Federated Optimization with Doubly Regularized Drift Correction

Xiaowen Jiang [1 2] **Anton Rodomanov** [1] **Sebastian U. Stich** [1]

Abstract

Federated learning is a distributed optimization paradigm that allows training machine learning models across decentralized devices while keeping the data localized. The standard method, FedAvg, suffers from client drift which can hamper performance and increase communication costs over centralized methods. Previous works proposed various strategies to mitigate drift, yet none have shown consistently improved communication-computation trade-offs over vanilla gradient descent across all standard function classes. In this work, we revisit DANE, an established method in distributed optimization. We show that (i) DANE can achieve the desired communication reduction under Hessian similarity constraints. Furthermore, (ii) we present an extension, DANE+, which supports arbitrary inexact local solvers and has more freedom to choose how to aggregate the local updates. We propose (iii) a novel method, FedRed, which has improved local computational complexity and retains the same communication complexity compared to DANE/DANE+. This is achieved by doubly regularized drift correction.

1. Introduction

With the growing scale of datasets and the complexity of models, distributed optimization plays an increasingly important role in large-scale machine learning. Federated learning has emerged as an essential modern distributed learning paradigm where clients (e.g. phones and hospitals) collaboratively train a model without sharing their data, thereby ensuring a certain level of privacy (McMahan et al., 2017; Kairouz et al., 2021). However, privacy leaks can still occur (Nasr et al., 2019).

Tackling communication bottlenecks is one of the key challenges in modern federated optimization (Konečný et al., 2016; McMahan et al., 2017; Stich, 2019). The relatively slow and unstable internet connections of participating clients often make communication highly expensive, which might impact the effectiveness of the overall training process. Therefore, a standard metric to evaluate the efficiency of a federated optimization algorithm is the total number of communication rounds required to reach a certain accuracy.

FedAvg (McMahan et al., 2017) as the pioneering algorithm improves the communication efficiency by doing multiple local stochastic gradient descent updates before communicating to the server. This strategy has demonstrated significant success in practice. However, when the data is heterogeneous, FedAvg suffers from *client drift*, which might result in slow and unstable convergence (Karimireddy et al., 2020). Subsequently, some advanced techniques, including drift correction and regularization, have been proposed to address this issue (Karimireddy et al., 2020; Li et al., 2020; Acar et al., 2021; Mitra et al., 2021; Mishchenko et al., 2022). The communication complexity established in these methods often depends on the smoothness constant L of each function in general, which shows no advantage over the centralized gradient-based methods (see the first five rows in Table 1, Scaffnew has a similar complexity to the centralized fast gradient method under strong-convexity (Nesterov, 2018)).

Instead of proving the dependency on the smoothness constant L, a recent line of research tries to develop algorithms with guarantees that rely on a potentially smaller constant. Karimireddy et al. (2020) first demonstrated that when minimizing quadratics, Scaffold can exploit the hidden similarity to reduce communication. Specifically, the communication complexity only depends on a measure δ_B, which measures the maximum dissimilarity of Hessians among individual functions. Notably, δ_B is always at most L in order but can often be much smaller in practice. Subsequently, SONATA (Sun et al., 2022), its accelerated version (Tian et al., 2022), and Accelerated ExtraGradient sliding (Kovalev et al., 2022) show explicit communica-

[1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany [2]Universität des Saarlandes, Saarbrücken, Germany. Correspondence to: <{xiaowen.jiang, anton.rodomanov, stich}@cispa.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Algorithm	μ-strongly convex		General convex		Non-convex		Guarantee
	# comm rounds	# local steps at round r	# comm rounds	# local steps at round r	# comm rounds	# local steps at round r	
Centralized GD (Nesterov, 2018) [a]	$\mathcal{O}\big(\frac{L}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	1	$\mathcal{O}\big(\frac{LR_0^2}{\varepsilon}\big)$	1	$\mathcal{O}\big(\frac{LF_0}{\varepsilon}\big)$	1	deterministic
Scaffold (Karimireddy et al., 2020) [b]	$\mathcal{O}\big(\frac{L}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	$\forall K \geq 1$ [c]	$\mathcal{O}\big(\frac{LR_0^2}{\varepsilon}\big)$	$\forall K \geq 1$	$\mathcal{O}\big(\frac{LF_0}{\varepsilon}\big)$	$\forall K \geq 1$	deterministic
FedDyn (Acar et al., 2021) [d]	$\mathcal{O}\big(\frac{L}{\mu}\log(\frac{R_0^2+A^2/L}{\varepsilon})\big)$	-	$\mathcal{O}\big(\frac{LR_0^2+A^2}{\varepsilon}\big)$	-	$\mathcal{O}\big(\frac{LF_0}{\varepsilon}\big)$	-	deterministic
InexactDane (Reddi et al., 2016) [e]	$\mathcal{O}\big(\frac{L}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	m [f]	$\mathcal{O}\big(\frac{LR_0^2}{\varepsilon}\big)$	m [f]	$\mathcal{O}\big(\frac{LF_0}{\varepsilon}\big)$	m [f]	deterministic
Scaffnew (Mishchenko et al., 2022) [g]	$\mathcal{O}\big(\sqrt{\frac{L}{\mu}}\log(\frac{R_0^2+\frac{H_0^2}{\mu L}}{\varepsilon})\big)$	$\sqrt{\frac{L}{\mu}}$	$\mathcal{O}\big(\frac{pLR_0^2+\frac{H_0^2}{Lp}}{\varepsilon}\big)$	$\frac{1}{p}$	unknown	unknown	in expectation
MimeMVR (Karimireddy et al., 2021) [h]	unknown [i]	unknown	unknown	unknown	$\mathcal{O}\big(\frac{\delta_B G F_0}{\sqrt{n}\varepsilon^{3/2}}+\frac{G^2}{n\varepsilon}+\frac{\delta_B F_0}{\varepsilon}\big)$	$\frac{L}{\delta_B}$	deterministic
SONATA (Sun et al., 2022) [d]	$\mathcal{O}\big(\frac{\delta_B}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	-	unknown	unknown	unknown	unknown	deterministic
CE-LGD (Patel et al., 2022) [j]	unknown	unknown	unknown	unknown	$\mathcal{O}\big(\frac{\delta_B F_0}{\varepsilon}\big)$	$\frac{L}{\delta_B}$	in expectation
SVRP-GD (Khaled and Jin, 2023) [k]	$\mathcal{O}\big((n+\frac{\delta_A^2}{\mu^2})\log(\frac{R_0^2}{\varepsilon})\big)$	$\mathcal{O}\Big(\frac{L\delta_A^2+\mu}{\mu\delta_A^2+\mu}\log\big(\frac{\max\{\mu^2 n/\delta_A^2,1\}}{\varepsilon}\big)\Big)$	unknown	unknown	unknown	unknown	in expectation
DANE (Shamir et al., 2014) (**this work**)	$\mathcal{O}\big(\frac{\delta_A}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	- [l]	$\mathcal{O}\big(\frac{\delta_A R_0^2}{\varepsilon}\big)$	-	-	-	deterministic
DANE+-GD (**this work**, Alg. 1)	$\mathcal{O}\big(\frac{\delta_A}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	$\mathcal{O}\Big(\frac{L}{\mu+\delta_A}\log\big(\frac{L}{\mu+\delta_A}(r+1)\big)\Big)$	$\mathcal{O}\big(\frac{\delta_A R_0^2}{\varepsilon}\big)$	$\mathcal{O}\Big(\frac{L}{\delta_A}\log\big(\frac{L}{\delta_A}(r+1)\big)\Big)$	$\mathcal{O}\big(\frac{\delta_B F_0}{\varepsilon}\big)$	$\mathcal{O}\Big(\frac{L}{\delta_B}R\Big)$ [m]	deterministic
FedRed-GD (**this work**, Alg. 3) [g]	$\mathcal{O}\big(\frac{\delta_A+\mu}{\mu}\log(\frac{R_0^2}{\varepsilon})\big)$	$\frac{L}{\delta_A}$	$\mathcal{O}\big(\frac{\delta_A R_0^2}{\varepsilon}\big)$	$\frac{L}{\delta_A}$	$\mathcal{O}\big(\frac{\delta_B F_0}{\varepsilon}\big)$	$\frac{L}{\delta_B}$	in expectation

