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

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

Abstract

Federated sampling algorithms have recently gained great popularity in the community of machine learning and statistics. This paper studies variants of such algorithms 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, respectively, primal, dual, and bidirectional compressors. We analyze the proposed methods under Log-Sobolev inequality and provide non-asymptotic convergence guarantees.

1. Introduction

Sampling from high-dimensional distributions holds immense significance in modern statistics and machine learning. This challenge is particularly relevant in Bayesian inference (Robert, 2007), where sampling from high-dimensional distributions poses difficulties. This paper focuses specifically on sampling from posteriors that arise in Bayesian federated learning (Kassab & Simeone, 2022; Vono et al., 2022; Liu & 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 its own data and maintains a (limited) internet connection with the server (Konečnỳ et al., 2016; McMahan et al., 2017). Consequently, the communication complexity becomes a computational bottleneck in most cases. The objective is to train a global model by performing local updates while minimizing the amount of information communicated.

[1]AI Initiative, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Correspondence to: Avetik Karagulyan <avetik.karagulyan@kaust.edu.sa>
The full version of the paper can be found at arXiv:2303.04622.

Workshop of Federated Learning and Analytics in Practice, colocated with 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

Mathematically, our problem can be formulated as follows. The target distribution π is a continuous distribution defined on the Euclidean space $\mathbb{R}^d$. For convenience, we will use π to refer to both the target distribution and its density function, given by:

$$\pi(x) \propto \exp(-F(x)), \quad (1)$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is the potential function. In the general Bayesian setting, F represents the logarithm of the posterior distribution. In the federated setting, the potential function is assumed to be sum-decomposable, with each component being 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, which is 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;
- P-ELF: LMC with primal compression;
- B-ELF: LMC with bidirectional compression.

The first algorithm, D-ELF, focuses on client-to-server (uplink) compression to reduce communication complexity. This approach was initially proposed in early federated learning papers like (Konečnỳ et al., 2016), where the assumption was made that uplink communication is more costly compared to server-to-client communication. However, more recent reports, such as one from Speedtest.net[1], indicate that the difference between uploading and downloading speeds is negligible (Philippenko & 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

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

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 both uplink and downlink compression, earning the term "bidirectional." Bidirectional federated optimization has been explored by several authors (Liu et al., 2020; Philippenko & Dieuleveut, 2020; Gruntkowska et al., 2022). However, this setting has not yet been extensively developed and studied for sampling problems. In this work, we focus on analyzing the first federated sampling algorithm that incorporates bidirectional compression.

1.1. Langevin sampling

A common way to solve this problem is based on discretizing a stochastic differential equation (SDE) called Langevin diffusion (LD). LD was initially designed to model the movement of particles in an environment with friction (Risken, 1996). Mathematically, it is written as

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

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). 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. When the score function is Lipschitz continuous, and the target satisfies Log-Sobolev inequality, the distribution of the K-th iterate converges to π (Vempala & Wibisono, 2019). See Appendix B for more context on the LMC.

1.2. EF21 and EF21-P

The Error Feedback algorithm (EF) was initially introduced as a stabilization mechanism for supervised learning using contractive compressors (Seide et al., 2014). However, it had limitations, including its inability to work in the distributed setting required for federated learning and the reliance on unrealistic assumptions for convergence analysis (Alistarh et al., 2018; Stich et al., 2018; Horváth & Richtárik, 2020).

The EF21 algorithm is an improved version of EF proposed by Richtarik et al. (Richtárik et al., 2021). It addresses the limitations of the original EF by compressing gradients before transmitting them to the server, making it suitable for the distributed setting. EF21 is considered state of the art in both theory and practice among error feedback methods (Fatkhullin et al., 2021).

EF21-P is a primal error-feedback method inspired by EF21. It acts as a reparametrization of the original method and compresses algorithm iterates instead of gradients. This approach reduces downlink communication complexity, which is important for large models. EF21-P can also be seen as an iteration perturbation method and finds applications in machine learning for generalization and smoothing (Gruntkowska et al., 2022; Orvieto et al., 2022; Duchi et al., 2012).

