draws $Z_k \sim \mathcal{N}(0, I_d)$;
  $\circ\ x_{k+1} = x_k - \gamma g_k + \sqrt{2\gamma} Z_k$;
  broadcasts x_{k+1};
  The devices in parallel:
  $\circ\ g_{k+1}^i = g_k^i + \mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i)$;
  broadcast $\mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i)$;
  The server:
  $\circ\ g_{k+1} = g_k + \frac{1}{n}\sum_{i=1}^n \mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i)$.
end for
Return: x_K

Since the minimum of our functional is equal to zero, the Log-Sobolev inequality (LSI) becomes the analog of P inequality.

Assumption 2 (Log-Sobolev inequality)**.** The target π satisfies the Log-Sobolev inequality (LSI) with parameter μ. That is for every probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$ we have

$$H_\pi(\nu) \leq \frac{1}{2\mu} J_\pi(\nu).$$

Bakry and Émery [1985] have shown that strongly log-concave distributions satisfy LSI. Furthermore, from Holley-Stroock's theorem we know that sufficiently small perturbations of strongly concave distributions still satisfy LSI [Holley and Stroock, 1986]. The latter distributions can be non log-concave, which means that we deal with a strictly larger class of probability measures using LSI.

Analyzing the sampling problems as an optimization problem on the Wasserstein space has been strongly influenced by the seminal paper of Jordan et al. [1998]. It has later been developed in subsequent work; see e.g. [Wibisono, 2018, Durmus et al., 2019]. We use Log-Sobolev inequality to derive bounds on the convergence error in KL divergence.

3 THE ELF ALGORITHMS

In this section, we present two federated Langevin Monte-Carlo algorithms, combining EF21 and EF21-P with LMC. We replace the gradient term $\nabla F(x_k)$ at each iteration with the gradient estimator g_k from the corresponding error feedback method, and add independent Gaussian noise. Details can be found in Algorithm 1 and Algorithm 2. The pseudocode distinguishes between optimization and sampling methods with a wave symbol.

Algorithm 2 P-ELF

Input: Starting point $x_0 = w_0 \sim \rho_0$, step-size h, number of iterations K
for $k = 0, 1, 2, \cdots, K-1$ **do**
  The server:
  draws $Z_k \sim \mathcal{N}(0, I_d)$;
  $\circ\ \nabla F(w_k) = \frac{1}{n}\sum_{i=1}^n \nabla F_i(w_k)$;
  $\circ\ x_{k+1} = x_k - \gamma \nabla F(w_k) + \sqrt{2\gamma} Z_k$;
  $\circ\ w_{k+1} = w_k + \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k)$;
  broadcasts in parallel $\mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k)$.
  The devices in parallel:
  $\circ\ w_{k+1} = w_k + \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k)$;
  $\circ\ \nabla F_i(w_{k+1})$;
  broadcast $\nabla F_i(w_{k+1})$;
end for
Return: x_K

3.1 DUAL COMPRESSION: D-ELF

The gradient estimator g_k of the dual method is defined as the average of the vectors g_k^i, where each g_k^i is computed on the i-th node and estimates the gradients $\nabla F_i(x_k)$. The key component of this estimator is the contractive compression operator $\mathcal{Q}^{\mathrm{D}} \in \mathbb{B}(\alpha^{\mathrm{D}})$. At the zeroth iteration, $g_0 = \nabla F(x_0)$. Then at iteration k, the server computes the new iterate $x_{k+1} = x_k - \gamma g_k + \sqrt{2\gamma} Z_k$ and broadcasts it parallelly to all the nodes. Each node updates g_k^i with the formula:

$$g_{k+1}^i = g_k^i + \mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i),$$

and broadcasts the compressed term to the server. The server aggregates the received information and computes the estimator of $\nabla F_i(x_{k+1})$:

$$g_{k+1} = g_k + \frac{1}{n}\sum_{i=1}^n \mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i).$$

For the pseudocode of the D-ELF, please refer to Algorithm 1.

3.2 PRIMAL COMPRESSION: P-ELF

The construction of the P-ELF algorithm is similar to the D-ELF. In particular, we take the EF21-P algorithm by Gruntkowska et al. [2022] and add only the independent Gaussian term. See Algorithm 2 for the complete definition. To better understand the comparison of the D-ELF and the P-ELF let us look at the simple one-node setting of the latter:

$$\begin{cases} w_0 := \mathcal{Q}^{\mathrm{P}}(x_0) \\ w_{k+1} = w_k + \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k) \\ x_{k+1} = x_k - \gamma \nabla F(w_k) + \sqrt{2\gamma} Z_k. \end{cases} \tag{3}$$

Here, $x_0 \sim \rho_0$ is a random starting point, $\mathcal{Q}^{\mathrm{P}} \in \mathbb{B}(\alpha^{\mathrm{P}})$, and $(Z_k)_k$ is a sequence of i.i.d. standard Gaussians on $\mathbb{R}^d$.

If we remove the additive Gaussian noise Z_k, then we recover the P-ELF algorithm, which is known to converge to the minimum of the potential function F [Gruntkowska et al., 2022]. The auxiliary sequence w_k is meant to estimate the iterate x_k. We then use its gradient as the minimizing direction. The important difference with the EF21 is that we apply the compressor $\mathcal{Q}^{\mathrm{P}}$ on the term $x_{k+1}-w_k$, instead of the gradient and its estimator. Hence, the letter "P"-primal in the name of the algorithm.

3.3 BIDIRECTIONAL COMPRESSION: B-ELF

This section focuses on the bidirectional setting. We propose the B-ELF algorithm. The algorithm uses EF21 for the uplink and EF21-P for the downlink compression. We use the same notation as for the previous methods and the details are presented in Algorithm 3.

Algorithm 3 B-ELF

Input: Starting point $x_0 = w_0 \sim \rho_0$, step-size γ, number of iterations K, $g_0 = \nabla F(x_0)$, $g_0^i = \nabla F_i(x_0)$.
for $k = 0, 1, 2, \cdots, K-1$ **do**
  The server:
  draws a Gaussian vector $Z_k \sim \mathcal{N}(0, I_d)$;
  computes $x_{k+1} = x_k - \gamma g_k + \sqrt{2\gamma} Z_k$;
  computes $v_k := \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k)$;
  computes $w_{k+1} = w_k + v_k$;
  broadcasts v_k in parallel to the devices;
  The device i (in parallel for all $i = 1, \ldots, n$):
  computes $w_{k+1} = w_k + v_k$;
  computes $h_{k+1}^i = \mathcal{Q}^{\mathrm{D}}(\nabla F_i(w_{k+1}) - g_k^i)$;
  computes $g_{k+1}^i = g_k^i + h_{k+1}^i$;
  broadcasts h_i^{k+1};
  The server:
  computes $g_{k+1} = g_k + \frac{1}{n}\sum_{i=1}^n h_{k+1}^i$;
end for
Return: x_K

4 CONVERGENCE OF THE METHODS

4.1 A UNIFIED ANALYSIS OF D-ELF AND P-ELF

The key component of the analysis of both methods is defining proper a Lyapunov-type function. For the D-ELF algorithm we define by $\mathbf{G}_k^{\mathrm{D}}$ the average squared estimation error of the vectors g_k^i:

$$\mathbf{G}_k^{\mathrm{D}} := \frac{1}{n}\sum_{i}^{n} \mathbb{E}\left[\left\|g_k^i - \nabla F_i(x_k)\right\|^2\right]. \tag{4}$$

As we will later in Appendix B.1, this quantity arises in the proof of the convergence rates. Important property of the sequence $\mathbf{G}_k$ is the following recurrent identity.

Proposition 4.1. *Let x_k be the iterates of the D-ELF, g_k^i be the EF21 estimators and $\mathbf{G}_k^{\mathrm{D}}$ be defined as (4). Then the following recurrent inequality is true:*

$$\mathbf{G}_{k+1}^{\mathrm{D}} \le (1-p)\mathbf{G}_k^{\mathrm{D}} + (1-p)\beta_{\mathrm{D}}\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right], \tag{5}$$

where $p := 1 - (1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}}) > 0$

$$\bar{L} := \frac{1}{n}\sum_{i=1}^{n} L_i^2 \quad \textit{and} \quad \beta_{\mathrm{D}} := \frac{1 + s_{\mathrm{D}}^{-1}}{1+s_{\mathrm{D}}}\bar{L},$$

for some $s_{\mathrm{D}} > 0$.

The Lyapunov term associated to the P-ELF is a simple upper bound on $\mathbf{G}^{\mathrm{D}}$. We denote it by $\mathbf{G}_k^{\mathrm{P}}$ and define with the formula below:

$$\mathbf{G}_k^{\mathrm{P}} := \bar{L}\mathbb{E}\left[\left\|w_k - x_k\right\|^2\right], \quad \text{where} \quad \bar{L} := \frac{1}{n}\sum_{i=1}^{n} L_i^2. \tag{6}$$

Indeed, $\mathbf{G}_k^{\mathrm{D}} \le \mathbf{G}_k^{\mathrm{P}}$ due to L_i smoothness of each component function F_i. See (26) in Appendix B.3 for the proof. The following proposition proves a recurrent identity similar to (5).

Proposition 4.2. *Let x_k and w_k be defined as in P-ELF and $\mathbf{G}^{\mathrm{P}}$ be its Lyapunov term. Then the following recurrent inequality is true:*

$$\mathbf{G}_{k+1}^{\mathrm{P}} \le (1-p)\mathbf{G}_k^{\mathrm{P}} + (1-p)\beta_{\mathrm{P}}\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right],$$

where $p := 1-(1-\alpha_{\mathrm{P}})(1+s_{\mathrm{P}}) > 0$, and $\beta_{\mathrm{P}} := \frac{1+s_{\mathrm{P}}^{-1}}{1+s_{\mathrm{P}}}\bar{L}$, for some $s_{\mathrm{P}} > 0$.

The next theorem gives a unified bound for both D-ELF and P-ELF. For the sake of space we use a general notation M-ELF, where M $\in$ {D,P}. This means, for example, that the M-ELF refers to the D-ELF when M = D.

Theorem 4.3. *Let x_k be the iterates of the M-ELF algorithm, where M $\in$ {D,P}. We denote by $\rho_k := \mathcal{L}(x_k)$ for every $k \in \mathbb{N}$. Under Assumptions 1 and 2, if*

$$0 < \gamma \le \min\left\{\frac{1}{14}\sqrt{\frac{p}{(1+\beta_{\mathrm{M}})}}, \frac{p}{6\mu}, \frac{1}{2\sqrt{2}L}\right\},$$

then the following is true for the KL error of the M-ELF algorithm:

$$H_\pi(\rho_K) \le e^{-\mu K\gamma}\Psi + \frac{\tau}{\mu},$$

where $p := 1-(1-\alpha_{\mathrm{M}})(1+s_{\mathrm{M}}) > 0$, $\tau = \left(2L^2 + C(1-p)\beta_{\mathrm{M}}\right)\left(16\gamma^2 d + 4d\gamma\right)$,

$$\Psi = H_\pi(\rho_0) + \frac{1-e^{-\mu\gamma}}{\mu}C\mathbf{G}_0^{\mathrm{M}},$$

$$C = \frac{8L^2\gamma^2 + 2}{e^{-\mu\gamma} - (1-p)\left(4\gamma^2\beta_{\mathrm{M}} + 1\right)}.$$

Table 1: In this table we compare error-feedback methods in optimization and sampling. The rates are computed in the case when $\alpha_{\mathrm{D}} = \alpha_{\mathrm{P}} = \alpha$.