[a] The smoothness parameter L for centralized GD can be as small as the Lipschitz constant of the global function f.
[b] To achieve $||\nabla f(\mathbf{x}^R)||^2 \leq \varepsilon$, Scaffold requires $\frac{L+\delta_B K}{\mu K}\log(\frac{R_0^2}{\varepsilon})$ communication rounds for strongly-convex quadratics, and $\frac{(L+\delta_B K)F_0}{K\varepsilon}$ communication rounds for convex quadratics, where K is the number of local steps.
[c] Scaffold allows to use any number of local steps K by choosing the stepsize to be inversely proportional to K.
[d] FedDyn and SONATA assume that the local subproblem can be solved exactly. '-' means there is no definition of local steps.
[e] InexactDane achieves communication reduction w.r.t δ_B when minimizing quadratics.
[f] We refer to the accuracy condition written in Algorithm 1 and the corresponding requirement of γ in the original paper Reddi et al. (2016).
[g] For Scaffnew and FedRed-GD, the column '# comm rounds' represents the expected number of total communications required to reach ε accuracy. The column '# of local steps at round r' is replaced with the expected number of local steps between two communications. The general convex result of Scaffnew is established in Theorem 11 in the RandProx paper (Condat and Richtárik, 2023).
[h] Mime assumes $\frac{1}{n}\sum_{i=1}^n||\nabla f_i(\mathbf{x})-\nabla f(\mathbf{x})||^2 \leq G^2$ for any $\mathbf{x}\in\mathbb{R}^d$ as it considers client sampling.
[i] 'unknown' means no theoretical results are established so far.
[j] The communication and computational complexity for the non-convex case is min-max optimal for minimizing twice-continuously differentiable smooth functions under δ_B-BHD using first-order gradient methods.
[k] SVRP, SVRS (Lin et al., 2024) and SABER (Mishchenko et al., 2024) aim at minimizing a different measure which is the total amount of information transmitted between the server and the clients.
[l] DANE uses exact local solvers.
[m] R is the number of communication rounds.

Table 1. Summary of convergence behaviors of the considered **non-accelerated** distributed algorithms where L and μ stand for the smoothness and strong-convexity parameters of each function f_i, δ_A, δ_B are defined in (2) and (1), $R_0^2 := ||\mathbf{x}^0-\mathbf{x}^\star||^2$, $F_0 := f(\mathbf{x}^0)-f^\star$, $H_0^2 := \frac{1}{n}\sum_{i=1}^n||\mathbf{h}_{i,0}-\nabla f_i(\mathbf{x}^\star)||^2$, and $A^2 := \frac{1}{Ln}\sum_{i=1}^n||\nabla f_i(\mathbf{x}^\star)||^2$. The suboptimality ε is defined via $||\hat{\mathbf{x}}^R-\mathbf{x}^\star||^2$, $f(\hat{\mathbf{x}}^R)-f^\star$ and $||\nabla f(\hat{\mathbf{x}}^R)||^2$ respectively for strongly-convex, general convex, and non-convex functions ($\hat{\mathbf{x}}^R$ is a certain output produced by the algorithm after R communications.)

tion reduction in terms of δ_B. For instance, the required communication rounds for SONATA to reach ε accuracy is $\mathcal{O}\big(\frac{\delta_B}{\mu}\log(R_0^2/\varepsilon)\big)$ which is $\frac{L}{\delta_B}$ smaller than that achieved by centralized GD (see line 7 of Table 1). Later, CE-LGD (Patel et al., 2022) also achieves improved communication complexity for smooth non-convex functions.

More recently, Khaled and Jin (2023) introduced the averaged Hessian dissimilarity constant δ_A, which can be even smaller than δ_B. They established communication complexity with respect to δ_A for SVRP under strong-convexity. Compared to SONATA, the communication rounds turn into $\mathcal{O}\big((n+\frac{\delta_A^2}{\mu^2})\log(R_0^2/\varepsilon)\big)$ (see line 9 of Table 1). However, the aim of SVRP is to reduce the total amount of information transmitted between the server and clients rather than to reduce the total number of communication rounds. This is orthogonal to the focus of this work.

So far, there are no methods that achieve provably good communication complexity with respect to δ_A for solving strongly-convex and general convex problems. In this work, we revisit DANE, an established method in distributed optimization. Previously, its communication efficiency has only been proved when minimizing quadratic functions (Shamir et al., 2014).

We show that this historically first foundational algorithm already achieves good theoretical bounds in terms of δ_A under function similarity. We further utilize this fact to design simple algorithms that are efficient in terms of both local computation and total communication complexity for minimizing more general functions.

Contributions. We make the following contributions:

- We identify the key mechanism that allows communication reduction for arbitrary functions when they share certain similarities. This technique, which we call *regularized drift correction*, has already been implicitly used in DANE (Shamir et al., 2014) and other algorithms.
- We establish improved communication complexity for DANE in terms of δ_A under both general convexity and strong convexity.
- We present a slightly extended framework DANE+ for DANE. DANE+ allows the use of inexact local solvers and arbitrary control variates. In addition, it has more freedom to choose how to aggregate the local updates. We provide a tighter analysis of the communication and local computation complexity for this framework, for arbitrary continuously differentiable functions. We show that DANE+ achieves deterministic communication re-

duction across all the cases.