2. Problem setup

2.1. Preliminaries

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 previously, 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 $f(t) \leq Mg(t)$, when t is large enough. Similarly, $f(t) = \tilde{\mathcal{O}}(g(t))$, if $f(t)\log(t) = \mathcal{O}(g(t))$.

2.2. Mathematical framework

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

Definition 1 (L-smoothness)**.** *We say that a function is L-smooth, if $F(y) \leq F(x) + \langle\nabla F(x), y - x\rangle + \frac{L}{2}\|x-y\|^2$.*

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

Definition 2 (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 $\mathcal{Q} \in \mathbb{B}(\alpha)$.

We observe that unbiased compressors with bounded variance are commonly used in many federated learning algorithms (Konečný et al., 2016; Mishchenko et al., 2019; Gorbunov et al., 2021). However, it is worth noting that the class of contractive compressors is more extensive. For example the Top-k compressor (Alistarh et al., 2017), which selects the k coordinates with the largest absolute values from the input vector is a biased contractive compressor.

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 and Fisher information.

Definition 3 (KL divergence, Fisher information)**.** *The KL divergence and, respectively, Fisher information between two probability measures ν and π are defined as*

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

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

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 $\nu = \pi$. To solve this problem we borrow another well-known notion from optimization: PL-inequality (Polyak, 1963; Lojasiewicz, 1963). 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. Since the minimum of our functional is equal to zero, the Log-Sobolev inequality (LSI) becomes the analog of PL inequality.

Definition 4 (Log-Sobolev inequality)**.** *We say that π satisfies the Log-Sobolev inequality (LSI) with parameter μ, if for every probability measure ν we have $H_\pi(\nu) \leq \frac{1}{2\mu} J_\pi(\nu)$.*

(Bakry & 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 & 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.

3. The ELF algorithms

In this section, we describe the general scheme that we follow to construct our algorithms. Generally, stochastic optimization algorithms can be described such as EF21 and EF21-P can written as $x_{k+1} = x_k - \gamma g_k$, where g_k is an estimator of $\nabla F(x_k)$. 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 Z_k: $x_{k+1} = x_k - \gamma g_k + \sqrt{2\gamma} Z_k$. For the sake of space, we defer the pseudocodes and other details of all three algorithms to Appendix A.

3.1. 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:

$$\mathbf{G}_k^{\mathrm{D}} := \frac{1}{n}\sum_{i}^{n} \mathbb{E}\left[\left\|g_k^i - \nabla F_i(x_k)\right\|^2\right]. \tag{3}$$

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

Proposition 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 (3). Then the following recurrent inequality is true:*

$$\mathbf{G}_{k+1}^{\mathrm{D}} \leq (1-p)\mathbf{G}_k^{\mathrm{D}} + (1-p)\beta_{\mathrm{D}}\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right], \tag{4}$$

where $p := 1 - (1-\alpha_{\mathrm{D}})(1+s_{\mathrm{D}}) > 0$, $\bar{L} := \frac{1}{n}\sum_{i=1}^n L_i^2$ and $\beta_{\mathrm{D}} := \frac{1+s_{\mathrm{D}}^{-1}}{1+s_{\mathrm{D}}}\bar{L}$, 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:

$$\mathbf{G}_k^{\mathrm{P}} := \bar{L}\mathbb{E}\left[\|w_k - x_k\|^2\right], \quad \text{where} \quad \bar{L} := \frac{1}{n}\sum_{i=1}^n L_i^2. \tag{5}$$

Indeed, $\mathbf{G}_k^{\mathrm{D}} \leq \mathbf{G}_k^{\mathrm{P}}$ due to L_i smoothness of each component function F_i. See (26) in Appendix E.2 for the proof. The following proposition proves a recurrent identity similar to (4).