METHOD	ASSUMPTION	COMPLEXITY	REFERENCE
GD	μ-S.C.	$\tilde{\mathcal{O}}\left(\frac{dL}{\mu\varepsilon}\right)$	NESTEROV [2013]
EF21	μ-S.C.	$\tilde{\mathcal{O}}\left(\frac{L}{\alpha\mu\varepsilon}\right)$	RICHTÁRIK ET AL. [2021]
EF21-P	μ-S.C.	$\tilde{\mathcal{O}}\left(\frac{L}{\alpha\mu\varepsilon}\right)$	GRUNTKOWSKA ET AL. [2022]
LMC	μ-LSI	$\tilde{\mathcal{O}}\left(\frac{L^2 d}{\mu^2\varepsilon}\right)$	VEMPALA AND WIBISONO [2019]
D-ELF	μ-LSI	$\tilde{\mathcal{O}}\left(\frac{\bar{L}d}{\alpha^2\mu^2\varepsilon}\right)$	COROLLARY 4.6
P-ELF	μ-LSI	$\tilde{\mathcal{O}}\left(\frac{\bar{L}d}{\alpha^2\mu^2\varepsilon}\right)$	COROLLARY 4.6
B-ELF	μ-LSI	$\tilde{\mathcal{O}}\left(\frac{\bar{L}d}{\alpha^4\mu^2\varepsilon}\right)$	COROLLARY 4.7

We refer the reader to Appendix B.3 for the proof of the theorem. The right-hand side consists of two terms. The first term corresponds to the convergence error, while the second term is the bias that comes from the discretization. To make the error small, one would first need to choose γ small enough so that $\tau/\mu < \varepsilon$. Then, the number of iterations are chosen to be of order $\tilde{\mathcal{O}}(1/\mu\gamma)$. See Section 4.3 for more on the complexity of D-ELF and P-ELF.

These bounds can also be extended to other probability distance metrics, such as TV and W_2. The relation of TV and KL is established with Pinsker's inequality: $\mathsf{TV}(\nu_1, \nu_2) \leq \sqrt{\frac{1}{2} H_{\nu_2}(\nu_1)}$. Thus, the convergence in KL divergence implies convergence in TV. Similar result is true for the Wasserstein-2 distance. It is known that LSI implies Talagrand's inequality [Otto and Villani, 2000]. The latter bounds the W_2 distance with KL divergence: $W_2(\nu, \pi) \leq \sqrt{\frac{2H_\pi(\nu)}{\mu}}$ for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Again, from the convergence in KL we can deduce convergence in W_2.

4.2 CONVERGENCE ANALYSIS OF THE B-ELF

The Lyapunov term for the B-ELF algorithm is the same as for the D-ELF, that is $\mathbf{G}_k^{\mathrm{D}}$. However, the recurrent identity of Proposition 4.1 is not valid in this case. Instead, another bound is true which includes the term $\mathbf{G}_k^{\mathrm{P}}$. The latter arises because of the downlink compression. We present *informally* the new recurrent inequality. We refer the reader to Proposition A.1 in the Appendix for the complete statement.

Proposition 4.4 (Informal). *If x_k are the iterations of Algorithm 3, $\mathbf{G}_k^{\mathrm{D}}$ and $\mathbf{G}_k^{\mathrm{P}}$ are defined as in (4) and (6), then*

$$\mathbf{G}_{k+1}^{\mathrm{D}} \leq \lambda_1 \mathbf{G}_k^{\mathrm{D}} + \lambda_2 \mathbb{E}\left[\|x_k - x_{k+1}\|^2\right] + \lambda_3 \mathbf{G}_k^{\mathrm{P}},$$

where λ_1, λ_2 and λ_3 are positive numbers.

Theorem 4.5. *Let x_k be the iterates of the B-ELF algorithm. We denote by $\rho_k := \mathcal{L}(x_k)$ for every $k \in \mathbb{N}$. Under Assumptions 1 and 2, if*

$$\gamma \leq \min\left\{\frac{\alpha_{\mathrm{D}}}{4\mu}, \frac{\alpha_{\mathrm{P}}}{4\mu}, \frac{\alpha_{\mathrm{D}}\alpha_{\mathrm{P}}}{495\sqrt{\left(1 - \frac{\alpha_{\mathrm{D}}}{2}\right)\left(1 - \frac{\alpha_{\mathrm{P}}}{2}\right)}\,\bar{L}}\right\}.$$

Then, for every $K \in \mathbb{N}$,

$$H_\pi(\nu_K) \leq e^{-\mu\gamma K}\left[H_\pi(\rho_0) + \frac{1}{\mu}\left(C\mathbf{G}_0^{\mathrm{D}} + D\mathbf{G}_0^{\mathrm{P}}\right)\right] + \frac{\tau}{\mu},$$

where $C, D > 0$ are constants depending on the parameters of the algorithm and

$$C = \frac{2.125}{e^{-\mu\gamma} - \lambda_1}, \quad D = \frac{C\lambda_3}{e^{-\mu\gamma} - (1 - \alpha_{\mathrm{P}})(1 + w)},$$
$$\tau = \left(2L^2 + \frac{5C\lambda_2}{\alpha_{\mathrm{P}}}\right)\left(16\gamma^2 dL + 4d\gamma\right).$$

The exact definitions of the undefined constants are written in the proof of the theorem, which is postponed to Appendix B.4.

4.3 DISCUSSION ON THE COMMUNICATION COMPLEXITY

Doing the computations as mentioned at the end of Section 4.1, we can deduce the following.

Corollary 4.6. *Under the assumptions of Theorem 4.3 and $\gamma = \mathcal{O}\left(\frac{\mu p \varepsilon}{\beta_{\mathrm{M}} d}\right)$, $K = \mathcal{O}\left(\frac{(1+\beta_{\mathrm{M}})d}{\mu^2 p \varepsilon} \log\left(\frac{\Psi}{\varepsilon}\right)\right)$, the primal and dual ELF algorithms satisfy $H_\pi(\rho_K) \leq \varepsilon$.*

Similarly, for the bidirectional ELF we have the below.

Corollary 4.7. *If $\alpha_{\mathrm{P}} = \alpha_{\mathrm{D}} = \alpha < 1/2$, under the conditions of Theorem 4.5, the iteration complexity for the B-ELF is $\tilde{\mathcal{O}}(d\bar{L}/\alpha^4\mu^2\varepsilon)$.*

The proof of Corollary 4.7 is in Appendix B.5. When $1/\alpha = O(1)$, the rate of the LMC algorithm is recovered for all three algorithms. In particular, the scaled unbiased compressors, such as $\frac{8}{9}\mathcal{Q}^{\mathrm{nat}}$, have a contractive coefficient of $\frac{8}{9}$. Our analysis may not match the usual LMC for other compressors, as the communication complexity is $\tilde{\mathcal{O}}(d^2/\varepsilon)$ for LMC, while both the iteration and communication complexity is $\tilde{\mathcal{O}}(d^5/\varepsilon)$ for B-ELF with Top-1. However, in the next section we will see, that these theoretical bounds are conservative and that the performance of the proposed methods on simple classification tasks match the performance of LMC.

5 EXPERIMENTS

In this section, we conduct numerical experiments to compare {B,D,P}-ELF with the LMC. The code for the experiments can be found in https://anonymous.4open.

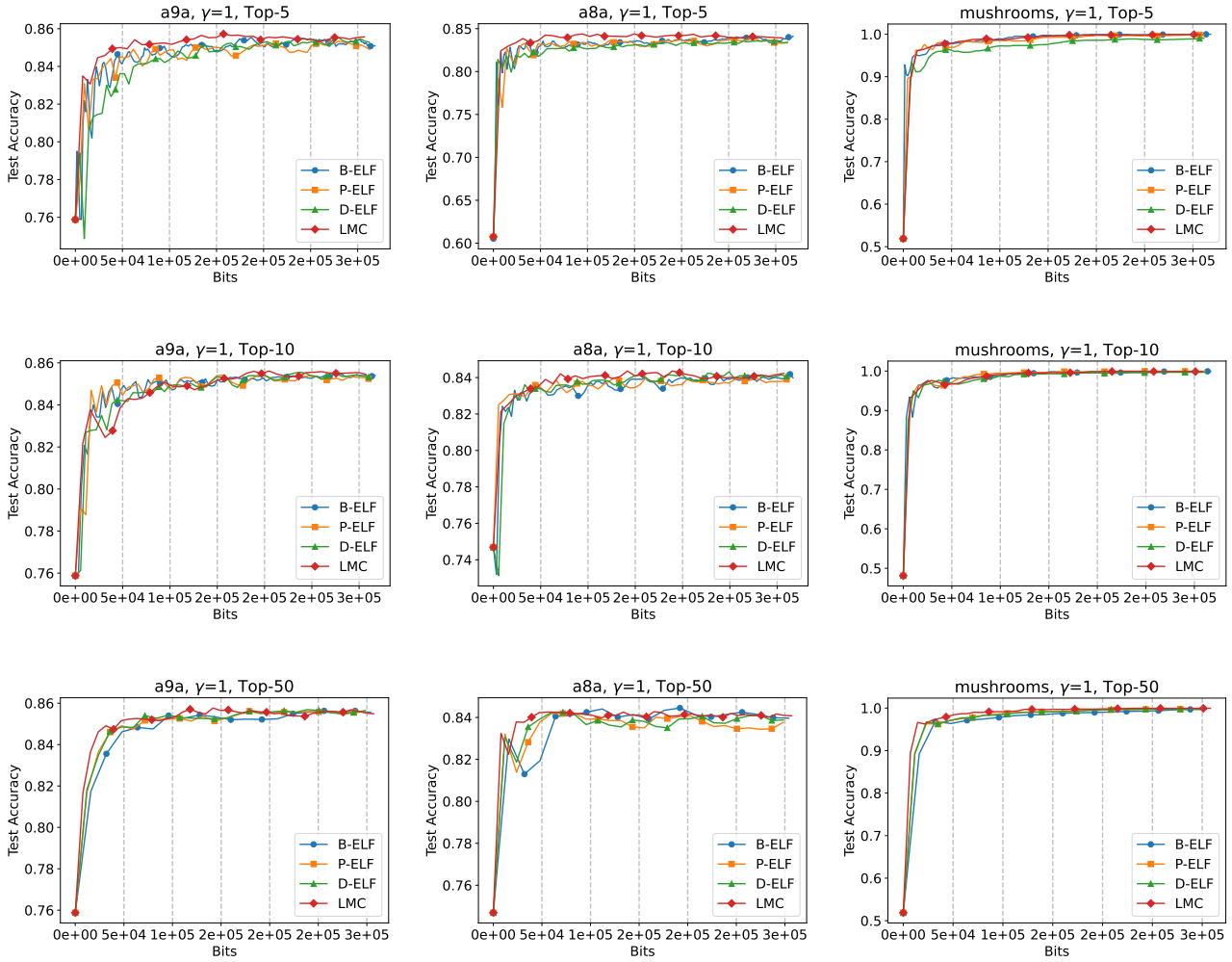

Figure 1: Bayesian logistic regression with a Gaussian prior performed on three different datasets from LibSVM [Chang and Lin, 2011]. The X-axis represents the number of bits communicated, while the Y-axis represents the accuracy of the final estimator on the test set.