- We propose a novel framework FedRed which employs *doubly regularized drift correction*. We prove that FedRed enjoys the same communication reduction as DANE+ but has improved local computational complexity. Specifically, when specifying GD as a local solver, FedRed may require fewer communication rounds than vanilla GD without incurring additional computational overhead. Consequently, FedRed demonstrates the effectiveness of taking standard local gradient steps for the minimization of smooth functions.

Table 1 summarizes our main complexity results. Compared with the first 8 algorithms, DANE, DANE+ and FedRed achieve communication complexity that depends on δ_A instead of L and δ_B for minimizing convex problems. Compared with SVRP, our rates for strongly-convex problems is $\mathcal{O}(\frac{\delta_A}{\mu})$ instead of $\mathcal{O}(\frac{\delta_A^2}{\mu^2}+n)$. Moreover, DANE+ and FedRed also achieve the rate of $\mathcal{O}(\frac{\delta_B F_0}{\varepsilon})$ for solving non-convex problems, which is the same as CE-LGD and is better than all the other methods. Furthermore, when specifying GD as a local solver, FedRed-GD is strictly better than DANE+-GD in terms of total local computations.

Related work. The seminal work DANE (Shamir et al., 2014) is the first distributed algorithm that shows communication reduction for minimizing quadratics when the number of clients is large. Follow-up works interpret DANE as a preconditioning method. They provide better analysis and derive faster rates for quadratic losses (Zhang and Lin, 2015; Yuan and Li, 2019; Wang et al., 2018). Hendrikx et al. (2020) propose SPAG, an accelerated method, and prove a better uniform concentration bound of the conditioning number when solving strongly-convex problems. All these methods assume the iteration-subproblems can be solved exactly or with certain second-order optimizers. Reddi et al. (2016) first analyze DANE under arbitrary local accuracy conditions with the guarantee generally depending on L. They also proposed an accelerated method.

On the other hand, federated optimization algorithms lean towards solving the subproblems inexactly by taking local updates. The standard FL method FedAvg has been shown ineffectiveness for convex optimization in the heterogeneous setting (Woodworth et al., 2020; Koloskova et al., 2020). The celebrated Scaffold (Karimireddy et al., 2020) adds control variate to cope with the client drift issues in heterogeneous networks. It is also the first work to quantify the usefulness of local-steps for quadratics. Afterwards, drift correction is employed in many other works such as FedDC (Gao et al., 2022) and Adabest (Varno et al., 2022). FedPVR (Li et al., 2023) propose to apply drift correction only to the last layers of neural networks. Scaffnew (Mishchenko et al., 2022) first illustrates the usefulness of taking standard local gradient steps under strongy-convexity by using a special choice of control variate. More advanced methods with refined analysis and features such as client sampling and compression have been proposed (Hu and Huang, 2023; Condat et al., 2023; Grudzień et al., 2023; Condat and Richtárik, 2023). Sadiev et al. (2022) proposed APDA that retains the same communication complexity as Scaffnew, but provably reduces the local computation complexity. FedProx (Li et al., 2020) adds a proximal regularization term to the local functions to control the deviation between client and server models. FedDyn (Acar et al., 2021) proposes dynamic regularization to improve the convergence of FedProx. FedPD (Zhang et al., 2021) applies both drift correction and regularization and design the algorithm from the primal-dual perspective.

A more closely related line of research focuses on the study of convergence guarantee under relaxed function similarity conditions. SONATA (Sun et al., 2022) and Accelerated SONATA (Tian et al., 2022) provide better communication guarantee than vanilla GD under strong convexity. Accelerated ExtraGradient sliding improves the rate of (Tian et al., 2022) and allows the use of inexact local solvers. Khaled and Jin (2023) work with a more relaxed similarity assumption and propose a centralized stochastic proximal point variance-reduced method called SVRP. Methods with better complexity have been proposed by Lin et al. (2024) and Mishchenko et al. (2024) where the latter works also in the non-convex settings. Mime combines drift correction with momentum for solving non-convex problems. CE-LGD (Patel et al., 2022) proposed to use recursive gradients together with momentum for solving non-convex distributed optimization problems.

2. Problem formulation and background

In this work, we consider the following distributed optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}f_i(\mathbf{x})\right].$$

We focus on the standard federated learning setting where n clients jointly train a model and a central server coordinates the global learning procedure. Specifically, during each communication round $r=0,1,2...$,

(i) the server broadcasts $\mathbf{x}^r$ to the clients;

(ii) each client $i\in[n]$ calls a local solver to compute the next iterate $\mathbf{x}_{i,r+1}$;

(iii) the server aggregates $\mathbf{x}^{r+1}$ using $(\mathbf{x}_{i,r+1})_{i=1}^n$.

Goal: We assume that establishing the connection between the server and the clients is expensive. Therefore, the main

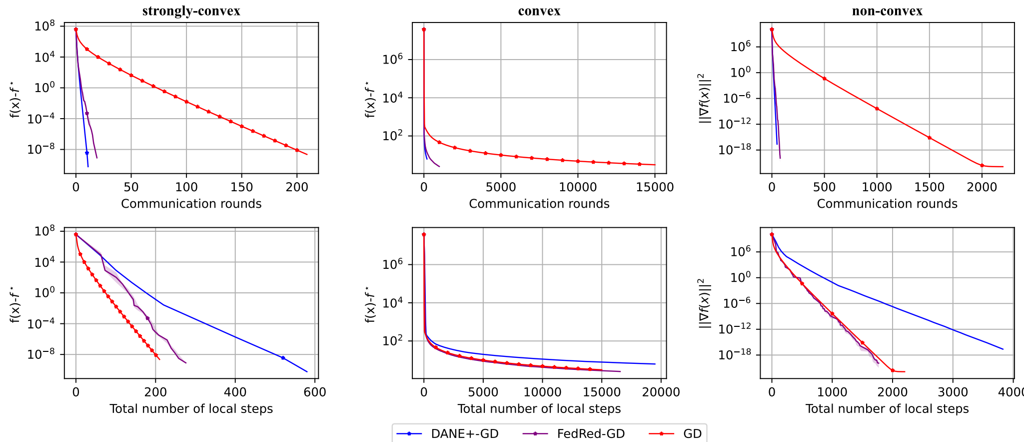

Figure 1. Illustrating communication reduction for DANE+-GD and FedRed-GD on synthetic dataset using quadratic loss with $\frac{L}{\delta_A} \approx \frac{L}{\delta_B} \gtrsim 20$. DANE+-GD and FedRed-GD require roughly 20 times fewer communication rounds to reach the same suboptimality as GD while the total number of local computations of FedRed-GD is at the same scale as GD. (Repeated 3 times for FedRed-GD. The solid lines and the shaded area represent the mean and the region between the minimum and the maximum values.)

goal is to reduce the total number of communication rounds required to reach a desired accuracy.

Notations: We use $\mathbf{x}^\star$ to denote $\arg\min_x\{f(\mathbf{x})\}$ and we assume it exists under convexity. We assume f is lower bounded and denote the infimum of f by $f^\star$.

