Faster Rates For Federated Variational Inequalities

Guanghui Wang*
Georgia Institute of Technology
gwang369@gatech.edu

Satyen Kale
Apple
satyen@apple.com

Abstract

In this paper, we study federated optimization for solving stochastic variational inequalities (VIs), a problem that has attracted growing attention in recent years. Despite substantial progress, a significant gap remains between existing convergence rates and the state-of-the-art bounds known for federated convex optimization. In this work, we address this limitation by establishing a series of improved convergence rates. First, we show that, for general smooth and monotone variational inequalities, the classical Local Extra SGD algorithm admits tighter guarantees under a refined analysis. Next, we identify an inherent limitation of Local Extra SGD, which can lead to excessive client drift. Motivated by this observation, we propose a new algorithm, the Local Inexact Proximal Point Algorithm with Extra Step (LIPPAX), and show that it mitigates client drift and achieves improved guarantees in several regimes, including bounded Hessian, bounded operator, and low-variance settings. Finally, we extend our results to federated composite variational inequalities and establish improved convergence guarantees.

1 Introduction

Federated learning [1, 2] is a powerful framework for solving large-scale machine learning problems. Unlike traditional approaches [3, 4] that aggregate all training data on a central server, federated optimization keeps the data on local devices, such as mobile phones or organizational silos, and performs training locally. Periodically, these devices communicate only model updates (e.g., gradients or parameters), which are then aggregated to update a global model. This decentralized design enhances privacy and reduces communication costs, making federated learning particularly attractive for modern applications where data are sensitive and massive in scale.

The central problem of interest is federated empirical risk minimization. The canonical algorithm is *Local Stochastic Gradient Descent* (LSGD), also known as *Federated Averaging* [2, 5, 6]: In this scheme, each of the M participating devices performs multiple SGD steps locally with its own data, followed by synchronizing its model with the central server every K local steps. By leveraging these *local updates*, LSGD achieves faster convergence than its centralized counterpart, mini-batch SGD, particularly in settings with many participating machines [7]. Building on this foundational work, numerous extensions have been proposed, from advanced strategies to mitigate data heterogeneity across clients [8, 9] to the integration of adaptive optimizers [10, 11].

To date, the theory of federated optimization is primarily restricted to smooth and convex loss functions. A natural question is whether these results can be extended to more general problems. In this work, we investigate federated optimization for *smooth and monotone variational inequalities* [15–17], a unifying framework that not only recovers convex and smooth optimization as a special case, but also encompasses a broader range of complex optimization problems, such as convex-concave optimization, equilibrium computation, and fixed point computation. A particular application of interest is smooth convex-concave optimization. Such problems have arisen in many modern

*This work was done during an internship at Apple.

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

Table 1: Summary of convergence rates for federated optimization and variational inequalities. FO: Federated Optimization. FVI: Federated VI. The red terms denote sub-optimal factors compared to the LSGD-rate. The displayed bounds only show the dependence on K, R, σ, M, d; dependence on other constants is suppressed and absorbed in $O(\cdot)$. The $\widetilde{O}(\cdot)$ notation further hides logarithmic factors in K and R. See the referenced theorems for full expressions.

Setting	Algo.	Reference	Bound	Assumption
FO	LSGD	[7]	$O\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$	–
FVI	LESGD	[12]	$O\left(\frac{1}{\sqrt{K}R} + \frac{\sigma}{\sqrt{MKR}} + \sqrt{\frac{\sigma}{\sqrt{K}R} + \frac{\sigma^2}{R}}\right)$	–
		Theorem 1	$O\left(\frac{1}{\sqrt{K}R} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$	–
		Corollary 1 Theorem 2	$O\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$	$\sigma = \Omega\left(\frac{1}{\sqrt{R}K^{1/4}}\right)$ affine operator
	LIPPAX	Theorem 3	$\widetilde{O}\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{\sigma}{\sqrt{KR}}\right)$	Bounded operator
		Corollary 2 Theorem 4	$\widetilde{O}\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$	$\sigma = O\left(\frac{1}{\sqrt{KR}}\right)$ Bounded Hessian
	SLIPPAX	Theorem 5	$\widetilde{O}\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{d^{\frac{1}{8}}\sigma^{\frac{1}{2}}}{K^{\frac{3}{4}}R^{\frac{3}{4}}} + \frac{d^{\frac{1}{10}}\sigma^{\frac{2}{5}}}{K^{\frac{2}{5}}R^{\frac{4}{5}}}\right)$	Bounded operator
	LSGD	Theorem 6	$O\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$	Co-coercive
FCO	LDA	[13]	$O\left(\frac{1}{KR} + \frac{1}{R^{\frac{2}{3}}} + \frac{\sigma}{\sqrt{MKR}}\right)$	Bounded gradient
FCVI	LDE	[14]	$O\left(\frac{1}{KR} + \frac{1}{\sqrt{R}} + \frac{\sigma}{\sqrt{MKR}} + \frac{1}{K^{\frac{1}{4}}R^{\frac{3}{4}}}\right)$	Bounded operator
		Theorem 7	$O\left(\frac{1}{KR} + \frac{1}{R^{\frac{2}{3}}} + \frac{\sigma}{\sqrt{MKR}}\right)$	Bounded operator

machine learning applications, including the training of generative adversarial networks [18], robust reinforcement learning [19], and multi-distribution machine learning [20]. In these settings, data and computation may naturally be distributed across agents or organizations, which motivate study of solving VI in federated environments.

The study of federated variational inequalities has gained increasing attention in recent years [21–24]. Most of these efforts have focused on LSGD-type methods, the natural extension of LSGD to the VI setting. However, it is well known that vanilla SGD may diverge on general smooth monotone problems. As a result, these works have mostly been restricted to the narrower case of smooth and *strongly* monotone VIs. To deal with general smooth and monotone VIs, researchers have adapted the Extra SGD algorithm [17], the classical method for general smooth monotone VIs, to the federated context, leading to the LESGD method [12]. The resulting convergence rate is $O\left(\frac{1}{\sqrt{KR}} + \frac{\sigma}{\sqrt{MKR}} + \sqrt{\frac{\sigma}{\sqrt{KR}} + \frac{\sigma^2}{R}}\right)$. However, the last term does not decay with the number of local steps K or number of machines M, which brings into question the utility of local steps. In fact, if we simply run Extra SGD on 1 machine (ignoring all other machines) for KR steps, then we obtain a much better convergence rate of $O\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{KR}}\right)$ [17], with no communication needed, and $\frac{1}{M}$th of the computation of LESGD!

In this paper, we revisit the foundational problem of federated optimization for *general smooth and monotone variational inequalities*. We note that solving federated variational inequalities is inherently challenging: Prior theoretical analyses in federated convex optimization often rely on the smoothness of the objective, which does *not* extend to min–max problems or general variational inequalities, since min–max objectives are typically non-smooth when viewed as functions of the primal variable (even when the objective is jointly smooth), due to the inner maximization. More essentially, convex and smooth objectives naturally satisfy *co-coercivity*, a property that is crucial in most analyses of federated optimization. However, co-coercivity generally fails to hold for smooth and monotone variational inequalities. As a result, it remains unclear whether improved convergence rates are achievable in this setting.

In this paper, we provide a positive answer to this question by establishing a series of faster convergence rates. To ground these rates, our primary point of comparison is the tight rate for LSGD [25],

for federated *convex and smooth optimization* (a special case of the federated VI problem studied in this paper), which is $O\left(\frac{1}{KR} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{2/3}}{K^{1/3}R^{2/3}}\right)$. This is the best known unaccelerated rate for this special case which shows improvement with the number of local steps, K, in all terms. We will refer to this latter rate as the LSGD-rate for short, and aim to achieve this rate for federated VI problems. We have the following convergence rates for federated VI problems:

- **LESGD**: For LESGD, we show that a faster $O\left(\frac{1}{\sqrt{KR}} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$ rate can be obtained (Theorem 1). Our rate matches the LSGD-rate up to an additional $\sqrt{K}$ factor in the first term. In Corollary 1, we show that the LSGD-rate is achieved when $\sigma = \Omega\left(\frac{1}{\sqrt{R}K^{\frac{1}{4}}}\right)$, i.e., the variance is not too small. We also show that the LSGD-rate can be achieved if the operator is co-coercive (Theorem 6).

- **Composite VI**: We extend our analysis to *composite* VI problems, where the smooth and monotone VI is regularized by a potentially non-smooth function. For this harder problem, [14] provide the following bound for a generalized version of LESGD: $O\left(\frac{1}{KR} + \frac{1}{\sqrt{R}} + \frac{\sigma}{\sqrt{MKR}} + \frac{1}{K^{\frac{1}{4}}R^{\frac{3}{4}}}\right)$. In Theorem 7, we improve this to $O\left(\frac{\sigma}{\sqrt{MKR}} + \frac{1}{R^{\frac{2}{3}}} + \frac{1}{KR}\right)$, which exactly matches the best known rate for federated composite convex optimization [13], which is a special case of federated composite VI problems.

Our main technical novelty for obtaining the results above lies in a novel analysis of the convergence bound. We note that in the analysis of federated learning algorithms, an important term is the client drift, which measures the divergence between local models across different clients. Traditional analysis of federated optimization usually establishes a connection between the convergence upper bound and the squared norm of the client drift. However, in federated VI analysis, previous work establishes connections to both the norm and its square, due to the more complicated nature of VIs compared with convex optimization. In this paper, we show that only the squared client drift matters, which leads to improved convergence rates.

Note that our general bound for LESGD (Theorem 1) contains a $\frac{1}{\sqrt{KR}}$ term, which is a factor of $\sqrt{K}$ worse than the $\frac{1}{KR}$ term in the LSGD-rate. This term is also present in the results of prior work [12], although it is difficult to discern from their paper. We find that this term stems from the lack of co-coercivity (and, in the worst case, even anti-cocoercivity) of the extra-gradient operation [26], which leads to looser control of client drift across local updates and reveals a fundamental limitation of this class of methods. We refer to Remark 1 for a detailed discussion. In Theorem 2, we further show that the LSGD-rate can be achieved when the operator is affine, since the extra-gradient operation becomes co-coercive in this case. However, co-coercivity of the extra-gradient operation does not hold in general. This shortcoming is addressed in our next contribution:

- **New Algorithm, LIPPAX**: We propose a new algorithm called the Local Inexact Proximal Point Algorithm with Extra Step (LIPPAX, Algorithm 2). Our key observation is that the proximal point operation incurs far less client drift than the extra-gradient step. While the exact proximal point is impractical to compute, we approximate it using multiple SGD steps on a regularized objective, followed by one extra gradient step from the starting point. This approach successfully eliminates the $\frac{1}{\sqrt{KR}}$ term, but introduces an additional $\frac{\sigma}{\sqrt{KR}}$ variance term (Theorem 3). This extra term arises from the variance accumulated during the inner SGD loop, and a direct analysis suggests it is irreducible by M. Nevertheless, through a novel bias-variance decomposition, we show that this term can be eliminated assuming a second-order boundedness condition on the operator (Theorem 4). Finally, we propose a Gaussian-smoothing variant called SLIPPAX (Algorithm 3), which removes the second-order boundedness assumption and achieves, up to logarithmic factors, the LSGD-rate plus an $O\left(\frac{d^{\frac{1}{8}}\sigma^{\frac{1}{2}}}{K^{\frac{3}{4}}R^{\frac{3}{4}}} + \frac{d^{\frac{1}{10}}\sigma^{\frac{2}{5}}}{K^{\frac{2}{5}}R^{\frac{4}{5}}}\right)$ lower order factor that depends on d, under only the assumption that the operator is bounded (Theorem 5).

This paper primarily focuses on the homogeneous setting of federated learning where all machines sample from the same distribution. Appendix H extends the LESGD analysis to bounded heterogeneity, with additional drift terms, and recovers Theorem 1 when heterogeneity is zero.

2 Related Work

In this section, we briefly review related work in federated optimization and VI problems.

Variational Inequalities. The study of VIs dates back to the 1960s [15, 27, 28], with more modern classical monographs in [16, 29]. Two of the most important algorithms for smooth and monotone VIs are the Extra Gradient (EG) method [30] and the Optimistic Gradient (OG) method [31]. For stochastic smooth and monotone VIs, [17] proposed the Mirror-Prox algorithm (we refer to it as Extra SGD in the ℓ_2-space), which can be viewed as the stochastic and mirror descent variant of EG. They show that for stochastic smooth and monotone VIs, Mirror-Prox achieves an $O\Big(\frac{L}{T} + \frac{\sigma}{\sqrt{T}}\Big)$ convergence rate, where L is the smoothness parameter and T is the total number of iterations. This matches the convergence rate of SGD for stochastic convex and smooth optimization [32]. In distributed optimization, a mini-batch version of Extra SGD (*without local updates*) has been studied for saddle-point problems [24], achieving an $O\Big(\frac{L}{R} + \frac{\sigma}{\sqrt{MRK}}\Big)$ rate, where R is the number of communication rounds, M is the number of machines, and K is the batch size. We refer readers to the recent survey [33] for a detailed review of stochastic smooth and monotone inequalities.

Federated Optimization. Federated optimization has been extensively studied [34]. The most related foundational algorithm to our work is by [2], whose Federated Averaging algorithm coincides with the parallel SGD algorithm with identical clients [5]. The LSGD algorithm has been analyzed by a long line of research [6, 7, 35, 36], with various sharp upper and lower bounds. [13] studied federated optimization for composite problems, where the objective consists of a convex and smooth function plus a convex but non-smooth regularizer. They proposed the local dual averaging algorithm, which achieves an $O\Big(\frac{1}{KR} + \frac{L}{R^{2/3}} + \frac{\sigma}{\sqrt{MRK}}\Big)$ rate under the bounded gradients assumption.

Federated Variational Inequalities. For federated VIs, as mentioned in the introduction, there is a growing body of work [21–23], but these studies focus only on smooth and *strongly* monotone VIs and employ SGDA-type methods, which are known to diverge in the merely monotone setting. More recent work [12, 14] has considered the LESGD algorithm. However, as discussed above, their results still exhibit a gap compared to the LSGD-rate.

3 Preliminaries

Let $V : \mathbb{R}^d \mapsto \mathbb{R}^d$ be an operator of interest. In this section, we review some basic concepts.

Definition 1 (Monotonicity)**.** An operator $V : \mathbb{R}^d \mapsto \mathbb{R}^d$ is monotone iff $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$, we have:

$$\langle V(\mathbf{z}) - V(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0. \tag{1}$$

Definition 2 (L-Smoothness)**.** An operator $V : \mathbb{R}^d \mapsto \mathbb{R}^d$ is L-smooth, iff $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$:

$$\|V(\mathbf{z}) - V(\mathbf{z}')\| \leq L\|\mathbf{z} - \mathbf{z}'\|. \tag{2}$$

We assume that we have access to an unbiased noisy oracle $\widetilde{V} : \mathbb{R}^d \times \Xi \mapsto \mathbb{R}^d$, where Ξ is some sample space with an associated distribution that samples $\xi \in \Xi$ are drawn from, such that for any $\mathbf{z} \in \mathbb{R}^d$, we have $\mathbb{E}_\xi[\widetilde{V}(\mathbf{z}, \xi)] = V(\mathbf{z})$. Moreover, we assume the variance is upper bounded: $\mathbb{E}_\xi\|V(\mathbf{z}) - \widetilde{V}(\mathbf{z}, \xi)\|^2 \leq \sigma^2$. For readability, we drop ξ and use $\widetilde{V}(\mathbf{z})$ when the context is clear. Each oracle evaluation uses a fresh sample, independent across clients, and independent of the history before that call.

Goal and Performance Metric. In this work, we focus on solving variational inequalities, that is, we would like to find a solution $\mathbf{z}^* \in \mathbb{R}^d$, such that:

$$\sup_{\mathbf{z} \in \mathbb{R}^d} \langle V(\mathbf{z}), \mathbf{z}^* - \mathbf{z} \rangle \leq 0. \tag{3}$$

Algorithm 1 Local Extra SGD (LESGD)

Initialization: $\mathbf{z}_0^m = \mathbf{0}$, for $m = 1, \ldots, M$.
for $t = 1$ to T **do**
  for each machine $m = 1, \ldots, M$ **do**
    $\mathbf{x}_t^m = \begin{cases} \mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{z}_{t-1}^m), & \text{if } \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M}\sum_{m=1}^M \left(\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{z}_{t-1}^m)\right), & \text{if } \mathrm{mod}(t, K) = 0. \end{cases}$
    $\mathbf{z}_t^m = \begin{cases} \mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m), & \text{if } \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M}\sum_{m=1}^M \left(\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m)\right), & \text{if } \mathrm{mod}(t, K) = 0. \end{cases}$
  end for
end for
return $\mathbf{x}_o = \frac{1}{MT}\sum_{m=1}^M \sum_{t=1}^T \mathbf{x}_t^m$

Without loss of generality, we initialize all algorithms at the origin. For any constant $D > 0$, let $\mathcal{Z}_D = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| \leq D\}$, and we define the corresponding restricted error metric of an algorithm output $\overline{\mathbf{z}} \in \mathbb{R}^d$ as:

$$\mathrm{err}_D(\overline{\mathbf{z}}) = \sup_{\mathbf{z} \in \mathcal{Z}_D} \langle V(\mathbf{z}), \overline{\mathbf{z}} - \mathbf{z} \rangle .$$

We note that, here only the comparator in the error metric is restricted to Z_D; the output and all intermediate algorithmic iterates remain unconstrained. This error metric specifies the class of comparators against which the output is evaluated, and, unless mentioned otherwise, all our results are stated in terms of this radius D. The comparison radius D also enters the step-size choices in our bounds. When a bound on the distance from initialization to a solution is available, D can be chosen so that the comparison ball contains that solution.

4 Faster Rates For Federated VIs

In this section, we provide a series of faster rates for federated variational inequalities. In Section 4.1, we establish tighter bounds for the LESGD algorithm. In Section 4.2, we introduce our proposed LIPPAX and SLIPPAX algorithms and demonstrate that they achieve improved rates in different regimes. Finally, in Section 4.3, we prove that LSGD attains faster rates for co-coercive operators.

4.1 Faster Rates for LESGD

In this part, we study the classical LESGD algorithm for solving smooth and monotone variational inequalities, which has been examined in prior work [12, 14], and we establish faster convergence rates. The algorithm is outlined in Algorithm 1. Intuitively, each client $m \in [M]$ first executes a stochastic gradient descent step at the current point $\mathbf{z}_{t-1}^m$ to obtain $\mathbf{x}_t^m$, and then updates $\mathbf{z}_{t-1}^m$ to $\mathbf{z}_t^m$ using the gradient at $\mathbf{x}_t^m$. The clients synchronize with each other every K steps. We write $T = KR$, where R is the number of synchronization periods. We then obtain the following result. The proof is in Appendix A.

Theorem 1. *Suppose V is L-smooth and monotone. Let $\mathbf{x}_o = \frac{1}{MT}\sum_{m=1}^M \sum_{t=1}^T \mathbf{x}_t^m$ be the output of Algorithm 1. Set: $\eta = \min\left\{\frac{1}{\sqrt{14}KL}, \frac{D\sqrt{M}}{\sigma\sqrt{6KR}}, \frac{D^{\frac{2}{3}}}{936^{\frac{1}{3}} e^{\frac{2}{3}} K^{\frac{2}{3}} R^{\frac{1}{3}} \sigma^{\frac{2}{3}} L^{\frac{1}{3}}}\right\}$. Then we have*

$$\mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] \leq \frac{12LD^2}{\sqrt{KR}} + \frac{6D\sigma}{\sqrt{MKR}} + \frac{26D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}. \tag{4}$$

Theorem 1 shows an $O\left(\frac{L}{\sqrt{KR}} + \frac{\sigma}{\sqrt{MKR}} + \frac{\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right)$ convergence for LESGD, which nearly matches the LSGD-rate in the classical federated optimization setting up to an additional $\sqrt{K}$ factor in the first term. Compared with the bound for LESGD established in prior work [12], $O\left(\frac{L}{\sqrt{KR}} + \frac{\sigma}{\sqrt{MKR}} + \sqrt{\frac{L\sigma}{\sqrt{KR}} + \frac{\sigma^2}{R}}\right)$, our result eliminates the undesirable last term.

Remark 1 (Discussion on the $\frac{L}{\sqrt{KR}}$ term)**.** *As discussed above, compared with the bound for LSGD [7], the bound for LESGD contains an additional $\sqrt{K}$ factor in the first term. This arises because,*

Algorithm 2 Local Inexact Proximal Point with Extra Gradient (LIPPAX)

Initialization: $\mathbf{z}_0^m = \mathbf{0}$, for $m = 1, \ldots, M$.
for $t = 1$ to T **do**
  for each machine $m = 1, \ldots, M$ **do**
    $\mathbf{x}_{t,0}^m = \mathbf{z}_{t-1}^m$
    for $\ell = 1$ to H **do**
      $\mathbf{x}_{t,\ell}^m = \mathbf{x}_{t,\ell-1}^m - \gamma \left(\widetilde{V}(\mathbf{x}_{t,\ell-1}^m) + \frac{1}{\eta} \left(\mathbf{x}_{t,\ell-1}^m - \mathbf{z}_{t-1}^m \right) \right)$
    end for
    $\mathbf{x}_t^m = \mathbf{x}_{t,H}^m$
    $\mathbf{z}_t^m = \begin{cases} \mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m), & \text{if } \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M} \sum_{m=1}^M \left(\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m) \right), & \text{if } \mathrm{mod}(t, K) = 0. \end{cases}$
  end for
end for
return $\mathbf{x}_o = \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \mathbf{x}_t^m$

as shown in above, LESGD requires a smaller step size of order $\eta = O(\frac{1}{\sqrt{K}L})$, whereas classical LSGD uses $\eta = O(\frac{1}{L})$. The smaller step size is necessary due to an inherent limitation of the extra-gradient method. In particular, a key quantity in the analysis of federated optimization is the client drift*, measured by $\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2$ for any two clients $m, m' \in [M]$. Under LSGD, this drift naturally contracts thanks to the* co-coercivity *of the SGD operator. However, as recently shown by [26], the extra-gradient operator can be non-co-coercive, which makes $\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2$ expansive instead of contractive. In fact, [26] show that there is a particular* co-coercive *operator V and points $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$ such that $\|\mathbf{z} - V(\mathbf{z} - V(\mathbf{z})) - (\mathbf{z}' - V(\mathbf{z}' - V(\mathbf{z})))\|^2 \geq (1 + \eta^4)\|\mathbf{z} - \mathbf{z}'\|^2$. To control this expansion, a smaller step size is therefore required.*

Next, we show the convergence rate can improve to the LSGD-rate under suitable assumptions.

Corollary 1 (Large Variance)**.** Suppose V is L-smooth and monotone. Assume $\sigma \geq \frac{14^{\frac{3}{4}} D L}{6e\sqrt{26}\sqrt{R} K^{1/4}} = \Omega\left(\frac{DL}{R^{\frac{1}{2}} K^{\frac{1}{4}}}\right)$. Let $\eta = \min\left\{ \frac{D\sqrt{M}}{\sigma\sqrt{6KR}}, \frac{D^{\frac{2}{3}}}{936^{\frac{1}{3}} e^{\frac{2}{3}} K^{\frac{2}{3}} R^{\frac{1}{3}} \sigma^{\frac{2}{3}} L^{\frac{1}{3}}} \right\}$. Then we have

$$\mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] \leq \frac{26 D^{\frac{4}{3}} \sigma^{\frac{2}{3}} L^{\frac{1}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{6 D \sigma}{\sqrt{MKR}}.$$

The corollary above shows that the LSGD-rate can be achieved when σ is not too small. This naturally occurs, as the undesirable small step size $\eta = O(\frac{1}{\sqrt{K}L})$ (as discussed in Remark 1) is dominated by the rest of the terms in the configuration of the step size, which leads to the LSGD-rate. Finally, we note that, when the operator is also affine and monotone, the VI problem reduces to a standard quadratic convex optimization problem. In this case, the algorithm given in Algorithm 1 achieves the LSGD-rate with a different step size. For completeness, we include this result as an illustration of the algorithm's performance in this special case.

Lemma 2 (Affine Operator)**.** *Suppose V is affine, L-smooth and monotone. Then consider the algorithm given in Algorithm 1 with $\eta = \min\left\{ \frac{1}{L}, \frac{D\sqrt{M}}{\sigma\sqrt{6KR}}, \frac{D^{\frac{2}{3}}}{180^{\frac{1}{3}} K^{\frac{2}{3}} R^{\frac{1}{3}} \sigma^{\frac{2}{3}} L^{\frac{1}{3}}} \right\}$, and we get*

$$\mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] \leq \frac{20 D^{\frac{4}{3}} \sigma^{\frac{2}{3}} L^{\frac{1}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{2\sqrt{6} D \sigma}{\sqrt{MKR}} + \frac{2LD^2}{KR}. \tag{5}$$

Comparing Theorem 2 with Theorem 1, the key difference lies in the step size: η can be larger ($O\left(\frac{1}{\sqrt{K}L}\right)$ vs. $O(\frac{1}{L})$), which enables the LSGD-rate. This improvement is due to the non-expansive property of the extra-gradient operator for affine VIs [26], which leads to a tighter bound.

4.2 Faster Rates By Inexact PPA With Extra Step

Corollary 1 shows that the LESGD algorithm achieves the LSGD-rate when the variance is not too small. However, for the small variance regime, i.e., $\sigma = O\left(\frac{1}{R^{\frac{1}{2}} K^{\frac{1}{4}}}\right)$, the first term in (4) will

dominate the rest, making the overall convergence bound become the sub-optimal rate $O\left(\frac{L}{\sqrt{KR}}\right)$, *even when* $\sigma = 0$. As discussed in Remark 1, this is caused by the expansive property of the extra SGD operator, and can be a fundamental limitation of this kind of method.