`science/r/elf_code-DE51/`. We implemented all four algorithms to solve a Bayesian logistic regression problem. The datasets are `a8a`, `a9a`, `mushrooms` for the LibSVM repository Chang and Lin [2011].

In Figure 1, we observe that all the methods have similar communication complexity on the abovementioned problem. In particular, this means that despite the theoretical results obtained above, the performance of ELF is not worse than LMC. Thus, in practice we achieve compression for free.

6 CONCLUSION

In this paper we proposed three error feedback based federated Langevin algorithms with dual, primal and bidirectional compression. The first two are analyzed with one theorem and have similar theoretical performance. The third algorithm uses bidirectional compression which is slower due to the fact that EF21 and EF21-P do not couple. To the best of our knowledge, this is the first study of the federated sampling algorithms with bidirectional compression. Our theoretical findings show that the communication complexity of this algorithm is worse than the one for the standard LMC, nonetheless, simple experiments show that the theoretical analysis is rather conservative and that it can still be improved. This phenomenon is not surprising, as it was also observed for the original EF21 algorithm.

6.1 FUTURE WORK

An immediate continuation of our paper would be to conduct more thorough experimental analysis of the ELF algorithms with other federated sampling techniques on high-

dimensional data. Another possible direction is the theoretical analysis of the Langevin algorithm combined with EF21-P+DIANA. The latter is a bidirectional federated optimization algorithm that uses DIANA gradient estimator for the uplink compression instead of EF21. This method matches the performance of the GD due to the coupling of two methods [Gruntkowska et al., 2022].

Finally, there are yet many important algorithms of optimization that are relevant to our setting. Adaptation of these methods to the sampling setting can lead to fruitful results.

Acknowledgements The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) CRG Grant ORFS-CRG12-2024-6460, iii) Center of Excellence for Generative AI, under award number 5940, and iv) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient SGD via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

Herbert Amann. *Ordinary differential equations: an introduction to nonlinear analysis*, volume 13. Walter de Gruyter, 2011.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Seminaire de probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

RN Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *The Annals of Probability*, pages 541–553, 1978.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Longbing Cao, Hui Chen, Xuhui Fan, Joao Gama, Yew-Soon Ong, and Vipin Kumar. Bayesian federated learning: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 7233–7242, 2023.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

Hong-You Chen and Wei-Lun Chao. FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. In *International Conference on Learning Representations*, 2021.

X. Cheng and P. L Bartlett. Convergence of Langevin MCMC in KL-divergence. *PMLR 83*, (83):186–211, 2018.

Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

Cary Coglianese and David Lehr. Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ*, 105:1147, 2016.

Miles Cranmer, Daniel Tamayo, Hanno Rein, Peter Battaglia, Samuel Hadden, Philip J Armitage, Shirley Ho, and David N Spergel. A Bayesian neural network predicts the dissolution of compact planetary systems. *Proceedings of the National Academy of Sciences*, 118 (40):e2026053118, 2021.

A. S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Arnak S. Dalalyan, Lionel Riou-Durand, and Avetik Karagulyan. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. 2019.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale

distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

Wei Deng, Yi-An Ma, Zhao Song, Qian Zhang, and Guang Lin. On convergence of federated averaging Langevin dynamics. *arXiv preprint arXiv:2112.05120*, 2021.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797. PMLR, 2018.

Meherwar Fatima, Maruf Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Nicholas Geneva and Nicholas Zabaras. Modeling the dynamics of PDE systems with physics-constrained deep auto-regressive networks. *Journal of Computational Physics*, 403:109056, 2020.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik. MARINA: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. EF21-P and Friends: Improved Theoretical Communication Complexity for Distributed Optimization with Bidirectional Compression, 2022. URL https://arxiv.org/abs/2209.15218.

Richard Holley and Daniel W Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. 1986.

Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. *arXiv preprint arXiv:2006.11077*, 2020.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Avetik Karagulyan and Arnak Dalalyan. Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets. *Advances in Neural Information Processing Systems*, 33, 2020.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *ECML-PKDD*, pages 795–811. Springer, 2016.

Rahif Kassab and Osvaldo Simeone. Federated generalized Bayesian learning via distributed Stein variational gradient descent. *IEEE Transactions on Signal Processing*, 2022.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Jiajun Liang, Qian Zhang, Wei Deng, Qifan Song, and Guang Lin. Bayesian federated learning with Hamiltonian Monte-Carlo: Algorithm and theory. *Journal of Computational and Graphical Statistics*, pages 1–10, 2024.

Dongzhu Liu and Osvaldo Simeone. Wireless federated Langevin Monte Carlo: Repurposing channel noise for Bayesian sampling and privacy. *IEEE Transactions on Wireless Communications*, 2022.

Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for efficient distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 133–143. PMLR, 2020.

Xun Liu, Wei Xue, Lei Xiao, and Bo Zhang. PBODL: Parallel Bayesian online deep learning for click-through rate prediction in tencent advertising system. *arXiv preprint arXiv:1707.00802*, 2017.

Stanislaw Łojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 2259–2293. PMLR, 2019.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. URL http://arxiv.org/abs/1602.05629.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

Y. E Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. Anticorrelated noise injection for improved generalization. *arXiv preprint arXiv:2202.02831*, 2022.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

Constantin Philippenko and Aymeric Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees. *arXiv preprint arXiv:2006.14591*, 2020.

Vincent Plassier, Alain Durmus, and Eric Moulines. Federated Averaging Langevin Dynamics: Toward a unified theory and new algorithms. *arXiv preprint arXiv:2211.00100*, 2022.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

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

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, pages 63–95. Springer, 1996.

Christian Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media, 2007.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Adil Salim, Dmitry Kovalev, and Peter Richtárik. Stochastic proximal Langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6649–6661, 2019.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth annual conference of the international speech communication association*, 2014.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

Lukang Sun, Adil Salim, and Peter Richtárik. Federated learning with a sampling algorithm under isoperimetry, 2022. URL https://arxiv.org/abs/2206.00920.

Elahe Vedadi, Joshua V Dillon, Philip Andrew Mansfield, Karan Singhal, Arash Afkanpour, and Warren Richard Morningstar. Federated variational inference: Towards improved personalization and generalization. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 323–327, 2024.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8092–8104, 2019.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Maxime Vono, Vincent Plassier, Alain Durmus, Aymeric Dieuleveut, and Eric Moulines. QLSD: Quantised Langevin Stochastic Dynamics for Bayesian Federated Learning . In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6459–6500. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/vono22a.html.

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

A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory (COLT)*, pages 2093–3027, 2018.

A. Wibisono. Proximal Langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*, 2019.

Andrew Gordon Wilson, Pavel Izmailov, Matthew D Hoffman, Yarin Gal, Yingzhen Li, Melanie F Pradier, Sharad Vikram, Andrew Foong, Sanae Lotfi, and Sebastian Farquhar. Evaluating approximate inference in Bayesian deep learning. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 113–124. PMLR, 2022.

Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019.

ELF: Federated Langevin Algorithms with Primal, Dual and Bidirectional Compression (Supplementary Material)

Avetik Karagulyan[1] **Peter Richtárik**[2]

[1]CNRS, CentraleSupélec, Université Paris-Saclay, Laboratoire des Signaux et Systèmes, France
[2]King Abdullah University of Science and Technology, Saudi Arabia

A PROOFS OF THE PROPOSITIONS

A.1 PROOF OF PROPOSITION 4.1

From the definition

$$\begin{aligned}\mathbf{G}_{k+1}^{\mathrm{D}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|g_{k+1}^{i}-\nabla F_i(x_{k+1})\right\|^2\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|g_{k}^{i}+\mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1})-g_k^i)-\nabla F_i(x_{k+1})\right\|^2 \mid x_1,\dots,x_{k+1}\right]\right]\\ &\le \frac{1-\alpha_{\mathrm{D}}}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|g_{k}^{i}-\nabla F_i(x_{k+1})\right\|^2\right].\end{aligned}$$

Applying Cauchy-Schwartz and the Lipschitz continuity of the function $\nabla F_i(\cdot)$, we obtain

$$\begin{aligned}\mathbf{G}_{k+1}^{\mathrm{D}} &\le \frac{(1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}})}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|g_{k}^{i}-\nabla F_i(x_{k})\right\|^2\right]\\ &\quad+\frac{(1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}}^{-1})}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla F_i(x_k)-\nabla F_i(x_{k+1})\right\|^2\right]\\ &\le (1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}})\mathbf{G}_k^{\mathrm{D}}+\frac{(1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}}^{-1})}{n}\sum_{i=1}^{n}L_i^2\mathbb{E}\left[\left\|x_k-x_{k+1}\right\|^2\right]\\ &\le (1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}})\mathbf{G}_k^{\mathrm{D}}+(1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}}^{-1})\bar{L}\mathbb{E}\left[\left\|x_k-x_{k+1}\right\|^2\right]\\ &\le (1-p_{\mathrm{D}})\mathbf{G}_k^{\mathrm{D}}+(1-p_{\mathrm{D}})\beta_{\mathrm{D}}\mathbb{E}\left[\left\|x_k-x_{k+1}\right\|^2\right].\end{aligned}$$

This concludes the proof.

A.2 PROOF OF PROPOSITION 4.2

From the definition

$$\begin{aligned}
\mathbf{G}_{k+1}^{\mathrm{P}} &= L^2\mathbb{E}\left[\|w_{k+1} - x_{k+1}\|^2\right] \\
&= L^2\mathbb{E}\left[\left\|w_k - x_{k+1} - \mathcal{Q}^{\mathrm{P}}(w_k - x_{k+1})\right\|^2\right] \\
&= (1-\alpha_{\mathrm{P}})L^2\mathbb{E}\left[\|w_k - x_{k+1}\|^2\right] \\
&= (1-\alpha_{\mathrm{P}})L^2\mathbb{E}\left[\|w_k - x_k + x_k - x_{k+1}\|^2\right] \\
&\leq (1-\alpha_{\mathrm{P}})(1+s_{\mathrm{P}})L^2\mathbb{E}\left[\|w_k - x_k\|^2\right] + (1-\alpha_{\mathrm{P}})(1+s_{\mathrm{P}}^{-1})L^2\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right].
\end{aligned} \tag{7}$$

Choosing s_{P} small enough, we can make the coefficient $(1-\alpha_{\mathrm{P}})(1+s_{\mathrm{P}})$ smaller than one. Thus, defining $p = 1-(1-\alpha_{\mathrm{P}})(1+s_{\mathrm{P}})$, we conclude the proof.