Vanilla/Centralized GD. Gradient descent is a standard method that solves the problem. At each communication round r, each device computes its gradient $\nabla f_i(\mathbf{x}^r)$, and performs one local gradient descent step: $\mathbf{x}_{i,r+1} = \mathbf{x}^r - \eta\nabla f_i(\mathbf{x}^r)$. Then the server updates the global model by: $\mathbf{x}^{r+1} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,r+1}$. The update rule can thus be written as: $\mathbf{x}^{r+1} = \mathbf{x}^r - \eta\nabla f(\mathbf{x}^r)$. The convergence of this method is well understood (Nesterov, 2018).

Client drift. FedAvg or Local-(S)GD approximate the global function f by f_i on each device. Each device i runs multiple (stochastic) gradient descent steps initialized at $\mathbf{x}^r$, i.e. $\mathbf{x}_{i,r+1} \approx \arg\min_x f_i(\mathbf{x})$. Then $\mathbf{x}^{r+1} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,r+1}$ is aggregated by the server.

This method suffers from client drift when $\{f_i\}$ are heterogeneous (Karimireddy et al., 2020; Khaled et al., 2019; Koloskova et al., 2020) because the local updates move towards the minimizers of each f_i. In general, $\frac{1}{n}\sum_{i=1}^n \arg\min_x f_i(\mathbf{x}) \neq \mathbf{x}^\star$.

Regularization. To address the issue of client drift, FedProx (Li et al., 2020) proposes to add regularization to each local function f_i to control the local iterates from deviating from the current global state, i.e. $\mathbf{x}_{i,r+1} \approx \arg\min_x\{f_i(\mathbf{x}) + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2\}$. However, the non-convergence issue still exists. Suppose we initialize at $\mathbf{x}^\star$ ($\mathbf{x}^0 = \mathbf{x}^\star$), and local solvers return the exact minimizers. We have $\nabla f_i(\mathbf{x}_{i,1}) + \lambda(\mathbf{x}_{i,1} - \mathbf{x}^\star) = 0$, which implies that $\mathbf{x}^1 = \mathbf{x}^\star - \frac{1}{\lambda}\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_{i,1})$. Since $\mathbf{x}^1$ is typically not equal to $\mathbf{x}^\star$, $\mathbf{x}^\star$ is not a stationary point for FedProx.

Drift correction. Scaffold (Karimireddy et al., 2020) addresses the non-convergence issues by using drift correction. It approximates the global function by shifting each f_i based on the current difference of gradients, i.e. $\mathbf{x}_{i,r+1} \approx \arg\min_x\{f_i(\mathbf{x}) + \langle\nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r), \mathbf{x}\rangle\}$, and uses (S)GD to solve the subproblem. However, Scaffold requires much smaller stepsize than GD to ensure convergence. Consequently, the overall communication complexity is no better than GD. Moreover, the subproblem itself is not well-defined. For instance, if some f_i is a linear function, then the subproblem has no solution. Indeed, Scaffold can be viewed as a composite gradient descent method with stepsize set to be infinity, which we state as below.

Regularized drift correction. The appropriate solution to the previously mentioned issues is to incorporate both drift correction and regularization, a strategy already applied in DANE. Here, we provide a more intuitive explanation and cast it as a composite optimization method. Note that f can be written as $f = f_i + \psi$ where $\psi := f - f_i$. As each client has no access to ψ, the standard approach is to linearize ψ at $\mathbf{x}^r$ and perform the composite gradient method: $\mathbf{x}_{i,r+1} \approx \arg\min\{f_i(\mathbf{x}) + \langle\nabla\psi(\mathbf{x}^r), \mathbf{x}-\mathbf{x}^r\rangle + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2\}$, where $\lambda > 0$ should approximately be set to $||\nabla\psi^2(\mathbf{x}^r)||$. If $||\nabla\psi^2(\mathbf{x}^r)||$ is much smaller than the smoothness parameter of f, or equivalently if f_i and f have similar Hessians, then this approach can provably converge faster than GD. This motivates us to study the following algorithms which incorporate regularized drift correction to distributed optimization settings.

Algorithm 1 DANE+

Input: $\lambda \geq 0$, $\mathbf{x}^0 \in \mathbb{R}^d$
for $r = 0, 1, 2...$ **do**
  for each client $i \in [n]$ **in parallel do**
    Update $\mathbf{h}_{i,r}$
    Set $\mathbf{x}_{i,r+1} \approx \arg\min_{\mathbf{x}\in\mathbb{R}^d}\{F_{i,r}(\mathbf{x})\}$, where
    $F_{i,r}(\mathbf{x}) := f_i(\mathbf{x}) - \langle \mathbf{x}, \mathbf{h}_{i,r}\rangle + \frac{\lambda}{2}||\mathbf{x} - \mathbf{x}^r||^2$.
  end for
  Set $\mathbf{x}^{r+1} = \text{Average}(\mathbf{x}_{i,r+1})_{i=1}^n$,
  where Average$(\cdot)$ denotes an averaging strategy.
end for

3. DANE+ with regularized drift correction

In this section, we first describe a framework that generalizes DANE (Algorithm 1) for distributed optimization with regularized drift correction. We then discuss how it achieves communication reduction when the functions of clients exhibit certain second-order similarities.

Method. During each communication round r of DANE+, each client i is assigned a local objective function $F_{i,r}$ and returns an approximate solution for it by running a local solver. The function $F_{i,r}$ has three standard components: i) the individual loss function f_i, ii) an inner product term for drift correction, and iii) a regularizer that controls the deviation of the local iterates from the current global model. After that, the server applies an averaging strategy to aggregate all the iterates returned by the clients.

To run the algorithm, we need to choose a control variate $\mathbf{h}_{i,r}$ and an averaging method.

Karimireddy et al. (2020) proposed two choices of control variates for drift correction. In this work, we study the properties of the standard one which is defined as follows:

$$\mathbf{h}_{i,r} := \nabla f_i(\mathbf{x}^r) - \nabla f(\mathbf{x}^r) \ . \tag{1}$$

Another choice of control variate is also used in Scaffnew. In this work, we focus on Hessian Dissimilarity and defer the studies of that one for DANE+ to Appendix B.3.

There exist various ways of averaging vectors. Let $(\mathbf{x}_i)_{i=1}^n$ be a set of vectors where $\mathbf{x}_i \in \mathbb{R}^d$. The standard averaging is defined as follows.

$$\text{Avg}(\mathbf{x}_i)_{i=1}^n := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \ .$$

In this work, we also work on randomized averaging:

$$\text{Rand}(\mathbf{x}_i)_{i=1}^n := \mathbf{x}_{\hat{j}} \ .$$

where $\hat{j}$ denotes a random variable that follows a fixed probability distribution at each iteration.

Suppose that control variate (1) is used, each local solver provides the exact solution, and that standard averaging is applied, then DANE+ is equivalent to DANE.