In this section, we provide an alternative approach, which achieves the LSGD-rate up to logarithmic factors under proper assumptions. The algorithm is given in Algorithm 2, which we referred to as Local Inexact Proximal Point Algorithm with Extra Gradient (LIPPAX). To motivate our algorithm, we first recall the classical proximal point algorithm, which can handle general smooth and monotone VIs and obtain optimal results. Suppose at round t, the decision of machine m is $\mathbf{z}_{t-1}^m$. Then one step of the "exact" proximal point algorithm conducts the following implicit update to obtain a decision $\mathbf{x}_t^{*,m}$: $\mathbf{x}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m})$. However, this update is hard to implement, as both sides involves $\mathbf{x}_t^{*,m}$. To avoid this, notice that $\mathbf{x}_t^{*,m}$ is the optimizer of the variational inequality defined by the following "regularized" strongly monotone operator: $F(\mathbf{x}) = V(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$. Then, we can conduct multiple LSGD steps to approximate $\mathbf{x}_t^{*,m}$ (Steps 4-7), and then perform one step of inexact proximal point step using the approximation (Steps 8-9). The H inner steps require no communication between clients. This use of proximal regularization is related to FedProx [37], which is an algorithm for convex optimization, not VI. Both LIPPAX and FedProx uses the proximal term, while the extra-gradient step after proximal optimization in LIPPAX is new and is introduced specifically to handle VI. We have the following conclusion for Algorithm 2. The proof is given in Appendix C.

Theorem 3. *Suppose V is L-smooth and monotone. Moreover, assume $\|V(\mathbf{x})\| \le G$ for $\mathbf{x} \in \mathbb{R}^d$. Let $\eta = \min\left\{\frac{1}{L}, \frac{D\sqrt{M}}{\sigma\sqrt{KR}}, \frac{D^{\frac{2}{5}}}{(60e)^{\frac{1}{5}}K^{\frac{3}{5}}R^{\frac{1}{5}}\sigma^{\frac{2}{5}}L^{\frac{2}{5}}}, \frac{D^{\frac{2}{3}}}{(54e)^{\frac{1}{3}}K^{\frac{2}{3}}R^{\frac{1}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}, \frac{D}{\sigma\sqrt{15KR}}\right\}$, and let $\gamma = \frac{1}{\eta\left(L+\frac{1}{\eta}\right)^2}$. Then Algorithm 2 ensures that:*

$$\begin{aligned}\mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] \le & \frac{LD^2}{KR} + \frac{4\sigma D}{\sqrt{KR}} + \frac{2(60e)^{\frac{1}{5}}D^{\frac{8}{5}}\sigma^{\frac{2}{5}}L^{\frac{2}{5}}}{K^{\frac{2}{5}}R^{\frac{4}{5}}} + \frac{2(54e)^{\frac{1}{3}}D^{\frac{4}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} \\ & + \frac{4D\sigma}{\sqrt{MKR}} + \frac{60eG^2K^2L + 30G^2}{4^H}.\end{aligned} \tag{6}$$

Remark 2. *We comment on the bounded operator over $\mathbb{R}^d$ assumption from a few perspectives: **(1)** Boundedness of the comparison set $\mathcal{Z}_D$ does not imply boundedness of the operator along the optimization trajectory. **(3)** Almost all existing algorithms on federated optimization (including but not limited to VI) do not involve* projection to the decision set*, and consequently assumptions such as bounded gradients (operators) and smoothness are* imposed over all of $\mathbb{R}^d$*. This is a long-standing question for general federated optimization, instead of this specific work. **(2)** In our proof, the assumption is actually only required to hold along the optimization trajectory. This is spiritually consistent with how bounded-gradient assumptions are applied in existing federated learning works. **(4)** Section 5 of our paper studies composite VIs with potentially non-smooth regularizers; by setting the regularizer to be the indicator function of a constrained domain, we obtain results in the constrained setting as well.*

To see the merit of this bound, firstly note that, the first term becomes $O\left(\frac{L}{KR}\right)$, instead of $O\left(\frac{L}{\sqrt{KR}}\right)$. Moreover, the last term decreases exponentially with H, and will be dominated by other terms as long as $H = O(\ln KR)$. However, there is a price paid in the second term, which is on the order of $O\left(\frac{\sigma}{\sqrt{KR}}\right)$. We have the following corollary.

Corollary 2. Assume $\sigma \le \frac{DL}{\sqrt{KR}} = O\left(\frac{DL}{\sqrt{KR}}\right)$, then

$$\mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] \le \frac{5LD^2}{KR} + \frac{4D\sigma}{\sqrt{MKR}} + \frac{2(54e)^{\frac{1}{3}}D^{\frac{4}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{2(60e)^{\frac{1}{5}}D^{\frac{8}{5}}\sigma^{\frac{2}{5}}L^{\frac{2}{5}}}{K^{\frac{2}{5}}R^{\frac{4}{5}}} + \frac{60eG^2K^2L + 30G^2}{4^H}.$$

Here, the first three terms match the LSGD-rate, and the remaining terms are of lower order. While Theorem 1 indicates that LESGD achieves the LSGD rate when $\sigma = \Omega\left(\frac{DL}{R^{\frac{1}{2}}K^{\frac{1}{4}}}\right)$, and Corollary 2 implies that LIPPAX (Algorithm 2) achieves the LSGD-rate when $\sigma = O\left(\frac{DL}{\sqrt{RK}}\right)$, it is still unclear

Algorithm 3 LIPPAX with Gaussian-smoothing (SLIPPAX)

Initialization: $\mathbf{z}_0^m = \mathbf{0}$, for $m = 1, \dots, M$.
for $t = 1$ to T **do**
  for each machine $m = 1, \dots, M$ **do**
    $\mathbf{x}_{t,0}^m = \mathbf{z}_{t-1}^m$
    for $\ell = 1$ to H **do**
      Sample $\mathbf{s}_{t,\ell}^m \sim \mathcal{N}(\mathbf{0}, I_d)$, and $\xi_{t,\ell}^m$
      $\mathbf{x}_{t,\ell}^m = \mathbf{x}_{t,\ell-1}^m - \gamma \left(V(\mathbf{x}_{t,\ell-1}^m + \delta \mathbf{s}_{t,\ell}^m; \xi_{t,\ell}^m) + \frac{1}{\eta} \left(\mathbf{x}_{t,\ell-1}^m - \mathbf{z}_{t-1}^m \right) \right)$
    end for
    $\mathbf{x}_t^m = \mathbf{x}_{t,H}^m$
    $\mathbf{z}_t^m = \begin{cases} \mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m), & \text{if } \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M} \sum_{m=1}^M \left(\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m) \right), & \text{if } \mathrm{mod}(t, K) = 0. \end{cases}$
  end for
end for
return $\mathbf{x}_o = \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \mathbf{x}_t^m$

what happens in the small gap between $\sigma = \Omega\left(\frac{DL}{\sqrt{RK}}\right)$ and $\sigma = O\left(\frac{DL}{\sqrt{R}K^{\frac{1}{4}}}\right)$. Moreover, due to the $O\left(\frac{\sigma}{\sqrt{KR}}\right)$ term in (6), the upper bound in Theorem 3 degrades when σ is large, as this term is not accelerated by M. Next, we show that the LSGD-rate can be achieved under the assumption that the operator satisfies a second-order boundedness condition.

Assumption 1 (Second-order Boundedness)**.** Assume V satisfies that there exists $\Lambda > 0$ s.t. $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|V(\mathbf{x}) - V(\mathbf{y}) - J_V(\mathbf{y})(\mathbf{x} - \mathbf{y})\| \leq \Lambda \|\mathbf{x} - \mathbf{y}\|^2$, where $J_V(\mathbf{y})$ denotes the Jacobian of V at $\mathbf{y}$.

We have the following conclusion based on Assumption 1. The proof is given in Appendix D.

Theorem 4. *Suppose V is L-smooth and monotone, and $\|V(\mathbf{x})\| \leq G$ for all $\mathbf{x} \in \mathbb{R}^d$. Furthermore, suppose Assumption 1 hold. Consider running Algorithm 2 with $\gamma = \frac{1}{\eta(L + \frac{1}{\eta})^2}$, $\eta = \min\left\{ \frac{1}{L}, \frac{D^{\frac{2}{5}}}{K^{\frac{3}{5}} R^{\frac{1}{5}} \sigma^{\frac{2}{5}} L^{\frac{2}{5}}}, \frac{D^{\frac{2}{3}}}{K^{\frac{2}{3}} R^{\frac{1}{3}} \sigma^{\frac{2}{3}} L^{\frac{1}{3}}}, \frac{D^{\frac{1}{2}}}{K^{\frac{1}{4}} R^{\frac{1}{4}} \sigma^{\frac{1}{2}} L^{\frac{1}{2}}}, \frac{D^{\frac{1}{3}}}{\Lambda^{\frac{1}{3}} \sigma^{\frac{2}{3}} R^{\frac{1}{6}} K^{\frac{1}{6}}}, \frac{D\sqrt{M}}{K^{\frac{1}{2}} R^{\frac{1}{2}} \sigma} \right\}$. Then:*

$$\begin{aligned} \mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] &\leq 240 \left[\frac{D^2 L}{KR} + \frac{D^{\frac{4}{3}} L^{\frac{1}{3}} \sigma^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{D\sigma}{\sqrt{MKR}} \right] \\ &+ 240 \left[\frac{D^{\frac{8}{5}} \sigma^{\frac{2}{5}} L^{\frac{2}{5}}}{K^{\frac{2}{5}} R^{\frac{4}{5}}} + \frac{D^{\frac{3}{2}} \sigma^{\frac{1}{2}} L^{\frac{1}{2}}}{K^{\frac{3}{4}} R^{\frac{3}{4}}} + \frac{D^{\frac{5}{3}} \Lambda^{\frac{1}{3}} \sigma^{\frac{2}{3}}}{R^{\frac{5}{6}} K^{\frac{5}{6}}} \right] + \frac{10H^2 \Lambda^2 G^4 + 60G^2 + 120G^2 K^2}{4^H}. \end{aligned} \tag{7}$$

Note that, the bound given in Theorem 4 can be divided by three parts: The first part is exactly on the order of $O\left(\frac{1}{KR} + \frac{\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{\sigma}{\sqrt{MKR}}\right)$, which matches the LSGD-rate, while the second term contains only lower order terms that are dominated by the first term. The third term decreases with H exponentially, so setting $H = O(\ln KR)$ making this term become lower order.

Removing Assumption 1. Theorem 4 assumes the second-order boundedness of the operator V. For more general V, we propose using *Gaussian-smoothing* in the proximal point phase of LIPPAX to smooth the operator. The details of the algorithm is given in Algorithm 3. Compared with Algorithm 2, the main difference lies in Step 7: instead of directly querying $\widetilde{V}(\mathbf{x}_{t,\ell}^m)$, Algorithm 3 queries $\widetilde{V}(\mathbf{x}_{t,\ell}^m + \delta \mathbf{s}_{t,\ell}^m)$, where $\mathbf{s}_{t,\ell}^m \sim \mathcal{N}(0, I_d)$ is a Gaussian random variable. Let $\mathring{V}(\mathbf{x}) = \mathbb{E}[\widetilde{V}(\mathbf{x} + \delta \mathbf{s}_{t,\ell}^m)]$. Gaussian smoothing preserves monotonicity and L-smoothness while providing the second-order regularity needed by our bias analysis. The inner loop approximately solves the regularized VI associated with $\mathring{V}$. We have the following conclusion. The proof is in Appendix E.

Algorithm 4 Local Dual Extrapolation for Composite VI

Initialization: $\mathbf{z}_0^m = \mathbf{0}$, for $m = 1, \ldots, M$.
for $t = 1$ to T **do**
  for each machine $m = 1, \ldots, M$ **do**
    $\mathbf{u}_{t-1}^m = \nabla h_{t-1}^*(\mathbf{z}_{t-1}^m)$
    $\mathbf{x}_t^m = \begin{cases} \mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{u}_{t-1}^m), & \text{if } \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M} \sum_{m=1}^M \left(\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{u}_{t-1}^m) \right), & \text{if } \mathrm{mod}(t, K) = 0. \end{cases}$
    $\mathbf{v}_t^m = \nabla h_t^*(\mathbf{x}_t^m)$
    $\mathbf{z}_t^m = \begin{cases} \mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{v}_t^m), & \text{if } \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M} \sum_{m=1}^M \left(\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{v}_t^m) \right), & \text{if } \mathrm{mod}(t, K) = 0. \end{cases}$
  end for
end for
return $\mathbf{v}_o = \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \mathbf{v}_t^m$

Theorem 5. *Suppose V is L-smooth and monotone, and $\|V(\mathbf{x})\| \leq G$ for all $\mathbf{x} \in \mathbb{R}^d$. Let $\eta = \min\left\{ \frac{D^{\frac{1}{2}}}{K^{\frac{1}{4}} R^{\frac{1}{4}} L^{\frac{1}{2}} \sigma^{\frac{1}{2}} d^{\frac{1}{8}}}, \frac{D^{\frac{2}{5}}}{\sigma^{\frac{2}{5}} L^{\frac{2}{5}} d^{\frac{1}{10}} K^{\frac{3}{5}} R^{\frac{1}{5}}}, \frac{D^{\frac{2}{3}}}{K^{\frac{2}{3}} R^{\frac{1}{3}} \sigma^{\frac{2}{3}} L^{\frac{2}{3}}}, \frac{1}{L} \right\}$, $\delta = \frac{\eta\sigma}{\sqrt{d}}$, and $\gamma = \frac{1}{\eta\left(L + \frac{1}{\eta}\right)^2}$. Then:*

$$\mathbb{E}\left[\mathrm{err}_D(\mathbf{x}_o)\right] \leq O\Bigg(\left[\frac{D^2 L}{KR} + \frac{D\sigma}{\sqrt{MKR}} + \frac{D^{\frac{4}{3}} \sigma^{\frac{2}{3}} L^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} \right] + \left[\frac{D^{\frac{3}{2}} L^{\frac{1}{2}} \sigma^{\frac{1}{2}} d^{\frac{1}{8}}}{K^{\frac{3}{4}} R^{\frac{3}{4}}} + \frac{D^{\frac{8}{5}} \sigma^{\frac{2}{5}} L^{\frac{2}{5}} d^{\frac{1}{10}}}{R^{\frac{4}{5}} K^{\frac{2}{5}}} \right] + \frac{G^2 + K^2 L^2 G^2 + \frac{H^2 L^2 d^{\frac{1}{2}} G^4}{\sigma^2}}{4^H} \Bigg).$$

4.3 Faster Rates for Co-Coercive Operators via LSGD

In this section, we show when the operator is co-coercive, simple LSGD achieves the LSGD-rate.

Definition 3. V is β-co-coercive, if $\forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$, $\|V(\mathbf{z}) - V(\mathbf{z}')\|^2 \leq \beta \langle V(\mathbf{z}) - V(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle$.

For β-co-coercive operators, we consider the following classical LSGD algorithm:

$$\mathbf{x}_t^m = \begin{cases} \mathbf{x}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_{t-1}^m), & \mathrm{mod}(t, K) \neq 0; \\ \frac{1}{M} \sum_{m=1}^M \left(\mathbf{x}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_{t-1}^m) \right), & \mathrm{mod}(t, K) = 0, \end{cases} \tag{8}$$

and we have the following convergence guarantees. The proof is provided in Appendix F.

Theorem 6. *Suppose V is L-smooth, β-co-coercive and monotone. Consider the algorithm given in (8) with $\eta = \min\left\{ \frac{1}{\beta}, \frac{D\sqrt{M}}{\sqrt{KR}\sigma}, \frac{D^{\frac{2}{3}}}{2^{\frac{1}{2}} K^{\frac{2}{3}} R^{\frac{1}{3}} L^{\frac{1}{3}} \sigma^{\frac{2}{3}}} \right\}$. Then we have*

$$\mathbb{E}\left[\mathrm{err}_D\left(\frac{1}{TM} \sum_{t=1}^T \sum_{m=1}^M \mathbf{x}_t^m \right) \right] \leq \frac{4\beta D^2}{KR} + \frac{4\sigma D}{\sqrt{MKR}} + \frac{8 D^{\frac{4}{3}} L^{\frac{1}{3}} \sigma^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}}.$$

5 Faster Rates For Composite Variational Inequalities

In this section, we consider solving the following composite variational inequality problem [14, 38], where we would like to find a $\mathbf{z}^* \in \mathbb{R}^d$ such that $\sup_{\mathbf{z} \in \mathbb{R}^d} \langle V(\mathbf{z}), \mathbf{z}^* - \mathbf{z} \rangle + \phi(\mathbf{z}^*) - \phi(\mathbf{z}) \leq 0$, where the function ϕ is convex but potentially non-smooth. The performance measure is given by $\mathrm{err}_c(\overline{\mathbf{z}}) = \sup_{\mathbf{z}} \langle V(\mathbf{z}), \overline{\mathbf{z}} - \mathbf{z} \rangle + \phi(\overline{\mathbf{z}}) - \phi(\mathbf{z})$. This problem is a generalization of the smooth and monotone variational inequality problem, and it is more challenging due to the non-smooth regularizer. One important application for this problem is the composite saddle-point problem: $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) + g_1(\mathbf{x}) - g_2(\mathbf{y})$, where $f(\mathbf{x}, \mathbf{y})$ is smooth and convex-concave, while g_1, g_2 are convex but not necessarily smooth. For these problems, [14] proposed the local dual averaging algorithm, which can be considered as a mirror-descent style generalization of the LESGD algorithm.

For this algorithm, [14] show that an $O\left(\frac{L}{KR}+\frac{\sqrt{L}}{\sqrt{R}}+\frac{\sigma}{\sqrt{MKR}}+\frac{\sqrt{L}}{K^{\frac{1}{4}}R^{\frac{3}{4}}}\right)$ rate. By contrast, for composite optimization, [13] obtained an $O\left(\frac{\sigma}{\sqrt{MKR}}+\frac{1}{R^{\frac{2}{3}}}+\frac{1}{KR}\right)$ bound. Here, we show local dual averaging also achieves this bound.

To present the algorithm, we first define some notation. We refer to [14] for more technical details. Let $h:\mathbb{R}^d\mapsto\mathbb{R}_+$ be a distance generating function, and without loss of generality, we assume it is 1-strongly convex. Let $h_t=h+t\eta\phi$ be the general distance function. Let h_t^* be the convex conjugate of h_t, and $\nabla h_t^*:\mathbb{R}^d\mapsto\mathbb{R}^d$ defines a mirror map from the dual space to the primal space. The algorithm (in the dual space) is given in Algorithm 4. In each round $t\in[T]$, each machine $m\in[M]$ still conducts extra gradient step in the dual space, but using the corresponding operator value in the primal space. We have the following conclusion. The proof is postponed to Appendix G.

Theorem 7. *Suppose V is L-smooth, monotone, and $\|V(\mathbf{x})\|\leq G$ for all $\mathbf{x}\in\mathbb{R}^d$. Then Algorithm 4 with $\eta=\min\left\{\frac{D\sqrt{M}}{\sigma\sqrt{6KR}},\frac{D^{\frac{2}{3}}}{17^{\frac{1}{3}}KR^{\frac{1}{3}}L^{\frac{2}{3}}G^{\frac{2}{3}}},\frac{1}{\sqrt{10}L}\right\}$, guarantees that*

$$\mathbb{E}[\mathrm{err}_c(\mathbf{v}_o)]\leq\frac{2\sqrt{6}D\sigma}{\sqrt{MKR}}+\frac{17^{\frac{1}{3}}L^{\frac{2}{3}}D^{\frac{4}{3}}G^{\frac{2}{3}}}{R^{\frac{2}{3}}}+\frac{\sqrt{10}D^2L}{KR}.$$

Note that, this bound exactly matches the optimal results for federated composite optimization [13], and faster than the bound provided by [14] as the term $O\left(\frac{1}{\sqrt{R}}\right)$ is improved to $O\left(\frac{1}{R^{\frac{2}{3}}}\right)$.

6 Conclusion

In this paper, we study federated optimization for stochastic smooth and monotone VIs. We first show that the LESGD enjoys an $O\left(\frac{1}{\sqrt{KR}}+\frac{\sigma}{\sqrt{MKR}}+\frac{\sigma^{2/3}}{K^{1/3}R^{2/3}}\right)$ convergence rate. We then prove that the algorithm achieves the optimal rate when $\sigma=\Omega\Big(\frac{1}{\sqrt{R}K^{1/4}}\Big)$, or when the operator is affine. Next, we propose a new algorithm, called LIPPAX, which achieves the optimal rate when $\sigma=O\Big(\frac{1}{\sqrt{KR}}\Big)$, or when the Hessian of V is bounded. Then, we propose a Gauss-smoothing variant of LIPPAX, which achieves an LSGD-rate while only relying on V being bounded. Finally, we present faster rates for co-coercive smooth and monotone VIs as well as for composite VI problems.

Several open questions remain. The SLIPPAX algorithm achieves the LSGD-rate only up to a logarithmic factor and relies on the assumption that V is bounded. In future work, we plan to investigate whether these restrictions can be relaxed. Moreover, for federated composite VIs (and even federated composite optimization), existing results assume bounded gradients and include an $O\big(\frac{G}{R^{2/3}}\big)$ term in the best-known bounds, which is independent of the variance. This implies that the optimal $O\big(\frac{1}{KR}\big)$ rate cannot be achieved even when $\sigma=0$, and it remains unclear whether this limitation can be overcome. Finally, this paper focuses primarily on the homogeneous setting where all local clients share the same data distribution, whereas previous work [12] also considers the heterogeneous setting. Extending our improved rates to this setting remains an open problem; we refer to Appendix H for a more detailed discussion.

7 Statement of AI Use

Generative AI tools were used for language editing.

References

[1] Jakub Konecný, 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.

[2] 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*, 2017.

[3] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc' aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Ng. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems 25*, 2012.

[4] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[5] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems 23*, 2010.

[6] Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

[7] Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[8] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[9] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[10] Cong Xie, Oluwasanmi Koyejo, Indranil Gupta, and Haibin Lin. Local adaalter: Communication-efficient stochastic gradient descent with adaptive learning rates. *arXiv preprint arXiv:1911.09030*, 2019.

[11] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konecný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[12] Aleksandr Beznosikov, Pavel Dvurechenskii, Anastasiia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Advances in Neural Information Processing Systems*, 35:38116–38133, 2022.

[13] Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.

[14] Site Bai and Brian Bullins. Local composite saddle point optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Guido Stampacchia. Variational inequalities. In *Congres international des mathématiciens*, volume 2, pages 877–883, 1970.

[16] David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. 2000.

[17] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1), 2011.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[19] Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[20] Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35*, 2022.

[21] Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[22] Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*, 2021.

[23] Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[24] Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle point problems: lower bounds, near-optimal and robust algorithms. *Optimization Methods and Software*, pages 1–18, 2025.

[25] Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090, 2022.

[26] Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022.

[27] George J Minty. Monotone (nonlinear) operators in hilbert space. 1962.

[28] Guido Stampacchia. Formes bilineaires coercitives sur les ensembles convexes. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, 1964.

[29] Raymond Trémolieres, J-L Lions, and Roland Glowinski. *Numerical analysis of variational inequalities*, volume 8. Elsevier, 2011.

[30] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[31] Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 1980.

[32] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 2015.

[33] Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems: A survey. *European Mathematical Society Magazine*, 2023.

[34] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

[35] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

[36] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[37] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[38] Yurii Nesterov. High-order reduced-gradient methods for composite variational inequalities. *arXiv preprint arXiv:2311.15154*, 2023.

[39] Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. In *Conference on Learning Theory*, pages 831–875. PMLR, 2017.

[40] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020.

A Proof of Theorem 1

We start by defining the following shadow updates: $\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^M \mathbf{x}_t^m$, and $\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^M \mathbf{z}_t^m$. We have

$$\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^M \mathbf{x}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^M \widetilde{V}(\mathbf{z}_{t-1}^m) = \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{z},t-1}, \tag{9}$$

and

$$\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^M \mathbf{z}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^M \widetilde{V}(\mathbf{x}_t^m) = \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{x},t}. \tag{10}$$

where we define $\widetilde{V}_{\mathbf{z},t-1} = \frac{1}{M}\sum_{m=1}^M \widetilde{V}(\mathbf{z}_{t-1}^m)$, and $\widetilde{V}_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^M \widetilde{V}(\mathbf{x}_t^m)$. Moreover, let $V_{\mathbf{z},t-1} = \frac{1}{M}\sum_{m=1}^M V(\mathbf{z}_{t-1}^m)$, and $V_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^M V(\mathbf{x}_t^m)$. We first introduce the following basic lemma.