A.3 FULL STATEMENT OF PROPOSITION 4.4 AND ITS PROOF

We state now the complete version of Proposition 4.4.

Proposition A.1. *The Lyapunov term $\mathbf{G}_k^{\mathrm{D}}$ of the bidirectional Langevin algorithm satisfies the following recurrent inequality:*

$$\mathbf{G}_{k+1}^{\mathrm{D}} \leq \lambda_1\mathbf{G}_k^{\mathrm{D}} + \lambda_2\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right] + \lambda_3\mathbf{G}_k^{\mathrm{P}},$$

where $\mathbf{G}_k^{\mathrm{P}} := \bar{L}\mathbb{E}\left[\|w_k - x_k\|^2\right]$ is the Lyapunov term for P-ELF and

$$\begin{aligned}
\lambda_1 &= (1-\alpha_{\mathrm{D}})(1+s)(1+q); \\
\lambda_2 &= (1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u)\bar{L} \\
&\quad + \left((1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u^{-1}) + (1+s^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}; \\
\lambda_3 &= \left((1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u^{-1}) + (1+s^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w).
\end{aligned} \tag{8}$$

Here, s, q, u, w are any positive numbers.

Proof. From the definition of $\mathbf{G}_k^{\mathrm{D}}$ and Young's inequality we have

$$\begin{aligned}
\mathbf{G}_{k+1}^{\mathrm{D}} &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|g_{k+1}^i - \nabla F_i(x_{k+1})\right\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[\left\|g_k^i + \mathcal{Q}^{\mathrm{D}}(\nabla F_i(w_{k+1}) - g_k^i) - \nabla F_i(x_{k+1})\right\|^2 \mid x_1, \ldots, x_{k+1}\right]\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n \Big\{(1+s)\mathbb{E}\left[\mathbb{E}\left[\left\|g_k^i + \mathcal{Q}^{\mathrm{D}}(\nabla F_i(w_{k+1}) - g_k^i) - \nabla F_i(w_{k+1})\right\|^2 \mid x_1, \ldots, x_{k+1}\right]\right] \\
&\quad + (1+s^{-1})\mathbb{E}\left[\|\nabla F_i(w_{k+1}) - \nabla F_i(x_{k+1})\|^2\right]\Big\}.
\end{aligned}$$

The contractivity of $\mathcal{Q}^{\mathrm{D}}$ implies

$$\begin{aligned}
\mathbf{G}_{k+1}^{\mathrm{D}} &\leq \frac{1}{n}\sum_{i=1}^n (1-\alpha_{\mathrm{D}})(1+s)\mathbb{E}\left[\left\|g_k^i - \nabla F_i(w_{k+1})\right\|^2\right] + (1+s^{-1})\bar{L}\mathbb{E}\left[\|w_{k+1} - x_{k+1}\|^2\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n (1-\alpha_{\mathrm{D}})(1+s)(1+q)\mathbb{E}\left[\left\|g_k^i - \nabla F_i(x_k)\right\|^2\right] + (1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})\mathbb{E}\left[\|\nabla F_i(x_k) - \nabla F_i(w_{k+1})\|^2\right] \\
&\quad + (1+s^{-1})\bar{L}\mathbb{E}\left[\|w_{k+1} - x_{k+1}\|^2\right] \\
&\leq (1-\alpha_{\mathrm{D}})(1+s)(1+q)\mathbf{G}_k^{\mathrm{D}} + (1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})\bar{L}\mathbb{E}\left[\|x_k - w_{k+1}\|^2\right] + (1+s^{-1})\mathbf{G}_{k+1}^{\mathrm{P}}.
\end{aligned}$$

Applying Young's inequality to the second term, we deduce

$$\begin{aligned}\bar{L}\mathbb{E}\left[\|x_k - w_{k+1}\|^2\right] &\le (1+u)\bar{L}\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right] + (1+u^{-1})\bar{L}\mathbb{E}\left[\|x_{k+1} - w_{k+1}\|^2\right] \\ &= (1+u)\bar{L}\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right] + (1+u^{-1})\mathbf{G}^{\mathrm{P}}_{k+1}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{G}^{\mathrm{D}}_{k+1} \le{}& (1-\alpha_{\mathrm{D}})(1+s)(1+q)\mathbf{G}^{\mathrm{D}}_k + (1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u)\bar{L}\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right] \\ &+ (1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u^{-1})\mathbf{G}^{\mathrm{P}}_{k+1} + (1+s^{-1})\mathbf{G}^{\mathrm{P}}_{k+1}.\end{aligned}$$

Let us now bound the auxiliary term $\mathbf{G}^{\mathrm{P}}_{k+1}$. We notice that $\mathbf{G}^{\mathrm{P}}_k$ is the Lyapunov term of the P-ELF algorithm. Thus, from Proposition 4.2 we have

$$\begin{aligned}\mathbf{G}^{\mathrm{P}}_{k+1} &= \bar{L}\mathbb{E}\left[\|w_{k+1} - x_{k+1}\|^2\right] \\ &\le (1-\alpha_{\mathrm{P}})(1+w)\mathbf{G}^{\mathrm{P}}_k + (1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right].\end{aligned} \tag{9}$$

Recalling the definitions of $\lambda_1, \lambda_2, \lambda_3$ we deduce

$$\mathbf{G}^{\mathrm{D}}_{k+1} \le \lambda_1 \mathbf{G}^{\mathrm{D}}_k + \lambda_2 \mathbb{E}\left[\|x_k - x_{k+1}\|^2\right] + \lambda_3 \mathbf{G}^{\mathrm{P}}_k.$$

This concludes the proof of the proposition. □

B PROOFS OF THE MAIN THEOREMS

B.1 GENERAL SCHEME OF THE PROOFS

For all three algorithms the update of the LMC iteration is a stochastic estimator of the gradient $\nabla F(x_k)$. Generally, it depends on x_k and ξ_k, where ξ_k is a sequence of i.i.d. random variables defined on some probability space $(\Xi, \mathcal{F}, \mathcal{P})$. The sequence ξ_k comprises the randomness that arises at each step of the particular algorithm and it is independent of x_k. In order to prove convergence in KL divergence, we use the interpolation method proposed in [Vempala and Wibisono, 2019]. The method is based on the Fokker-Planck equation of the Langevin diffusion. We state a lemma for general LMC algorithms with stochastic drift terms. In particular, all our algorithms can be generally written as

$$x_{k+1} = x_k - \gamma f_{\xi_k}(x_k) + \sqrt{2\gamma} Z_k, \tag{10}$$

where ξ_k are i.i.d. random variables defined on some probability space $(\Xi, \mathcal{F}, \mathcal{P})$. On the other hand, each step can be seen as a realization of a Langevin diffusion with a constant drift term $f_{\xi_k}(x_k)$:

$$\mathrm{d}y_t = -f_{\xi_k}(x_k)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{11}$$

with $y_0 = x_k$ and $t \in [0, \gamma]$. Indeed,

$$\begin{aligned}y_\gamma &= y_0 - \int_0^\gamma f_{\xi_k}(y_0)\mathrm{d}t + \sqrt{2}(B_\gamma - B_0) \\ &= x_k - \gamma f_{\xi_k}(x_k) + \sqrt{2\gamma} Z_1 = x_{k+1}.\end{aligned}$$

The interpolation method is based on analyzing the Fokker-Planck equation of this diffusion. In particular, we will upper bound the time derivative of $H_\pi(\rho_t)$:

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} = \int_{\mathbb{R}^d} \frac{\partial \rho_t(z)}{\partial t} \log\left(\frac{\rho_t}{\pi}\right)(z)\mathrm{d}z. \tag{12}$$

Here, the first term of the product under the integral can be computed using the abovementioned Fokker-Planck equation. The following lemma is the cornerstone of our analysis.

Lemma B.1. *If y_t is the solution of the diffusion (11) and $\rho_t = \mathcal{L}(y_t)$, then for every $t \in [0, \gamma]$,*

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} \leq -\frac{3}{4} J_\pi(\rho_t) + \mathbb{E}\left[\left\|f_{\xi_k}(y_0) - \nabla F(y_t)\right\|^2\right]. \tag{13}$$

The bound (13) was initially derived by Vempala and Wibisono [2019] for the standard Langevin Monte-Carlo. Its current stochastic form was later proved in [Sun et al., 2022] for MARINA Langevin algorithm. The proof is postponed to Appendix C.1.

Lemma B.1 is valid for all our algorithms. We then insert the value of the gradient estimator for each method and bound the last term by $\mathbf{G}_k^{\mathrm{D}}$. Using the recurrent properties of the Lyapunov terms and replacing Fisher information term by Kullback-Leibler divergence with LSI inequality we conclude the proof.

B.2 SOME TECHNICAL LEMMAS

We will use repeatedly, sometimes without even mentioning, a simple inequality which is a consequence of Young's inequality. It goes as follows.

Lemma B.2. *For any two vectors $x, y \in \mathbb{R}^d$ and any $s > 0$*

$$\|x + y\|^2 \leq (1 + s)\|x\|^2 + (1 + s^{-1})\|y\|^2.$$

Proof.

$$\begin{aligned}\|x + y\|^2 &= \|x\|^2 + 2\langle x, y\rangle + \|y\|^2 \\ &\leq (1 + s)\|x\|^2 + (1 + s^{-1})\|y\|^2.\end{aligned}$$

The second passage is due to Young's inequality. □

We also use two lemmas from the literature, which we present below without proofs. The first one is an instance of Grönwall's inequality in its integral form. Its proof can be found in [Amann, 2011].

Lemma B.3 (Grönwall's Inequality)**.** *Assume $\phi, B : [0, T] \to \mathbb{R}$ are bounded non-negative measurable function and $C : [0, T] \to \mathbb{R}$ is a non-negative integrable function with the property that*

$$\phi(t) \leq B(t) + \int_0^t C(\tau)\phi(\tau)\mathrm{d}\tau \quad \text{for all } t \in [0, T]. \tag{14}$$

Then

$$\phi(t) \leq B(t) + \int_0^t B(s)C(s)\exp\left(\int_s^t C(\tau)\mathrm{d}\tau\right)\mathrm{d}s \quad \text{for all } t \in [0, T].$$

The second is a technical lemma borrowed from Chewi et al. [2021].