Communication Reduction. As discussed in Section 2, regularized drift correction can reduce the communication given that the functions $\{f_i\}$ are similar. A reasonable measure is to look at the second-order Hessian dissimilarity (Karimireddy et al., 2021; 2020; Zindari et al., 2023; Patel et al., 2024), which is defined as follows:

Definition 1 (Bounded Hessian Dissimilarity (Karimireddy et al., 2020)). *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$, then $\{f_i\}$ have δ_B-BHD if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and any $i \in [n]$,*

$$||\nabla h_i(\mathbf{x}) - \nabla h_i(\mathbf{y})|| \leq \delta_B ||\mathbf{x} - \mathbf{y}|| \ . \tag{2}$$

where $h_i := f_i - f$, and $f := \frac{1}{n}\sum_{i=1}^n f_i$.

The standard Bounded Hessian Dissimilarity defined in (Karimireddy et al., 2020) assumes $||\nabla h_i^2(\mathbf{x})|| \leq \delta_B$ for any $\mathbf{x} \in \mathbb{R}^d$, which implies Definition 1 and is thus slightly stronger. Here, we only need first-order differentiability.

We also work on a weaker Hessian Dissimilarity definition.

Definition 2 (Averaged Hessian Dissimilarity (Khaled and Jin, 2023)). *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$, then $\{f_i\}$ have δ_A-AHD if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$\frac{1}{n}\sum_{i=1}^n ||\nabla h_i(\mathbf{x}) - \nabla h_i(\mathbf{y})||^2 \leq \delta_A^2 ||\mathbf{x} - \mathbf{y}||^2 \ . \tag{3}$$

where $h_i := f_i - f$ and $f := \frac{1}{n}\sum_{i=1}^n f_i$.

Even if each f_i is twice-continuously differentiable, Definition 2 is still weaker than assuming $\frac{1}{n}\sum_{i=1}^n ||\nabla h_i^2(\mathbf{x})|| \leq \delta_A^2$ for any $\mathbf{x} \in \mathbb{R}^d$. The details can be found in Remark 18.

Suppose each f_i is L-smooth, then $\delta_A \leq 2L$ and $\delta_B \leq 2L$. If further assuming each f_i is convex, then $\delta_A \leq L$ and $\delta_B \leq L$. In practice, we expect the clients to be similar and thus $\delta_A \ll L$ and $\delta_B \ll L$ (Karimireddy et al., 2021). In principle, if each h_i is δ_i-smooth, then $\delta_A = (\frac{1}{n}\sum_i \delta_i^2)^{1/2}$ can potentially be much smaller (up to $\sqrt{n}$ times) than $\delta_B = \max_i\{\delta_i\}$.

3.1. Theory

We state the convergence rates of DANE+ in this section. All the proofs can be found in Appendix B.

Nonconvex. When the local functions $\{f_i\}$ are nonconvex, it becomes difficult to obtain an approximate minimizer of $F_{i,r}$. Instead, it is more realistic to assume that each local solver can converge to an approximate stationary point. In this case, the standard averaging may be less effective than randomized averaging, especially when $\{f_i\}$ are similar. The intuition is as follows.

Suppose that $f_i = f$ for any $i \in [n]$. Then $F_{i,r}(\mathbf{x}) \equiv f(\mathbf{x})$ if we set $\lambda = 0$. Assume that each local solver can com-

Algorithm 2 FedRed: **Fe**derated optimization framework with **d**oubly **Re**gularized **d**rift correction

1: **Input:** $p \in [0,1]$, $\lambda, \eta \geq 0$, $\mathbf{x}^0 \in \mathbb{R}^d$
2: Set $\mathbf{x}_{i,0} = \tilde{\mathbf{x}}_0 = \mathbf{x}^0$, $\forall i \in [n]$.
3: **for** $k = 0, 1, 2...$ **do**
4: **for each client** $i \in [n]$ **in parallel do**
5: Update $\mathbf{h}_{i,k}$
6: $$\mathbf{x}_{i,k+1} = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \left\{ F_{i,k}(\mathbf{x}) \right\} \; .^{a}$$
7: where $$F_{i,k}(\mathbf{x}) := f_i(\mathbf{x}) - \langle \mathbf{x}, \mathbf{h}_{i,k} \rangle + \frac{\eta}{2} ||\mathbf{x} - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2} ||\mathbf{x} - \tilde{\mathbf{x}}_k||^2 \; .$$
8: Sample $\theta_k \in \{0,1\}$ where $\text{Prob}(\theta_k = 1) = p$
9: $$\tilde{\mathbf{x}}_{k+1} = \begin{cases} \text{Average}(\mathbf{x}_{i,k+1})_{i=1}^n & \text{if } \theta_k = 1 \\ \tilde{\mathbf{x}}_k & \text{otherwise} \end{cases} \; .$$
10: **end for**
11: **end for**

[a] When $F_{i,k}$ is non-convex, it is sufficient to return a point such that $\nabla F_{i,k}(\mathbf{x}_{i,k+1}) = 0$ and $F_{i,k}(\mathbf{x}_{i,k+1}) \leq F_{i,k}(\mathbf{x}_{i,k})$.

Algorithm 3 FedRed-(S)GD

1: The same as Algorithm 2, except on line 6
$$F_{i,k}(\mathbf{x}) := f_i(\mathbf{x}_{i,k}) + \langle \mathbf{g}_i(\mathbf{x}_{i,k}), \mathbf{x} - \mathbf{x}_{i,k} \rangle - \langle \mathbf{x}, \mathbf{h}_{i,k} \rangle + \frac{\eta}{2} ||\mathbf{x} - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2} ||\mathbf{x} - \tilde{\mathbf{x}}_k||^2 \; ,$$
where $\mathbf{g}_i(\mathbf{x}, \xi_i)$ is a stochastic estimator of $\nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$.

pute an exact stationary point. Then it is more reasonable to update the global model by selecting an update $\mathbf{x}_{i,r+1}$ rather than the standardally averaging all the iterates, as the averaged stationary points can be even worse. That being said, the standard averaging can still be effective and more stable if each client uses the same local solver. We leave this possibility in the future work.

In the following, we show that it is sufficient to arbitrarily choose a model from $(\mathbf{x}_{i,r+1})_{i=1}^n$ to update the global model. For instance, $\mathbf{x}^{r+1} = \text{Rand}(\mathbf{x}_{i,r+1})_{i=1}^n$. In practice, this index set can be predetermined and during each round, only a single client is required to do local updates.