Lemma 8. *Let $\mathbf{z}^+ = \mathbf{z} - \eta\mathbf{v}$, where $\mathbf{z}, \mathbf{v} \in \mathbb{R}^d$ and $\eta \in \mathbb{R}$. Then, for any $\mathbf{z}' \in \mathbb{R}^d$, we have:*

$$\|\mathbf{z}^+ - \mathbf{z}'\|^2 = \|\mathbf{z} - \mathbf{z}'\|^2 - 2\eta\left\langle \mathbf{v}, \mathbf{z}^+ - \mathbf{z}'\right\rangle - \|\mathbf{z}^+ - \mathbf{z}\|^2.$$

Proof. We have

$$\begin{aligned}
\|\mathbf{z}^+ - \mathbf{z}'\|^2 =& \|\mathbf{z} - \eta\mathbf{v} - \mathbf{z}'\|^2 = \|\mathbf{z} - \mathbf{z}'\|^2 - 2\eta\left\langle \mathbf{v}, \mathbf{z} - \mathbf{z}'\right\rangle + \|\mathbf{z}^+ - \mathbf{z}\|^2 \\
=& \|\mathbf{z} - \mathbf{z}'\|^2 - 2\eta\left\langle \mathbf{v}, \mathbf{z}^+ - \mathbf{z}'\right\rangle + \|\mathbf{z}^+ - \mathbf{z}\|^2 + 2\eta\left\langle \mathbf{v}, \mathbf{z}^+ - \mathbf{z}\right\rangle \\
=& \|\mathbf{z} - \mathbf{z}'\|^2 - 2\eta\left\langle \mathbf{v}, \mathbf{z}^+ - \mathbf{z}'\right\rangle + \|\mathbf{z}^+ - \mathbf{z}\|^2 - 2\|\mathbf{z}^+ - \mathbf{z}\|^2 \\
=& \|\mathbf{z} - \mathbf{z}'\|^2 - 2\eta\left\langle \mathbf{v}, \mathbf{z}^+ - \mathbf{z}'\right\rangle - \|\mathbf{z}^+ - \mathbf{z}\|^2.
\end{aligned}$$

□

Apply this lemma for the update in (9) and (10), we have $\forall \mathbf{z} \in \mathbb{R}^d$,

$$\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 = \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_t - \mathbf{z}\right\rangle - \|\overline{\mathbf{z}}_t - \overline{\mathbf{z}}_{t-1}\|^2,$$

and

$$\|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_t\|^2 = \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{z},t-1}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_t\right\rangle - \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2.$$

Thus

$$\begin{aligned}
&\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_t\|^2 \\
=& \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_t - \mathbf{z}\right\rangle - 2\eta\left\langle \widetilde{V}_{\mathbf{z},t-1}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_t\right\rangle \\
=& \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle + 2\eta\left\langle \widetilde{V}_{\mathbf{x},t} - \widetilde{V}_{\mathbf{z},t-1}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_t\right\rangle \\
\leq& \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{x},t} - \widetilde{V}_{\mathbf{z},t-1}\right\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_t\|^2,
\end{aligned}$$

where the inequality is due to Young's inequality. In summary, the inequality above indicates that

$$\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 \leq \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 - 2\eta \left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle + \eta^2 \left\| \widetilde{V}_{\mathbf{x},t} - \widetilde{V}_{\mathbf{z},t-1} \right\|^2. \quad (11)$$

To proceed, we firstly focus on the last term of (11), and we have

$$\begin{aligned}
&\left\| \widetilde{V}_{\mathbf{x},t} - \widetilde{V}_{\mathbf{z},t-1} \right\|^2 \\
=&\left\| \widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t} + V_{\mathbf{x},t} - V(\overline{\mathbf{x}}_t) + V(\overline{\mathbf{x}}_t) - V(\overline{\mathbf{z}}_{t-1}) + V(\overline{\mathbf{z}}_{t-1}) - V_{\mathbf{z},t-1} + V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1} \right\|^2 \\
\leq& 5\left\| \widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t} \right\|^2 + 5\left\| V_{\mathbf{x},t} - V(\overline{\mathbf{x}}_t) \right\|^2 + 5\left\| V(\overline{\mathbf{x}}_t) - V(\overline{\mathbf{z}}_{t-1}) \right\|^2 + 5\left\| V(\overline{\mathbf{z}}_{t-1}) - V_{\mathbf{z},t-1} \right\|^2 \\
&+ 5\left\| V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1} \right\|^2 \\
\leq& 5\left\| \widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t} \right\|^2 + \frac{5L^2}{M} \sum_{m=1}^{M} \|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 + 5L^2 \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 + \frac{5L^2}{M} \sum_{m=1}^{M} \|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2 \\
&+ 5\left\| V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1} \right\|^2,
\end{aligned} \quad (12)$$

where the first inequality is based on Cauchy-Schwarz and Young's inequality, and the second inequality is based on the L-smoothness of V, Cauchy-Schwarz and Young's inequality. Next, we bound the third term of (11). We have

$$\begin{aligned}
&-2\eta \left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle = -\frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{z} \rangle + 2\eta \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle \\
=& -\frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m + \mathbf{x}_t^m - \mathbf{z} \rangle + 2\eta \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle \\
=& -\frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z} \rangle - \frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m \rangle + 2\eta \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle \\
=& -\frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z} \rangle - \underbrace{\frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m \rangle}_{=0} + 2\eta \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle \\
&+ \frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m \rangle.
\end{aligned} \quad (13)$$

Finally, note that, based on the L-smoothness of V and Cauchy-Schwarz, we have

$$\frac{2\eta}{M} \sum_{m=1}^{M} \langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m \rangle \leq \frac{2\eta}{M} \sum_{m=1}^{M} \|V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^m)\| \|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\| \leq \frac{2\eta L}{M} \sum_{m=1}^{M} \|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|^2. \quad (14)$$

Plugging (12), (13) and (14) into (11) and rearrange, we get:

$$\begin{aligned}
\frac{1}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z} \rangle \leq& \frac{\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{z}}_t - \mathbf{z}\|^2}{2\eta} + \frac{(5L^2\eta^2 - 1)}{2\eta} \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 \\
&+ \Gamma_t + \frac{5L^2\eta}{2M} \sum_{m=1}^{M} \|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2 + \frac{(5L^2\eta + 2L)}{2M} \sum_{m=1}^{M} \|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2,
\end{aligned} \quad (15)$$

where

$$\Gamma_t = \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z} \right\rangle + \frac{5\eta}{2} \left[\left\| \widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t} \right\|^2 + \left\| V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1} \right\|^2 \right],$$

which includes the noisy term. Sum it over $T = KR$, divide both sides by T, taking expectation on both sides, with $\eta \leq \frac{1}{\sqrt{14}L}$, and using the monotonicity, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle V(\mathbf{z}^*), \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq \mathbb{E}\left[\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}^*\rangle\right] \\
\leq &\frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{\mathbb{E}[\sum_{t=1}^{T}\Gamma_t]}{T} + \frac{5L^2\eta}{2MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}[\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2] + \frac{(5L^2\eta + 2L)}{2MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2] \\
\leq &\frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{\mathbb{E}[\sum_{t=1}^{T}\Gamma_t]}{T} + \frac{6L}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\left[\mathbb{E}[\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2] + \mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2]\right].
\end{aligned} \tag{16}$$

where we let

$$\mathbf{z}^* = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_D}\left\langle V(\mathbf{z}), \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}\right\rangle.$$

To proceed, we bound each term in (16) respectively. For the second term, recall that

$$\Gamma_t = \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}^*\right\rangle + \frac{5\eta}{2}\left[\left\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\right\|^2 + \left\|V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1}\right\|^2\right].$$

We start from the first term. Note that $\mathbf{z}^*$ is a random variable that depends on all data. Therefore, this term is not necessarily 0 in expectation. Nevertheless, there exist standard tricks for dealing this term [17, 24]. Let $\Delta_t = V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}$, and $\mathbf{p}^{t+1} = \mathbf{p}^t - \eta\Delta_t$, with $\mathbf{p}^1 = \overline{\mathbf{z}}_0$. Then we have

$$\begin{aligned}
\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}^*\right\rangle &= \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{p}^t + \mathbf{p}^t - \mathbf{z}^*\right\rangle \\
&= \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{p}^t\right\rangle + \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \mathbf{p}^t - \mathbf{z}^*\right\rangle,
\end{aligned}$$

Now the first term is zero in expectation as it does not depend on $\mathbf{z}^*$, and we mainly focus on the second term. Based on the definition of $\mathbf{p}^t$, we have

$$\begin{aligned}
\langle\eta\Delta_t, \mathbf{p}^t - \mathbf{z}^*\rangle &= \langle\eta\Delta_t, \mathbf{p}^t - \mathbf{p}^{t+1}\rangle + \langle\eta\Delta_t, \mathbf{p}^{t+1} - \mathbf{z}^*\rangle \\
&= \langle\eta\Delta_t, \mathbf{p}^t - \mathbf{p}^{t+1}\rangle + \langle\mathbf{p}^{t+1} - \mathbf{p}^t, \mathbf{z}^* - \mathbf{p}^{t+1}\rangle \\
&= \langle\eta\Delta_t, \mathbf{p}^t - \mathbf{p}^{t+1}\rangle + \frac{1}{2}\|\mathbf{p}^t - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 \\
&\leq \frac{\eta^2}{2}\|\Delta_t\|^2 + \frac{1}{2}\|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 + \frac{1}{2}\|\mathbf{p}^t - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 \\
&= \frac{\eta^2}{2}\|\Delta_t\|^2 + \frac{1}{2}\|\mathbf{p}^t - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2.
\end{aligned}$$

Here, the second equality is because

$$\|\mathbf{p}^t - \mathbf{z}^*\|^2 = \|\mathbf{p}^t - \mathbf{p}^{t+1} + \mathbf{p}^{t+1} - \mathbf{z}^*\|^2 = \|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 + \|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2 + 2\langle\mathbf{p}^{t+1} - \mathbf{p}^t, \mathbf{z}^* - \mathbf{p}^{t+1}\rangle,$$

and the inequality is based on Cauchy-Schwarz. Finally, note that since $\widetilde{V}(\mathbf{x}_t^m)$ are i.i.d. for each $m \in [M]$, we have

$$\mathbb{E}\left[\|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2\right] \leq \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|V(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^m)\|^2\right] \leq \frac{\sigma^2}{M}, \tag{17}$$

and

$$\mathbb{E}\left[\|V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1}\|^2\right] \leq \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}[\|V(\mathbf{z}_{t-1}^m) - \widetilde{V}(\mathbf{z}_{t-1}^m)\|^2] \leq \frac{\sigma^2}{M}. \tag{18}$$

Combining the above inequalities, we get

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Gamma_t\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{6\sigma^2\eta}{M}. \tag{19}$$

Finally, we deal with the two drift terms in (16), and we introduce the following lemma. The proof is given in A.1.

Lemma 9 (Client Drift). *For $\eta \leq \frac{1}{\sqrt{14KL}}$, we have*

$$\frac{6L}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\left[\mathbb{E}[\|\mathbf{z}_{t-1}^m-\overline{\mathbf{z}}_{t-1}\|^2]+\mathbb{E}[\|\mathbf{x}_t^m-\overline{\mathbf{x}}_t\|^2]\right]\leq 936eL\eta^2\sigma^2K.$$

Combining (19), Lemma 9 with (16), we finally get:

$$\mathbb{E}\left[\left\langle V(\mathbf{z}^*),\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m-\mathbf{z}^*\right\rangle\right]\leq\frac{\mathbb{E}[\|\mathbf{z}_0-\mathbf{z}^*\|^2]}{\eta KR}+\frac{6\sigma^2\eta}{M}+936eL\eta^2\sigma^2K.$$

Let $\eta=\min\left\{\frac{1}{\sqrt{14KL}},\frac{D\sqrt{M}}{\sigma\sqrt{6KR}},\frac{D^{\frac{2}{3}}}{936^{\frac{1}{3}}e^{\frac{2}{3}}K^{\frac{2}{3}}R^{\frac{1}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}\right\}$, we get

$$\mathbb{E}\left[\left\langle V(\mathbf{z}^*),\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m-\mathbf{z}^*\right\rangle\right]\leq\frac{20D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}+\frac{2\sqrt{6}D\sigma}{\sqrt{MKR}}+\frac{2\sqrt{14}LD^2}{\sqrt{K}R}.$$

A.1 Proof of Lemma 9

We first introduce the following lemma about the drift term in local updates.

Lemma 10 (Lemma 4 of [7]). *For any $t\in[T]$ and $m'\neq m$, we have*

$$\mathbb{E}[\|\mathbf{x}_t^m-\overline{\mathbf{x}}_t\|^2]\leq\mathbb{E}[\|\mathbf{x}_t^m-\mathbf{x}_t^{m'}\|^2],\quad \textit{and},\quad \mathbb{E}[\|\mathbf{z}_{t-1}^m-\overline{\mathbf{z}}_{t-1}\|^2]\leq\mathbb{E}[\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\|^2].$$

We first focus on the drift of $\mathbf{x}$. We have:

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_t^m-\mathbf{x}_t^{m'}\|^2]=\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}-\eta\left(\widetilde{V}(\mathbf{z}_{t-1}^m)-\widetilde{V}(\mathbf{z}_{t-1}^{m'})\right)\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+\eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{z}_{t-1}^m)-\widetilde{V}(\mathbf{z}_{t-1}^{m'})\right\|^2\right]-\mathbb{E}\left[2\eta\left\langle\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'},\widetilde{V}(\mathbf{z}_{t-1}^m)-\widetilde{V}(\mathbf{z}_{t-1}^{m'})\right\rangle\right]\\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+\eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{z}_{t-1}^m)-\widetilde{V}(\mathbf{z}_{t-1}^{m'})\right\|^2\right]-\mathbb{E}\left[2\eta\underbrace{\left\langle\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'},V(\mathbf{z}_{t-1}^m)-V(\mathbf{z}_{t-1}^{m'})\right\rangle}_{\geq 0}\right]\\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+\eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{z}_{t-1}^m)-\widetilde{V}(\mathbf{z}_{t-1}^{m'})\right\|^2\right]\\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+2\eta^2\mathbb{E}\left[\left\|V(\mathbf{z}_{t-1}^m)-\widetilde{V}(\mathbf{z}_{t-1}^m)\right\|^2\right]+2\eta^2\mathbb{E}\left[\left\|V(\mathbf{z}_{t-1}^{m'})-\widetilde{V}(\mathbf{z}_{t-1}^{m'})\right\|^2\right]\\
&+\eta^2\mathbb{E}[\|V(\mathbf{z}_{t-1}^m)-V(\mathbf{z}_{t-1}^{m'})\|^2]\\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+4\eta^2\sigma^2+\eta^2\mathbb{E}[\|V(\mathbf{z}_{t-1}^m)-V(\mathbf{z}_{t-1}^{m'})\|^2]\\
\leq&\left(1+\eta^2L^2\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+4\eta^2\sigma^2\leq 2\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m-\mathbf{z}_{t-1}^{m'}\right\|^2\right]+4\eta^2\sigma^2,
\end{aligned}\tag{20}$$

where the first inequality is due to monotonicity, and the last inequality is based on $\eta\leq\frac{1}{L}$. Next, we bound the draft for $\mathbf{z}$. Let t_0, $t-t_0\leq K$, be the communication round such that $\mathbf{z}_{t_0}^m=\mathbf{z}_{t_0}^{m'}$. Then

we have

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2] = \mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} - \eta\left(\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta\mathbb{E}\left[\left\langle \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}, \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\rangle\right] \\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta\underbrace{\mathbb{E}\left[\left\langle \mathbf{x}_t^m - \mathbf{x}_t^{m'}, \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\rangle\right]}_{\geq 0} \\
&- 2\eta\mathbb{E}\left[\left\langle \mathbf{z}_{t-1}^m - \mathbf{x}_t^m + \mathbf{x}_t^{m'} - \mathbf{z}_{t-1}^{m'}, \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\rangle\right] \\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta^2\mathbb{E}\left[\left\langle \widetilde{V}(\mathbf{z}_{t-1}^m) - \widetilde{V}(\mathbf{z}_{t-1}^{m'}), \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\rangle\right] \\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 4\eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{z}_{t-1}^m) - \widetilde{V}(\mathbf{z}_{t-1}^{m'})\right\|^2\right] + 5\eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right] \\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 4\eta^2\mathbb{E}\left[L^2\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2 + 4\sigma^2\right] + 5\eta^2\mathbb{E}\left[L^2\left\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\right\|^2 + 4\sigma^2\right] \\
\leq&\left(1 + 14\eta^2L^2\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 37\eta^2\sigma^2 \leq \left(1 + \frac{1}{K}\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 37\eta^2\sigma^2 \\
\leq&37e\eta^2\sigma^2K.
\end{aligned}
\tag{21}
$$

where in the fourth inequality we used (20), and the last two inequalities are based on the fact that $\eta \leq \frac{1}{\sqrt{14K}L}$ and $\left(1 + \frac{1}{K}\right)^K \leq e$. Combining with (20), we have

$$
\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] \leq 78e\eta^2\sigma^2K.
$$

B Proof of Lemma 2

We first introduce the following lemma, which is Lemma D.1 of [26].

Lemma 11. *Suppose the operator V is affine and L-smooth, and let operator $F(\mathbf{z}) = V(\mathbf{z} - \eta V(\mathbf{z}))$. Then F is $\frac{2}{\eta}$-cocoercive, for $\eta \leq \frac{1}{L}$. That is, for any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$,*

$$
\|F(\mathbf{z}) - F(\mathbf{z}')\|^2 \leq \frac{2}{\eta}\left\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}'\right\rangle.
$$

Next, we extend this conclusion to the stochastic setting. The proof can be found in section B.1.

Proposition 12 (Stochastic co-coercivity of EG)**.** *Assume V is affine and L-smooth, $\mathbb{E}[\widetilde{V}(\mathbf{z})] = V(\mathbf{z})$, and $\mathbb{E}\|\widetilde{V}(\mathbf{z}) - V(\mathbf{z})\|^2 \leq \sigma^2$ for $\mathbf{z} \in \mathbb{R}^d$. Let*

$$
F_{SEG}(\mathbf{z}) = V\big(\mathbf{z} - \eta\,\widetilde{V}(\mathbf{z})\big),
$$

and suppose $\eta \leq 1/L$. Then for any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$,

$$
\mathbb{E}\|F_{SEG}(\mathbf{z}) - F_{SEG}(\mathbf{z}')\|^2 \leq \frac{2}{\eta}\mathbb{E}\big\langle F_{SEG}(\mathbf{z}) - F_{SEG}(\mathbf{z}'),\, \mathbf{z} - \mathbf{z}'\big\rangle + 4L^2\eta^2\sigma^2.
$$

By using this lemma, we provide an alternative proof for upper bound of the drift term in (21). We have

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2] = \mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} - \eta\left(\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta\mathbb{E}\left[\left\langle \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}, \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\rangle\right] \\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|V(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta\mathbb{E}\left[\left\langle \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}, \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\rangle\right] + 4\eta^2\sigma^2 \\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 8\eta^2\sigma^2 \leq 8\eta^2\sigma^2 K,
\end{aligned} \tag{22}$$

where for the second inequality we used Proposition 12 and $\eta \leq \frac{1}{L}$. Combining with (20), it also implies that

$$\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] \leq 20\eta^2\sigma^2 K.$$

Therefore, we can immediately obtain the following bound for the client drift.

Lemma 13 (Client Drift for Affine Operator)**.** *For $\eta \leq \frac{1}{L}$, we have*

$$\frac{6L}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\left[\mathbb{E}[\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2] + \mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2]\right] \leq 180L\eta^2\sigma^2 K. \tag{23}$$

we finally get:

$$\mathbb{E}\left[\left\langle V(\mathbf{z}^*), \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq \frac{\mathbb{E}[\|\mathbf{z}_0 - \mathbf{z}^*\|^2]}{\eta KR} + \frac{6\sigma^2\eta}{M} + 180L\eta^2\sigma^2 K. \tag{24}$$

Let $\eta = \min\left\{\frac{1}{L}, \frac{D\sqrt{M}}{\sigma\sqrt{6KR}}, \frac{D^{\frac{2}{3}}}{180^{\frac{1}{3}}K^{\frac{2}{3}}R^{\frac{1}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}\right\}$, we get

$$\mathbb{E}\left[\left\langle V(\mathbf{z}^*), \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq \frac{20D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{2\sqrt{6}D\sigma}{\sqrt{MKR}} + \frac{2LD^2}{KR}. \tag{25}$$

B.1 Proof of Proposition 12

Let $F(\mathbf{z}) = V\big(\mathbf{z} - \eta V(\mathbf{z})\big)$. Since V is affine, let $V(\mathbf{z}) = A\mathbf{z} + b$. We have

$$\begin{aligned}
F_{\text{SEG}}(\mathbf{z}) - F_{\text{SEG}}(\mathbf{z}') &= V(\mathbf{z} - \eta\widetilde{V}(\mathbf{z})) - V(\mathbf{z}' - \eta\widetilde{V}(\mathbf{z}')) \\
&= A[(\mathbf{z} - \eta V(\mathbf{z})) - (\mathbf{z}' - \eta V(\mathbf{z}'))] - \eta A\Big[(\widetilde{V}(\mathbf{z}) - V(\mathbf{z})) - (\widetilde{V}(\mathbf{z}') - V(\mathbf{z}'))\Big].
\end{aligned} \tag{26}$$

Taking squared norms and expectation, the cross term vanishes by $\mathbb{E}[\widetilde{V}(\cdot) - V(\cdot)] = 0$, hence

$$\begin{aligned}
&\mathbb{E}\|F_{\text{SEG}}(\mathbf{z}) - F_{\text{SEG}}(\mathbf{z}')\|^2 = \|F(\mathbf{z}) - F(\mathbf{z}')\|^2 + \eta^2\,\mathbb{E}\big\|A\big((\widetilde{V}(\mathbf{z}) - V(\mathbf{z})) - (\widetilde{V}(\mathbf{z}') - V(\mathbf{z}'))\big)\big\|^2 \\
\leq&\|F(\mathbf{z}) - F(\mathbf{z}')\|^2 + L^2\eta^2\,\mathbb{E}\|(\widetilde{V}(\mathbf{z}) - V(\mathbf{z})) - (\widetilde{V}(\mathbf{z}') - V(\mathbf{z}'))\|^2 \\
\leq&\|F(\mathbf{z}) - F(\mathbf{z}')\|^2 + 4L^2\eta^2\sigma^2 \\
\leq&\frac{2}{\eta}\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}'\rangle + 4L^2\eta^2\sigma^2 = \frac{2}{\eta}\mathbb{E}\big\langle F_{\text{SEG}}(\mathbf{z}) - F_{\text{SEG}}(\mathbf{z}'), \mathbf{z} - \mathbf{z}'\big\rangle + 4L^2\eta^2\sigma^2.
\end{aligned}$$

C Proof of Theorem 3

Similar to previous proof, we start by defining the following shadow updates: Let $\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m$, and $\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m$, and

$$\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{x}_t^m) = \overline{\mathbf{z}}_{t-1} - \widetilde{V}_{\mathbf{x},t}.$$

where we define $\widetilde{V}_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{x}_t^m)$. and let $V_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^{M}V(\mathbf{x}_t^m)$.

Next, we define another series of shadow updates for the "exact" proximal update. Note that these updates are only used for analysis and are not required to be updated when implementing the algorithm. Specifically, at round t, for each machine $m = 1, \ldots, M$, define

$$\mathbf{x}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m}),$$

and let $\overline{\mathbf{x}}_t^* = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^{*,m}$, and $V_{\mathbf{x}^*,t} = \frac{1}{M}\sum_{m=1}^{M}V(\mathbf{x}_t^{*,m})$. We first introduce the following lemma about the properties of $\mathbf{x}_t^{*,m}$. The proof is given in Appendix C.1.

Lemma 14. *Let* $\mathbf{x}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m})$, $\gamma = \frac{1}{\eta\left(L+\frac{1}{\eta}\right)^2}$, *and* $\eta \le \frac{1}{L}$. *then*

$$\mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2\right] \le 0.25^H\eta^2G^2 + \eta^2\sigma^2.$$

Moreover, we have

$$\|\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\| \le \eta G.$$

We now begin the proof by first bounding the following potential function.

$$\begin{aligned}
&\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 = \left\|\overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{x},t} - \mathbf{z}\right\|^2 = \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle\widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2\\
=&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle\\
=&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle + 2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle.
\end{aligned} \tag{27}$$

Next, we bound each term in (27) respectively. For the second term, we have

$$\begin{aligned}
&-2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle = -2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle = -2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m + \mathbf{x}_t^m - \mathbf{z}\right\rangle\\
=&-2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\right\rangle - 2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\right\rangle\\
=&\underbrace{-2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\right\rangle}_{=0} - 2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\right\rangle + 2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\right\rangle\\
\le&2\eta L\frac{1}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 - 2\eta\frac{1}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\right\rangle,
\end{aligned} \tag{28}$$

where the inequality is based on the L-smoothness of V and the Cauchy-Schwarz inequality. For the third term in (27), we have

$$2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle = \underbrace{2\eta\left\langle V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle}_{A_1} + \underbrace{2\eta\left\langle V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_{t-1}\right\rangle}_{A_2} + \underbrace{2\eta\left\langle V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\right\rangle}_{A_3}. \tag{29}$$

For A_1, we have

$$\|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\| \le \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + \|\overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_{t-1}\| = \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + \eta\|V_{\mathbf{x}^*,t}\| \tag{30}$$

therefore

$$\begin{aligned}
A_1 \le&2\eta\|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + 2\eta^2\|V_{\mathbf{x}^*,t}\|\|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|\\
\le&\frac{\eta(L+9)}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + \frac{\eta^2}{4}\|V_{\mathbf{x}^*,t}\|^2.
\end{aligned}$$

We also have

$$A_2 = -2\eta^2\|V_{\mathbf{x}^*,t}\|^2,$$

and

$$A_3 \leq \frac{\eta^2}{4}\|V_{\mathbf{x}^*,t}\|^2 + \frac{8}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2.$$

Therefore, we have

$$2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle \leq \frac{15}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 - 1.5\eta^2\|V_{\mathbf{x}^*,t}\|^2$$

Finally, for the fourth term of (27), we have

$$\begin{aligned}
&\|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 = \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{x}}_t^* + \overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_t\|^2 \leq 1.5\|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{x}}_t^*\|^2 + 3\|\overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_t\|^2 \\
=&1.5\eta^2\|V_{\mathbf{x}^*,t}\|^2 + 3\eta^2\|V_{\mathbf{x}^*,t} - \widetilde{V}_{\mathbf{x},t}\|^2 \\
\leq&1.5\eta^2\|V_{\mathbf{x}^*,t}\|^2 + \frac{6\eta^2L^2}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^{*,m} - \mathbf{x}_t^m\|^2 + 6\eta^2\|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2
\end{aligned} \tag{31}$$

Therefore, combining all the above inequalities, for (27) the upper bound becomes

$$\begin{aligned}
\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 \leq&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 + 2\eta L\frac{1}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 - 2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle \\
&+ \frac{30}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + 6\eta^2\|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle
\end{aligned}$$

Rearrange, and taking expectation on both sides, we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}\|^2]}{2\eta T} + \frac{L}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\left[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2\right] \\
&+ \frac{15}{\eta MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Gamma_t\right],
\end{aligned} \tag{32}$$

where

$$\Gamma_t = 3\eta\|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2 + \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle,$$

is the noise term. To proceed, we need to bound the drift term, and we have the following lemma. The proof is given in Appendix C.2.