Lemma B.4. *Suppose that ∇F is L-Lipschitz. Then for any probability measure ν, the following inequality is satisfied:*

$$\mathbb{E}_\nu\left[\|\nabla F\|^2\right] \leq \mathbb{E}_\nu\left[\left\|\nabla \log\left(\frac{\nu}{\pi}\right)\right\|^2\right] + 2dL = J_\pi(\nu) + 2dL.$$

B.3 PROOF OF THEOREM 4.3

We follow the scheme described in Appendix B.1. Let us recall the initial setting first. The update rule of both D-ELF and P-ELF can be abstractly defined by

$$x_{k+1} = x_k - \gamma g_k + \sqrt{2\gamma} Z_k.$$

The vector g_k is a stochastic estimator of the potential function's gradient at the k-th iterate: $\nabla F(x_k)$. On the other hand, for each k the next iteration can be computed using the following SDE:

$$\mathrm{d}y_t = -g_k \mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{15}$$

with $y_0 = x_k$ and $t \in [0, \gamma]$. Then, as shown in Appendix B.1, $y_\gamma = x_{k+1}$. Denote by ρ_t the distribution of y_t. Lemma B.1 yields:

$$\begin{aligned}\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} &\leq -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\left\|f_{\xi_k}(y_0) - \nabla F(y_t)\right\|^2\right] \\ &\leq -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\left\|g_k - \nabla F(y_t)\right\|^2\right].\end{aligned} \tag{16}$$

The proof for D-ELF: The Lyapunov term for the D-ELF algorithm is defined as

$$\mathbf{G}_k^{\mathrm{D}} := \frac{1}{n}\sum_i^n \mathbb{E}\left[\left\|g_k^i - \nabla F_i(x_k)\right\|^2\right].$$

Next lemma bounds the second term in (16) using $\mathbf{G}_k^{\mathrm{D}}$.

Lemma B.5. *If $f_{\xi_k}(x_k)$ is the gradient estimator g_k from Algorithm 1, then ρ_t satisfies*

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} \leq -\frac{3}{4}J_\pi(\rho_t) + 2L^2\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{D}}. \tag{17}$$

Let us now add $C\mathbf{G}_{k+1}^{\mathrm{D}}$ to both sides of the inequality (17), where $C > 0$ is a constant to be determined later:

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} \leq -\frac{3}{4}J_\pi(\rho_t) + 2L^2\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{D}} + C\mathbf{G}_{k+1}^{\mathrm{D}}.$$

Combining Proposition 4.1 and Lemma B.5 we deduce

$$\begin{aligned}\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} &\leq -\frac{3}{4}J_\pi(\rho_t) + 2L^2\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{D}} \\ &\quad + C\left((1-p)\mathbf{G}_k^{\mathrm{D}} + (1-p)\beta_{\mathrm{D}}\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right]\right) \\ &= -\frac{3}{4}J_\pi(\rho_t) + \left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right)\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] + \left(2 + C(1-p)\right)\mathbf{G}_k^{\mathrm{D}}.\end{aligned}$$

The lemma below bounds the term $\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right]$.

Lemma B.6. *If $\gamma \leq \frac{1}{2\sqrt{2}L}$, then the iterates of the stochastic LMC algorithm (10) satisfy the following inequality, where $\mathbf{G}_k^{\mathrm{D}}$ is the Lyapunov term of D-ELF algorithm defined in (4):*

$$\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] \leq 8\gamma^2\mathbb{E}\left[\left\|\nabla F(y_t)\right\|^2\right] + 4\gamma^2\mathbf{G}_k^{\mathrm{D}} + 4d\gamma. \tag{18}$$

Lemma B.6 yields the following

$$\begin{aligned}\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} &\leq -\frac{3}{4}J_\pi(\rho_t) + \left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right)\left(8\gamma^2\mathbb{E}\left[\left\|\nabla F(y_t)\right\|^2\right] + 4\gamma^2\mathbf{G}_k^{\mathrm{D}} + 4d\gamma\right) \\ &\quad + \left(2 + C(1-p)\right)\mathbf{G}_k^{\mathrm{D}}.\end{aligned}$$

Let us now apply Lemma B.4 to the right-hand side. We obtain

$$\begin{aligned}\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} &\leq -\frac{3}{4}J_\pi(\rho_t) + \left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right)\left(8\gamma^2\left(J_\pi(\rho_t) + 2dL\right) + 4\gamma^2\mathbf{G}_k^{\mathrm{D}} + 4d\gamma\right) \\ &\quad + \left(2 + C(1-p)\right)\mathbf{G}_k^{\mathrm{D}} \\ &= -\left(\frac{3}{4} - 8\gamma^2\left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right)\right)J_\pi(\rho_t) \\ &\quad + \left(8L^2\gamma^2 + C(1-p)\left(4\gamma^2\beta_{\mathrm{D}} + 1\right) + 2\right)\mathbf{G}_k^{\mathrm{D}} \\ &\quad + \left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right)\left(16L\gamma^2 d + 4d\gamma\right).\end{aligned}$$

From the definition of τ we obtain the following:

$$\begin{aligned}\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} \leq &-\left(\frac{3}{4} - 8\gamma^2\left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right)\right) J_\pi(\rho_t) \\ &+ \left(8L^2\gamma^2 + C(1-p)\left(4\gamma^2\beta_{\mathrm{D}}+1\right)+2\right)\mathbf{G}_k^{\mathrm{D}} + \tau.\end{aligned} \tag{19}$$

Let $C = \left(8L^2\gamma^2 + C(1-p)\left(4\gamma^2\beta_{\mathrm{D}}+1\right)+2\right)e^{\mu\gamma}$. Solving this linear equation w.r.t. C, we get

$$C = \frac{8L^2\gamma^2+2}{e^{-\mu\gamma} - (1-p)\left(4\gamma^2\beta_{\mathrm{D}}+1\right)}. \tag{20}$$

Without loss of generality we may assume that $\mu\gamma < 1$ and thus we have $e^{\mu\gamma} \leq 1 + 2\mu\gamma$. In order for C to be positive, we need to assure that

$$1 - (1-p)\left(4\beta_{\mathrm{D}}\gamma^2+1\right)(1+2\mu\gamma) > 0.$$

The latter is equivalent to

$$\frac{1-p}{p}8\mu\beta_{\mathrm{D}}\gamma^3 + \frac{1-p}{p}4\beta_{\mathrm{D}}\gamma^2 + \frac{1-p}{p}2\mu\gamma < 1.$$

A simple solution to this inequality is to make all three terms smaller than $1/3$. The latter is equivalent to

$$\gamma < \min\left\{\left(\frac{p}{24\mu\beta_{\mathrm{D}}(1-p)}\right)^{1/3}, \left(\frac{p}{12\beta_{\mathrm{D}}(1-p)}\right)^{1/2}, \frac{p}{6\mu(1-p)}\right\}. \tag{21}$$

On the other hand, we will require the coefficient of $J_\pi(\rho_t)$ in (19) to be negative. This is to ensure contraction. That means

$$8\gamma^2\left(2L^2 + C(1-p)\beta_{\mathrm{D}}\right) = 8\gamma^2\left(2L^2 + \frac{(8L^2\gamma^2+2)(1-p)\beta_{\mathrm{D}}}{e^{-\mu\gamma} - (1-p)\left(4\gamma^2\beta_{\mathrm{D}}+1\right)}\right) \leq \frac{1}{4}.$$

Solving this inequality we get

$$\gamma \leq \frac{1}{2}\sqrt{\frac{1-(1-p)e^{\mu\gamma}}{(16+(1-p)(17\beta_{\mathrm{D}}-16)e^{\mu\gamma})}}. \tag{22}$$

From (21), we know that $\gamma < \frac{p}{6\mu(1-p)}$, so $e^{\mu\gamma} \leq 1 + 2\mu\gamma \leq 1 + \frac{p}{3(1-p)}$. Inserting this upper bound into (22), we get a lower bound on the right hand side. That is

$$\begin{aligned}\frac{1}{2}\sqrt{\frac{2p}{[17\beta_{\mathrm{D}}(3-2p)+32p]}} &= \frac{1}{2}\sqrt{\frac{1-(1-p)(1+\frac{p}{3(1-p)})}{\left(16+(1-p)(17\beta_{\mathrm{D}}-16)(1+\frac{p}{3(1-p)})\right)}} \\ &\leq \frac{1}{2}\sqrt{\frac{1-(1-p)e^{\mu\gamma}}{(16+(1-p)(17\beta_{\mathrm{D}}-16)e^{\mu\gamma})}}.\end{aligned}$$

So we need

$$\gamma < \min\left\{\frac{1}{2}\sqrt{\frac{2p}{[17\beta_{\mathrm{D}}(3-2p)+32p]}}, \left(\frac{p}{24\mu\beta_{\mathrm{D}}(1-p)}\right)^{1/3}, \left(\frac{p}{12\beta_{\mathrm{D}}(1-p)}\right)^{1/2}, \frac{p}{6\mu(1-p)}\right\}.$$

We can further simplify this inequality. The first and third terms are larger than $a := \frac{1}{14}\sqrt{\frac{p}{(1+\beta_{\mathrm{D}})}}$, while as the fourth term is larger than $b := \frac{p}{6\mu}$. On the other hand, $\min\{a,b\}$ is less than the second term. Indeed,

$$\min\{a,b\} \leq a^{2/3}b^{1/3} = \left(\frac{p^2}{1176\mu(1+\beta_{\mathrm{D}})}\right)^{1/3} \leq \left(\frac{p}{24\mu\beta_{\mathrm{D}}(1-p)}\right)^{1/3}.$$

Summing up, we obtain the following bound on the step-size that guarantees $C \geq 0$ and (22):

$$\gamma \leq \min\left\{\frac{1}{14}\sqrt{\frac{p}{(1+\beta_{\mathrm{D}})}}, \frac{p}{6\mu}\right\}.$$

Therefore, the above the conditions are satisfies. This yields the following:

$$\frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} \leq -\frac{1}{2}J_\pi\left(\rho_t\right) + e^{-\mu\gamma}C\mathbf{G}_k^{\mathrm{D}} + C\tau. \tag{23}$$

Since π satisfies Log-Sobolev inequality, we deduce

$$\frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} \leq -\mu H_\pi\left(\rho_t\right) + e^{-\mu\gamma}C\mathbf{G}_k^{\mathrm{D}} + C\tau. \tag{24}$$

One may check that the equivalent integral form of (24) satisfies (14) with $\phi(t) = H_\pi\left(\rho_t\right)$, $B(t) = \left(e^{-\mu\gamma}C\mathbf{G}_k^{\mathrm{D}} - C\mathbf{G}_{k+1}^{\mathrm{D}} + \tau\right)t + H_\pi\left(\rho_{k\gamma}\right)$, $C(t) = -\mu$. Therefore, from Lemma B.3 we deduce

$$H_\pi\left(\rho_t\right) \leq e^{-\mu t}H_\pi\left(\rho_{k\gamma}\right) + \frac{1-e^{-\mu t}}{\mu}\left(e^{-\mu\gamma}C\mathbf{G}_k^{\mathrm{D}} - C\mathbf{G}_{k+1}^{\mathrm{D}} + C\tau\right),$$

let $t = \gamma$ and $\beta = e^{\mu\gamma}$, then we have

$$\begin{aligned} H_\pi\left(\rho_{(k+1)\gamma}\right) + \frac{1-e^{-\mu\gamma}}{\mu}C\mathbf{G}_{k+1}^{\mathrm{D}} &\leq e^{-\mu\gamma}\left(H_\pi\left(\rho_{k\gamma}\right) + e^{\mu\gamma}\frac{1-e^{-\mu\gamma}}{\mu}\beta^{-1}C\mathbf{G}_k^{\mathrm{D}}\right) + \frac{1-e^{-\mu\gamma}}{\mu}C\tau \\ &= e^{-\mu\gamma}\left(H_\pi\left(\rho_{k\gamma}\right) + \frac{1-e^{-\mu\gamma}}{\mu}C\mathbf{G}_k^{\mathrm{D}}\right) + \frac{1-e^{-\mu\gamma}}{\mu}C\tau. \end{aligned} \tag{25}$$

Repeating this step for $k = 0, 1, 2, \cdots, K-1$, we obtain

$$\mathbf{H}_K \leq e^{-K\mu\gamma}\mathbf{H}_0 + \frac{1-e^{-K\mu\gamma}}{\mu}\tau.$$

This proves Theorem 4.3 for D-ELF.