Theorem 1. *Consider Algorithm 1 with control variate (1). Let the global model be updated by choosing an arbitrary local model for each communication round. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Suppose that the solutions returned by local solvers satisfy $F_{i,r}(\mathbf{x}_{i,r+1}) \leq F_{i,r}(\mathbf{x}^r)$ and $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \leq e_{r+1}$ with $e_{r+1} \geq 0$ for any $r \geq 0$ and any $i \in [n]$. Let $\lambda = a\delta_B$ with $a > 1$ Then after R communication rounds, we have:*
$$||\nabla f(\bar{\mathbf{x}}^R)||^2 \leq \frac{4(a+1)^2}{(a-1)} \frac{\delta_B(f(\mathbf{x}^0) - f^\star)}{R} + 2\frac{1}{R}\sum_{r=1}^{R} e_r^2 \; .$$
where $\bar{\mathbf{x}}^R = \arg\min_{\mathbf{x} \in \{\mathbf{x}^r\}_{r=0}^R} \{||\nabla f(\mathbf{x})||\}$.

Theorem 1 gives a convergence guarantee of DANE+ for arbitrary functions (not necessarily with Lipschitz gradient) with any local solvers. The minimal squared gradient norm decreases at a rate of $\mathcal{O}(\frac{\delta_B(f(\mathbf{x}^0)-f^\star)}{R})$. To reach arbitrary accuracy ε, the average errors should be at most ε. We next establish the local computational complexity assuming that the gradient of each f_i is Lipschitz.

Corollary 2. *Consider Algorithm 1 with control variate (1). Let the global model be updated by choosing an arbitrary local model for each communication round. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Suppose each local solver runs the standard gradient descent starting from $\mathbf{x}^r$ for $\Theta\left(\frac{L}{\delta_B} R\right)$ local steps, for any $r \geq 0$, and returns the point with the minimum gradient norm. Let $\lambda = 2\delta_B$. After R communication rounds, we have:*
$$||\nabla f(\bar{\mathbf{x}}^R)||^2 \leq \frac{72\delta_B(f(\mathbf{x}^0) - f^\star)}{R} \; .$$
where $\bar{\mathbf{x}}^R = \arg\min_{\mathbf{x} \in \{\mathbf{x}^r\}_{r=0}^R} \{||\nabla f(\mathbf{x})||\}$.

The communication complexity of DANE+-GD is reduced by a factor of $\frac{L}{\delta_B}$ compared with GD. However, the total gradient computations to reach ε accuracy is $R = \Theta\left(\frac{\delta_B(f(\mathbf{x}^0)-f^\star)}{\varepsilon}\right)$ times worse than GD.

Convex. In contrast to non-convex problems, convex optimization can benefit from the standard averaging. This allows us to obtain the following convergence guarantee depending on the Averaged Hessian Dissimilarity.

Theorem 3. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. In general, suppose that the solutions returned by local solvers satisfy $\frac{1}{n}\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq e_{r+1}^2$ for any $r \geq 0$ with $e_{r+1} \geq 0$. Let $\lambda \geq 2\delta_A$. After R communication rounds, we have:*

$$f(\bar{\mathbf{x}}^R) - f(\mathbf{x}^\star) + \frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,R} - \mathbf{x}^\star||^2 \leq \frac{\mu}{(1+\frac{\mu}{\lambda})^R - 1}||\mathbf{x}^0 - \mathbf{x}^\star||^2 + \frac{2}{\mu+\lambda}\sum_{r=1}^R e_r^2 \ .$$

where $\bar{\mathbf{x}}^R := \arg\min_{\mathbf{x}\in\{\mathbf{x}^r\}_{r=1}^R} f(\mathbf{x})$.

Theorem 3 provides the convergence guarantee for DANE+ when the local functions are convex (and not necessarily smooth). Each local solver can be chosen arbitrarily, provided that it can return a solution that satisfies a specific accuracy condition. The convergence rate is a continuous estimate both in μ and λ, since $\frac{\mu}{(1+\frac{\mu}{\lambda})^R - 1} \leq \frac{\lambda}{R}$. To reach a certain accuracy ε, the errors should satisfy $\sum_{r=1}^R e_r^2 \leq \frac{\varepsilon(\mu+\lambda)}{2}$. We next show that if the solutions returned by local solvers satisfy more specific conditions, then the additional error term $\sum_{r=1}^R e_r^2$ can be removed.

Theorem 4. *Under the same conditions as Theorem 3, suppose that the solutions returned by local solvers satisfy $\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq e_r^2 \sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$ for any $r \geq 0$ and $e_r \geq 0$. Let $\lambda \geq 2\delta_A$ and let $\sum_{r=0}^{+\infty} e_r^2 \leq \frac{\lambda(\mu+\lambda)}{8}$. After R communication rounds:*

$$f(\bar{\mathbf{x}}^R) - f(\mathbf{x}^\star) + \frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,R} - \mathbf{x}^\star||^2 \leq \frac{\mu}{[(1+\frac{\mu}{\lambda})^R - 1]}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \leq \frac{\lambda}{R}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \ .$$

where $\bar{\mathbf{x}}^R := \arg\min_{\mathbf{x}\in\{\mathbf{x}^r\}_{r=1}^R} f(\mathbf{x})$.

In practice, it is sufficient to stop running the local solver on each device once the solutions satisfy $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \leq e_r ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||$ with $e_r^2 := \frac{\lambda(\mu+\lambda)}{8(r+1)(r+2)}$. As a direct corollary of Theorem 4, we can estimate the required number of local iterations for standard first-order methods assuming that each function f_i is L-smooth.

Corollary 5. *Consider Algorithm 1 with control variate (1) and standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$, and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD with $L \geq \delta_A + \mu$. Suppose for any $r \geq 0$, each device uses the standard (fast) gradient descent initialized at $\mathbf{x}^r$ until $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq \frac{\lambda(\mu+\lambda)}{8(r+1)(r+2)}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$ is satisfied. Let $\lambda = 2\delta_A$. To have the same convergence guarantee as in Theorem 4, the total number of local steps K_r required at each round r is no more than:*

DANE+-GD:

$$K_r \leq \Theta\left(\frac{L}{\delta_A + \mu}\ln\left(\frac{L}{\delta_A+\mu}(r+1)\right)\right) .$$

DANE+-FGD:

$$K_r \leq \Theta\left(\sqrt{\frac{L}{\delta_A + \mu}}\ln\left(\frac{L}{\delta_A+\mu}(r+1)\right)\right) .$$

The total computational complexity of DANE+-GD is equivalent to that of the centralized gradient descent method, up to a logarithmic factor that depends on r.

4. FedRed: an optimization framework with Doubly Regularized Drift Correction

While DANE+ can achieve communication speed-up, the local computations are inefficient. This is because, for every communication round, a difficult subproblem has to be approximately solved. In this section, we present a general framework that improves the overall computational efficiency while maintaining the communication reduction.

Key idea: Let us add an additional regularizer to improve the conditioning of the subproblem. FedRed 2 retains two sequences throughout the iteration: $\{\mathbf{x}_{i,k}\}$ denote the iterates stored on each device, and $\{\tilde{\mathbf{x}}_k\}$ are the reference points. At each iteration k, FedRed adds an additional regularizer to the proxy function defined in DANE+. By setting η to be a large number (e.g. $2L$ where L is the Lipschitz constant), $F_{i,k}$ can become a function with a good condition number (an absolute constant) and thus can be solved efficiently in a constant (up to the logarithmic factor) number of steps.