Proposition 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:*

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

where $p := 1 - (1-\alpha_{\mathrm{P}})(1+s_{\mathrm{P}}) > 0$, $\beta_{\mathrm{P}} := \frac{1+s_{\mathrm{P}}^{-1}}{1+s_{\mathrm{P}}}\bar{L}$, 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 1. *Assume that LSI holds with constant $\mu > 0$ and 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 $k \in \mathbb{N}$. If*

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

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 $p := 1-(1-\alpha_{\mathrm{M}})(1+s_{\mathrm{M}}) > 0$, $\tau = \left(2L^2 + C(1-p)\beta_{\mathrm{M}}\right)\left(16\gamma^2 d + 4d\gamma\right)$, $\Psi = H_\pi(\rho_0) + \frac{1-e^{-\mu\gamma}}{\mu} C\mathbf{G}_0^{\mathrm{M}}$, *and* $C = \frac{8L^2\gamma^2+2}{e^{-\mu\gamma}-(1-p)(4\gamma^2\beta_{\mathrm{M}}+1)}$.

We refer the reader to Appendix E.2 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 3.3 for more on the complexity of D-ELF and P-ELF.

3.2. Convergence analysis of the B-ELF

The Lyapunov term for the B-ELF algorithm is the as for the D-ELF, that is $\mathbf{G}_k^{\mathrm{D}}$. However, the recurrent identity of Proposition 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 4 in the Appendix for the complete statement.

Proposition 3 (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* (3) *and* (5)*, then*

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

where λ_1, λ_2 *and* λ_3 *are positive numbers.*

Theorem 2. *Assume that LSI holds with constant* $\mu > 0$ *and let* x_k *be the iterates of the B-ELF algorithm. We denote by* $\rho_k := \mathcal{L}(x_k)$ *for every* $k \in \mathbb{N}$*. Let the step-size the following condition:*

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

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

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

where $C, D > 0$ *are constants depending on the parameters of the algorithm and* $C = \frac{2.125}{e^{-\mu\gamma}-\lambda_1}$, $D = \frac{C\lambda_3}{e^{-\mu\gamma}-(1-\alpha_{\mathrm{P}})(1+w)}$ *and* $\tau = \left(2L^2 + \frac{5C\lambda_2}{\alpha_{\mathrm{P}}}\right)\left(16\gamma^2 dL + 4d\gamma\right)$.

The exact definitions of the undefined constants are written in the proof of the theorem, which is postponed to Appendix E.3.

3.3. Discussion on the communication complexity

Doing the computations as mentioned at the end of Section 3.1, we can deduce the following.

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

Similarly, for the bidirectional ELF we have the below.

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

The proof of Corollary 2 is in Appendix E.4. When α is $O(1)$, 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. Despite its higher theoretical complexity, error feedback optimization methods outperform gradient descent in practice (Richtárik et al., 2021). ELF is expected to be superior to the usual LMC in practice.

4. 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, in practice error feedback based methods outperform other compression methods (Fatkhullin et al., 2021). We believe that this phenomenon shall also transfer to the sampling case.

4.1. Future work

An immediate continuation of our paper would be to conduct an experimental analysis of the ELF algorithms with other federated sampling techniques on real data. One would expect that one would observe the same behavior as in the optimization case. That is, in practice the ELF algorithms outperform the other methods, despite the theoretical analysis.

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.

References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

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

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

Bakry, D. and Émery, M. Diffusions hypercontractives. In *Seminaire de probabilités XIX 1983/84*, pp. 177–206. Springer, 1985.

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

Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P., and Jordan, M. On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pp. 764–773. PMLR, 2018.

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

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., and Zhang, M. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

Dalalyan, A. S. 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.

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

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

Deng, W., Ma, Y.-A., Song, Z., Zhang, Q., and Lin, G. On convergence of federated averaging langevin dynamics. *arXiv preprint arXiv:2112.05120*, 2021.

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

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

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

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory*, pp. 793–797. PMLR, 2018.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Gorbunov, E., Burlachenko, K. P., Li, Z., and Richtárik, P. Marina: Faster non-convex distributed learning with compression. In *International Conference on Machine Learning*, pp. 3788–3798. PMLR, 2021.