Lemma 15. *We have*

$$\mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2] \leq \frac{60eG^2K^2L\eta^4 + 6\eta^2G^2}{4^H} + 60e\eta^4K^2L\sigma^2 + 54e\eta^2K\sigma^2.$$

Finally, we have the following bound for the noise term Γ_t. The proof is given in Appendix C.3.

Lemma 16. *We have*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Gamma_t\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{2\sigma^2\eta}{M}.$$

We get

$$\begin{aligned}
\mathbb{E}\left[\left\langle V(\mathbf{z}^*), \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq&\frac{\mathbb{E}[\|\mathbf{z}_0 - \mathbf{z}^*\|^2]}{\eta KR} + \frac{2\sigma^2\eta}{M} + 60eK^2L^2\sigma^2\eta^4 + 54eKL\sigma^2\eta^2 \\
&+ 15\eta\sigma^2 + \frac{60eG^2K^2L\eta^4 + 30\eta G^2}{4^H}
\end{aligned} \tag{33}$$

Finally, we let

$$\eta = \min\left\{\frac{1}{L}, \frac{D\sqrt{M}}{\sigma\sqrt{KR}}, \frac{D^{\frac{2}{5}}}{(60e)^{\frac{1}{5}}K^{\frac{3}{5}}R^{\frac{1}{5}}\sigma^{\frac{2}{5}}L^{\frac{2}{5}}}, \frac{D^{\frac{2}{3}}}{(54e)^{\frac{1}{3}}K^{\frac{2}{3}}R^{\frac{1}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}, \frac{D}{\sigma\sqrt{15KR}}\right\}.$$

C.1 Proof of Lemma 14

Define $F(\mathbf{x}) = V(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$. Firstly, it is easy to show that $F(x)$ is $\frac{1}{\eta}$ strongly monotone, and $L + \frac{1}{\eta}$ smooth. More specifically, for any $\mathbf{x}, \mathbf{y}$, we have

$$\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle = \langle V(\mathbf{x}) - V(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{1}{\eta}\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle \geq \frac{1}{\eta}\|\mathbf{x} - \mathbf{y}\|^2. \tag{34}$$

On the other hand,

$$\|F(\mathbf{x}) - F(\mathbf{y})\|^2 \leq L^2\|\mathbf{x} - \mathbf{y}\|^2 + \frac{2L}{\eta}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\eta}\|\mathbf{x} - \mathbf{y}\|^2 = \left(L + \frac{1}{\eta}\right)^2 \|\mathbf{x} - \mathbf{y}\|^2. \tag{35}$$

Next, we show the convergence. Let $\widetilde{F}(\mathbf{x}) = \widetilde{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$. Note that $\|\widetilde{F}(\mathbf{x}) - F(\mathbf{x})\|^2 \leq \sigma^2$. Firstly, note that:

$$F(\mathbf{x}_t^{*,m}) = V(\mathbf{x}_t^{*,m}) + \frac{1}{\eta}(\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m) = \mathbf{0}.$$

Moreover, since $\gamma = \frac{1}{\eta\left(L + \frac{1}{\eta}\right)^2}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_{t,\ell}^m - \mathbf{x}_t^{*,m}\|^2\right] = \mathbb{E}\left[\|\mathbf{x}_{t,\ell-1}^m - \gamma\widetilde{F}(\mathbf{x}_{t,\ell-1}^m) - \mathbf{x}_t^{*,m} + \gamma F(\mathbf{x}_t^{*,m})\|^2\right] \\
=&\mathbb{E}[\|\mathbf{x}_{t,\ell-1}^m - \mathbf{x}_t^{*,m}\|^2] - 2\gamma\mathbb{E}\left[\left\langle F(\mathbf{x}_{t,\ell-1}^m) - F(\mathbf{x}_t^{*,m}), \mathbf{x}_{t,\ell-1}^m - \mathbf{x}_t^{*,m}\right\rangle\right] + \gamma^2\mathbb{E}[\|\widetilde{F}(\mathbf{x}_{t,\ell-1}^m) - F(\mathbf{x}_t^{*,m})\|^2] \\
\leq&\mathbb{E}[\|\mathbf{x}_{t,\ell-1}^m - \mathbf{x}_t^{*,m}\|^2] - \frac{2\gamma}{\eta}\mathbb{E}[\|\mathbf{x}_{t,\ell-1}^m - \mathbf{x}_t^{*,m}\|^2] + \gamma^2\mathbb{E}[\|F(\mathbf{x}_{t,\ell-1}^m) - F(\mathbf{x}_t^{*,m})\|^2 + \gamma^2\mathbb{E}[\|F(\mathbf{x}_{t,\ell-1}^m) - \widetilde{F}(\mathbf{x}_{t,\ell-1}^m)\|^2 \\
=&\left(1 - \frac{2\gamma}{\eta} + \gamma^2\left(L + \frac{1}{\eta}\right)^2\right)\mathbb{E}[\|\mathbf{x}_{t,\ell-1}^m - \mathbf{x}_t^{*,m}\|^2] + \gamma^2\sigma^2 \\
=&\left(1 - \frac{1}{(\eta L + 1)^2}\right)\mathbb{E}[\|\mathbf{x}_{t,\ell-1}^m - \mathbf{x}_t^{*,m}\|^2] + \gamma^2\sigma^2 \leq \left(1 - \frac{1}{(\eta L + 1)^2}\right)^H \mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m}\|^2] + \frac{\sigma^2}{\left(L + \frac{1}{\eta}\right)^2} \\
\leq&\left(1 - \frac{1}{(\eta L + 1)^2}\right)^H \mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m}\|^2] + \eta^2\sigma^2 \leq 0.25^H\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m}\|^2] + \eta^2\sigma^2.
\end{aligned} \tag{36}$$

Since $F(\mathbf{x}) = V(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$ is $\frac{1}{\eta}$-strongly monotone, we have

$$\left\langle F(\mathbf{x}_t^{*,m}) - F(\mathbf{z}_{t-1}^m), \mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\right\rangle \geq \frac{1}{\eta}\|\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\|^2.$$

Also, we have $F(\mathbf{x}_t^{*,m}) = 0$, and $F(\mathbf{z}_{t-1}^m) = V(\mathbf{z}_{t-1}^m)$, so

$$G\|\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\| \geq \left\langle F(\mathbf{x}_t^{*,m}) - F(\mathbf{z}_{t-1}^m), \mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\right\rangle \geq \frac{1}{\eta}\|\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\|^2,$$

so $\|\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m\| \leq \eta G$.

C.2 Proof of Lemma 15

We have

$$\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2 \leq 3\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + 3\|\mathbf{x}_t^{*,m} - \mathbf{x}_t^{*,m'}\|^2 + 3\|\mathbf{x}_t^{m'} - \mathbf{x}_t^{*,m'}\|^2, \tag{37}$$

where the inequality is based on Cauchy-Schwarz and Young's inequalities. Note that, the first and third terms are small based on Lemma 14. Therefore, in the following we mainly focus on the middle

term. We have

$$
\begin{aligned}
&\|\mathbf{x}_t^{*,m} - \mathbf{x}_t^{*,m'}\|^2 = \|\mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m}) - \mathbf{z}_{t-1}^m + \eta V(\mathbf{x}_t^{*,m'})\|^2 \\
=&\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2 - 2\eta \left\langle V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle + \eta^2 \|V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'})\|^2 \\
=&\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2 - 2\eta \left\langle V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'}), \mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m} + \mathbf{x}_t^{*,m} - \mathbf{x}_t^{*,m'} + \mathbf{x}_t^{*,m'} - \mathbf{z}_{t-1}^{m'} \right\rangle \\
&+ \eta^2 \|V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'})\|^2 \\
=&\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2 - \eta^2 \|V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'})\|^2 - 2\eta \underbrace{\left\langle V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'}), \mathbf{x}_t^{*,m} - \mathbf{x}_t^{*,m'} \right\rangle}_{\geq 0} \\
\leq&\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2.
\end{aligned}
$$

The inequality above indicates that the drift of $\mathbf{x}_t^{*,m}$ is bounded by the drift of $\mathbf{z}_{t-1}^m$. Next, we turn to bound the drift of $\mathbf{z}_{t-1}^m$. We have

$$
\begin{aligned}
&\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2 = \|\mathbf{z}_{t-1}^m - \eta \widetilde{V}(\mathbf{x}_t^m) - \mathbf{z}_{t-1}^{m'} + \eta \widetilde{V}(\mathbf{x}_t^{m'})\|^2 \\
=&\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2 - 2\eta \left\langle \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle + \eta^2 \left\| \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'}) \right\|^2
\end{aligned} \tag{38}
$$

We deal with each term separately. For the third term of (38), we have

$$
\begin{aligned}
&\eta^2 \left\| \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'}) \right\|^2 \\
=&\eta^2 \left\| \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^m) + V(\mathbf{x}_t^m) - V(\mathbf{x}_t^{*,m}) + V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'}) + V(\mathbf{x}_t^{*,m'}) - V(\mathbf{x}_t^{m'}) + V(\mathbf{x}_t^{m'}) - \widetilde{V}(\mathbf{x}_t^{m'}) \right\|^2 \\
\leq&8\eta^2 \left\| \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^m) \right\|^2 + 8\eta^2 \left\| V(\mathbf{x}_t^m) - V(\mathbf{x}_t^{*,m}) \right\|^2 + 2\eta^2 \left\| V(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m'}) \right\|^2 \\
&+ 8\eta^2 \left\| V(\mathbf{x}_t^{*,m'}) - V(\mathbf{x}_t^{m'}) \right\|^2 + 8\eta^2 \left\| V(\mathbf{x}_t^{m'}) - \widetilde{V}(\mathbf{x}_t^{m'}) \right\|^2.
\end{aligned} \tag{39}
$$

For the second term in (38), we have

$$
\begin{aligned}
&- 2\eta \left\langle \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle = -2\eta \left\langle V(\mathbf{x}_t^{m,*}) - V(\mathbf{x}_t^{m',*}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle \\
&- 2\eta \left\langle \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m,*}) + V(\mathbf{x}_t^{m',*}) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle \\
=&- 2\eta \left\langle V(\mathbf{x}_t^{m,*}) - V(\mathbf{x}_t^{m',*}), \mathbf{z}_{t-1}^m - \mathbf{x}_t^{m,*} + \mathbf{x}_t^{m,*} - \mathbf{x}_t^{m',*} + \mathbf{x}_t^{m',*} - \mathbf{z}_{t-1}^{m'} \right\rangle \\
&- 2\eta \left\langle \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m,*}) + V(\mathbf{x}_t^{m',*}) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle \\
=&- 2\eta^2 \left\| V(\mathbf{x}_t^{m,*}) - V(\mathbf{x}_t^{m',*}) \right\|^2 - \underbrace{2\eta \left\langle V(\mathbf{x}_t^{m,*}) - V(\mathbf{x}_t^{m',*}), \mathbf{x}_t^{m,*} - \mathbf{x}_t^{m',*} \right\rangle}_{\geq 0} \\
&- 2\eta \left\langle \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m,*}) + V(\mathbf{x}_t^{m',*}) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} \right\rangle
\end{aligned} \tag{40}
$$

Note that the first term here in (40) cancels the third term in (39). Plugging (39) and (40) into (38) and taking expectation on both sides, we get

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2\right] \leq \mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] - \mathbb{E}\left[2\eta\left\langle \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m,*}) + V(\mathbf{x}_t^{m',*}) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\rangle\right] \\
&+ \mathbb{E}\left[8\eta^2\left\|\widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^m)\right\|^2 + 8\eta^2\left\|V(\mathbf{x}_t^m) - V(\mathbf{x}_t^{*,m})\right\|^2 + 8\eta^2\left\|V(\mathbf{x}_t^{*,m'}) - V(\mathbf{x}_t^{m'})\right\|^2\right] \\
&+ 8\eta^2\mathbb{E}\left[\left\|V(\mathbf{x}_t^{m'}) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right] \\
\leq&\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] + \mathbb{E}\left[2\eta^2 KL\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + 2\eta^2 KL\|\mathbf{x}_t^{m'} - \mathbf{x}_t^{*,m'}\|^2 + \frac{1}{K}\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2\right] \\
&+ 16\eta^2\sigma^2 + 8\eta^2 L\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2] + 8\eta^2 L\mathbb{E}[\|\mathbf{x}_t^{m'} - \mathbf{x}_t^{*,m'}\|^2] \\
\leq&\left(1 + \frac{1}{K}\right)\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] + 20\eta^2 KL\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2] + 16\eta^2\sigma^2
\end{aligned}$$

Based on Lemma 14, we have

$$\mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2\right] \leq 0.25^H\eta^2 G^2 + \eta^2\sigma^2.$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2\right] \leq&\left(1 + \frac{1}{K}\right)\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] + 20\eta^2 KL\left(0.25^H\eta^2 G^2 + \eta^2\sigma^2 + \eta^2\sigma^2\right) + 16\eta^2\sigma^2 \\
\leq&\frac{20eG^2K^2L\eta^4}{4^H} + 20e\eta^4K^2L\sigma^2 + 16e\eta^2K\sigma^2.
\end{aligned}$$

Plugging it back to (37), and again apply Lemma 14, and use the fact that $\eta \leq \frac{1}{L} \leq 1$, we get:

$$\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] \leq \frac{60eG^2K^2L\eta^4 + 6\eta^2G^2}{4^H} + 60e\eta^4K^2L\sigma^2 + 54e\eta^2K\sigma^2.$$

C.3 Proof of Lemma 16

let

$$\mathbf{z}^* = \operatorname*{argmax}_{\mathbf{z}}\left\langle V(\mathbf{z}), \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}\right\rangle.$$

Note that $\mathbf{z}^*$ is a random variable that depends on all data. To proceed, we bound each term respectively. For the second term, recall that

$$\Gamma_t = \frac{2\eta}{M}\sum_{m=1}^{M}\|V(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^m)\|^2 + \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle,$$

We start from the second term. Let $\Delta_t = V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}$, and $\mathbf{p}_{t+1} = \mathbf{p}_t - \eta\Delta_t$, with $\mathbf{p}_1 = \overline{\mathbf{z}}_0$. Then we have

$$\begin{aligned}
\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}^*\right\rangle =&\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{p}_t + \mathbf{p}_t - \mathbf{z}^*\right\rangle \\
=&\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{p}_t\right\rangle + \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \mathbf{p}_t - \mathbf{z}^*\right\rangle,
\end{aligned}$$

Note that the first term is zero in expectation as it does not depend on $\mathbf{z}^*$, and we mainly focus on the second term. Based on the definition of $\mathbf{p}_t$, we have

$$\begin{aligned}
\langle\eta\Delta_t, \mathbf{p}^t - \mathbf{z}^*\rangle =&\langle\eta\Delta_t, \mathbf{p}^t - \mathbf{p}^{t+1}\rangle + \langle\mathbf{p}^{t+1} - \mathbf{p}^t, \mathbf{z}^* - \mathbf{p}^{t+1}\rangle \\
=&\langle\eta\Delta_t, \mathbf{p}^t - \mathbf{p}^{t+1}\rangle + \frac{1}{2}\|\mathbf{p}^t - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 \\
\leq&\frac{\eta^2}{2}\|\Delta_t\|^2 + \frac{1}{2}\|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 + \frac{1}{2}\|\mathbf{p}^t - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 \\
=&\frac{\eta^2}{2}\|\Delta_t\|^2 + \frac{1}{2}\|\mathbf{p}^t - \mathbf{z}^*\|^2 - \frac{1}{2}\|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2.
\end{aligned}$$

Here, the second equality is because
$\|\mathbf{p}^t - \mathbf{z}^*\|^2 = \|\mathbf{p}^t - \mathbf{p}^{t+1} + \mathbf{p}^{t+1} - \mathbf{z}^*\|^2 = \|\mathbf{p}^t - \mathbf{p}^{t+1}\|^2 + \|\mathbf{p}^{t+1} - \mathbf{z}^*\|^2 + 2\langle \mathbf{p}^{t+1} - \mathbf{p}^t, \mathbf{z}^* - \mathbf{p}^{t+1}\rangle$,
and the inequality is based on Cauchy-Schwarz. Finally, note that since $\widetilde{V}(\mathbf{x}_t^m)$ are i.i.d. for each $m \in [M]$, we have

$$\mathbb{E}\left[\|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2\right] \leq \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\|V(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^m)\|^2\right] \leq \frac{\sigma^2}{M},$$

Combining the above inequalities, we get

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Gamma_t\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{2\sigma^2\eta}{M}.$$

D Proof of Theorem 4

The proof is similar to that of Theorem 3, and the main difference is that we bound the potential function differently to make it depend on $\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2$, for which we introduce a tight bound based on Assumption 1. We refer to Remark 3 for a detailed comparison and discussion. We start by recalling the following notations. Let $\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m$, and $\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m$, and

$$\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{x}_t^m) = \overline{\mathbf{z}}_{t-1} - \widetilde{V}_{\mathbf{x},t}.$$

where $\widetilde{V}_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{x}_t^m)$. Moreover, we have $V_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^{M}V(\mathbf{x}_t^m)$. Also recall the definition of the shadow proximal update:

$$\mathbf{x}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m}),$$

and let $\overline{\mathbf{x}}_t^* = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^{*,m}$, and $V_{\mathbf{x}^*,t} = \frac{1}{M}\sum_{m=1}^{M}V(\mathbf{x}_t^{*,m})$. We first introduce the following lemma about the properties of $\mathbf{x}_t^{*,m}$, apart from Lemma 14. It is the key for obtaining the improved results. The proof is given in Appendix D.1.

Lemma 17. *We have*

$$\mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2] \leq 6\Lambda^2\eta^6\sigma^4 + \frac{\eta^2\sigma^2}{M} + 0.25^H \cdot 4\eta^2G^2 + 0.25^H \cdot H^2\Lambda^2\eta^6G^4$$

We now begin the proof by first bounding the following potential function.

$$\begin{aligned}
&\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 = \left\|\overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{x},t} - \mathbf{z}\right\|^2 = \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle\widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 \\
=&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\langle V_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle \\
=&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\rangle + 2\eta\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle.
\end{aligned} \tag{41}$$

Next, we bound each term in (41) respectively. For the second term, we have

$$\begin{aligned}
&-2\eta\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\rangle = -2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{z}\rangle = -2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m + \mathbf{x}_t^m - \mathbf{z}\rangle \\
=&-2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle - 2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle \\
=&\underbrace{-2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle}_{=0} - 2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle + 2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle \\
\leq&2\eta L\frac{1}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 - 2\eta\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle,
\end{aligned} \tag{42}$$

where the inequality is based on the L-smoothness of V and the Cauchy-Schwarz inequality. For the third term in (41), we have

$$2\eta \langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1} \rangle = \underbrace{2\eta \langle V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1} \rangle}_{A_1} + \underbrace{2\eta \langle V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_{t-1} \rangle}_{A_2} + \underbrace{2\eta \langle V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^* \rangle}_{A_3}. \tag{43}$$

For A_1, we have

$$\|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\| \leq \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + \|\overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_{t-1}\| = \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + \eta \|V_{\mathbf{x}^*,t}\| \tag{44}$$

therefore

$$\begin{aligned} A_1 \leq & 2\eta \|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\| \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + 2\eta^2 \|V_{\mathbf{x}^*,t}\| \|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\| \\ \leq & \eta^2 \|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2 + \frac{\eta^2}{4} \|V_{\mathbf{x}^*,t}\|^2 + 4\eta^2 \|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|^2 \\ \leq & \frac{5\eta^2 L^2}{M} \sum_{m=1}^{M} \|\mathbf{x}_t^m - \mathbf{x}_t^*\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^{*,m}\|^2 + \frac{\eta^2}{4} \|V_{\mathbf{x}^*,t}\|^2 \end{aligned} \tag{45}$$

We also have

$$A_2 = -2\eta^2 \|V_{\mathbf{x}^*,t}\|^2,$$

and

$$A_3 \leq \frac{\eta^2}{4} \|V_{\mathbf{x}^*,t}\|^2 + 8\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2$$

Combining plugging into (43), we have

$$2\eta \langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1} \rangle \leq 9\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2 + \frac{5\eta^2 L^2}{M} \sum_{m=1}^{M} \|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 - 1.5\eta^2 \|V_{\mathbf{x}^*,t}\|^2$$

Finally, for the fourth term of (41), we have

$$\begin{aligned} & \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 = \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{x}}_t^* + \overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_t\|^2 \leq 1.5\|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{x}}_t^*\|^2 + 3\|\overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_t\|^2 \\ = & 1.5\eta^2 \|V_{\mathbf{x}^*,t}\|^2 + 3\eta^2 \|V_{\mathbf{x}^*,t} - \widetilde{V}_{\mathbf{x},t}\|^2 \\ \leq & 1.5\eta^2 \|V_{\mathbf{x}^*,t}\|^2 + \frac{6\eta^2 L}{M} \sum_{m=1}^{M} \|\mathbf{x}_t^{*,m} - \mathbf{x}_t^m\|^2 + 6\eta^2 \|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2 \end{aligned} \tag{46}$$

Therefore, combining all the above inequalities, for (41) the upper bound becomes

$$\begin{aligned} & \|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 \leq \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 + 2\eta L \frac{1}{M} \sum_{m=1}^{M} \|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 - 2\eta \frac{1}{M} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z} \rangle \\ & + 20\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2 + \frac{10\eta^2 L^2}{M} \sum_{m=1}^{M} \|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + 4\eta^2 \|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2 + 2\eta \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z} \right\rangle \end{aligned} \tag{47}$$

Rearrange, and taking expectation on both sides, we get

$$\begin{aligned} & \mathbb{E}\left[\frac{1}{MT} \sum_{t=1}^{T} \sum_{m=1}^{M} \langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z} \rangle \right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}\|^2]}{2\eta T} + \frac{L}{MT} \sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{E}\left[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 \right] \\ & + \frac{5\eta L^2}{MT} \sum_{t=1}^{T} \sum_{m=1}^{M} \mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 \right] + \frac{10}{\eta T} \sum_{t=1}^{T} \mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2] + \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \Gamma_t \right], \end{aligned} \tag{48}$$

where $\Gamma_t = 2\eta \|V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}\|^2 + \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z} \right\rangle$.

Remark 3. Compared with (32), we could observe two differences: Firstly, the term

$$\frac{15}{\eta MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\left[\|\mathbf{x}_t^m-\mathbf{x}_t^{*,m}\|^2\right]$$

in (32) is improved to

$$\frac{5\eta L^2}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\left[\|\mathbf{x}_t^m-\mathbf{x}_t^{*,m}\|^2\right],$$

and the extra η^2 factor making this term much smaller. On the other hand, there is an extra $\frac{10}{\eta T}\sum_{t=1}^{T}\mathbb{E}[\|\overline{\mathbf{x}}_t-\overline{\mathbf{x}}_t^*\|^2]$ term, which we show is small by using Lemma 17.