The proof for P-ELF: The gradient estimator $\nabla f_{\xi_k}(x_k)$ in this case is equal to

$$\nabla f_{\xi_k}(x_k) = \nabla F(w_k) = \frac{1}{n}\sum_{i=1}^{n}\nabla F_i(w_k).$$

From L_i-smoothness of the i-th component function F_i we deduce the following relation:

$$\begin{aligned} \mathbf{G}_k^{\mathrm{D}} &= \frac{1}{n}\sum_{i}^{n}\mathbb{E}\left[\left\|\nabla F_i(w_k) - \nabla F_i(x_k)\right\|^2\right] \\ &\leq \frac{1}{n}\sum_{i}^{n}\mathbb{E}\left[L_i^2\left\|w_k - x_k\right\|^2\right] \\ &= \mathbf{G}_k^{\mathrm{P}}. \end{aligned} \tag{26}$$

Therefore, combining this inequality with Lemma B.5 we obtain

$$\begin{aligned} \frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} &\leq -\frac{3}{4}J_\pi\left(\rho_t\right) + 2L^2\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{D}} \\ &\leq -\frac{3}{4}J_\pi\left(\rho_t\right) + 2L^2\mathbb{E}\left[\left\|x_{k+1} - x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{P}}. \end{aligned}$$

The latter means that we can repeat exactly the rest of the proof of D-ELF by replacing $\mathbf{G}_k^{\mathrm{D}}$ with $\mathbf{G}_k^{\mathrm{P}}$ and using Proposition 4.2 instead of Proposition 4.1. Therefore,

$$\mathbf{H}_K \leq e^{-K\mu\gamma}\mathbf{H}_0 + \frac{1-e^{-K\mu\gamma}}{\mu}\tau.$$

This concludes the proof of Theorem 4.3.

B.4 PROOF OF THEOREM 4.5

We recall the definition of the Lyapunov term $\mathbf{G}_k^{\mathrm{D}}$:

$$\mathbf{G}_k^{\mathrm{D}} := \frac{1}{n}\sum_i^n \mathbb{E}\left[\left\|g_k^i - \nabla F_i(x_k)\right\|^2\right].$$

As described in Appendix B.1, we use the interpolation proof scheme. That is for the k-th iteration we define the process y_t as in (11). Thus, from Lemma B.1 we have

$$\begin{aligned}\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} &\leq -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\left\|f_{\xi_k}(y_0) - \nabla F(y_t)\right\|^2\right]\\ &= -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\left\|g_0 - \nabla F(y_t)\right\|^2\right].\end{aligned}$$

Combining this with Proposition A.1 and (9), we obtain

$$\begin{aligned}&\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} + D\mathbf{G}_{k+1}^{\mathrm{P}}\\ &\quad\leq -\frac{3}{4}J_\pi(\rho_t) + 2L^2\mathbb{E}\left[\left\|x_{k+1}-x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{D}} + C\mathbf{G}_{k+1}^{\mathrm{D}} + D\mathbf{G}_{k+1}^{\mathrm{P}}\\ &\quad\leq -\frac{3}{4}J_\pi(\rho_t) + 2L^2\mathbb{E}\left[\left\|x_{k+1}-x_k\right\|^2\right] + 2\mathbf{G}_k^{\mathrm{D}} + C\left(\lambda_1\mathbf{G}_k^{\mathrm{D}} + \lambda_2\mathbb{E}\left[\left\|x_k - x_{k+1}\right\|^2\right] + \lambda_3\mathbf{G}_k^{\mathrm{P}}\right)\\ &\quad + D\left((1-\alpha_{\mathrm{P}})(1+w)\mathbf{G}_k^{\mathrm{P}} + (1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\mathbb{E}\left[\left\|x_k - x_{k+1}\right\|^2\right]\right)\\ &\quad= -\frac{3}{4}J_\pi(\rho_t) + \left(2L^2 + C\lambda_2 + D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\mathbb{E}\left[\left\|x_k - x_{k+1}\right\|^2\right]\\ &\quad+ (2 + C\lambda_1)\mathbf{G}_k^{\mathrm{D}} + \left(C\lambda_3 + D(1-\alpha_{\mathrm{P}})(1+w)\right)\mathbf{G}_k^{\mathrm{P}}.\end{aligned}$$

Lemma B.6 yields

$$\mathbb{E}\left[\left\|x_{k+1}-x_k\right\|^2\right] \leq 8\gamma^2\mathbb{E}\left[\left\|\nabla F(y_t)\right\|^2\right] + 4\gamma^2\mathbf{G}_k^{\mathrm{D}} + 4d\gamma,$$

for $\gamma < 1/8L$. The latter condition on the step-size is a consequence of our assumptions from the statement of Theorem 4.5. Therefore,

$$\begin{aligned}&\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} + D\mathbf{G}_{k+1}^{\mathrm{P}}\\ &\quad\leq -\frac{3}{4}J_\pi(\rho_t) + \left(2L^2 + C\lambda_2 + D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\left(8\gamma^2\mathbb{E}\left[\left\|\nabla F(y_t)\right\|^2\right] + 4\gamma^2\mathbf{G}_k^{\mathrm{D}} + 4d\gamma\right)\\ &\quad+ (2 + C\lambda_1)\mathbf{G}_k^{\mathrm{D}} + \left(C\lambda_3 + D(1-\alpha_{\mathrm{P}})(1+w)\right)\mathbf{G}_k^{\mathrm{P}}.\end{aligned}$$

Applying Lemma B.4 we deduce

$$\begin{aligned}&\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}_{k+1}^{\mathrm{D}} + D\mathbf{G}_{k+1}^{\mathrm{P}}\\ &\quad\leq -\frac{3}{4}J_\pi(\rho_t) + \left(2L^2 + C\lambda_2 + D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\left(8\gamma^2\left[J_\pi(\rho_t) + 2dL\right] + 4\gamma^2\mathbf{G}_k^{\mathrm{D}} + 4d\gamma\right)\\ &\quad+ (2 + C\lambda_1)\mathbf{G}_k^{\mathrm{D}} + \left(C\lambda_3 + D(1-\alpha_{\mathrm{P}})(1+w)\right)\mathbf{G}_k^{\mathrm{P}}\\ &\quad= \left(-\frac{3}{4} + 8\gamma^2\left(2L^2 + C\lambda_2 + D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\right)J_\pi(\rho_t)\\ &\quad+ \left\{2 + C\lambda_1 + 4\gamma^2\left(2L^2 + C\lambda_2 + D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\right\}\mathbf{G}_k^{\mathrm{D}} + \left(C\lambda_3 + D(1-\alpha_{\mathrm{P}})(1+w)\right)\mathbf{G}_k^{\mathrm{P}}\\ &\quad+ \left(2L^2 + C\lambda_2 + D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\left(16\gamma^2 dL + 4d\gamma\right).\end{aligned}$$

Let us choose C and D to satisfy

$$C = \frac{2.125}{e^{-\mu\gamma} - \lambda_1} \quad \text{and} \quad D = \frac{2.125\lambda_3}{\left(e^{-\mu\gamma} - \lambda_1\right)\left(e^{-\mu\gamma} - (1-\alpha_{\mathrm{P}})(1+w)\right)}, \tag{27}$$

where μ is the constant from Log-Sobolev inequality. In order for C and D to be positive we need λ_1 and $(1-\alpha_{\mathrm{P}})(1+w)$ to be smaller than $e^{-\mu\gamma}$. We will choose w and $q=s$ as solutions to the following equations:

$$\begin{aligned}\lambda_1 = (1-\alpha_{\mathrm{D}})(1+q)^2 = 1-\frac{\alpha_{\mathrm{D}}}{2};\\ (1-\alpha_{\mathrm{P}})(1+w) = 1-\frac{\alpha_{\mathrm{P}}}{2}.\end{aligned} \tag{28}$$

Then,

$$e^{-\mu\gamma} > 1-\mu\gamma > \max\left\{1-\alpha_{\mathrm{D}}/4, 1-\alpha_{\mathrm{P}}/4\right\} \tag{29}$$

thus the denominators are positive. Furthermore,

$$D = \frac{2.125\lambda_3}{\left(e^{-\mu\gamma}-\lambda_1\right)\left(e^{-\mu\gamma}-(1-\alpha_{\mathrm{P}})(1+w)\right)} \leq \frac{4C\lambda_3}{\alpha_{\mathrm{P}}}.$$

Recall that the definitions of λ_2 and λ_3 are given in (8). Since $(1-\alpha_{\mathrm{P}})(1+w)<1$, from the definition of λ_3 we have

$$\begin{aligned}\lambda_3 &= \left(2(1-\alpha_{\mathrm{D}})(1+q)(1+q^{-1})+(1+q^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w)\\ &\leq \left(2(1-\alpha_{\mathrm{D}})(2+q+q^{-1})+(1+q^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w)\\ &\leq \left(2(1-\alpha_{\mathrm{D}})(2+q+q^{-1})+(1+q^{-1})\right).\end{aligned}$$

Therefore, (8) implies

$$\lambda_3(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L} = \left(2(1-\alpha_{\mathrm{D}})(2+q+q^{-1})+(1+q^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L} \leq \lambda_2.$$

Thus,

$$\begin{aligned}\gamma^2\left(2L^2+C\lambda_2+D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right) &\leq \gamma^2\left(2L^2+C\lambda_2+\frac{4C\lambda_3}{\alpha_{\mathrm{P}}}(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\\ &\leq \gamma^2\left(2L^2+C\lambda_2+\frac{4C\lambda_2}{\alpha_{\mathrm{P}}}\right)\\ &\leq \gamma^2\left(2L^2+\frac{5C\lambda_2}{\alpha_{\mathrm{P}}}\right).\end{aligned}$$

The next lemma bounds the right hand side of the previous inequality by a constant. This will allow us to get a negative coefficient for the $J_\pi(\rho_t)$ term.