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

Holley, R. and Stroock, D. W. Logarithmic Sobolev inequalities and stochastic Ising models. 1986.

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

Jain, P., Kar, P., et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

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

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

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

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

Liu, X., Li, Y., Tang, J., and Yan, M. A double residual compression algorithm for efficient distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 133–143. PMLR, 2020.

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

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

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. 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.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

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

Orvieto, A., Kersting, H., Proske, F., Bach, F., and Lucchi, A. Anticorrelated noise injection for improved generalization. *arXiv preprint arXiv:2202.02831*, 2022.

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

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

Plassier, V., Durmus, A., and Moulines, E. Federated averaging langevin dynamics: Toward a unified theory and new algorithms. *arXiv preprint arXiv:2211.00100*, 2022.

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

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

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34:4384–4396, 2021.

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

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

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

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

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

Salim, A., Kovalev, D., and Richtárik, P. Stochastic proximal Langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6649–6661, 2019.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 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.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

Sun, L., Salim, A., and Richtárik, P. Federated learning with a sampling algorithm under isoperimetry, 2022. URL https://arxiv.org/abs/2206.00920.

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

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

Vono, M., Plassier, V., Durmus, A., Dieuleveut, A., and Moulines, E. QLSD: Quantised Langevin Stochastic Dynamics for Bayesian Federated Learning . In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6459–6500. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/vono22a.html`.

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

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

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., and Girolami, M. Langevin diffusions and the metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.

Contents

A. Formulations of the 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.

A.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.

Algorithm 1 D-ELF

Input: Initialization $x_0 \sim \rho_0$, $g_k^i = g_k = \nabla F(x_0)$, 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:
  $\circ$ $\mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i)$.
  $\circ$ $g_{k+1}^i = g_k^i + \mathcal{Q}^{\mathrm{D}}(\nabla F_i(x_{k+1}) - g_k^i)$;
  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

A.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:

$$\begin{cases} w_0 := \mathcal{Q}^{\mathrm{P}}(x_0) \\ w_{k+1} = w_k + \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k) \\ x_{k+1} = x_k - \gamma \nabla F(w_k). \end{cases} \tag{6}$$

Here, $x_0 \sim \rho_0$ is a random starting point and $(Z_k)_k$ is a sequence of i.i.d. standard Gaussians on $\mathbb{R}^d$. The auxiliary sequence w_k is meant to estimate to 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.

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 $Z_k \sim \mathcal{N}(0, I_d)$;
    $\circ\ \nabla F(w_k) = \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(w_k)$;
    $\circ\ x_{k+1} = x_k - \gamma \nabla F(w_k) + \sqrt{2\gamma} Z_k$;
    $\circ\ \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k)$;
    $\circ\ w_{k+1} = w_k + \mathcal{Q}^{\mathrm{P}}(x_{k+1} - w_k)$;
    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

A.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. The details are presented in Algorithm 3.

Algorithm 3 B-ELF

Input: Starting point $x_0 = w_0 \sim \rho_0$, step-size h, 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

B. Related work

In their seminal paper, Roberts & Tweedie (1996) study the convergence properties of the LMC algorithm. They argued that a bias occurs when discretizing the continuous SDE. Thus, Langevin Monte-Carlo generates a homogeneous Markov chain whose stationary distribution differs from the target π. They solve this issue with a Metropolis-Hastings adjustment step at each iteration of the LMC, which modifies the chain to have π as its stationary distribution. The resulting algorithm is called Metropolis Adjusted Langevin Algorithm (MALA), and it was studied by many (Roberts & Rosenthal, 1998; Roberts & Stramer, 2002; Xifara et al., 2014; Dwivedi et al., 2018).