Plugging Lemmas 14, 15, 16, and 17 into (48), and using the fact that $\eta\le 1/L$, we get

$$\begin{aligned}\mathbb{E}\left[\left\langle V(\mathbf{z}^*),\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m-\mathbf{z}^*\right\rangle\right]&\le\frac{10H^2\Lambda^2G^4+60G^2+120G^2K^2}{4^H}\\&+120\left[\frac{\mathbb{E}[\|\mathbf{z}_0-\mathbf{z}^*\|^2]}{\eta KR}+L^2K^2\sigma^2\eta^4+K\sigma^2\eta^2L+\eta^3\sigma^2L^2+\Lambda^2\sigma^4\eta^5+\frac{\eta\sigma^2}{M}\right].\end{aligned}\tag{49}$$

The proof is finished by setting

$$\eta=\min\left\{\frac{1}{L},\frac{D^{\frac{2}{5}}}{K^{\frac{3}{5}}R^{\frac{1}{5}}\sigma^{\frac{2}{5}}L^{\frac{2}{5}}},\frac{D^{\frac{2}{3}}}{K^{\frac{2}{3}}R^{\frac{1}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}},\frac{D^{\frac{1}{2}}}{K^{\frac{1}{4}}R^{\frac{1}{4}}\sigma^{\frac{1}{2}}L^{\frac{1}{2}}},\frac{D^{\frac{1}{3}}}{\Lambda^{\frac{1}{3}}\sigma^{\frac{2}{3}}R^{\frac{1}{6}}K^{\frac{1}{6}}},\frac{D\sqrt{M}}{K^{\frac{1}{2}}R^{\frac{1}{2}}\sigma}\right\}.$$

D.1 Proof of Lemma 17

Note that conditioned on $\mathbf{x}_{t,0}^m$ for $m=1,\dots,M$, we have

$$\begin{aligned}\mathbb{E}[\|\overline{\mathbf{x}}_t-\overline{\mathbf{x}}_t^*\|^2]&=\|\mathbb{E}[\overline{\mathbf{x}}_t]-\overline{\mathbf{x}}_t^*\|^2+\mathrm{Var}(\overline{\mathbf{x}}_t-\overline{\mathbf{x}}_t^*)=\|\mathbb{E}[\overline{\mathbf{x}}_t]-\overline{\mathbf{x}}_t^*\|^2+\frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Var}(\mathbf{x}_t^m-\mathbf{x}_t^{*,m})\\&\le\frac{1}{M}\sum_{m=1}^{M}\|\mathbb{E}[\mathbf{x}_t^m]-\mathbf{x}_t^{*,m}\|^2+\frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}[\|\mathbf{x}_t^m-\mathbf{x}_t^{*,m}\|^2].\end{aligned}$$

Here, Var denotes the variance of the random variable. Next, we bound each term respectively. For the first term, recall $F(\mathbf{x})=V(\mathbf{x})+\frac{1}{\eta}(\mathbf{x}-\mathbf{z}_{t-1}^m)$, $F(\mathbf{x}_t^{*,m})=0$. Let $G(\mathbf{x})=\mathbf{x}-\gamma F(\mathbf{x})$, we have $G(\mathbf{x}_t^{*,m})=\mathbf{x}_t^{*,m}$, i.e., $\mathbf{x}_t^{*,m}$ is a fixed point of G, and let

$$\widetilde{G}(\mathbf{x})=\mathbf{x}-\gamma\widetilde{F}(\mathbf{x})=G(\mathbf{x})+\gamma(V(\mathbf{x})-\widetilde{V}(\mathbf{x})).$$

Then, based on the update rule in Algorithm 2, we have $\mathbf{x}_{t,\ell+1}^m=\widetilde{G}(\mathbf{x}_{t,\ell}^m)$, so

$$\mathbb{E}[\mathbf{x}_{t,\ell+1}^m]=\mathbb{E}[\widetilde{G}(\mathbf{x}_{t,\ell}^m)]=\mathbb{E}[G(\mathbf{x}_{t,\ell}^m)].$$

Moreover, we can expand G as

$$G(\mathbf{x})\;=\;\mathbf{x}-\gamma\Big(V(\mathbf{x})+\tfrac{1}{\eta}(\mathbf{x}-\mathbf{z}_{t-1}^m)\Big)=\Big(1-\tfrac{\gamma}{\eta}\Big)\mathbf{x}+\tfrac{\gamma}{\eta}\mathbf{z}_{t-1}^m-\gamma V(\mathbf{x}).$$

Therefore, its Jacobian is

$$J_G(\mathbf{y})\;=\;\Big(1-\tfrac{\gamma}{\eta}\Big)I-\gamma J_V(\mathbf{y}).$$

Thus

$$\begin{aligned}&G(\mathbf{x})-G(\mathbf{y})-J_G(\mathbf{y})(\mathbf{x}-\mathbf{y})=\;\Big[\Big(1-\tfrac{\gamma}{\eta}\Big)\mathbf{x}+\tfrac{\gamma}{\eta}\mathbf{z}_{t-1}^m-\gamma V(\mathbf{x})\Big]-\Big[\Big(1-\tfrac{\gamma}{\eta}\Big)\mathbf{y}+\tfrac{\gamma}{\eta}\mathbf{z}_{t-1}^m-\gamma V(\mathbf{y})\Big]\\&-\Big[\Big(1-\tfrac{\gamma}{\eta}\Big)I-\gamma J_V(\mathbf{y})\Big](\mathbf{x}-\mathbf{y})=\;-\gamma\Big(V(\mathbf{x})-V(\mathbf{y})-J_V(\mathbf{y})(\mathbf{x}-\mathbf{y})\Big).\end{aligned}$$

Taking the norm and applying Assumption 1, we obtain

$$\|G(\mathbf{x})-G(\mathbf{y})-J_G(\mathbf{y})(\mathbf{x}-\mathbf{y})\|\;\le\gamma\Lambda\|\mathbf{x}-\mathbf{y}\|^2.$$

Let

$$R(\mathbf{x}_{t,\ell}^m) = G(\mathbf{x}_{t,\ell}^m) - G(\mathbb{E}[\mathbf{x}_{t,\ell}^m]) - \left\langle J_G(\mathbb{E}[\mathbf{x}_{t,\ell}]), \mathbf{x}_{t,\ell}^m - \mathbb{E}[\mathbf{x}_{t,\ell}^m] \right\rangle,$$

we know that $\mathbb{E}[R(\mathbf{x}_{t,\ell}^m)] = \mathbb{E}[G(\mathbf{x}_{t,\ell}^m)] - G(\mathbb{E}[\mathbf{x}_{t,\ell}^m])$. Thus, combining with Assumption 1, we have

$$\|\mathbb{E}[G(\mathbf{x}_{t,\ell}^m)] - G(\mathbb{E}[\mathbf{x}_{t,\ell}^m])\| = \|\mathbb{E}[R(\mathbf{x}_{t,\ell}^m)]\| \leq \mathbb{E}\|R(\mathbf{x}_{t,\ell}^m)\| \leq \lambda\Lambda\mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbb{E}[\mathbf{x}_{t,\ell}^m]\|^2] = \lambda\Lambda\text{Var}(\mathbf{x}_{t,\ell}^m).$$

To proceed, we have

$$\begin{aligned}
&\|\mathbb{E}[\mathbf{x}_{t,\ell+1}^m] - \mathbf{x}_t^{*,m}\| = \|\mathbb{E}[G(\mathbf{x}_{t,\ell}^m)] - \mathbf{x}_t^{*,m}\| \leq \|G(\mathbb{E}[\mathbf{x}_{t,\ell}^m]) - \mathbf{x}_t^{*,m}\| + \|\mathbb{E}[G(\mathbf{x}_{t,\ell}^m)] - G(\mathbb{E}[\mathbf{x}_{t,\ell}^m])\| \\
\leq &\|G(\mathbb{E}[\mathbf{x}_{t,\ell}^m]) - \mathbf{x}_t^{*,m}\| + \gamma\Lambda\text{Var}(\mathbf{x}_{t,\ell}^m) \leq 0.5\|\mathbb{E}[\mathbf{x}_{t,\ell}^m] - \mathbf{x}_t^{*,m}\| + \gamma\Lambda\text{Var}(\mathbf{x}_{t,\ell}^m)
\end{aligned} \tag{50}$$

Here, the last inequality is based on the fact that G is a contraction map, and $\mathbf{x}_t^{*,m}$ is the fixed point. More specifically, for every $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

$$\begin{aligned}
&\|G(\mathbf{x}) - G(\mathbf{x}')\|^2 = \|\mathbf{x} - \mathbf{x}' - \gamma\left(F(\mathbf{x}) - F(\mathbf{x}')\right)\|^2 \\
=&\|\mathbf{x} - \mathbf{x}'\|^2 - 2\gamma\left\langle \mathbf{x} - \mathbf{x}', F(\mathbf{x}) - F(\mathbf{x}')\right\rangle + \gamma^2\|F(\mathbf{x}) - F(\mathbf{x}')\|^2 \\
\leq &\left(1 - \frac{2\gamma}{\eta} + \gamma^2\left(L + \frac{1}{\eta}\right)^2\right)\|\mathbf{x} - \mathbf{x}'\|^2 = \left(1 - \frac{1}{(\eta L + 1)^2}\right)\|\mathbf{x} - \mathbf{x}'\|^2 \leq 0.25\|\mathbf{x} - \mathbf{x}'\|^2.
\end{aligned}$$

Next, we would like to bound $\text{Var}(\mathbf{x}_{t,\ell}^m)$. We have

$$\begin{aligned}
&\text{Var}(\mathbf{x}_{t,\ell}^m) = \mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbf{x}_t^{*,m}\|^2] - \|\mathbb{E}[\mathbf{x}_{t,\ell}^m] - \mathbf{x}_t^{*,m}\|^2 \leq \mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbf{x}_t^{*,m}\|^2] \\
\leq &0.25^\ell\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m}\|^2] + \eta^2\sigma^2 \leq 0.25^\ell\eta^2G^2 + \eta^2\sigma^2.
\end{aligned}$$

So, based on (50), and the conclusions above, we have

$$\begin{aligned}
&\|\mathbb{E}[\mathbf{x}_{t,H}^m] - \mathbf{x}_t^{*,m}\| \leq 0.5\|\mathbb{E}[\mathbf{x}_{t,H-1}^m] - \mathbf{x}_t^{*,m}\| + \gamma\Lambda 0.25^H\eta^2G^2 + \gamma\Lambda\eta^2\sigma^2 \\
\leq &0.5^2\|\mathbb{E}[\mathbf{x}_{t,H-2}^m] - \mathbf{x}_t^{*,m}\| + 0.5\cdot\gamma\Lambda 0.25^{H-1}\eta^2G^2 + 0.5\cdot\gamma\Lambda\eta^2\sigma^2 + \gamma\Lambda 0.25^H\eta^2G^2 + \gamma\Lambda\eta^2\sigma^2 \\
\leq &0.5^H\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m}\|] + H0.5^H\gamma\Lambda\eta^2G^2 + 2\gamma\Lambda\eta^2\sigma^2 \\
\leq &0.5^H\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{x}_t^{*,m}\|] + H0.5^H\Lambda\eta^3G^2 + 2\Lambda\eta^3\sigma^2 \\
\leq &0.5^H\eta G + H0.5^H\Lambda\eta^3G^2 + 2\Lambda\eta^3\sigma^2.
\end{aligned}$$

Thus,

$$\|\mathbb{E}[\mathbf{x}_{t,H}^m] - \mathbf{x}_t^{*,m}\|^2 \leq 0.25^H\cdot 3\eta^2G^2 + H^20.25^H\Lambda^2\eta^6G^4 + 6\Lambda^2\eta^6\sigma^4.$$

Finally, we also have

$$\frac{1}{M^2}\sum_{m=1}^M\mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2\right] \leq 0.25^H\eta^2G^2 + \frac{\eta^2\sigma^2}{M}.$$

Combining all conclusions, we get

$$\mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2] \leq 6\Lambda^2\eta^6\sigma^4 + \frac{\eta^2\sigma^2}{M} + 0.25^H\cdot 4\eta^2G^2 + 0.25^H\cdot H^2\Lambda^2\eta^6G^4.$$

E Proof of Theorem 5

We start by recalling the following notations. Let $\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^M\mathbf{x}_t^m$, and $\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^M\mathbf{z}_t^m$, and

$$\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^M\mathbf{z}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^M\widetilde{V}(\mathbf{x}_t^m) = \overline{\mathbf{z}}_{t-1} - \widetilde{V}_{\mathbf{x},t}.$$

where we define $\widetilde{V}_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^M\widetilde{V}(\mathbf{x}_t^m)$. and let $V_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^M V(\mathbf{x}_t^m)$. Also recall the definition of the shadow proximal update:

$$\mathbf{x}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m}),$$

and let $\overline{\mathbf{x}}_t^* = \frac{1}{M}\sum_{m=1}^M\mathbf{x}_t^{*,m}$, and $V_{\mathbf{x}^*,t} = \frac{1}{M}\sum_{m=1}^M V(\mathbf{x}_t^{*,m})$. We first introduce the following lemma. The proof is given in Appendix E.1.

Lemma 18. *Let* $\mathbf{x}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta V(\mathbf{x}_t^{*,m})$, $\gamma = \frac{1}{\eta\left(L+\frac{1}{\eta}\right)^2}$, *and* $\eta \le \frac{1}{L}$. *Then we have*

$$\mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2] \le 30\left[\frac{\eta^2 G^2 + \frac{H^2 L^2 d\eta^6 G^4}{\delta^2}}{4^H} + \frac{\eta^2\sigma^2}{M} + \delta^2\eta^2 L^2 d + \frac{L^2\eta^6\sigma^4}{\delta^2}\right].$$

Moreover, and

$$\mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbf{x}_t^{*,m}\|^2] \le 4\left[\frac{\eta^2 G^2}{4^H} + \eta^2\sigma^2 + \eta^2 L^2\delta^2 d\right].$$

Comparing with Lemma 17, we can observe that, Gaussian-smoothing introduces extra terms related to δ. We now begin the proof by first bounding the following potential function.

$$\begin{aligned}
&\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 = \left\|\overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{x},t} - \mathbf{z}\right\|^2 = \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle\widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 \\
=&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle \\
=&\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - 2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle + 2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle + \|\overline{\mathbf{z}}_{t-1} - \overline{\mathbf{z}}_t\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{z}}_{t-1} - \mathbf{z}\right\rangle.
\end{aligned} \tag{51}$$

Next, we bound each term in respectively. For the second term, we have

$$\begin{aligned}
&-2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle = -2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle = -2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m + \mathbf{x}_t^m - \mathbf{z}\right\rangle \\
=&-2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\right\rangle - 2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\right\rangle \\
=&\underbrace{-2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\right\rangle}_{=0} - 2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\right\rangle + 2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\right\rangle \\
\le&2\eta L\frac{1}{M}\sum_{m=1}^M\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 - 2\eta\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\right\rangle,
\end{aligned}$$

where the inequality is based on the L-smoothness of V and the Cauchy-Schwarz inequality. For the third term, we have

$$2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle = \underbrace{2\eta\left\langle V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle}_{A_1} + \underbrace{2\eta\left\langle V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_{t-1}\right\rangle}_{A_2} + \underbrace{2\eta\left\langle V_{\mathbf{x}^*,t}, \overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\right\rangle}_{A_3}.$$

For A_1, we have

$$\|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\| \le \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + \|\overline{\mathbf{x}}_t^* - \overline{\mathbf{z}}_{t-1}\| = \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + \eta\|V_{\mathbf{x}^*,t}\|$$

therefore

$$\begin{aligned}
A_1 \le&2\eta\|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\| + 2\eta^2\|V_{\mathbf{x}^*,t}\|\|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\| \\
\le&\eta^2\|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2 + \frac{\eta^2}{4}\|V_{\mathbf{x}^*,t}\|^2 + 4\eta^2\|V_{\mathbf{x},t} - V_{\mathbf{x}^*,t}\|^2 \\
\le&\frac{5\eta^2 L^2}{M}\sum_{m=1}^M\|\mathbf{x}_t^m - \mathbf{x}_t^*\|^2 + \|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^{*,m}\|^2 + \frac{\eta^2}{4}\|V_{\mathbf{x}^*,t}\|^2
\end{aligned}$$

We also have $A_2 = -2\eta^2\|V_{\mathbf{x}^*,t}\|^2$, and $A_3 \le \frac{\eta^2}{4}\|V_{\mathbf{x}^*,t}\|^2 + 8\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2$. Thus

$$2\eta\left\langle V_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\right\rangle \le 9\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2 + \frac{5\eta^2 L^2}{M}\sum_{m=1}^M\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 - 1.5\eta^2\|V_{\mathbf{x}^*,t}\|^2$$

Finally, we have

$$\begin{aligned}&\|\overline{\mathbf{z}}_{t-1}-\overline{\mathbf{z}}_t\|^2=\|\overline{\mathbf{z}}_{t-1}-\overline{\mathbf{x}}_t^*+\overline{\mathbf{x}}_t^*-\overline{\mathbf{z}}_t\|^2\leq 1.5\|\overline{\mathbf{z}}_{t-1}-\overline{\mathbf{x}}_t^*\|^2+3\|\overline{\mathbf{x}}_t^*-\overline{\mathbf{z}}_t\|^2\\ =&1.5\eta^2\|V_{\mathbf{x}^*,t}\|^2+3\eta^2\|V_{\mathbf{x}^*,t}-\widetilde{V}_{\mathbf{x},t}\|^2\\ \leq&1.5\eta^2\|V_{\mathbf{x}^*,t}\|^2+\frac{6\eta^2L}{M}\sum_{m=1}^M\|\mathbf{x}_t^{*,m}-\mathbf{x}_t^m\|^2+6\eta^2\|V_{\mathbf{x},t}-\widetilde{V}_{\mathbf{x},t}\|^2\end{aligned}$$

Therefore, combining all the above inequalities, we get

$$\begin{aligned}&\|\overline{\mathbf{z}}_t-\mathbf{z}\|^2\leq\|\overline{\mathbf{z}}_{t-1}-\mathbf{z}\|^2+2\eta L\frac{1}{M}\sum_{m=1}^M\|\mathbf{x}_t^m-\overline{\mathbf{x}}_t\|^2-2\eta\frac{1}{M}\sum_{m=1}^M\langle V(\mathbf{x}_t^m),\mathbf{x}_t^m-\mathbf{z}\rangle\\ &+20\|\overline{\mathbf{x}}_t-\overline{\mathbf{x}}_t^*\|^2+\frac{10\eta^2L^2}{M}\sum_{m=1}^M\|\mathbf{x}_t^m-\mathbf{x}_t^{*,m}\|^2+4\eta^2\|V_{\mathbf{x},t}-\widetilde{V}_{\mathbf{x},t}\|^2+2\eta\left\langle V_{\mathbf{x},t}-\widetilde{V}_{\mathbf{x},t},\overline{\mathbf{z}}_{t-1}-\mathbf{z}\right\rangle\end{aligned}$$

Rearrange, and taking expectation on both sides, we get

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{MT}\sum_{t=1}^T\sum_{m=1}^M\langle V(\mathbf{x}_t^m),\mathbf{x}_t^m-\mathbf{z}\rangle\right]\leq\frac{\mathbb{E}[\|\overline{\mathbf{z}}_0-\mathbf{z}\|^2]}{2\eta T}+\frac{L}{MT}\sum_{t=1}^T\sum_{m=1}^M\mathbb{E}\left[\|\mathbf{x}_t^m-\overline{\mathbf{x}}_t\|^2\right]\\ &+\frac{5\eta L^2}{MT}\sum_{t=1}^T\sum_{m=1}^M\mathbb{E}\left[\|\mathbf{x}_t^m-\mathbf{x}_t^{*,m}\|^2\right]+\frac{10}{\eta T}\sum_{t=1}^T\mathbb{E}[\|\overline{\mathbf{x}}_t-\overline{\mathbf{x}}_t^*\|^2]+\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T\Gamma_t\right],\end{aligned}$$

where $\Gamma_t=2\eta\|V_{\mathbf{x},t}-\widetilde{V}_{\mathbf{x},t}\|^2+\left\langle V_{\mathbf{x},t}-\widetilde{V}_{\mathbf{x},t},\overline{\mathbf{z}}_{t-1}-\mathbf{z}\right\rangle$. Next, we have the following lemma. The proof is given in Appendix E.2.

Lemma 19. *We have*

$$\mathbb{E}[\|\mathbf{x}_t^m-\overline{\mathbf{x}}_t\|^2]\leq 480\left[\frac{\eta^4K^2LG^2+\eta^2G^2}{4^H}+\eta^4K^2L\sigma^2+\eta^4K^2L^3\delta^2d+\eta^2K\sigma^2+\eta^2L^2\delta^2d\right]$$

we get

$$\begin{aligned}&\mathbb{E}\left[\left\langle V(\mathbf{z}^*),\frac{1}{TM}\sum_{t=1}^T\sum_{m=1}^M\mathbf{x}_t^m-\mathbf{z}^*\right\rangle\right]\leq C\left[\frac{\eta G^2+\eta^4K^2L^2G^2+\frac{H^2L^2d\eta^5G^4}{\delta^2}}{4^H}\right.\\ &\left.+\delta^2(\eta L^2d+\eta^4K^2L^4d)+\frac{L^2\eta^5\sigma^4}{\delta^2}+\eta^4K^2L^2\sigma^2+\eta^2LK\sigma^2+\eta^3L^2\sigma^2+\frac{\eta\sigma^2}{M}\right],\end{aligned}\tag{52}$$

where $C\leq 1440$ is a constant. Let $\delta=\frac{\eta\sigma}{d^{\frac{1}{4}}}$, we get

$$\begin{aligned}&\mathbb{E}\left[\left\langle V(\mathbf{z}^*),\frac{1}{TM}\sum_{t=1}^T\sum_{m=1}^M\mathbf{x}_t^m-\mathbf{z}^*\right\rangle\right]\leq 2C\left[\frac{\eta G^2+\eta^4K^2L^2G^2+\frac{H^2L^2d\eta^5G^4}{\delta^2}}{4^H}\right.\\ &\left.+\eta^3L^2\sigma^2\sqrt{d}+\eta^4\sigma^2K^2L^2\sqrt{d}+\eta^2L^2K\sigma^2+\frac{\eta\sigma^2}{M}+\frac{D^2}{\eta T}\right].\end{aligned}\tag{53}$$

Setting

$$\eta=\min\left\{\frac{D^{\frac{1}{2}}}{K^{\frac{1}{4}}R^{\frac{1}{4}}L^{\frac{1}{2}}\sigma^{\frac{1}{2}}d^{\frac{1}{8}}},\frac{D^{\frac{2}{5}}}{\sigma^{\frac{2}{5}}L^{\frac{2}{5}}d^{\frac{1}{10}}K^{\frac{3}{5}}R^{\frac{1}{5}}},\frac{D^{\frac{2}{3}}}{K^{\frac{2}{3}}R^{\frac{1}{3}}\sigma^{\frac{2}{3}}L^{\frac{2}{3}}},\frac{1}{L}\right\},$$

we get

$$\begin{aligned}&\mathbb{E}\left[\sup_{\mathbf{z}\in\mathcal{Z}_D}\langle V(\mathbf{z}),\mathbf{x}_o-\mathbf{z}\rangle\right]\\ &\leq O\left(\frac{G^2+K^2L^2G^2+\frac{H^2L^2d^{1/2}G^4}{\sigma^2}}{4^H}+\frac{D^{\frac{3}{2}}L^{\frac{1}{2}}\sigma^{\frac{1}{2}}d^{\frac{1}{8}}}{K^{\frac{3}{4}}R^{\frac{3}{4}}}+\frac{D^{\frac{8}{5}}\sigma^{\frac{2}{5}}L^{\frac{2}{5}}d^{\frac{1}{10}}}{R^{\frac{4}{5}}K^{\frac{2}{5}}}+\frac{D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}+\frac{D\sigma}{\sqrt{MKR}}+\frac{D^2L}{KR}\right).\end{aligned}$$

E.1 Proof of Lemma 18

Define $\widehat{V}(\mathbf{x}) = V(\mathbf{x}+\delta\mathbf{s};\xi)$, and

$$\mathring{V}(\mathbf{x}) = \mathbb{E}_{\mathbf{s},\xi}[V(\mathbf{x}+\delta\mathbf{s};\xi)] = \mathbb{E}[\widehat{V}(\mathbf{x})].$$

It is easy to verify that $\mathring{V}(\mathbf{x})$ is also monotone, since for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$,

$$\left\langle \mathring{V}(\mathbf{x}) - \mathring{V}(\mathbf{y}), \mathbf{x}-\mathbf{y}\right\rangle = \mathbb{E}\left[\langle V(\mathbf{x}+\delta\mathbf{s};\xi) - V(\mathbf{y}+\delta\mathbf{s};\xi), (\mathbf{y}+\delta\mathbf{s}) - (\mathbf{y}+\delta\mathbf{s})\rangle\right] \geq 0.$$

Moreover, $\mathring{V}(\mathbf{x})$ is also L-smooth. Define $\mathring{F}(\mathbf{x}) = \mathring{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x}-\mathbf{z}_{t-1}^m)$. It can be seen that, the first part of Algorithm 3 is essentially optimizing the VI defined by $\mathring{F}$ with LSGD. Similar to previous proof, we can observe that $\mathring{F}(\mathbf{x})$ is $\frac{1}{\eta}$ strongly monotone, and $L+\frac{1}{\eta}$ smooth. Let

$$\mathring{\mathbf{x}}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta\mathring{V}(\mathring{\mathbf{x}}_t^{*,m}),$$

We know $\mathring{F}(\mathring{\mathbf{x}}_t^{*,m}) = 0$. Moreover, let $\overline{\mathring{\mathbf{x}}_t^*} = \frac{1}{M}\sum_{m=1}^M \mathring{\mathbf{x}}_t^{*,m}$. Let $\widehat{F}(\mathbf{x}) = \widehat{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x}-\mathbf{z}_{t-1}^m)$. Note that we have the variance

$$\mathbb{E}\|\widehat{F}(\mathbf{x}) - \mathring{F}(\mathbf{x})\|^2 \leq L^2\delta^2 d + \sigma^2.$$

In the following, we will firstly show that $\overline{\mathbf{x}}_t$ is close to $\overline{\mathring{\mathbf{x}}_t^*}$, and then prove that $\overline{\mathring{\mathbf{x}}_t^*}$ is close to $\overline{\mathbf{x}}_t^*$. Therefore, $\overline{\mathbf{x}}_t$ is also close to $\overline{\mathbf{x}}_t^*$. Note that conditioned on $\mathbf{x}_{t,0}^m$, for $m=1,\ldots,M$,

$$\mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathring{\mathbf{x}}_t^*}\|^2] = \|\mathbb{E}[\overline{\mathbf{x}}_t] - \overline{\mathring{\mathbf{x}}_t^*}\|^2 + \mathrm{Var}(\overline{\mathbf{x}}_t - \overline{\mathring{\mathbf{x}}_t^*}) = \|\mathbb{E}[\overline{\mathbf{x}}_t] - \overline{\mathring{\mathbf{x}}_t^*}\|^2 + \frac{1}{M^2}\sum_{m=1}^M \mathrm{Var}(\mathbf{x}_t^m - \mathring{\mathbf{x}}_t^{*,m})$$

$$\leq \frac{1}{M}\sum_{m=1}^M \|\mathbb{E}[\mathbf{x}_t^m] - \mathring{\mathbf{x}}_t^{*,m}\|^2 + \frac{1}{M^2}\sum_{m=1}^M \mathbb{E}[\|\mathbf{x}_t^m - \mathring{\mathbf{x}}_t^{*,m}\|^2].$$

Next, we bound each term respectively. For the first term, recall $\mathring{F}(\mathbf{x}) = \mathring{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x}-\mathbf{z}_{t-1}^m)$, $\mathring{F}(\mathring{\mathbf{x}}_t^{*,m}) = 0$. Let $\mathring{G}(\mathbf{x}) = \mathbf{x} - \gamma\mathring{F}(\mathbf{x})$, we have $\mathring{G}(\mathring{\mathbf{x}}_t^{*,m}) = \mathring{\mathbf{x}}_t^{*,m}$, so $\mathring{\mathbf{x}}_t^{*,m}$ is a fixed point of $\mathring{G}$, and let

$$\widehat{G}(\mathbf{x}) = \mathbf{x} - \gamma\widehat{F}(\mathbf{x}) = G(\mathbf{x}) + \gamma(\mathring{V}(\mathbf{x}) - \widehat{V}(\mathbf{x})).$$

Then, based on the update rule in Algorithm 3, we have $\mathbf{x}_{t,\ell+1}^m = \widehat{G}(\mathbf{x}_{t,\ell}^m)$, so

$$\mathbb{E}[\mathbf{x}_{t,\ell+1}^m] = \mathbb{E}[\widehat{G}(\mathbf{x}_{t,\ell}^m)] = \mathbb{E}[\mathring{G}(\mathbf{x}_{t,\ell}^m)].$$

Moreover, we can expand $\mathring{G}$ as

$$\mathring{G}(\mathbf{x}) \;=\; \mathbf{x} - \gamma\Big(\mathring{V}(\mathbf{x}) + \tfrac{1}{\eta}(\mathbf{x}-\mathbf{z}_{t-1}^m)\Big) = \Big(1-\tfrac{\gamma}{\eta}\Big)\mathbf{x} + \tfrac{\gamma}{\eta}\mathbf{z}_{t-1}^m - \gamma\mathring{V}(\mathbf{x}).$$

Therefore, its Jacobian is

$$J_{\mathring{G}}(\mathbf{x}) \;=\; \Big(1-\tfrac{\gamma}{\eta}\Big)I - \gamma J_{\mathring{V}}(\mathbf{x}).$$

Thus

$$\mathring{G}(\mathbf{x}) - \mathring{G}(\mathbf{y}) - J_{\mathring{G}}(\mathbf{y})(\mathbf{x}-\mathbf{y}) = \Big[\Big(1-\tfrac{\gamma}{\eta}\Big)\mathbf{x} + \tfrac{\gamma}{\eta}\mathbf{z}_{t-1}^m - \gamma\mathring{V}(\mathbf{x})\Big] - \Big[\Big(1-\tfrac{\gamma}{\eta}\Big)\mathbf{y} + \tfrac{\gamma}{\eta}\mathbf{z}_{t-1}^m - \gamma\mathring{V}(\mathbf{y})\Big]$$
$$- \Big[\Big(1-\tfrac{\gamma}{\eta}\Big)I - \gamma J_{\mathring{V}}(\mathbf{y})\Big](\mathbf{x}-\mathbf{y}) = -\gamma\Big(\mathring{V}(\mathbf{x}) - \mathring{V}(\mathbf{y}) - J_{\mathring{V}}(\mathbf{y})(\mathbf{x}-\mathbf{y})\Big).$$

To proceed, we provide the following lemma, which shows the second-order smoothness of $\mathring{V}$.