However, strong regularization prevents aggressive progress of the algorithm. Therefore, the communication is only triggered with probability p, with the reference points being adjusted to the averaged $(\mathbf{x}_{i,k+1})_{i=1}^n$. Lastly, the control variates $\{\mathbf{h}_{i,k}\}$ are updated accordingly. We study the same control variate as defined in (1):

$$\mathbf{h}_{i,k} := \nabla f_i(\tilde{\mathbf{x}}_k) - \nabla f(\tilde{\mathbf{x}}_k) \ . \tag{4}$$

Note that DANE+ becomes a special case of FedRed when $p = 1$ and $\eta = 0$. We next show the convergence rates for FedRed using exact local solvers.

4.1. FedRed with Exact Solvers

Theorem 6. *Consider Algorithm 2 with control variate (4) and randomized averaging with random index set $\{i_k\}_{k=0}^{+\infty}$.*

Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Let $\lambda = \delta_B$, $p = \frac{\lambda}{\eta}$ and $\eta \geq 4\delta_B$. For any $K \geq 1$, it holds that:

$$\mathbb{E}\Big[||\nabla f(\bar{\mathbf{x}}_K)||^2\Big] \leq \frac{150\eta(f(\mathbf{x}^0) - f^\star)}{K} ,$$

where $\bar{\mathbf{x}}_K$ is uniformly sampled from $(\mathbf{x}_{i_k,k})_{k=0}^{K-1}$.

To reach ε-accuracy, the communication complexity is $p\mathcal{O}(\frac{\eta}{K}) = \mathcal{O}(\frac{\delta_B}{\varepsilon})$ in expectation, which is the same as DANE+. However, due to the additional regularization, the subproblem becomes easier to solve. The same conclusion can also be derived for the convex settings.

Theorem 7. *Consider Algorithm 2 with control variate (4) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. By choosing $p = \frac{\lambda+\mu/2}{\eta+\mu/2}$ and $\eta \geq \lambda \geq \delta_A$, for any $K \geq 1$, it holds that:*

$$\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] + \frac{\mu}{4}\mathbb{E}\Big[||\tilde{\mathbf{x}}_K - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,K} - \mathbf{x}^\star||^2\Big]$$
$$\leq \frac{\mu}{2}\frac{||\mathbf{x}_0 - \mathbf{x}^\star||^2}{(1+\frac{\mu}{2\eta})^K - 1} \leq \frac{\eta||\mathbf{x}_0 - \mathbf{x}^\star||^2}{K} .$$

where $\bar{\bar{\mathbf{x}}}_K := \sum_{k=1}^K \frac{1}{q^k}\bar{\mathbf{x}}_k / \sum_{k=1}^K \frac{1}{q^k}$, $\bar{\mathbf{x}}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,k}$, and $q := 1 - \frac{\mu}{2\eta+\mu}$.

Comparison to existing algorithms. We now compare FedRed with two similar algorithms. SVRP (Khaled and Jin, 2023) is a stochastic proximal point variance-reduced method. As discussed in the introduction, the main focus of SVRP is different in this paper. The main algorithmic differences between SVRP and FedRed are that: 1) FedRed uses double regularization and thus the regularized centering points are different 2) For SVRP, communication happens at each iteration , while FedRed skips communication with probability $1 - p$, and 3) FedRed allows standard averaging for convex optimization. Another similar algorithm is FedPD (Zhang et al., 2021) where regularized drift correction and the strategy of skipping communication with probability are used. However, due to the different regularizations, FedPD is less efficient in local computations.

4.2. FedRed-(S)GD

Previous theorems assume that the exact minimizers (or stationary points) of the subproblems are returned at each iteration. In this section, we show that, in fact, it suffices to make only one local step for minimizing smooth functions.

Note the derivation of Algorithm 3 is to simply linearize f_i at $\mathbf{x}_{i,k}$. Therefore, the solution of the subproblem has a closed form: $\mathbf{x}_{i,k+1} = \frac{\eta}{\eta+\lambda}\mathbf{x}_{i,k} + \frac{\lambda}{\eta+\lambda}\tilde{\mathbf{x}}_k - \frac{1}{\eta+\lambda}(\mathbf{g}_i(\mathbf{x}_{i,k}) - \mathbf{h}_{i,k})$, which is a convex combination of $\mathbf{x}_{i,k}$ and $\tilde{\mathbf{x}}_k$ followed by a gradient descent step.

Here we allow the use of a stochastic gradient $\mathbf{g}_i(\mathbf{x}, \xi_i)$ instead of the full gradient $\nabla f_i(\mathbf{x})$. While more extensions could be made such as approximating $\mathbf{h}_{i,k}$ using stochastic gradients, we do not attempt to be exhaustive here as the goal is to illustrate the communication reduction. We make the following standard assumptions for $\mathbf{g}_i(\mathbf{x}, \xi_i)$.

Assumption 1. *For any $i \in [n]$, $\mathbf{g}_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic estimator of $\nabla f_i(\mathbf{x})$ with bounded variance σ^2 such that for any $\mathbf{x} \in \mathbb{R}^d$, it holds that:*

$$\mathbb{E}[\mathbf{g}_i(\mathbf{x})] = \nabla f_i(\mathbf{x}), \quad \mathbb{E}[||\mathbf{g}_i(\mathbf{x}) - \nabla f_i(\mathbf{x})||^2] \leq \sigma^2. \quad (5)$$

Theorem 8. *Consider Algorithm 3 with control variate (4) and randomized averaging with random index set $\{i_k\}_{k=0}^{+\infty}$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD with $\delta_B \leq L$ and that Assumption 1 holds. By choosing $\eta = 3L + \sqrt{9L^2 + \frac{L\sigma^2 K}{f(\mathbf{x}^0) - f^\star}}$, $\lambda = \delta_B$ and $p = \frac{\delta_B}{L}$, for any $K \geq 1$, it holds that:*

$$\mathbb{E}\big[||\nabla f(\bar{\mathbf{x}}_K)||^2\big] \leq \frac{96L(f(\mathbf{x}^0) - f^\star)}{K} + 24\sqrt{\frac{L(f(\mathbf{x}^0) - f^\star)}{K}}\sigma . \quad (6)$$

where $\bar{\mathbf{x}}_K$ is uniformly sampled from $(\mathbf{x}_{i_k,k})_{k=0}^{K-1}$.

According to Theorem 8, FedRed-GD (using full-batch gradient) requires the same total gradient computations as gradient descent to reach a certain accuracy. However, in expectation, it can skip communication for every $\Theta(\frac{L}{\delta_B})$ steps without additional cost.