The bias of the LMC, however, depends on the discretization step γ. Dalalyan (2017) proved a bound on this error. Thus, similar to the analysis of the SGD, controlling the step-size and taking enough iterations, one can make the error of the LMC algorithm smaller than any ε. Later, different properties of the LMC were studied by many (Durmus & Moulines, 2017; Cheng et al., 2018; Cheng & Bartlett, 2018; Dalalyan & Karagulyan, 2019; Durmus & Moulines, 2019; Vempala & Wibisono, 2019).

Looking closely at (2), we observe its similarity with the gradient descent (GD) algorithm. In fact, (2) is an instance of the stochastic gradient descent (SGD) with a Gaussian noise independent of the iterate. This similarity has been repetitively exploited in various settings for sampling problems (see e.g. (Raginsky et al., 2017; Chatterji et al., 2018; Wibisono, 2019; Salim et al., 2019; Karagulyan & Dalalyan, 2020)). In particular, a line of research has been initiated on federated sampling Langevin algorithms, which combine LMC with existing optimization mechanisms: LMC+FedAvg (McMahan et al., 2017; Deng et al., 2021; Plassier et al., 2022), LMC+MARINA (Gorbunov et al., 2021; Sun et al., 2022), LMC+QSGD (Alistarh et al., 2017; Vono et al., 2022). Our work continues the logic of these papers by adding the error-feedback mechanisms EF21 and EF21-P to the classic LMC algorithm in the federated setting.

As in the case of optimization, the strong convexity of the potential function plays an important role in the analysis of Langevin Monte-Carlo. Non-convex optimization, however, has long been a central topic in the domain. We refer the reader to (Jain et al., 2017) for an overview of non-convex optimization in machine learning. In comparison, sampling from non-strongly log-concave distributions is less studied. Cheng et al. (2018) studied convergence of the LMC when the potential strongly convex outside a ball. Dalalyan et al. (2019) and (Karagulyan & Dalalyan, 2020) proposed a penalization of the convex potential to make it strongly convex and gave convergence bounds depending on the penalty. The analysis of MALA in the non-convex regime can be found in (Mangoubi & Vishnoi, 2019). However, these results either do not cover the general non-convex case or they require some conditions that scale poorly with the dimension. A more efficient approach relies on isoperimetric inequalities. It is known that isoperimetry implies rapid mixture of the continuous stochastic processes (Villani, 2008). Thus, one would assume that this property could be extended to their discretizations. Vempala & Wibisono (2019) proved the convergence of the LMC under Log-Sobolev inequality. Later, Sun et al. (2022) used this as a general scheme for LMC with stochastic gradient estimators in the context of federated Langevin sampling. We simplify their proof and adapt it to our setting.

C. General scheme of the proofs and comparison of rates

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 & 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

$$x_{k+1} = x_k - \gamma f_{\xi_k}(x_k) + \sqrt{2\gamma} Z_k, \tag{7}$$

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 $f_{\xi_k}(x_k)$:

$$\mathrm{d}y_t = -f_{\xi_k}(x_k)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{8}$$

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 $H_\pi(\rho_t)$:

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} = \int_{\mathbb{R}^d} \frac{\partial \rho_t(z)}{\partial t} \log\left(\frac{\rho_t}{\pi}\right)(z)\mathrm{d}z. \tag{9}$$

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 1. *If y_t is the solution of the diffusion* (8) *and $\rho_t = \mathcal{L}(y_t)$, then for every $t \in [0, \gamma]$,*

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} \leq -\frac{3}{4} J_\pi(\rho_t) + \mathbb{E}\left[\left\|f_{\xi_k}(y_0) - \nabla F(y_t)\right\|^2\right]. \tag{10}$$

The bound (10) was initially derived by Vempala & 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 F.1.

Lemma 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.

C.1. Table of comparison

Table 1: In this table we compare error-feedback methods in optimization and sampling. The rates are computed in the case when $\alpha_{\mathrm{D}} = \alpha_{\mathrm{P}} = \alpha$. The constant α for MARINA+LMC is defined differently. However, it takes the role of the compression coefficient of our setting and often has the same order.