Lemma 20. *We have*

$$\|\mathring{V}(\mathbf{x}) - \mathring{V}(\mathbf{y}) - J_{\mathring{V}}(\mathbf{y})(\mathbf{x}-\mathbf{y})\| \;\leq\; \frac{L\sqrt{d}}{\delta}\|\mathbf{x}-\mathbf{y}\|^2.$$

Proof. Let $\mathring{V}(\mathbf{x}) = [\mathring{V}_1(\mathbf{x}), \ldots, \mathring{V}_d(\mathbf{x})]$. Then we have

$$\mathring{V}_i(\mathbf{x}) = \int_{\mathbb{R}^d} \mathbb{E}_\xi[V_i(\mathbf{x}+\delta\mathbf{s};\xi)] \cdot \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right) d\mathbf{s} = \frac{1}{\delta}\int_{\mathbb{R}^d} \mathbb{E}_\xi[V_i(\mathbf{z};\xi)] \cdot \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{z}-\mathbf{x}\|^2}{2\delta^2}\right) d\mathbf{z},$$

where we let $\mathbf{z} = \mathbf{x} + \delta\mathbf{s}$. Therefore,

$$\begin{aligned}\nabla\mathring{V}_i(\mathbf{x}) =& \frac{1}{\delta}\int_{\mathbb{R}^d} \mathbb{E}_\xi[V_i(\mathbf{z};\xi)] \frac{\mathbf{z}-\mathbf{x}}{(2\pi)^{d/2}\delta^2} \exp\left(-\frac{\|\mathbf{z}-\mathbf{x}\|^2}{2\delta^2}\right) d\mathbf{z} \\ =& \frac{1}{\delta}\int_{\mathbb{R}^d} \mathbb{E}_\xi[V_i(\mathbf{x}+\delta\mathbf{s};\xi)] \frac{\mathbf{s}}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{s}\|^2}{2}\right) d\mathbf{s}.\end{aligned}$$

Thus, the Jacobian of $\mathring{V}$ can be written in the following form:

$$J_{\mathring{V}}(\mathbf{x}) = \frac{1}{\delta}\mathbb{E}[V(\mathbf{x}+\delta\mathbf{s};\xi)\mathbf{s}^\top] = \frac{1}{\delta}\mathbb{E}[(V(\mathbf{x}+\delta\mathbf{s};\xi) - V(\mathbf{x}))\mathbf{s}^\top],$$

where the second equality is based on the fact that $\mathbb{E}[\mathbf{s}] = \mathbf{0}$. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, We have

$$\begin{aligned}\|J_{\mathring{V}}(\mathbf{x}) - J_{\mathring{V}}(\mathbf{y})\| &= \frac{1}{\delta}\left\|\mathbb{E}\left[(V(\mathbf{x}+\delta\mathbf{s};\xi) - V(\mathbf{y}+\delta\mathbf{s};\xi) - V(\mathbf{x};\xi) + V(\mathbf{y};\xi))\mathbf{s}^\top\right]\right\| \\ &\le \frac{1}{\delta}\mathbb{E}\left[\|V(\mathbf{x}+\delta\mathbf{s};\xi) - V(\mathbf{y}+\delta\mathbf{s};\xi)\|\|\mathbf{s}\|\right] + \frac{1}{\delta}\mathbb{E}\left[\|V(\mathbf{x};\xi) - V(\mathbf{y};\xi)\|\|\mathbf{s}\|\right] \\ &\le \frac{1}{\delta}\mathbb{E}[L\|\mathbf{x}-\mathbf{y}\|\|\mathbf{s}\|] + \frac{1}{\delta}\mathbb{E}[L\|\mathbf{x}-\mathbf{y}\|\|\mathbf{s}\|] = \frac{2L}{\delta}\,\mathbb{E}\|\mathbf{s}\|\,\|\mathbf{x}-\mathbf{y}\| \le \frac{2L\sqrt{d}}{\delta}\|\mathbf{x}-\mathbf{y}\|,\end{aligned}$$

where we consider the spectral norm for matrices. Therefore, $J_{\mathring{V}}$ is $\frac{2L\sqrt{d}}{\delta}$-Lipschitz, which finishes the proof. □

Taking the norm and applying the lemma above, we obtain

$$\|\mathring{G}(\mathbf{x}) - \mathring{G}(\mathbf{y}) - J_{\mathring{G}}(\mathbf{y})(\mathbf{x}-\mathbf{y})\| \;\le\; \frac{\gamma L\sqrt{d}}{\delta}\|\mathbf{x}-\mathbf{y}\|^2.$$

Let

$$\mathring{R}(\mathbf{x}_{t,\ell}^m) = \mathring{G}(\mathbf{x}_{t,\ell}^m) - \mathring{G}(\mathbb{E}[\mathbf{x}_{t,\ell}^m]) - \left\langle J_{\mathring{G}}(\mathbb{E}[\mathbf{x}_{t,\ell}^m]), \mathbf{x}_{t,\ell}^m - \mathbb{E}[\mathbf{x}_{t,\ell}^m]\right\rangle,$$

we know that $\mathbb{E}[\mathring{R}(\mathbf{x}_{t,\ell}^m)] = \mathbb{E}[\mathring{G}(\mathbf{x}_{t,\ell}^m)] - \mathring{G}(\mathbb{E}[\mathbf{x}_{t,\ell}^m])$. Thus

$$\|\mathbb{E}[\mathring{G}(\mathbf{x}_{t,\ell}^m)] - \mathring{G}(\mathbb{E}[\mathbf{x}_{t,\ell}^m])\| = \|\mathbb{E}[\mathring{R}(\mathbf{x}_{t,\ell}^m)]\| \le \mathbb{E}\|\mathring{R}(\mathbf{x}_{t,\ell}^m)\| \le \frac{\gamma L\sqrt{d}}{\delta}\mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbb{E}[\mathbf{x}_{t,\ell}^m]\|^2] = \frac{\gamma L\sqrt{d}}{\delta}\mathrm{Var}(\mathbf{x}_{t,\ell}^m).$$

To proceed, we have

$$\begin{aligned}&\|\mathbb{E}[\mathbf{x}_{t,\ell+1}^m] - \mathring{\mathbf{x}}_t^{*,m}\| \le \|\mathring{G}(\mathbb{E}[\mathbf{x}_{t,\ell}^m]) - \mathring{\mathbf{x}}_t^{*,m}\| + \|\mathbb{E}[\mathring{G}(\mathbf{x}_{t,\ell}^m)] - \mathring{G}(\mathbb{E}[\mathbf{x}_{t,\ell}^m])\| \\ \le& \|\mathring{G}(\mathbb{E}[\mathbf{x}_{t,\ell}^m]) - \mathring{\mathbf{x}}_t^{*,m}\| + \frac{\gamma L\sqrt{d}}{\delta}\mathrm{Var}(\mathbf{x}_{t,\ell}^m) \le 0.5\|\mathbb{E}[\mathbf{x}_{t,\ell}^m] - \mathring{\mathbf{x}}_t^{*,m}\| + \frac{\gamma L\sqrt{d}}{\delta}\mathrm{Var}(\mathbf{x}_{t,\ell}^m)\end{aligned}$$

Here, the last inequality is based on the fact that $\mathring{G}$ is also a contraction map, more specifically, for every $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

$$\begin{aligned}&\|\mathring{G}(\mathbf{x}) - \mathring{G}(\mathbf{x}')\|^2 = \left\|\mathbf{x} - \mathbf{x}' - \gamma\left(\mathring{F}(\mathbf{x}) - \mathring{F}(\mathbf{x}')\right)\right\|^2 \\ =& \|\mathbf{x}-\mathbf{x}'\|^2 - 2\gamma\left\langle \mathbf{x}-\mathbf{x}', \mathring{F}(\mathbf{x}) - \mathring{F}(\mathbf{x}')\right\rangle + \gamma^2\|\mathring{F}(\mathbf{x}) - \mathring{F}(\mathbf{x}')\|^2 \\ \le& \left(1 - \frac{2\gamma}{\eta} + \gamma^2\left(L + \frac{1}{\eta}\right)^2\right)\|\mathbf{x}-\mathbf{x}'\|^2 = \left(1 - \frac{1}{(\eta L+1)^2}\right)\|\mathbf{x}-\mathbf{x}'\|^2 \le 0.25\|\mathbf{x}-\mathbf{x}'\|^2.\end{aligned}$$

Next, we would like to bound $\mathrm{Var}(\mathbf{x}_{t,\ell}^m)$. We have

$$\mathrm{Var}(\mathbf{x}_{t,\ell}^m) = \mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2] - \|\mathbb{E}[\mathbf{x}_{t,\ell}^m] - \mathring{\mathbf{x}}_t^{*,m}\|^2 \le \mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2] \le 0.25^\ell\eta^2G^2 + \eta^2(L^2\delta^2 d + \sigma^2).$$

where the last inequality is based on the following lemma. The proof is given in Appendix E.1.1.

Lemma 21. *We have*

$$\mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2] \leq 0.25^\ell \eta^2 G^2 + \eta^2(L^2\delta^2 d + \sigma^2).$$

Moreover, we have $\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\| \leq \eta G$*, and*

$$\mathbb{E}[\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2] \leq \eta^2 L^2 \delta^2 d,$$

and

$$\mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbf{x}_t^{*,m}\|^2] \leq \frac{2\eta^2 G^2}{4^\ell} + 2\eta^2\sigma^2 + 4\eta^2 L^2\delta^2 d.$$

Therefore, we have

$$\begin{aligned}
&\|\mathbb{E}[\mathbf{x}_{t,H}^m] - \mathring{\mathbf{x}}_t^{*,m}\| \leq 0.5\|\mathbb{E}[\mathbf{x}_{t,H-1}^m] - \mathring{\mathbf{x}}_t^{*,m}\| + \frac{0.25^H \gamma L\sqrt{d}\eta^2 G^2}{\delta} + \frac{\gamma L\sqrt{d}\eta^2(L^2\delta^2 d + \sigma^2)}{\delta} \\
\leq& 0.5^2\|\mathbb{E}[\mathbf{x}_{t,H-2}^m] - \mathring{\mathbf{x}}_t^{*,m}\| + 0.5\frac{0.25^{H-1}\gamma L\sqrt{d}\eta^2 G^2}{\delta} \\
&+ 0.5\frac{\gamma L\sqrt{d}\eta^2(L^2\delta^2 d + \sigma^2)}{\delta} + \frac{0.25^H \gamma L\sqrt{d}\eta^2 G^2}{\delta} + \frac{\gamma L\sqrt{d}\eta^2(L^2\delta^2 d + \sigma^2)}{\delta} \\
\leq& 0.5^H \mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|] + H0.5^H \frac{\gamma L\sqrt{d}\eta^2 G^2}{\delta} + \frac{2\gamma L\eta^2(L^2\delta^2 d + \sigma^2)}{\delta} \\
\leq& 0.5^H \mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|] + H0.5^H \frac{L\sqrt{d}\eta^3 G^2}{\delta} + \frac{2L\eta^3(L^2\delta^2 d + \sigma^2)}{\delta} \\
\leq& 0.5^H \eta G + H0.5^H \frac{L\sqrt{d}\eta^3 G^2}{\delta} + \frac{2L\eta^3(L^2\delta^2 d + \sigma^2)}{\delta}.
\end{aligned}$$

Thus,

$$\|\mathbb{E}[\mathbf{x}_{t,H}^m] - \mathbf{x}_t^{*,m}\|^2 \leq 0.25^H \cdot 3\eta^2 G^2 + 3H^2 0.25^H \frac{L^2 d\eta^6 G^4}{\delta^2} + \frac{6L^2\eta^6(L^2\delta^2 d + \sigma^2)^2}{\delta^2}$$

Finally, we also have

$$\frac{1}{M^2}\sum_{m=1}^M \mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2\right] \leq 0.25^H \eta^2 G^2 + \frac{\eta^2(L^2\delta^2 d + \sigma^2)}{M}.$$

Combining all conclusions, we get

$$\mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathring{\mathbf{x}}}_t^*\|^2] \leq 0.25^H \cdot 4\eta^2 G^2 + \frac{\eta^2(L^2\delta^2 d + \sigma^2)}{M} + 3H^2 0.25^H \frac{L^2 d\eta^6 G^4}{\delta^2} + \frac{6L^2\eta^6(L^2\delta^2 d + \sigma^2)^2}{\delta^2}.$$

Thus combining with Lemma 21,

$$\begin{aligned}
\mathbb{E}[\|\overline{\mathbf{x}}_t - \overline{\mathbf{x}}_t^*\|^2] \leq& 0.25^H \cdot 8\eta^2 G^2 + \frac{2\eta^2(L^2\delta^2 d + \sigma^2)}{M} + 6H^2 0.25^H \frac{L^2 d\eta^6 G^4}{\delta^2} + \frac{12L^2\eta^6(L^2\delta^2 d + \sigma^2)^2}{\delta^2} + 2\eta^2 L^2\delta^2 d \\
\leq& \frac{8\eta^2 G^2 + \frac{6H^2 L^2 d\eta^6 G^4}{\delta^2}}{4^H} + \frac{2\eta^2\sigma^2}{M} + 30\delta^2\eta^2 L^2 d + \frac{24L^2\eta^6\sigma^4}{\delta^2}
\end{aligned}$$

E.1.1 Proof of Lemma 21

Recall that

$$\mathring{V}(\mathbf{x}) = \mathbb{E}[V(\mathbf{x} + \delta\mathbf{s}; \xi)] = \mathbb{E}[\widehat{V}(\mathbf{x})].$$

Moreover, $\mathring{V}(\mathbf{x})$ is monotone, L-smooth. Recall $\mathring{F}(\mathbf{x}) = \mathring{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$, which is $\frac{1}{\eta}$ strongly monotone, and $L + \frac{1}{\eta}$ smooth. Let

$$\mathring{\mathbf{x}}_t^{*,m} = \mathbf{z}_{t-1}^m - \eta\mathring{V}(\mathring{\mathbf{x}}_t^{*,m}),$$

We know $\mathring{F}(\mathbf{x}_t^{*,m}) = 0$. Next, we show the convergence. Let $\widehat{F}(\mathbf{x}) = \widehat{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$. Note that we have the variance

$$\mathbb{E}\|\widehat{F}(\mathbf{x}) - \mathring{F}(\mathbf{x})\|^2 \leq L^2\delta^2 d + \sigma^2.$$

since $\gamma = \frac{1}{\eta(L+\frac{1}{\eta})^2}$, we have

$$\begin{aligned}
&\mathbb{E}\big[\|\mathbf{x}_{t,\ell}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2\big] = \mathbb{E}\big[\|\mathbf{x}_{t,\ell-1}^m - \gamma\widehat{F}(\mathbf{x}_{t,\ell-1}^m) - \mathring{\mathbf{x}}_t^{*,m} + \gamma\mathring{F}(\mathring{\mathbf{x}}_t^{*,m})\|^2\big] \\
&= \mathbb{E}\big[\|\mathbf{x}_{t,\ell-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2\big] - 2\gamma\,\mathbb{E}\Big[\Big\langle \mathring{F}(\mathbf{x}_{t,\ell-1}^m) - \mathring{F}(\mathring{\mathbf{x}}_t^{*,m}),\ \mathbf{x}_{t,\ell-1}^m - \mathring{\mathbf{x}}_t^{*,m}\Big\rangle\Big] \\
&\quad + \gamma^2\,\mathbb{E}\big[\|\widehat{F}(\mathbf{x}_{t,\ell-1}^m) - \mathring{F}(\mathring{\mathbf{x}}_t^{*,m})\|^2\big] \\
&\le \mathbb{E}\big[\|\mathbf{x}_{t,\ell-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2\big] - \frac{2\gamma}{\eta}\,\mathbb{E}\big[\|\mathbf{x}_{t,\ell-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2\big] + \gamma^2\,\mathbb{E}\big[\|\mathring{F}(\mathbf{x}_{t,\ell-1}^m) - \mathring{F}(\mathring{\mathbf{x}}_t^{*,m})\|^2\big] \\
&\quad + \gamma^2\,\mathbb{E}\big[\|\widehat{F}(\mathbf{x}_{t,\ell-1}^m) - \mathring{F}(\mathbf{x}_{t,\ell-1}^m)\|^2\big] \\
&\le \Big(1 - \tfrac{2\gamma}{\eta} + \gamma^2\Big(L + \tfrac{1}{\eta}\Big)^2\Big)\,\mathbb{E}\big[\|\mathbf{x}_{t,\ell-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2\big] + \gamma^2\big(L^2\delta^2 d + \sigma^2\big) \\
&= \left(1 - \frac{1}{(\eta L+1)^2}\right)\,\mathbb{E}\big[\|\mathbf{x}_{t,\ell-1}^m - \mathring{\mathbf{x}}_t^{*,m}\|^2\big] + \gamma^2\big(L^2\delta^2 d + \sigma^2\big).
\end{aligned} \tag{54}$$

Since $\mathring{F}(\mathbf{x}) = \mathring{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$ is $\frac{1}{\eta}$-strongly monotone, we have

$$\Big\langle \mathring{F}(\mathring{\mathbf{x}}_t^{*,m}) - \mathring{F}(\mathbf{z}_{t-1}^m), \mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\Big\rangle \ge \frac{1}{\eta}\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\|^2.$$

Also, we have $\mathring{F}(\mathbf{x}_t^{*,m}) = 0$, and $\mathring{F}(\mathbf{z}_{t-1}^m) = V(\mathbf{z}_{t-1}^m)$, so

$$G\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\| \ge \Big\langle \mathring{F}(\mathring{\mathbf{x}}_t^{*,m}) - \mathring{F}(\mathbf{z}_{t-1}^m), \mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\Big\rangle \ge \frac{1}{\eta}\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\|^2,$$

so $\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{z}_{t-1}^m\| \le \eta G$. Finally, note that

$$\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2] \le 2\mathbb{E}[\|\mathbf{x}_t^m - \mathring{\mathbf{x}}_t^{*,m}\|^2] + 2\mathbb{E}[\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2].$$

In the following, we show that $\mathring{\mathbf{x}}_t^{*,m}$ and $\mathbf{x}_t^{*,m}$ are close. Since $\mathring{F}(\mathbf{x}) = \mathring{V}(\mathbf{x}) + \frac{1}{\eta}(\mathbf{x} - \mathbf{z}_{t-1}^m)$ are $\frac{1}{\eta}$-strongly monotone, we have

$$\begin{aligned}
\frac{1}{\eta}\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2 &\le \langle \mathring{F}(\mathring{\mathbf{x}}_t^{*,m}) - \mathring{F}(\mathbf{x}_t^{*,m}),\ \mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\rangle = -\langle \mathring{F}(\mathbf{x}_t^{*,m}),\ \mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\rangle \\
&\le \|\mathring{F}(\mathbf{x}_t^{*,m})\|\,\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\| \le \frac{\eta}{2}\|\mathring{F}(\mathbf{x}_t^{*,m})\|^2 + \frac{1}{2\eta}\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2
\end{aligned}$$

Therefore, $\frac{1}{\eta}\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2 \le \eta\|\mathring{F}(\mathbf{x}_t^{*,m})\|^2$, so

$$\frac{1}{\eta}\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2 \le \eta\|\mathring{F}(\mathbf{x}_t^{*,m})\|^2 = \eta\left\|\mathring{V}(\mathbf{x}_t^{*,m}) - \frac{1}{\eta}(\mathbf{x}_t^{*,m} - \mathbf{z}_{t-1}^m)\right\|^2 = \eta\|\mathring{V}(\mathbf{x}_t^{*,m}) - V(\mathbf{x}_t^{*,m})\|^2.$$

Since V is L-Lipschitz and $\mathring{V}(\mathbf{x}) = \mathbb{E}_s[V(\mathbf{x} + \delta\mathbf{s})]$, it holds that $\|\mathring{V}(\mathbf{x}) - V(\mathbf{x})\| \le L\delta\,\mathbb{E}\|\mathbf{s}\|$. Thus,

$$\mathbb{E}[\|\mathring{\mathbf{x}}_t^{*,m} - \mathbf{x}_t^{*,m}\|^2] \le \eta^2 L^2\delta^2\,\mathbb{E}\|\mathbf{s}\|^2 \le \eta^2 L^2\delta^2 d.$$

So

$$\mathbb{E}[\|\mathbf{x}_{t,\ell}^m - \mathbf{x}_t^{*,m}\|^2] \le \frac{2\eta^2 G^2}{4^\ell} + 2\eta^2\sigma^2 + 4\eta^2 L^2\delta^2 d.$$

E.2 Proof of Lemma 19

We have

$$\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2 \le 3\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + 3\|\mathbf{x}_t^{*,m} - \mathbf{x}_t^{*,m'}\|^2 + 3\|\mathbf{x}_t^{m'} - \mathbf{x}_t^{*,m'}\|^2,$$

Note that, the first and third terms are small. Similar to the proof of Lemma 15, we get

$$\|\mathbf{x}_t^{*,m} - \mathbf{x}_t^{*,m'}\|^2 \le \|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2;$$

Next, we turn to bound the drift of $\mathbf{z}_{t-1}^m$. We have

$$\begin{aligned}&\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2 = \|\mathbf{z}_{t-1}^m - \eta\widetilde{V}(\mathbf{x}_t^m) - \mathbf{z}_{t-1}^{m'} + \eta\widetilde{V}(\mathbf{x}_t^{m'})\|^2\\ =&\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2 - 2\eta\left\langle \widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\rangle + \eta^2\left\|\widetilde{V}(\mathbf{x}_t^m) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\end{aligned}$$

and similarly

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2\right] \le \mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] - \mathbb{E}\left[2\eta\left\langle \widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m,*}) + V(\mathbf{x}_t^{m',*}) - \widetilde{V}(\mathbf{x}_t^{m'}), \mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\rangle\right]\\ &+ \mathbb{E}\left[8\eta^2\left\|\widetilde{V}(\mathbf{x}_t^m) - V(\mathbf{x}_t^m)\right\|^2 + 8\eta^2\left\|V(\mathbf{x}_t^m) - V(\mathbf{x}_t^{*,m})\right\|^2 + 8\eta^2\left\|V(\mathbf{x}_t^{*,m'}) - V(\mathbf{x}_t^{m'})\right\|^2\right]\\ &+ 8\eta^2\mathbb{E}\left[\left\|V(\mathbf{x}_t^{m'}) - \widetilde{V}(\mathbf{x}_t^{m'})\right\|^2\right]\\ \le&\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] + \mathbb{E}\left[2\eta^2KL\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2 + 2\eta^2KL\|\mathbf{x}_t^{m'} - \mathbf{x}_t^{*,m'}\|^2 + \frac{1}{K}\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2\right]\\ &+ 16\eta^2\sigma^2 + 8\eta^2L\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2] + 8\eta^2L\mathbb{E}[\|\mathbf{x}_t^{m'} - \mathbf{x}_t^{*,m'}\|^2]\\ \le&\left(1 + \frac{1}{K}\right)\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] + 20\eta^2KL\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2] + 16\eta^2\sigma^2\end{aligned}$$