Lemma B.7. *Suppose $u=1$, $q=s$, C and D are defined as in (27). Let (28) and (29) also be true. Under the assumptions of Theorem 4.5, the step-size satisfies the following inequality:*

$$\gamma^2\left(2L^2+\frac{5C\lambda_2}{\alpha_{\mathrm{P}}}\right) < \frac{1}{32}.$$

The proof is postponed to Appendix C.4. Applying Lemma B.7 to the first term we finally obtain the following recurrent inequality

$$\begin{aligned}&\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}^{\mathrm{D}}_{k+1} + D\mathbf{G}^{\mathrm{P}}_{k+1}\\ &\quad\leq -\frac{1}{2}J_\pi(\rho_t) + (2.125+C\lambda_1)\mathbf{G}^{\mathrm{D}}_k + (C\lambda_3+D(1-\alpha_{\mathrm{P}})(1+w))\mathbf{G}^{\mathrm{P}}_k\\ &\quad+ \left(2L^2+C\lambda_2+D(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\right)\left(16\gamma^2 dL+4d\gamma\right)\\ &\quad\leq -\frac{1}{2}J_\pi(\rho_t) + (2.125+C\lambda_1)\mathbf{G}^{\mathrm{D}}_k + (C\lambda_3+D(1-\alpha_{\mathrm{P}})(1+w))\mathbf{G}^{\mathrm{P}}_k\\ &\quad+ \underbrace{\left(2L^2+\frac{5C\lambda_2}{\alpha_{\mathrm{P}}}\right)\left(16\gamma^2 dL+4d\gamma\right)}_{:=\tau}.\end{aligned}$$

Then, inserting the values of C and D, we get

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}^{\mathrm{D}}_{k+1} + D\mathbf{G}^{\mathrm{P}}_{k+1} \leq -\frac{1}{2}J_\pi(\rho_t) + e^{-\mu\gamma}C\mathbf{G}^{\mathrm{D}}_k + e^{-\mu\gamma}D\mathbf{G}^{\mathrm{P}}_k + \tau.$$

Let us now apply LSI:

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} + C\mathbf{G}^{\mathrm{D}}_{k+1} + D\mathbf{G}^{\mathrm{P}}_{k+1} \quad \leq -\mu H_\pi(\rho_t) + e^{-\mu\gamma}C\mathbf{G}^{\mathrm{D}}_k + e^{-\mu\gamma}D\mathbf{G}^{\mathrm{P}}_k + \tau.$$

Hence, the derivative of the function $H_\pi(\rho_t)$ is bounded by itself plus a term that does not depend on t. Lemma B.3 yields the following:

$$H_\pi(\rho_t) \leq e^{-\mu t}H_\pi(\rho_0) + \frac{1-e^{-\mu t}}{\mu}\left(e^{-\mu\gamma}C\mathbf{G}^{\mathrm{D}}_k + e^{-\mu\gamma}D\mathbf{G}^{\mathrm{P}}_k - C\mathbf{G}^{\mathrm{D}}_{k+1} - D\mathbf{G}^{\mathrm{P}}_{k+1} + \tau\right).$$

In particular, for $t=\gamma$, we have

$$\begin{aligned} H_\pi(\rho_\gamma) + \frac{1-e^{-\mu\gamma}}{\mu}\left(C\mathbf{G}^{\mathrm{D}}_{k+1} + D\mathbf{G}^{\mathrm{P}}_{k+1}\right) &\leq e^{-\mu\gamma}H_\pi(\rho_0) + \frac{1-e^{-\mu\gamma}}{\mu}\left(e^{-\mu\gamma}C\mathbf{G}^{\mathrm{D}}_k + e^{-\mu\gamma}D\mathbf{G}^{\mathrm{P}}_k + \tau\right) \\ &= e^{-\mu\gamma}\left[H_\pi(\rho_0) + \frac{1-e^{-\mu\gamma}}{\mu}\left(C\mathbf{G}^{\mathrm{D}}_k + D\mathbf{G}^{\mathrm{P}}_k\right)\right] + \frac{1-e^{-\mu\gamma}}{\mu}\tau. \end{aligned}$$

We first recall that $\rho_\gamma = \nu_{K+1}$ and $\rho_0 = \nu_K$. Repeating this inequality recurrently we deduce the following bound:

$$H_\pi(\nu_K) + \frac{1-e^{-\mu\gamma}}{\mu}\left(C\mathbf{G}^{\mathrm{D}}_K + D\mathbf{G}^{\mathrm{P}}_K\right) \leq e^{-\mu\gamma K}\left[H_\pi(\rho_0) + \frac{1-e^{-\mu\gamma}}{\mu}\left(C\mathbf{G}^{\mathrm{D}}_0 + D\mathbf{G}^{\mathrm{P}}_0\right)\right] + \frac{\tau}{\mu}.$$

This concludes the proof of Theorem 4.5.

Remark B.8. One may check, that repeating the analysis for the case when one of the compressor operators ($\alpha = 1$) is the identity, we will recover the previously known algorithms.

B.5 PROOF OF CORROLLARY 4.7

First let us upper bound τ. Similar to the proof of Corrollary 4.6, $\left(16\gamma^2 dL + 4d\gamma\right) < 5d\gamma$. Thus,

$$\begin{aligned} \tau &\leq \left(2L^2 + \frac{5C\lambda_2}{\alpha_{\mathrm{P}}}\right)5d\gamma \leq \frac{45\lambda_2}{\alpha_{\mathrm{D}}\alpha_{\mathrm{P}}}5d\gamma \\ &= \mathcal{O}\left(\frac{\left(1-\frac{\alpha_{\mathrm{D}}}{2}\right)\left(1-\frac{\alpha_{\mathrm{P}}}{2}\right)}{qw\alpha_{\mathrm{D}}\alpha_{\mathrm{P}}\left(1-\alpha_{\mathrm{P}}\right)\left(1-\alpha_{\mathrm{D}}\right)}\bar{L}d\gamma\right) \\ &= \mathcal{O}\left(\frac{\bar{L}d\gamma}{qw\alpha_{\mathrm{D}}\alpha_{\mathrm{P}}}\right). \end{aligned}$$

C PROOFS OF THE LEMMAS

C.1 PROOF OF LEMMA B.1

Let ρ_{0t} denote the joint distribution of (y_0, ξ, y_t), which we write in terms of the conditionals and marginals as

$$\rho_{0t}(z, y_0, \xi) = \rho_0(y_0, \xi)\,\rho_{t|0}(z \mid y_0, \xi) = \rho_t(z)\,\rho_{0|t}(y_0, \xi \mid z).$$

Conditioning on (y_0, ξ), the drift vector field $f_{\xi_k}(y_0)$ is a constant, so the Fokker-Planck formula for the conditional density $\rho_{t|0}(z \mid y_0, \xi)$ is given by

$$\frac{\partial\rho_{t|0}(z \mid y_0, \xi)}{\partial t} = \nabla_z \cdot \left(\rho_{t|0}(z \mid y_0, \xi)\, f_\xi(y_0)\right) + \Delta\rho_{t|0}(z \mid y_0, \xi). \tag{30}$$

To derive the evolution of ρ_t, we integrate w.r.t. $(y_0, \xi) \sim \rho_0$:

$$\begin{aligned} \frac{\partial\rho_t(z)}{\partial t} &= \int_{\mathbb{R}^d\times\Xi} \frac{\partial\rho_{t|0}(z \mid y_0, \xi)}{\partial t}\rho_0(y_0, \xi)\,\mathrm{d}y_0\mathrm{d}\xi \\ &\overset{(30)}{=} \int_{\mathbb{R}^d\times\Xi} \left(\nabla_z \cdot \left(\rho_{t|0}(z \mid y_0, \xi)\, f_\xi(y_0)\right) + \Delta\rho_{t|0}(z \mid y_0, \xi)\right)\rho_0(y_0, \xi)\,\mathrm{d}y_0\mathrm{d}\xi. \end{aligned} \tag{31}$$

Using the definition of conditional densities and Fubini's theorem we deduce

$$\begin{aligned}\frac{\partial \rho_t(z)}{\partial t} &= \int_{\mathbb{R}^d \times \Xi} \left(\nabla_z \cdot \left(\rho_{0t}\left(z, y_0, \xi\right) f_\xi\left(y_0\right)\right) + \Delta \rho_{0t}\left(z, y_0, \xi\right)\right) \mathrm{d}y_0 \mathrm{d}\xi \\ &= \nabla_z \cdot \left(\rho_t(z) \int_{\mathbb{R}^d \times \Xi} \rho_{0|t}\left(y_0, \xi \mid z\right) f_\xi\left(y_0\right) \mathrm{d}y_0 \mathrm{d}\xi\right) + \Delta \rho_t(z) \\ &= \nabla_z \cdot \left(\rho_t(z) \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right]\right) + \Delta \rho_t(z).\end{aligned} \tag{32}$$

Writing down the definition of KL divergence and using Fubini's theorem, we deduce

$$\begin{aligned}\frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} &= \int_{\mathbb{R}^d} \frac{\partial \rho_t(z)}{\partial t} \log\left(\frac{\rho_t}{\pi}\right)(z) \mathrm{d}z \\ &= \int_{\mathbb{R}^d} \left(\nabla_z \cdot \left(\rho_t(z) \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right]\right) + \Delta \rho_t(z)\right) \log\left(\frac{\rho_t}{\pi}\right)(z) \mathrm{d}z \\ &= -\int_{\mathbb{R}^d} \left\langle \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right] + \nabla \log(\rho_t)(z), \nabla \log\left(\frac{\rho_t}{\pi}\right)(z) \right\rangle \rho_t(z) \mathrm{d}z \\ &= -\int_{\mathbb{R}^d} \left(\nabla \log\left(\frac{\rho_t}{\pi}\right)(z) - \nabla \log\left(\frac{\rho_t}{\pi}\right)(z) + \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right] + \nabla \log(\rho_t)(z)\right)^\top \\ &\qquad \times \nabla \log\left(\frac{\rho_t}{\pi}\right)(z) \rho_t(z) \mathrm{d}z \\ &= -\int_{\mathbb{R}^d} \left\langle \nabla \log\left(\frac{\rho_t}{\pi}\right)(z) + \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right] - \nabla F(z), \nabla \log\left(\frac{\rho_t}{\pi}\right)(z) \right\rangle \rho_t(z) \mathrm{d}z.\end{aligned} \tag{33}$$

We recall the definition of Fisher information to bound the first term of the scalar product:

$$\frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} \leq -J_\pi\left(\rho_t\right) - \int_{\mathbb{R}^d} \left\langle \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right] - \nabla F(z), \nabla \log\left(\frac{\rho_t}{\pi}\right)(z) \right\rangle \rho_t(z) \mathrm{d}z. \tag{34}$$

From the Cauchy-Schwartz inequality, we deduce

$$\begin{aligned}\frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} &\leq -J_\pi\left(\rho_t\right) + \frac{1}{4} J_\pi\left(\rho_t\right) + \int_{\mathbb{R}^d} \left\| \mathbb{E}_{\rho_{0|t}}\left[f_\xi\left(y_0\right) \mid y_t = z\right] - \nabla F(z) \right\|^2 \rho_t(z) \mathrm{d}z \\ &= -\frac{3}{4} J_\pi\left(\rho_t\right) + \mathbb{E}\left[\left\| \mathbb{E}\left[f_{\xi_k}(y_0) - \nabla F(y_t) \mid y_t\right] \right\|^2\right] \\ &\leq -\frac{3}{4} J_\pi\left(\rho_t\right) + \mathbb{E}\left[\mathbb{E}\left[\left\| f_{\xi_k}(y_0) - \nabla F(y_t) \right\|^2 \mid y_t\right]\right] \\ &= -\frac{3}{4} J_\pi\left(\rho_t\right) + \mathbb{E}\left[\left\| f_{\xi_k}(y_0) - \nabla F(y_t) \right\|^2\right].\end{aligned} \tag{35}$$

This concludes the proof of the lemma.