Method	Error	Assumption	Complexity	Reference
GD	L_2	μ-s.c.	$\tilde{O}\left(\frac{dL}{\mu\varepsilon}\right)$	(Nesterov, 2013)
EF21	L_2	μ-s.c.	$\tilde{O}\left(\frac{L}{\alpha\mu\varepsilon}\right)$	(Richtárik et al., 2021)
EF21-P	L_2	μ-s.c.	$\tilde{O}\left(\frac{L}{\alpha\mu\varepsilon}\right)$	(Gruntkowska et al., 2022)
LMC	KL	μ-LSI	$\tilde{O}\left(\frac{L^2 d}{\mu^2\varepsilon}\right)$	(Vempala & Wibisono, 2019)
Marina+LMC	KL	μ-LSI	$\tilde{O}\left(\frac{Ld}{\alpha\mu^2\varepsilon}\right)$	(Sun et al., 2022)
D-ELF	KL	μ-LSI	$\tilde{O}\left(\frac{\bar{L}d}{\alpha^2\mu^2\varepsilon}\right)$	Corollary 1
P-ELF	KL	μ-LSI	$\tilde{O}\left(\frac{\bar{L}d}{\alpha^2\mu^2\varepsilon}\right)$	Corollary 1
B-ELF	KL	μ-LSI	$\tilde{O}\left(\frac{\bar{L}d}{\alpha^4\mu^2\varepsilon}\right)$	Corollary 2

D. Proofs of the propositions

D.1. Proof of Proposition 1

From the definition

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

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

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

This concludes the proof.

D.2. Proof of Proposition 2

From the definition

$$\begin{aligned}
\mathbf{G}_{k+1}^{\mathrm{P}} &= L^2\mathbb{E}\left[\|w_{k+1}-x_{k+1}\|^2\right] \\
&= L^2\mathbb{E}\left[\left\|w_k-x_{k+1}-\mathcal{Q}^{\mathrm{P}}(w_k-x_{k+1})\right\|^2\right] \\
&= (1-\alpha_{\mathrm{P}})L^2\mathbb{E}\left[\|w_k-x_{k+1}\|^2\right] \\
&= (1-\alpha_{\mathrm{P}})L^2\mathbb{E}\left[\|w_k-x_k+x_k-x_{k+1}\|^2\right] \\
&\le (1-\alpha_{\mathrm{P}})(1+s)L^2\mathbb{E}\left[\|w_k-x_k\|^2\right]+(1-\alpha_{\mathrm{P}})(1+s^{-1})L^2\mathbb{E}\left[\|x_k-x_{k+1}\|^2\right].
\end{aligned} \tag{11}$$

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

D.3. Full statement of Proposition 3 and its proof

We state now the complete version of Proposition 3.

Proposition 4. *The Lyapunov term $\mathbf{G}_k^{\mathrm{D}}$ of the bidirectional Langevin algorithm satisfies the following recurrent inequality:*

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

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

$$\begin{aligned}
\lambda_1 &= (1-\alpha_{\mathrm{D}})(1+s)(1+q); \\
\lambda_2 &= (1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u)\bar{L} \\
&\quad+\left((1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u^{-1})+(1+s^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}; \\
\lambda_3 &= \left((1-\alpha_{\mathrm{D}})(1+s)(1+q^{-1})(1+u^{-1})+(1+s^{-1})\right)(1-\alpha_{\mathrm{P}})(1+w).
\end{aligned} \tag{12}$$

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

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

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

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

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

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

Therefore,

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

Let us now bound the auxiliary term $\mathbf{G}_{k+1}^{\mathrm{P}}$. We notice that $\mathbf{G}_k^{\mathrm{P}}$ is the Lyapunov term of the P-ELF algorithm. Thus, from Proposition 2 we have

$$\begin{aligned}\mathbf{G}_{k+1}^{\mathrm{P}} &= \bar{L}\mathbb{E}\left[\|w_{k+1} - x_{k+1}\|^2\right]\\ &\le (1-\alpha_{\mathrm{P}})(1+w)\mathbf{G}_k^{\mathrm{P}} + (1-\alpha_{\mathrm{P}})(1+w^{-1})\bar{L}\mathbb{E}\left[\|x_k - x_{k+1}\|^2\right].\end{aligned} \tag{13}$$

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

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

This concludes the proof of the proposition. □

E. Proofs of the main theorems

E.1. 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 2. *For any two vectors $x, y \in \mathbb{R}^d$ and any $s > 0$*

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

Proof.