Based on Lemma 21, we have

$$\mathbb{E}\left[\|\mathbf{x}_t^m - \mathbf{x}_t^{*,m}\|^2\right] \le 4\left[\frac{\eta^2G^2}{4^H} + \eta^2\sigma^2 + \eta^2L^2\delta^2d\right].$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2\right] \le&\left(1 + \frac{1}{K}\right)\mathbb{E}[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2] + 80\eta^2KL\left(\frac{\eta^2G^2}{4^H} + \eta^2\sigma^2 + \eta^2L^2\delta^2d\right) + 16\eta^2\sigma^2\\ \le&160\left[\frac{\eta^4K^2LG^2}{4^H} + \eta^4K^2L\sigma^2 + \eta^4K^2L^3\delta^2d + \eta^2K\sigma^2\right].\end{aligned}$$

Therefore, we get:

$$\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] \le 480\left[\frac{\eta^4K^2LG^2 + \eta^2G^2}{4^H} + \eta^4K^2L\sigma^2 + \eta^4K^2L^3\delta^2d + \eta^2K\sigma^2 + \eta^2L^2\delta^2d\right]$$

F Proof of Theorem 6

Similar to pervious proof, we define the shadow updates: $\overline{\mathbf{x}}_{t+1} = \frac{1}{M}\sum_{m=1}^M \mathbf{x}_{t+1}^m$. We have

$$\overline{\mathbf{x}}_{t+1} = \frac{1}{M}\sum_{m=1}^M \mathbf{x}_{t+1}^m = \overline{\mathbf{x}}_t - \eta\frac{1}{M}\sum_{m=1}^M \widetilde{V}(\mathbf{x}_{t-1}^m) = \overline{\mathbf{x}}_t - \eta\widetilde{V}_{\mathbf{x},t}. \tag{55}$$

Also, we define $V_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^M V(\mathbf{x}_t^m)$. Next, we start the proof by bounding the following potential function. For any $\mathbf{u} \in \mathbb{R}^d$, we have

$$\|\overline{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 = \|\overline{\mathbf{x}}_t - \eta\widetilde{V}_{\mathbf{x},t} - \mathbf{u}\|^2 = \|\overline{\mathbf{x}}_t - \mathbf{u}\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{x},t}\right\|^2. \tag{56}$$

Next, we bound each term in the above inequality respectively. For the third term of (56), we have

$$\begin{aligned}&\left\|\widetilde{V}_{\mathbf{x},t}\right\|^2 = \left\|\widetilde{V}_{\mathbf{x},t} - V(\mathbf{u}) + V(\mathbf{u})\right\|^2 = \left\|\widetilde{V}_{\mathbf{x},t} - V(\mathbf{u})\right\|^2 + 2\left\langle \widetilde{V}_{\mathbf{x},t} - V(\mathbf{u}), V(\mathbf{u})\right\rangle + \|V(\mathbf{u})\|^2\\ \le&\left\|\widetilde{V}_{\mathbf{x},t} - V(\mathbf{u})\right\|^2 + 2\left\langle \widetilde{V}_{\mathbf{x},t}, V(\mathbf{u})\right\rangle = \left\|\widetilde{V}_{\mathbf{x},t} - V(\mathbf{u})\right\|^2 + \frac{2}{M}\sum_{m=1}^M\left\langle \widetilde{V}\left(\mathbf{x}_t^{(m)}\right), V(\mathbf{u})\right\rangle\\ \le&2\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\|^2 + \frac{2}{M}\sum_{m=1}^M\|V(\mathbf{x}_t^{(m)}) - V(\mathbf{u})\|^2 + \frac{2}{M\eta}\sum_{m=1}^M\left\langle \mathbf{x}_t^{(m)} - \mathbf{x}_{t+1}^{(m)}, V(\mathbf{u})\right\rangle\end{aligned}$$

For the second term of (56), we have

$$
\begin{aligned}
&-2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle = -\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
=&-\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)} + \mathbf{x}_t^{(m)} - \mathbf{u}\right\rangle + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
=&-\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}), \mathbf{x}_t^{(m)} - \mathbf{u}\right\rangle - \frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)}\right\rangle + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
=&-\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}), \mathbf{x}_t^{(m)} - \mathbf{u}\right\rangle \underbrace{-\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)}\right\rangle}_{=0} + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
&+\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)}\right\rangle
\end{aligned}
\tag{57}
$$

Combining all inequalities above, we have

$$
\begin{aligned}
0 \leq&\|\overline{\mathbf{x}}_t - \mathbf{u}\|^2 - \|\overline{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 - \frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}), \mathbf{x}_t^{(m)} - \mathbf{u}\right\rangle + \frac{2\eta}{M}\sum_{m=1}^{M}\left\langle \mathbf{x}_t^{(m)} - \mathbf{x}_{t+1}^m, V(\mathbf{u})\right\rangle \\
&+\frac{2\eta^2}{M}\sum_{m=1}^{M}\|V(\mathbf{x}_t^{(m)}) - V(\mathbf{u})\|^2 + 2\eta^2\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
&+\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)}\right\rangle
\end{aligned}
\tag{58}
$$

Adding $\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{u}), \mathbf{x}_{t+1}^{(m)} - \mathbf{u}\right\rangle$ on both sides, we have

$$
\begin{aligned}
&\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{u}), \mathbf{x}_{t+1}^{(m)} - \mathbf{u}\right\rangle \leq \|\overline{\mathbf{x}}_t - \mathbf{u}\|^2 - \|\overline{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 - \frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\mathbf{x}_t^{(m)}) - V(\mathbf{u}), \mathbf{x}_t^{(m)} - \mathbf{u}\right\rangle \\
&+\frac{2\eta^2}{M}\sum_{m=1}^{M}\|V(\mathbf{x}_t^{(m)}) - V(\mathbf{u})\|^2 + 2\eta^2\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
&+\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)}\right\rangle \\
\leq&\|\overline{\mathbf{x}}_t - \mathbf{u}\|^2 - \|\overline{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 + 2\eta^2\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle \\
&+\frac{2\eta}{M}\sum_{m=1}^{M}\left\langle V(\overline{\mathbf{x}}_t) - V(\mathbf{x}_t^{(m)}), \overline{\mathbf{x}}_t - \mathbf{x}_t^{(m)}\right\rangle \\
\leq&\|\overline{\mathbf{x}}_t - \mathbf{u}\|^2 - \|\overline{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 + 2\eta^2\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\|^2 + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle + \frac{2\eta L}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2,
\end{aligned}
\tag{59}
$$

where in the second inequality we used the fact that V is β-co-coercive and $\eta \leq \frac{1}{\beta}$. Summing it up from 0 to $T-1$, dividing $2\eta T$ on both sides, letting $\mathbf{u} = \mathbf{z}^*$ and taking expecation on both sides, we have: we have

$$
\mathbb{E}\left[\frac{1}{MT}\left\langle V(\mathbf{u}), \frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{x}}_0 - \mathbf{z}^*\|^2]}{\eta T} + \frac{L}{M}\sum_{m=1}^{M}\mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2] + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\Gamma_t],
\tag{60}
$$

where

$$\Gamma_t = \eta\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\|^2 + \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{u}\right\rangle,$$

is the noise term. To proceed, we introduce the upper bound for the drift and Γ_t. For the drift term, we have the following bound. The proof can be found in Appendix F.1.

Lemma 22. *We have*

$$\mathbb{E}\left[\frac{L}{M}\sum_{m=1}^{M}\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2\right] \leq 4L\sigma^2 K\eta^2. \tag{61}$$

Finally, we provide the upper bound for Γ_t, which is exactly the same as Lemma 16.

Lemma 23. *We have*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Gamma_t\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{x}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{2\sigma^2\eta}{M}.$$

Combining all conclusions, we get:

$$\mathbb{E}\left[\frac{1}{MT}\left\langle V(\mathbf{u}), \frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq \frac{2D^2}{\eta T} + \frac{2\sigma^2\eta}{M} + 4L\sigma^2 K\eta^2.$$

Set

$$\eta = \min\left\{\frac{1}{\beta}, \frac{D\sqrt{M}}{\sqrt{KR}\sigma}, \frac{D^{\frac{2}{3}}}{2^{\frac{1}{2}}K^{\frac{2}{3}}R^{\frac{1}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}\right\},$$

we get

$$\mathbb{E}\left[\frac{1}{MT}\left\langle V(\mathbf{u}), \frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m - \mathbf{z}^*\right\rangle\right] \leq \frac{4\beta D^2}{KR} + \frac{4\sigma D}{\sqrt{MKR}} + \frac{8D^{\frac{4}{3}}L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}.$$

F.1 Proof of Lemma 22

We have

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] = \mathbb{E}\left[\left\|\mathbf{x}_{t-1}^m - \mathbf{x}_{t-1}^{m'} - \eta\left(\widetilde{V}(\mathbf{x}_{t-1}^m) - \widetilde{V}(\mathbf{x}_{t-1}^{m'})\right)\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\mathbf{x}_{t-1}^m - \mathbf{x}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}(\mathbf{x}_{t-1}^m) - \widetilde{V}(\mathbf{x}_{t-1}^{m'})\right\|^2\right] - 2\eta\mathbb{E}\left[\left\langle\mathbf{x}_{t-1}^m - \mathbf{x}_{t-1}^{m'}, \widetilde{V}(\mathbf{x}_{t-1}^m) - \widetilde{V}(\mathbf{x}_{t-1}^{m'})\right\rangle\right]\\
\leq&\mathbb{E}\left[\left\|\mathbf{x}_{t-1}^m - \mathbf{x}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|V(\mathbf{x}_{t-1}^m) - V(\mathbf{x}_{t-1}^{m'})\right\|^2\right] + 4\eta^2\sigma^2 - 2\eta\mathbb{E}\left[\left\langle\mathbf{x}_{t-1}^m - \mathbf{x}_{t-1}^{m'}, V(\mathbf{x}_{t-1}^m) - V(\mathbf{x}_{t-1}^{m'})\right\rangle\right]\\
\leq&4\eta^2\sigma^2 K,
\end{aligned}$$

where the last inequality is based on the co-coercivity of the operator.

G Proof of Theorem 7

We first recall the algorithm:

$$\begin{aligned}
\mathbf{x}_t^m &= \begin{cases}\mathbf{z}_{t-1}^m - \eta\widetilde{V}(\mathbf{u}_{t-1}^m), & \mathrm{mod}(t,K)\neq 0;\\ \frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{z}_{t-1}^m - \eta\widetilde{V}(\mathbf{u}_{t-1}^m)\right), & \mathrm{mod}(t,K)=0.\end{cases}\\
\mathbf{z}_t^m &= \begin{cases}\mathbf{z}_{t-1}^m - \eta\widetilde{V}(\mathbf{v}_t^m), & \mathrm{mod}(t,K)\neq 0;\\ \frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{z}_{t-1}^m - \eta\widetilde{V}(\mathbf{v}_t^m)\right), & \mathrm{mod}(t,K)=0.\end{cases}
\end{aligned} \tag{62}$$

Here, $\mathbf{u}_t^m = \nabla h_t^*(\mathbf{z}_t^m)$, and $\mathbf{v}_t^m = \nabla h_t^*(\mathbf{x}_t^m)$ are the primal variables. We start by defining the following shadow updates: $\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m$, and $\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m$. We have

$$\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{u}_{t-1}^m) = \overline{\mathbf{z}}_{t-1} - \widetilde{V}_{\mathbf{u},t-1},$$

and

$$\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{v}_t^m) = \overline{\mathbf{z}}_{t-1} - \widetilde{V}_{\mathbf{v},t}.$$

where we define $\widetilde{V}_{\mathbf{u},t-1} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{u}_{t-1}^m)$, and $\widetilde{V}_{\mathbf{v},t} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{V}(\mathbf{v}_t^m)$. Moreover, let $V_{\mathbf{u},t-1} = \frac{1}{M}\sum_{m=1}^{M}V(\mathbf{u}_{t-1}^m)$, and $V_{\mathbf{v},t} = \frac{1}{M}\sum_{m=1}^{M}V(\mathbf{v}_t^m)$. Finally, let $\overline{\mathbf{u}}_t = \nabla h_t^*(\overline{\mathbf{z}}_t)$, and $\overline{\mathbf{v}}_t = \nabla h_t^*(\overline{\mathbf{x}}_t)$. Note that, $\mathbf{x}$ and $\mathbf{z}$ are in dual, $\mathbf{u}$ and $\mathbf{v}$ are in primal. Define generalized Bregman divergence [14, 39]:

$$\widetilde{G}_{h_t}(\mathbf{u},\mathbf{z}) = h_t(\mathbf{u}) - h_t(\nabla h_t^*(\mathbf{z})) - \langle \mathbf{z}, \mathbf{u} - \nabla h_t^*(\mathbf{z})\rangle$$

Note that, here $\mathbf{u}$ is in the primal space, while $\mathbf{z}$ is in the dual space.

$$\begin{aligned}\widetilde{G}_{h_t}(\overline{\mathbf{u}}_t,\overline{\mathbf{x}}_t) =& h_t(\overline{\mathbf{u}}_t) - h_t(\overline{\mathbf{v}}_t) - \langle \overline{\mathbf{x}}_t, \overline{\mathbf{u}}_t - \overline{\mathbf{v}}_t\rangle\\ =& h_t(\overline{\mathbf{u}}_t) - h_t(\overline{\mathbf{v}}_t) - \left\langle \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{u},t-1}, \overline{\mathbf{u}}_t - \overline{\mathbf{v}}_t\right\rangle\\ =& h_{t-1}(\overline{\mathbf{u}}_t) - h_{t-1}(\overline{\mathbf{v}}_t) + \eta\left(\phi(\overline{\mathbf{u}}_t) - \phi(\overline{\mathbf{v}}_t)\right) - \left\langle \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{u},t-1}, \overline{\mathbf{u}}_t - \overline{\mathbf{v}}_t\right\rangle.\end{aligned} \tag{63}$$

Similarly, for any $\mathbf{u}$,

$$\begin{aligned}\widetilde{G}_{h_t}(\mathbf{u},\overline{\mathbf{z}}_t) =& h_t(\mathbf{u}) - h_t(\overline{\mathbf{u}}_t) - \langle \overline{\mathbf{z}}_t, \mathbf{u}_t - \overline{\mathbf{u}}_t\rangle\\ =& h_{t-1}(\mathbf{u}) - h_{t-1}(\overline{\mathbf{u}}_t) + \eta\left(\phi(\mathbf{u}) - \phi(\overline{\mathbf{u}}_t)\right) - \left\langle \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{v},t}, \mathbf{u} - \overline{\mathbf{u}}_t\right\rangle.\end{aligned} \tag{64}$$

Therefore,

$$\begin{aligned}&\widetilde{G}_{h_t}(\overline{\mathbf{u}}_t,\overline{\mathbf{x}}_t) + \widetilde{G}_{h_t}(\mathbf{u},\overline{\mathbf{z}}_t)\\ =& \underbrace{h_{t-1}(\mathbf{u}) - h_{t-1}(\overline{\mathbf{v}}_t) - \langle \overline{\mathbf{z}}_{t-1}, \mathbf{u} - \overline{\mathbf{v}}_t\rangle}_{A_1} + \eta(\phi(\mathbf{u}) - \phi(\overline{\mathbf{v}}_t))\\ &\underbrace{+\eta\left\langle \widetilde{V}_{\mathbf{u},t-1}, \overline{\mathbf{u}}_t - \overline{\mathbf{v}}_t\right\rangle + \eta\left\langle \widetilde{V}_{\mathbf{v},t}, \mathbf{u} - \overline{\mathbf{u}}_t\right\rangle}_{A_2}\end{aligned} \tag{65}$$

We have

$$\begin{aligned}A_1 =& h_{t-1}(\mathbf{u}) - h_{t-1}(\overline{\mathbf{u}}_{t-1}) - \langle \overline{\mathbf{z}}_{t-1}, \mathbf{u} - \overline{\mathbf{u}}_{t-1}\rangle - \left(h_{t-1}(\overline{\mathbf{v}}_t) - h_{t-1}(\overline{\mathbf{u}}_{t-1}) - \langle \overline{\mathbf{z}}_{t-1}, \overline{\mathbf{v}}_t - \overline{\mathbf{u}}_{t-1}\rangle\right)\\ =& \widetilde{G}_{h_{t-1}}(\mathbf{u},\overline{\mathbf{z}}_{t-1}) - \widetilde{G}_{h_{t-1}}(\overline{\mathbf{v}}_t,\overline{\mathbf{z}}_{t-1}),\end{aligned} \tag{66}$$

and

$$\begin{aligned}&\eta\left\langle \widetilde{V}_{\mathbf{u},t-1}, \overline{\mathbf{u}}_t - \overline{\mathbf{v}}_t\right\rangle + \eta\left\langle \widetilde{V}_{\mathbf{v},t}, \mathbf{u} - \overline{\mathbf{u}}_t\right\rangle = -\eta\left\langle \widetilde{V}_{\mathbf{v},t}, \overline{\mathbf{v}}_t - \mathbf{u}\right\rangle + \eta\left\langle \widetilde{V}_{\mathbf{v},t} - \widetilde{V}_{\mathbf{u},t-1}, \overline{\mathbf{v}}_t - \overline{\mathbf{u}}_t\right\rangle\\ \le& -\eta\left\langle \widetilde{V}_{\mathbf{v},t}, \overline{\mathbf{v}}_t - \mathbf{u}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{v},t} - \widetilde{V}_{\mathbf{u},t-1}\right\|_*^2 + \frac{1}{2}\left\|\overline{\mathbf{v}}_t - \overline{\mathbf{u}}_t\right\|^2\end{aligned} \tag{67}$$

To summarize, we have

$$\begin{aligned}&\eta\left(\phi(\overline{\mathbf{v}}_t) - \phi(\mathbf{u})\right) + \widetilde{G}_{h_t}(\overline{\mathbf{u}}_t,\overline{\mathbf{x}}_t) + \widetilde{G}_{h_t}(\mathbf{u},\overline{\mathbf{z}}_t)\\ \le& \widetilde{G}_{h_{t-1}}(\mathbf{u},\overline{\mathbf{z}}_{t-1}) - \widetilde{G}_{h_{t-1}}(\overline{\mathbf{v}}_t,\overline{\mathbf{z}}_{t-1}) - \eta\left\langle \widetilde{V}_{\mathbf{v},t}, \overline{\mathbf{v}}_t - \mathbf{u}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{v},t} - \widetilde{V}_{\mathbf{u},t-1}\right\|_*^2 + \frac{1}{2}\left\|\overline{\mathbf{v}}_t - \overline{\mathbf{u}}_t\right\|^2\\ \le& \widetilde{G}_{h_{t-1}}(\mathbf{u},\overline{\mathbf{z}}_{t-1}) - G_h(\overline{\mathbf{v}}_t,\overline{\mathbf{u}}_{t-1}) - \eta\left\langle \widetilde{V}_{\mathbf{v},t}, \overline{\mathbf{v}}_t - \mathbf{u}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{v},t} - \widetilde{V}_{\mathbf{u},t-1}\right\|_*^2 + \frac{1}{2}\left\|\overline{\mathbf{v}}_t - \overline{\mathbf{u}}_t\right\|^2\\ \le& \widetilde{G}_{h_{t-1}}(\mathbf{u},\overline{\mathbf{z}}_{t-1}) - \frac{1}{2}\left\|\overline{\mathbf{v}}_t - \overline{\mathbf{u}}_{t-1}\right\|^2 - \eta\left\langle \widetilde{V}_{\mathbf{v},t}, \overline{\mathbf{v}}_t - \mathbf{u}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{v},t} - \widetilde{V}_{\mathbf{u},t-1}\right\|_*^2 + \frac{1}{2}\left\|\overline{\mathbf{v}}_t - \overline{\mathbf{u}}_t\right\|^2.\end{aligned} \tag{68}$$

On the other hand, based on the strong convexity of h, we have

$$\widetilde{G}_{h_t}(\overline{\mathbf{u}}_t,\overline{\mathbf{x}}_t) \ge G_h(\overline{\mathbf{u}}_t,\overline{\mathbf{v}}_t) \ge \frac{1}{2}\|\overline{\mathbf{u}}_t - \overline{\mathbf{v}}_t\|^2. \tag{69}$$

Thus, we get

$$\eta\left(\phi(\overline{\mathbf{v}}_t)-\phi(\mathbf{u})\right)\leq \widetilde{G}_{h_{t-1}}(\mathbf{u},\overline{\mathbf{z}}_{t-1})-\widetilde{G}_{h_t}(\mathbf{u},\overline{\mathbf{z}}_t)-\eta\left\langle \widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle-\frac{1}{2}\left\|\overline{\mathbf{v}}_t-\overline{\mathbf{u}}_{t-1}\right\|^2+\eta^2\left\|\widetilde{V}_{\mathbf{v},t}-\widetilde{V}_{\mathbf{u},t-1}\right\|_*^2 \tag{70}$$

We first deal with the last term.

$$\begin{aligned}
&\left\|\widetilde{V}_{\mathbf{u},t}-\widetilde{V}_{\mathbf{v},t-1}\right\|_*^2\\
=&\left\|\widetilde{V}_{\mathbf{u},t}-V_{\mathbf{u},t}+V_{\mathbf{u},t}-V(\overline{\mathbf{u}}_t)+V(\overline{\mathbf{u}}_t)-V(\overline{\mathbf{v}}_{t-1})+V(\overline{\mathbf{v}}_{t-1})-V_{\mathbf{v},t-1}+V_{\mathbf{v},t-1}-\widetilde{V}_{\mathbf{v},t-1}\right\|_*^2\\
\leq&5\left\|\widetilde{V}_{\mathbf{u},t}-V_{\mathbf{u},t}\right\|_*^2+5\left\|V_{\mathbf{u},t}-V(\overline{\mathbf{u}}_t)\right\|_*^2+5\left\|V(\overline{\mathbf{u}}_t)-V(\overline{\mathbf{v}}_{t-1})\right\|_*^2+5\left\|V(\overline{\mathbf{v}}_{t-1})-V_{\mathbf{v},t-1}\right\|_*^2\\
&+5\left\|V_{\mathbf{v},t-1}-\widetilde{V}_{\mathbf{v},t-1}\right\|^2\\
\leq&5\left\|\widetilde{V}_{\mathbf{u},t}-V_{\mathbf{u},t}\right\|_*^2+\frac{5L^2}{M}\sum_{m=1}^M\left\|\mathbf{u}_t^m-\overline{\mathbf{u}}_t\right\|^2+5L^2\|\overline{\mathbf{u}}_t-\overline{\mathbf{v}}_{t-1}\|^2+\frac{5L^2}{M}\sum_{m=1}^M\left\|\mathbf{v}_{t-1}^m-\overline{\mathbf{v}}_{t-1}\right\|_*^2\\
&+5\left\|V_{\mathbf{v},t-1}-\widetilde{V}_{\mathbf{v},t-1}\right\|_*^2.
\end{aligned}$$

Next, we focus on the third term at the R.H.S. of (70). We have

$$\begin{aligned}
&-\eta\left\langle\widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle=-\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\mathbf{v}_t^m),\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle+\eta\left\langle V_{\mathbf{v},t}-\widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle\\
=&-\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\mathbf{v}_t^m),\overline{\mathbf{v}}_t-\mathbf{v}_t^m+\mathbf{v}_t^m-\mathbf{u}\right\rangle+\eta\left\langle V_{\mathbf{v},t}-\widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle\\
=&-\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\mathbf{v}_t^m),\mathbf{v}_t^m-\mathbf{u}\right\rangle-\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\mathbf{v}_t^m),\overline{\mathbf{v}}_t-\mathbf{v}_t^m\right\rangle+\eta\left\langle V_{\mathbf{v},t}-\widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle\\
=&-\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\mathbf{v}_t^m),\mathbf{v}_t^m-\mathbf{u}\right\rangle-\underbrace{\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\overline{\mathbf{v}}_t),\overline{\mathbf{v}}_t-\mathbf{v}_t^m\right\rangle}_{=0}+\eta\left\langle V_{\mathbf{v},t}-\widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle\\
&+\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\overline{\mathbf{v}}_t)-V(\mathbf{v}_t^m),\overline{\mathbf{v}}_t-\mathbf{v}_t^m\right\rangle.
\end{aligned}\tag{71}$$

Note that

$$\frac{\eta}{M}\sum_{m=1}^M\left\langle V(\overline{\mathbf{v}}_t)-V(\mathbf{v}_t^m),\overline{\mathbf{x}}_t-\mathbf{x}_t^m\right\rangle\leq\frac{\eta}{M}\sum_{m=1}^M\|V(\overline{\mathbf{v}}_t)-V(\mathbf{v}_t^m)\|_*\|\overline{\mathbf{v}}_t-\mathbf{v}_t^m\|\leq\frac{\eta L}{M}\sum_{m=1}^M\|\overline{\mathbf{v}}_t-\mathbf{v}_t^m\|^2. \tag{72}$$

Combining, and the fact that $\eta\leq\frac{1}{\sqrt{10}L}$, we get

$$\begin{aligned}
&\left(\phi(\overline{\mathbf{v}}_t)-\phi(\mathbf{u})\right)+\frac{1}{M}\sum_{m=1}^M\left\langle V(\mathbf{v}_t^m),\mathbf{v}_t^m-\mathbf{u}\right\rangle\leq\frac{\widetilde{G}_{h_{t-1}}(\mathbf{u},\overline{\mathbf{z}}_{t-1})-\widetilde{G}_{h_t}(\mathbf{u},\overline{\mathbf{z}}_t)}{\eta}+\frac{L}{M}\sum_{m=1}^M\|\overline{\mathbf{v}}_t-\mathbf{v}_t^m\|^2\\
&+\frac{5L^2\eta}{M}\sum_{m=1}^M\|\mathbf{u}_t^m-\overline{\mathbf{u}}_t\|^2+\frac{5L^2}{M}\sum_{m=1}^M\|\mathbf{v}_{t-1}^m-\overline{\mathbf{v}}_{t-1}\|^2+\Gamma_t
\end{aligned}\tag{73}$$

where

$$\Gamma_t=\left\langle V_{\mathbf{v},t}-\widetilde{V}_{\mathbf{v},t},\overline{\mathbf{v}}_t-\mathbf{u}\right\rangle+\left\|\widetilde{V}_{\mathbf{u},t}-V_{\mathbf{u},t}\right\|_*^2+5\left\|V_{\mathbf{v},t-1}-\widetilde{V}_{\mathbf{v},t-1}\right\|_*^2.$$

Note that we also have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\left((\phi(\overline{\mathbf{v}}_t)-\phi(\mathbf{u}))+\frac{1}{M}\sum_{m=1}^{M}\langle V(\mathbf{v}_t^m),\mathbf{v}_t^m-\mathbf{u}\rangle\right)\\ \geq&\phi\left(\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{v}_t^m\right)-\phi(\mathbf{u})+\left\langle V(\mathbf{u}),\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{v}_t^m-\mathbf{u}\right\rangle.\end{aligned}\tag{74}$$

With similar techniques, one can show that

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Gamma_t\right]\leq\frac{D^2}{2\eta T}+\frac{6\sigma^2\eta}{M}.\tag{75}$$

Therefore, combining with the bounded gradient assumption, it gives as

$$\begin{aligned}&\mathbb{E}\left\{\sup_{\mathbf{u}}\left[\phi\left(\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{v}_t^m\right)-\phi(\mathbf{u})+\left\langle V(\mathbf{u}),\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{v}_t^m-\mathbf{u}\right\rangle\right]\right\}\\ \leq&\frac{D^2}{\eta KR}+\frac{6\sigma^2\eta}{M}+17L^2G^2\eta^2K^2.\end{aligned}$$

setting

$$\eta=\min\left\{\frac{D\sqrt{M}}{\sigma\sqrt{6KR}},\frac{D^{\frac{2}{3}}}{17^{\frac{1}{3}}KR^{\frac{1}{3}}L^{\frac{2}{3}}G^{\frac{2}{3}}},\frac{1}{\sqrt{10}L}\right\},$$

yields

$$\begin{aligned}&\mathbb{E}\left\{\sup_{\mathbf{u}}\left[\phi\left(\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{v}_t^m\right)-\phi(\mathbf{u})+\left\langle V(\mathbf{u}),\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{v}_t^m-\mathbf{u}\right\rangle\right]\right\}\\ \leq&\frac{2\sqrt{6}D\sigma}{\sqrt{MKR}}+\frac{17^{\frac{1}{3}}L^{\frac{2}{3}}D^{\frac{4}{3}}G^{\frac{2}{3}}}{R^{\frac{2}{3}}}+\frac{\sqrt{10}D^2L}{KR}.\end{aligned}$$

H Extension to Heterogeneous Setting

In this section, we consider the heterogeneous case, where $V(\mathbf{x})=\frac{1}{M}\sum_{m=1}^{M}V_m(\mathbf{x})$. We introduce the following standard assumptions.