C.2 PROOF OF LEMMA B.5

If we replace $f_{\xi_k}(y_0)$ by g_0 in (13), we will have

$$\begin{aligned}\frac{\mathrm{d}H_\pi\left(\rho_t\right)}{\mathrm{d}t} &\leq -\frac{3}{4} J_\pi\left(\rho_t\right) + \mathbb{E}\left[\left\| \nabla F(y_t) - g_0 \right\|^2\right] \\ &\leq -\frac{3}{4} J_\pi\left(\rho_t\right) + 2\mathbb{E}\left[\left\| \nabla F(y_t) - \nabla F(y_0) \right\|^2\right] + 2\mathbb{E}\left[\left\| \nabla F(x_0) - g_0 \right\|^2\right] \\ &= -\frac{3}{4} J_\pi\left(\rho_t\right) + 2\mathbb{E}\left[\left\| \nabla F(y_t) - \nabla F(x_0) \right\|^2\right] + 2\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \left\{\nabla F_i(x_0) - g_0^i\right\} \right\|^2\right] \\ &\leq -\frac{3}{4} J_\pi\left(\rho_t\right) + 2\mathbb{E}\left[\left\| \nabla F(y_t) - \nabla F(x_0) \right\|^2\right] + 2\mathbf{G}_0^{\mathrm{D}}.\end{aligned}$$

Here the last implication is due to Jensen's inequality. Let us bound the second term. The smoothness of the gradient yields

$$\mathbb{E}\left[\left\| \nabla F(y_t) - \nabla F(x_0) \right\|^2\right] \leq L^2 \mathbb{E}\left[\left\| y_t - x_0 \right\|^2\right] = L^2 \mathbb{E}\left[\left\| t g_0 + \sqrt{2}\left(B_t - B_0\right) \right\|^2\right]. \tag{36}$$

Since the Brownian process has independent increments we get

$$\begin{aligned}\mathbb{E}\left[\|\nabla F(y_t)-\nabla F(x_0)\|^2\right] &\leq L^2t^2\|g_0\|^2+2tL^2d\\ &\leq L^2\gamma^2\|g_0\|^2+2hL^2d\\ &= L^2\mathbb{E}\left[\|x_1-x_0\|^2\right].\end{aligned}\tag{37}$$

This concludes the proof.

C.3 PROOF OF LEMMA B.6

Let us apply Lemma B.4 to bound the term $\mathbb{E}\left[\|x_{k+1}-x_k\|^2\right]$:

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1}-x_k\|^2\right] &= \gamma^2\mathbb{E}\left[\|g_k\|^2\right]+2d\gamma\\ &\leq 2\gamma^2\left(\mathbb{E}\left[\|\nabla F(x_k)\|^2\right]+\mathbb{E}\left[\|\nabla F(x_k)-g_k\|^2\right]\right)+2d\gamma\\ &\leq 2\gamma^2\mathbb{E}\left[\|\nabla F(x_k)\|^2\right]+2\gamma^2\mathbf{G}_k^{\mathrm{D}}+2d\gamma\\ &\leq 4\gamma^2\left(\mathbb{E}\left[\|\nabla F(y_t)\|\right]+\mathbb{E}\left[\|\nabla F(y_t)-\nabla F(x_k)\|^2\right]\right)+2\gamma^2\mathbf{G}_k^{\mathrm{D}}+2d\gamma\\ &\leq 4\gamma^2\mathbb{E}\left[\|\nabla F(y_t)\|\right]+4L^2\gamma^2\mathbb{E}\left[\|x_t-x_k\|^2\right]+2\gamma^2\mathbf{G}_k^{\mathrm{D}}+2d\gamma\\ &\leq 4\gamma^2\mathbb{E}\left[\|\nabla F(y_t)\|\right]+4L^2\gamma^2\mathbb{E}\left[\|x_{k+1}-x_k\|^2\right]+2\gamma^2\mathbf{G}_k^{\mathrm{D}}+2d\gamma.\end{aligned}$$

Regrouping the terms we obtain

$$(1-4L^2\gamma^2)\mathbb{E}\left[\|x_{k+1}-x_k\|^2\right]\leq 4\gamma^2\mathbb{E}\left[\|\nabla F(y_t)\|\right]+2\gamma^2\mathbf{G}_k^{\mathrm{D}}+2d\gamma.$$

Dividing both sides on $1-4L^2\gamma^2$ and recalling that $2\sqrt{2}L\gamma<1$, we conclude the proof.

C.4 PROOF OF LEMMA B.7

Is sufficient to show that

$$\gamma^2\leq\min\left\{\frac{1}{192L^2},\frac{\alpha_{\mathrm{P}}}{240C\lambda_2}\right\}.$$

From the assumption of the theorem, we know that $\gamma^2\leq\frac{1}{192L^2}$. Thus it remains to show that γ^2 is bounded by the minimum of the other two terms:

$$\gamma^2\leq\frac{\alpha_{\mathrm{P}}}{240C\lambda_2}=\frac{\alpha_{\mathrm{P}}\left(e^{-\mu\gamma}-\lambda_1\right)}{510\lambda_2}.$$

Since $u=1$ and $s=q$ we have the following bound on λ_2:

$$\begin{aligned}\lambda_2 &\leq \left[2(1+q)(1+q^{-1})+\left(2(1+q)(1+q^{-1})+(1+q^{-1})\right)(1+w^{-1})\right]\bar{L}\\ &= \left[2(2+q+q^{-1})+\left(2(2+q+q^{-1})+(1+q^{-1})\right)(1+w^{-1})\right]\bar{L}\\ &= \frac{1}{q}\left[2(2q+q^2+1)+\left(2(2q+q^2+1)+(q+1)\right)(1+w^{-1})\right]\bar{L}\\ &\leq \frac{1}{qw}5(q+1)^2(1+w)\bar{L}\\ &\leq \frac{5}{qw}\frac{\left(1-\frac{\alpha_{\mathrm{D}}}{2}\right)\left(1-\frac{\alpha_{\mathrm{P}}}{2}\right)}{\left(1-\alpha_{\mathrm{P}}\right)\left(1-\alpha_{\mathrm{D}}\right)}\bar{L}.\end{aligned}$$

Therefore, we have an upper bound on λ_2. This means that it is sufficient for us to prove

$$\gamma^2\leq\frac{\alpha_{\mathrm{P}}\left(e^{-\mu\gamma}-\lambda_1\right)}{510\frac{5}{qw}\frac{\left(1-\frac{\alpha_{\mathrm{D}}}{2}\right)\left(1-\frac{\alpha_{\mathrm{P}}}{2}\right)}{\left(1-\alpha_{\mathrm{P}}\right)\left(1-\alpha_{\mathrm{D}}\right)}\bar{L}}\quad=\frac{qw\alpha_{\mathrm{P}}\left(e^{-\mu\gamma}-\lambda_1\right)}{2550\bar{L}}\cdot\frac{\left(1-\alpha_{\mathrm{P}}\right)\left(1-\alpha_{\mathrm{D}}\right)}{\left(1-\frac{\alpha_{\mathrm{D}}}{2}\right)\left(1-\frac{\alpha_{\mathrm{P}}}{2}\right)}.$$

From $\mu\gamma < \min\{\alpha_{\rm D}, \alpha_{\rm P}\}/4$ and $e^t > 1+t$, we deduce $e^{-\mu\gamma} - \lambda_1 > \alpha_{\rm D}/4$. Combining these inequalities with (28), we deduce that it is sufficient to prove

$$\gamma^2 \leq \frac{qw\alpha_{\rm D}\alpha_{\rm P}\left(1-\alpha_{\rm P}\right)\left(1-\alpha_{\rm D}\right)}{10200\left(1-\frac{\alpha_{\rm D}}{2}\right)\left(1-\frac{\alpha_{\rm P}}{2}\right)\bar{L}}.$$

Finally, using (28) once again, we derive

$$qw \geq \frac{\alpha_{\rm P}\alpha_{\rm D}}{24(1-\alpha_{\rm P})(1-\alpha_{\rm D})}.$$

Therefore,

$$\gamma^2 \leq \frac{\alpha_{\rm D}^2\alpha_{\rm P}^2}{244800\left(1-\frac{\alpha_{\rm D}}{2}\right)\left(1-\frac{\alpha_{\rm P}}{2}\right)\bar{L}}.$$

Taking square root on both sides we obtain

$$\gamma \leq \frac{\alpha_{\rm D}\alpha_{\rm P}}{495\sqrt{\left(1-\frac{\alpha_{\rm D}}{2}\right)\left(1-\frac{\alpha_{\rm P}}{2}\right)\bar{L}}}.$$

This concludes the proof.

D DETAILS ON THE EXPERIMENTS

In this section, we describe the experimental setting in details. The code for the experiments can be found in https://anonymous.4open.science/r/elf_code-DE51/README.md.

D.1 THE SETTING

We are interested in the Bayesian logistic regression problem with a Gaussian prior. In particular, our goal is to sample from the posterior distribution, whose negative log-likelihood, that is the potential f, is given by

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x); \qquad f_i(x) = \frac{1}{m_i}\sum_{j=1}^{m_i} \log\left(1+e^{-b_{i,j}\cdot\langle a_{i,j},x\rangle}\right) + \frac{\lambda}{2}\|x\|^2,$$

where $x \in \mathbb{R}^d$ is the model, $(a_{i,j}, b_{i,j}) \in \mathbb{R}^d \times \{-1,1\}$ is one data point in the dataset of client i whose size is m_i. Here, the coefficient $\lambda > 0$ is the inverse variance of the prior distribution.

The datasets used in this study are chosen from the LibSVM repository [Chang and Lin, 2011]. Specifically, we implement the B, D, P-ELF algorithms, along with the LMC algorithm for the aforementioned target, to solve a classification problem on the datasets `a8a, a9a`, and `mushrooms`.

For each dataset, we partition the data points into 40 clients. Subsequently, we run all four methods with identical stepsizes selected from the set $0.01, 0.1, 0.5$. The compressor Top-τ is chosen for the ELF methods, where τ takes values from the set $1, 5, 10, 50, 100$. Given the stochastic nature of our algorithms, the final iterates are inherently random. To reduce variability in the finale estimate, we compute the average of the last 100 iterates for each method.

Each plot in Figure 1 features the communication complexity on the X-axis and the test accuracy on the Y-axis. Remarkably, across all plots, despite conservative theoretical expectations, the performance of all four algorithms appears nearly equivalent.