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

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 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*

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

Then

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

The second is a technical lemma borrowed from Chewi et al. (2021).

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

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

E.2. Proof of Theorem 1

We follow the scheme described in Appendix C. 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 C, $y_\gamma = x_{k+1}$. Denote by ρ_t the distribution of y_t. Lemma 1 yields:

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

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

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

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:

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

Combining Proposition 1 and (18) we deduce

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

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

Lemma 6. *If $\gamma \le \frac{1}{2\sqrt{2}L}$, then the iterates of the stochastic LMC algorithm* (7) *satisfy the following inequality, where $\mathbf{G}_k^{\mathrm{D}}$ is the Lyapunov term of D-ELF algorithm defined in* (3)*:*

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

Lemma 6 yields the following

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

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

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

From the definition of τ we obtain the following:

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

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 $e^{\mu\gamma} \le 1 + 2\mu\gamma$. 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

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

Solving this inequality we get

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

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

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

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,

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

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

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

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

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

Since π satisfies Log-Sobolev inequality, we deduce

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

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

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

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

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

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

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

This proves Theorem 1 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:

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

Therefore, combining this inequality with Lemma 5 we obtain

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

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 2 instead of Proposition 1. Therefore,

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

This concludes the proof of Theorem 1.

E.3. Proof of Theorem 2

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 C, we use the interpolation proof scheme. That is for the k-th iteration we define the process y_t as in (8). Thus, from Lemma 1 we have

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

Combining this with Proposition 4 and (13), we obtain

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

Lemma 6 yields

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

for $\gamma < 1/8L$. The latter condition on the step-size is a consequence of our assumptions from the statement of Theorem 2. Therefore,

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

Applying Lemma 4 we deduce

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

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,

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

Recall that the definitions of λ_2 and λ_3 are given in (12). Since $(1-\alpha_{\rm P})(1+w) < 1$, from the definition of λ_3 we have

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

Therefore, (12) implies

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

Thus,

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

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 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 2, the step-size satisfies the following inequality:*

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

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

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

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

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

Let us now apply LSI:

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

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

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

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

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

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

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

This concludes the proof of Theorem 2.

Remark 1. *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.*

E.4. Proof of Corollary 2

First let us upper bound τ. Similar to the proof of Corollary 1, $\left(16\gamma^2 dL+4d\gamma\right)<5d\gamma$. Thus,

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

F. Proofs of the lemmas

F.1. Proof of Lemma 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

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

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

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

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

$$\begin{aligned}
\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} &= \int_{\mathbb{R}^d} \frac{\partial \rho_t(z)}{\partial t} \log\left(\frac{\rho_t}{\pi}\right)(z)\mathrm{d}z \\
&= \int_{\mathbb{R}^d} \Big(\nabla_z \cdot \Big(\rho_t(z)\mathbb{E}_{\rho_{0|t}}\left[f_\xi(y_0) \mid y_t = z\right]\Big) + \Delta\rho_t(z)\Big) \log\left(\frac{\rho_t}{\pi}\right)(z)\mathrm{d}z \\
&= -\int_{\mathbb{R}^d} \left\langle \mathbb{E}_{\rho_{0|t}}\left[f_\xi(y_0) \mid y_t = z\right] + \nabla\log(\rho_t)(z), \nabla\log\left(\frac{\rho_t}{\pi}\right)(z)\right\rangle \rho_t(z)\mathrm{d}z \\
&= -\int_{\mathbb{R}^d} \left(\nabla\log\left(\frac{\rho_t}{\pi}\right)(z) - \nabla\log\left(\frac{\rho_t}{\pi}\right)(z) + \mathbb{E}_{\rho_{0|t}}\left[f_\xi(y_0) \mid y_t = z\right] + \nabla\log(\rho_t)(z)\right)^\top \\
&\qquad\qquad \times \nabla\log\left(\frac{\rho_t}{\pi}\right)(z)\rho_t(z)\mathrm{d}z \\
&= -\int_{\mathbb{R}^d} \left\langle \nabla\log\left(\frac{\rho_t}{\pi}\right)(z) + \mathbb{E}_{\rho_{0|t}}\left[f_\xi(y_0) \mid y_t = z\right] - \nabla F(z), \nabla\log\left(\frac{\rho_t}{\pi}\right)(z)\right\rangle \rho_t(z)\mathrm{d}z.
\end{aligned} \tag{31}$$