Assumption 2 (Bounded Heterogeneity)**.** There exists a constant $\xi>0$ such that for any $\mathbf{x}\in\mathbb{R}^d$ and $m\in[M]$, we have:

$$\|V_m(\mathbf{x})-V(\mathbf{x})\|\leq\xi.$$

Assumption 3 (Unbiased Stochastic Oracle)**.** Each client $m\in[M]$ has access to a stochastic oracle $\widetilde{V}_m(\mathbf{x})$ such that $\mathbb{E}[\widetilde{V}_m(\mathbf{x})]=V_m(\mathbf{x})$ and $\mathbb{E}[\|\widetilde{V}_m(\mathbf{x})-V_m(\mathbf{x})\|^2]\leq\sigma^2$ for all $\mathbf{x}\in\mathbb{R}^d$.

Under this setting, the LESGD is given by:

$$\mathbf{x}_t^m=\begin{cases}\mathbf{z}_{t-1}^m-\eta\widetilde{V}_m(\mathbf{z}_{t-1}^m), & \mathrm{mod}(t,K)\neq0;\\ \frac{1}{M}\sum_{m'=1}^{M}\left(\mathbf{z}_{t-1}^{m'}-\eta\widetilde{V}_{m'}(\mathbf{z}_{t-1}^{m'})\right), & \mathrm{mod}(t,K)=0.\end{cases}$$

$$\mathbf{z}_t^m=\begin{cases}\mathbf{z}_{t-1}^m-\eta\widetilde{V}_m(\mathbf{x}_t^m), & \mathrm{mod}(t,K)\neq0;\\ \frac{1}{M}\sum_{m'=1}^{M}\left(\mathbf{z}_{t-1}^{m'}-\eta\widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right), & \mathrm{mod}(t,K)=0.\end{cases}$$

We have the following conclusion for LESGD under the heterogeneous setting.

Theorem 24 (LESGD for Heterogeneous Setting)**.** *Suppose $V(\mathbf{x})=\frac{1}{M}\sum_{m=1}^{M}V_m(\mathbf{x})$, where each V_m is L-smooth and monotone. Suppose Assumptions 2 and 3 hold. Let $\mathcal{Z}_D=\{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}-\mathbf{z}_0\|\leq$*

$D\}$, where $D > 0$ is any constant picked by the user. Let $\mathbf{x}_o = \frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m$ be the output. where $T = KR$. Set:

$$\eta = \min\left\{\frac{1}{\sqrt{K}L}, \frac{D\sqrt{M}}{\sigma\sqrt{KR}}, \frac{D^{\frac{2}{3}}}{K^{\frac{2}{3}}R^{\frac{1}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}, \frac{D^{\frac{2}{3}}}{\xi^{\frac{2}{3}}KR^{\frac{1}{3}}L^{\frac{1}{3}}}, \frac{D}{(\xi\sigma)^{\frac{1}{2}}K^{\frac{3}{4}}R^{\frac{1}{2}}}, \frac{D}{\xi K\sqrt{R}}\right\}$$

Then we have

$$\mathbb{E}\left[\sup_{\mathbf{z}\in\mathcal{Z}_D}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle\right] = O\left(\frac{LD^2}{\sqrt{K}R} + \frac{D\sigma}{\sqrt{MKR}} + \frac{D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{D^{\frac{4}{3}}\xi^{\frac{2}{3}}L^{\frac{1}{3}}}{R^{\frac{2}{3}}} + \frac{D(\xi\sigma)^{\frac{1}{2}}}{K^{\frac{1}{4}}R^{\frac{1}{2}}} + \frac{D\xi}{\sqrt{R}}\right)$$

We note that when $\xi = 0$, the convergence rate reduces to the bound provided in Theorem 1. By contrast, the $O(\frac{\sigma}{\sqrt{R}})$ term in the bound established by previous work [12] is independent of ξ, and thus persists even when $\xi = 0$. On the other hand, for LSGD in federated optimization under the heterogeneous setting, [40] provided the following bound: $O\left(\frac{LD^2}{KR} + \frac{D\sigma}{\sqrt{MKR}} + \frac{D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{D^{\frac{4}{3}}\xi^{\frac{2}{3}}L^{\frac{1}{3}}}{R^{\frac{2}{3}}}\right)$. Compared with this bound, apart from the $\frac{1}{\sqrt{KR}}$ term which reflects the fundamental limitations of LESGD, the bound provided in Theorem 24 contains two additional terms: $\frac{D(\xi\sigma)^{\frac{1}{2}}}{K^{\frac{1}{4}}R^{\frac{1}{2}}} + \frac{D\xi}{\sqrt{R}}$. We believe this is due to the inherent difficulty of solving VIs in the federated setting, which we discuss briefly below. Specifically, in the proof of Theorem 24, we show that our objective can be decomposed as:

$$\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle = \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \mathbf{x}_t^m - \mathbf{z}\rangle + \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}) - V(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle.$$

Here, the first term is what we control in the homogeneous case, while the second term arises from heterogeneity (note that the above decomposition is equality). Since this term involves an arbitrary $\mathbf{z}$, we can only bound it using Cauchy-Schwarz as follows:

$$\begin{aligned}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle =& \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \mathbf{x}_t^m - \mathbf{z}\rangle + \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}) - V(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle \\ \leq& \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \mathbf{x}_t^m - \mathbf{z}\rangle + \frac{\xi}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|,\end{aligned}$$

which gives rise to the additional terms.

H.1 Proof of Theorem 24

We start by defining the following shadow updates: $\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m$, and $\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m$. We have

$$\overline{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^{M}\widetilde{V}_m(\mathbf{z}_{t-1}^m) = \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{z},t-1},$$

and

$$\overline{\mathbf{z}}_t = \frac{1}{M}\sum_{m=1}^{M}\mathbf{z}_t^m = \overline{\mathbf{z}}_{t-1} - \eta\frac{1}{M}\sum_{m=1}^{M}\widetilde{V}_m(\mathbf{x}_t^m) = \overline{\mathbf{z}}_{t-1} - \eta\widetilde{V}_{\mathbf{x},t}.$$

where we define $\widetilde{V}_{\mathbf{z},t-1} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{V}_m(\mathbf{z}_{t-1}^m)$, and $\widetilde{V}_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{V}_m(\mathbf{x}_t^m)$. Moreover, let $V_{\mathbf{z},t-1} = \frac{1}{M}\sum_{m=1}^{M}V_m(\mathbf{z}_{t-1}^m)$, and $V_{\mathbf{x},t} = \frac{1}{M}\sum_{m=1}^{M}V_m(\mathbf{x}_t^m)$. Let $\mathbf{x}_o = \frac{1}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbf{x}_t^m = \frac{1}{T}\sum_{t=1}^{T}\overline{\mathbf{x}}_t$ be the final output. We first establish a connection between the error and the client-specific

operators. We have

$$\begin{aligned}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle = \frac{1}{T}\sum_{t=1}^{T}\langle V(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{z}\rangle = \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{z}\rangle \\ =\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{x}_t^m + \mathbf{x}_t^m - \mathbf{z}\rangle \\ =\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \mathbf{x}_t^m - \mathbf{z}\rangle + \underbrace{\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle}_{=0} \\ + \frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}) - V(\mathbf{z}), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle \\ \leq\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{z}), \mathbf{x}_t^m - \mathbf{z}\rangle + \frac{\xi}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|,\end{aligned}$$

where the last inequality uses Assumption 2. To proceed, we introduce the following lemma.

Lemma 25. *We have*

$$\begin{aligned}\mathbb{E}\left[\frac{1}{TM}\sum_{t=1}^{T}\sum_{m=1}^{M}\langle V_m(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}^*\rangle\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{\mathbb{E}[\sum_{t=1}^{T}\Gamma_t]}{T} \\ + \frac{6L}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\left[\mathbb{E}[\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2] + \mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2]\right] + \frac{2\xi}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}[\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|],\end{aligned}$$

where

$$\begin{aligned}\Gamma_t =& \left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}^*\right\rangle + \frac{5\eta}{2}\left[\left\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\right\|^2 + \left\|V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1}\right\|^2\right], \\ \mathbf{z}^* =& \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_D}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle.\end{aligned}$$

Combining Lemma 25 with the above, and using the monotonicity of V, we get

$$\begin{aligned}\mathbb{E}\left[\sup_{\mathbf{z}\in\mathcal{Z}_D}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{\mathbb{E}[\sum_{t=1}^{T}\Gamma_t]}{T} \\ + \frac{6L}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\left[\mathbb{E}[\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2] + \mathbb{E}[\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2]\right] + \frac{3\xi}{MT}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}[\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|].\end{aligned}$$

For the second term, following the same argument as in (19), we have

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\Gamma_t\right] \leq \frac{\mathbb{E}[\|\overline{\mathbf{z}}_0 - \mathbf{z}^*\|^2]}{2\eta T} + \frac{6\sigma^2\eta}{M}.$$

It remains to bound the client drift terms. We have the following lemma.

Lemma 26 (Client Drift for Heterogeneous Setting)**.** *For $\eta \leq \frac{1}{\sqrt{150KL}}$, we have*

$$\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] \leq C\eta^2\sigma^2 K + C\eta^2\xi^2K^2,$$

and

$$\mathbb{E}[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2] \leq C\eta^2\sigma^2 K + C\eta^2\xi^2K^2,$$

where $C > 0$ is an absolute constant.

Applying Lemma 26 to the above, and using Jensen's inequality for expectation, we obtain

$$\mathbb{E}\left[\sup_{\mathbf{z}\in\mathcal{Z}_D}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle\right] = O\left(\frac{D^2}{\eta KR} + \frac{\sigma^2\eta}{M} + L\eta^2\sigma^2K + L\eta^2\xi^2K^2 + \xi\eta\sigma\sqrt{K} + \xi^2\eta K\right).$$

Setting

$$\eta = \min\left\{\frac{1}{\sqrt{K}L}, \frac{D\sqrt{M}}{\sigma\sqrt{KR}}, \frac{D^{\frac{2}{3}}}{K^{\frac{2}{3}}R^{\frac{1}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}, \frac{D^{\frac{2}{3}}}{\xi^{\frac{2}{3}}KR^{\frac{1}{3}}L^{\frac{1}{3}}}, \frac{D}{(\xi\sigma)^{\frac{1}{2}}K^{\frac{3}{4}}R^{\frac{1}{2}}}, \frac{D}{\xi K\sqrt{R}}\right\}$$

we get

$$\mathbb{E}\left[\sup_{\mathbf{z}\in\mathcal{Z}_D}\langle V(\mathbf{z}), \mathbf{x}_o - \mathbf{z}\rangle\right] = O\left(\frac{LD^2}{\sqrt{K}R} + \frac{D\sigma}{\sqrt{MKR}} + \frac{D^{\frac{4}{3}}\sigma^{\frac{2}{3}}L^{\frac{1}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}} + \frac{D^{\frac{4}{3}}\xi^{\frac{2}{3}}L^{\frac{1}{3}}}{R^{\frac{2}{3}}} + \frac{D(\xi\sigma)^{\frac{1}{2}}}{K^{\frac{1}{4}}R^{\frac{1}{2}}} + \frac{D\xi}{\sqrt{R}}\right)$$

H.2 Proof of Lemma 25

Following similar procedures as in Appendix A, we have for all $\mathbf{z} \in \mathbb{R}^d$,

$$\|\overline{\mathbf{z}}_t - \mathbf{z}\|^2 \le \|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 - 2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle + \eta^2\left\|\widetilde{V}_{\mathbf{x},t} - \widetilde{V}_{\mathbf{z},t-1}\right\|^2.$$

For the last term of the above, note that $V(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^M V_m(\mathbf{x})$ is L-smooth. We have

$$\begin{aligned}
&\left\|\widetilde{V}_{\mathbf{x},t} - \widetilde{V}_{\mathbf{z},t-1}\right\|^2 \\
\le& 5\left\|\widetilde{V}_{\mathbf{x},t} - V_{\mathbf{x},t}\right\|^2 + \frac{5L^2}{M}\sum_{m=1}^M\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 + 5L^2\|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 + \frac{5L^2}{M}\sum_{m=1}^M\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2 \\
&+ 5\left\|V_{\mathbf{z},t-1} - \widetilde{V}_{\mathbf{z},t-1}\right\|^2.
\end{aligned}$$

For the third term, we have

$$\begin{aligned}
&-2\eta\left\langle \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle = -\frac{2\eta}{M}\sum_{m=1}^M\langle V_m(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{z}\rangle + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle \\
=& -\frac{2\eta}{M}\sum_{m=1}^M\langle V_m(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle \underbrace{-\frac{2\eta}{M}\sum_{m=1}^M\langle V(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle}_{=0} + 2\eta\left\langle V_{\mathbf{x},t} - \widetilde{V}_{\mathbf{x},t}, \overline{\mathbf{x}}_t - \mathbf{z}\right\rangle \\
&+ \frac{2\eta}{M}\sum_{m=1}^M\langle V_m(\overline{\mathbf{x}}_t) - V_m(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle + \frac{2\eta}{M}\sum_{m=1}^M\langle V(\overline{\mathbf{x}}_t) - V_m(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle.
\end{aligned}$$

By the L-smoothness of each V_m and Cauchy-Schwarz, we have

$$\frac{2\eta}{M}\sum_{m=1}^M\langle V_m(\overline{\mathbf{x}}_t) - V_m(\mathbf{x}_t^m), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle \le \frac{2\eta L}{M}\sum_{m=1}^M\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|^2.$$

By Assumption 2,

$$\frac{2\eta}{M}\sum_{m=1}^M\langle V(\overline{\mathbf{x}}_t) - V_m(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t - \mathbf{x}_t^m\rangle \le \frac{2\eta\xi}{M}\sum_{m=1}^M\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|.$$

Plugging all bounds into the potential function inequality and rearranging, we get

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^M\langle V_m(\mathbf{x}_t^m), \mathbf{x}_t^m - \mathbf{z}\rangle \le \frac{\|\overline{\mathbf{z}}_{t-1} - \mathbf{z}\|^2 - \|\overline{\mathbf{z}}_t - \mathbf{z}\|^2}{2\eta} + \frac{(5L^2\eta^2 - 1)}{2\eta}\|\overline{\mathbf{x}}_t - \overline{\mathbf{z}}_{t-1}\|^2 \\
&+ \Gamma_t + \frac{5L^2\eta}{2M}\sum_{m=1}^M\|\mathbf{z}_{t-1}^m - \overline{\mathbf{z}}_{t-1}\|^2 + \frac{(5L^2\eta + 2L)}{2M}\sum_{m=1}^M\|\mathbf{x}_t^m - \overline{\mathbf{x}}_t\|^2 + \frac{2\xi}{M}\sum_{m=1}^M\|\overline{\mathbf{x}}_t - \mathbf{x}_t^m\|.
\end{aligned}$$

Summing over $t = 1, \ldots, T$, dividing both sides by T, taking expectation, and using $\eta \le \frac{1}{\sqrt{14}L}$, we obtain the desired result.

H.3 Proof of Lemma 26

We first bound the drift for $\mathbf{x}$. We have

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] = \mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} - \eta\left(\widetilde{V}_m(\mathbf{z}_{t-1}^m) - \widetilde{V}_{m'}(\mathbf{z}_{t-1}^{m'})\right)\right\|^2\right] \\
=&\mathbb{E}\Bigg[\Bigg\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} - \eta\Big(\widetilde{V}_m(\mathbf{z}_{t-1}^m) - V_m(\mathbf{z}_{t-1}^m) + V_m(\mathbf{z}_{t-1}^m) - V(\mathbf{z}_{t-1}^m) + V(\mathbf{z}_{t-1}^m) - V(\mathbf{z}_{t-1}^{m'}) \\
&+ V(\mathbf{z}_{t-1}^{m'}) - V_{m'}(\mathbf{z}_{t-1}^{m'}) + V_{m'}(\mathbf{z}_{t-1}^{m'}) - \widetilde{V}_{m'}(\mathbf{z}_{t-1}^{m'})\Big)\Bigg\|^2\Bigg] \\
\leq&(1+\eta^2L^2)\mathbb{E}\left[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2\right] + 12\eta^2\sigma^2 + 12\eta^2\xi^2 \\
\leq&2\mathbb{E}\left[\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\|^2\right] + 12\eta^2\sigma^2 + 12\eta^2\xi^2,
\end{aligned}
$$

where in the first inequality we used the monotonicity of V and Cauchy-Schwarz, and the last inequality follows from $\eta \leq \frac{1}{L}$. Next, we bound the drift for $\mathbf{z}$. We have

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2] = \mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'} - \eta\left(\widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right)\right\|^2\right] \\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta\mathbb{E}\left[\left\langle \mathbf{x}_t^m - \mathbf{x}_t^{m'}, \widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right\rangle\right] \\
&- 2\eta\mathbb{E}\left[\left\langle \mathbf{z}_{t-1}^m - \mathbf{x}_t^m + \mathbf{x}_t^{m'} - \mathbf{z}_{t-1}^{m'}, \widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right\rangle\right] \\
=&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \eta^2\mathbb{E}\left[\left\|\widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right\|^2\right] - 2\eta\mathbb{E}\left[\left\langle \mathbf{x}_t^m - \mathbf{x}_t^{m'}, V(\mathbf{x}_t^m) - V(\mathbf{x}_t^{m'})\right\rangle\right] \\
&- 2\eta\mathbb{E}\left[\left\langle \mathbf{x}_t^m - \mathbf{x}_t^{m'}, (V_m(\mathbf{x}_t^m) - V(\mathbf{x}_t^m)) - (V_{m'}(\mathbf{x}_t^{m'}) - V(\mathbf{x}_t^{m'}))\right\rangle\right] \\
&- 2\eta\underbrace{\mathbb{E}\left[\left\langle \mathbf{x}_t^m - \mathbf{x}_t^{m'}, (\widetilde{V}_m(\mathbf{x}_t^m) - V_m(\mathbf{x}_t^m)) - (\widetilde{V}_{m'}(\mathbf{x}_t^{m'}) - V_{m'}(\mathbf{x}_t^{m'}))\right\rangle\right]}_{=0} \\
&- 2\eta^2\mathbb{E}\left[\left\langle \widetilde{V}_m(\mathbf{z}_{t-1}^m) - \widetilde{V}_{m'}(\mathbf{z}_{t-1}^{m'}), \widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right\rangle\right] \\
\leq&\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 2\eta^2\mathbb{E}\left[\left\|\widetilde{V}_m(\mathbf{x}_t^m) - \widetilde{V}_{m'}(\mathbf{x}_t^{m'})\right\|^2\right] + \frac{1}{K}\mathbb{E}\left[\left\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\right\|^2\right] + 4K\eta^2\xi^2 \\
&+ \eta^2\mathbb{E}\left[\left\|\widetilde{V}_m(\mathbf{z}_{t-1}^m) - \widetilde{V}_{m'}(\mathbf{z}_{t-1}^{m'})\right\|^2\right],
\end{aligned}
$$

To proceed, note that for any $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^d$,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{V}_m(\mathbf{u}) - \widetilde{V}_{m'}(\mathbf{u}')\right\|^2\right] \\
\leq&5\mathbb{E}\left[\|\widetilde{V}_m(\mathbf{u}) - V_m(\mathbf{u})\|^2\right] + 5\mathbb{E}\left[\|V_m(\mathbf{u}) - V(\mathbf{u})\|^2\right] + 5\mathbb{E}\left[\|V(\mathbf{u}) - V(\mathbf{u}')\|^2\right] + 5\mathbb{E}\left[\|V(\mathbf{u}') - V_{m'}(\mathbf{u}')\|^2\right] \\
&+ 5\mathbb{E}\left[\|V_{m'}(\mathbf{u}') - \widetilde{V}_{m'}(\mathbf{u}')\|^2\right] \leq 5L^2\mathbb{E}\left[\|\mathbf{u} - \mathbf{u}'\|^2\right] + 10\sigma^2 + 10\xi^2.
\end{aligned}
$$

Thus,

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{z}_t^m - \mathbf{z}_t^{m'}\|^2] \le \mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \frac{1}{K}\mathbb{E}\left[\left\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\right\|^2\right] + 4K\eta^2\xi^2 \\
&\quad + \eta^2\mathbb{E}\left[5L^2\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2 + 10\sigma^2 + 10\xi^2\right] + 2\eta^2\mathbb{E}\left[5L^2\left\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\right\|^2 + 10\sigma^2 + 10\xi^2\right] \\
&\le \left(1 + 5\eta^2L^2\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + \left(\frac{1}{K} + 10\eta^2L^2\right)\mathbb{E}\left[\left\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\right\|^2\right] + 30\eta^2\sigma^2 + 30\eta^2\xi^2 + 4K\eta^2\xi^2 \\
&\le \left(1 + 5\eta^2L^2\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 30\eta^2\sigma^2 + 34K\eta^2\xi^2 \\
&\quad + \left(\frac{1}{K} + 10\eta^2L^2\right)\left(2\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 12\eta^2\sigma^2 + 12\eta^2\xi^2\right) \\
&\le \left(1 + \frac{3}{K} + 25\eta^2L^2\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 50\eta^2\sigma^2 + 50K\eta^2\xi^2 \\
&\le \left(1 + \frac{4}{K}\right)\mathbb{E}\left[\left\|\mathbf{z}_{t-1}^m - \mathbf{z}_{t-1}^{m'}\right\|^2\right] + 50\eta^2\sigma^2 + 50K\eta^2\xi^2 \le 50e^4\eta^2\sigma^2K + 50e^4\eta^2\xi^2K^2,
\end{aligned}$$

where the second-to-last inequality requires $\eta \le \frac{1}{\sqrt{25KL}}$, and the last inequality follows from unrolling the recursion and using $\mathbf{z}_0^m = \mathbf{z}_0^{m'}$ along with $(1 + \frac{4}{K})^K \le e^4$. Combining with the above, we have

$$\mathbb{E}[\|\mathbf{x}_t^m - \mathbf{x}_t^{m'}\|^2] \le 100e^4\eta^2\sigma^2K + 100e^4\eta^2\xi^2K^2.$$

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope. The faster rates for LESGD and its limitation are provided in Section 4.1, the proposed LIPPAX and SLIPPAX algorithms are given in Section 4.2, and the faster rates for federated composite variational inequalities are provided in Section 5.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitation is discussed in the end of Section 6.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions and detailed proof are clearly provided.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [N/A]

Justification: This paper does not include experiments.

Guidelines:

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

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

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

   Answer: [N/A]

   Justification: This paper does not include experiments requiring code.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: License: CC BY-NC-ND 4.0.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

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

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.