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

$$\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} \le -J_\pi(\rho_t) - \int_{\mathbb{R}^d} \left\langle \mathbb{E}_{\rho_{0|t}}\left[f_\xi(y_0) \mid y_t = z\right] - \nabla F(z), \nabla\log\left(\frac{\rho_t}{\pi}\right)(z)\right\rangle \rho_t(z)\mathrm{d}z. \tag{32}$$

From the Cauchy-Schwartz inequality, we deduce

$$\begin{aligned}
\frac{\mathrm{d}H_\pi(\rho_t)}{\mathrm{d}t} &\le -J_\pi(\rho_t) + \frac{1}{4}J_\pi(\rho_t) + \int_{\mathbb{R}^d} \left\|\mathbb{E}_{\rho_{0|t}}\left[f_\xi(y_0) \mid y_t = z\right] - \nabla F(z)\right\|^2 \rho_t(z)\mathrm{d}z \\
&= -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\left\|\mathbb{E}\left[f_{\xi_k}(y_0) - \nabla F(y_t) \mid y_t\right]\right\|^2\right] \\
&\le -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\mathbb{E}\left[\left\|f_{\xi_k}(y_0) - \nabla F(y_t)\right\|^2 \mid y_t\right]\right] \\
&= -\frac{3}{4}J_\pi(\rho_t) + \mathbb{E}\left[\left\|f_{\xi_k}(y_0) - \nabla F(y_t)\right\|^2\right].
\end{aligned} \tag{33}$$

This concludes the proof of the lemma.

F.2. Proof of Lemma 5

If we replace $f_{\xi_k}(y_0)$ by g_0 in (10), we will have

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

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

$$\mathbb{E}\left[\left\|\nabla F(y_t) - \nabla F(x_0)\right\|^2\right] \le L^2\mathbb{E}\left[\left\|y_t - x_0\right\|^2\right] = L^2\mathbb{E}\left[\left\|tg_0 + \sqrt{2}\left(B_t - B_0\right)\right\|^2\right]. \tag{34}$$

Since the Brownian process has independent increments we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla F(y_t) - \nabla F(x_0)\right\|^2\right] &\le L^2t^2\left\|g_0\right\|^2 + 2tL^2d \\
&\le L^2\gamma^2\left\|g_0\right\|^2 + 2hL^2d \\
&= L^2\mathbb{E}\left[\left\|x_1 - x_0\right\|^2\right].
\end{aligned} \tag{35}$$

This concludes the proof.

F.3. Proof of Lemma 6

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

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

Regrouping the terms we obtain

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

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

F.4. Proof of Lemma 7

Is sufficient to show that

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

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

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

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

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

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

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

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

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

Finally, using (28) once again, we derive

$$qw\ge\frac{\alpha_{\mathrm{P}}\alpha_{\mathrm{D}}}{24(1-\alpha_{\mathrm{P}})(1-\alpha_{\mathrm{D}})}.$$

Therefore,

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

Taking square root on both sides we obtain

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

This concludes the proof.