Theorem 9. *Consider Algorithm 3 with control variate (4) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$ and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD with $\delta_A + \mu \leq L$ and that Assumption 1 holds. Let $\lambda = \delta_A$, $p = \frac{\lambda+\mu/2}{\eta-\mu/2}$, and $\eta > L$. To reach ε-accuracy, i.e. $\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] \leq \varepsilon$, by choosing $\eta = \frac{\sigma^2}{\varepsilon} + L$, the total number of iterations is no more than:*

$$K \leq \left\lceil \Big(\frac{2L}{\mu} + \frac{2\sigma^2}{\mu\varepsilon}\Big)\ln\Big(1 + \frac{\mu||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon}\Big)\right\rceil .$$

where $\bar{\bar{\mathbf{x}}}_K := \sum_{k=1}^K \frac{1}{q^k}\bar{\mathbf{x}}_k / \sum_{k=1}^K \frac{1}{q^k}$, $\bar{\mathbf{x}}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,k}$, and $q := 1 - \frac{\mu}{2\eta-\mu}$.

Since $\Big(\frac{2L}{\mu} + \frac{2\sigma^2}{\mu\varepsilon}\Big)\ln\Big(1 + \frac{\mu||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon}\Big) \leq \frac{2L||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon} + \frac{2\sigma^2||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon^2}$, the estimate of K is continuous in μ. Let us

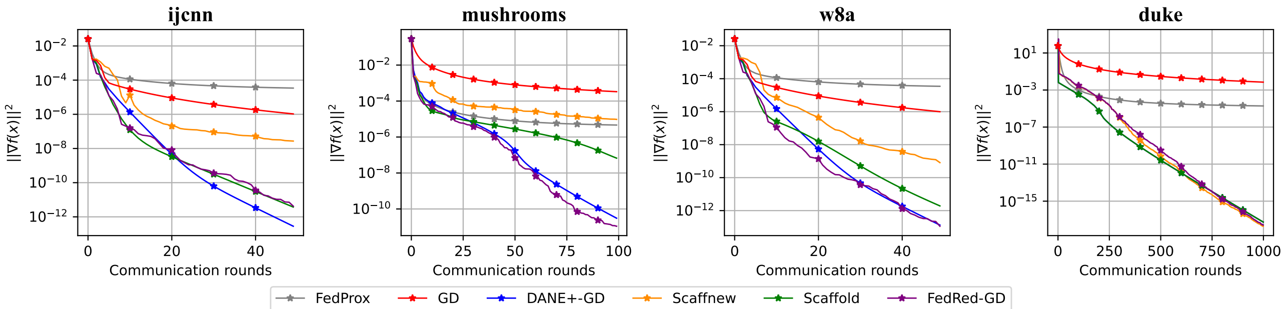

Figure 2. Comparison of DANE+-GD and FedRed-GD against four other distributed optimizers on four LIBSVM datasets using regularized logistic loss.

suppose $\sigma = 0$ and $\mu = 0$ for simplicity. FedRed-GD achieves the same computational complexity $K = \mathcal{O}(\frac{L}{\varepsilon})$ as GD. However, the communication complexity is $pK = \mathcal{O}(\frac{\delta_A}{\varepsilon})$ which is $\frac{L}{\delta_A}$ times faster than GD. The same conclusion remains effective where $\mu > 0$.

Discussion. Corollary 5 shows DANE+-FGD achieves a faster local convergence rate than DANE+-GD. We suspect the fast gradient method can also improve complexity for FedRed. We leave this potential as a future work.

5. Experiments

In this section, we illustrate the main theoretical properties of our studied methods in numerical experiments on both simulated and real datasets.

Synthetic data. We consider the minimization problem of the form: $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ where $f_i(\mathbf{x}) := \frac{1}{m}\sum_{j=1}^{m}\frac{1}{2}(\mathbf{x}-\mathbf{b}_{i,j})^T A_{i,j}(\mathbf{x}-\mathbf{b}_{i,j}) + \beta\sum_{k=1}^{d}\frac{[\mathbf{x}]_k^2}{1+[\mathbf{x}]_k^2}$, $\mathbf{b}_{i,j} \in \mathbb{R}^d$, $A_{i,j} \in \mathbb{R}^{d\times d}$, and $[\cdot]_k$ is an indexing operation of a vector. We set $\beta = 0$ for convex problems and $\beta = 400$ for the non-convex case. We further use $n = 5$, $m = 10$, and $d = 1000$. Note that $\delta_A = \sqrt{\frac{1}{n}\sum_{i=1}^{n}||\bar{A}_i - \bar{A}||^2}$ and $\delta_B = \max_i\{||\bar{A}_i - \bar{A}||\}$ where $\bar{A}_i := \frac{1}{m}\sum_{j=1}^{m} A_{i,j}$, and $\bar{A} := \frac{1}{n}\sum_{i=1}^{n}\bar{A}_i$. We generate $\{A_{i,j}\}$ such that $\max_{i,j}\{||A_{i,j}||\} = 100$ and $\delta_A \approx \delta_B \approx 5$. We generate a strongly-convex instance by further controlling the minimum eigenvalue of $A_{i,j}$ to be 1, and a general convex instance by setting some of the eigenvalues to small values close to zero (while ensuring that each $A_{i,j}$ is positive semi-definite). For the non-convex instance, we leave each $A_{i,j}$ as an indefinite matrix. By these constructions, we have that $\frac{L}{\delta_A} \approx \frac{L}{\delta_B} \gtrsim 20$.

We compare DANE+-GD and FedRed-GD against the vanilla gradient descent method. We compute approximate solutions for DANE+-GD by running local gradient descent until certain stopping criteria are reached. We use the constant probability (≈ 0.05) schedule for FedRed-GD. Lastly, we set the same step size for all three methods. In Figure 1, we can observe that both DANE+-GD and FedRed-GD require approximately 20 times fewer communication rounds than GD to reach the same accuracy. More importantly, the total cost of gradient computation for FedRed-GD is at the same scale as GD, demonstrating the usefulness of local steps and validating the theory.

Binary classification on LIBSVM datasets. We experiment with the binary classification task on four real-world LIBSVM datasets (Chang and Lin, 2011). We use the standard regularized logistic loss: $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ with $f_i(\mathbf{x}) := \frac{n}{M}\sum_{j=1}^{m_i}\log(1+\exp(-y_{i,j}\mathbf{a}_{i,j}^T\mathbf{x})) + \frac{1}{2M}||\mathbf{x}||^2$ where $(\mathbf{a}_{i,j}, y_{i,j}) \in \mathbb{R}^{d+1}$ are feature and labels and $M := \sum_{i=1}^{n} m_i$ is the total number of data points in the training dataset. We use $n = 5$ and split the dataset according to the Dirichlet distribution. We benchmark against popular distributed algorithms including Scaffold (Karimireddy et al., 2020), Scaffnew (Mishchenko et al., 2022), Fedprox (Li et al., 2020), and GD. We use control variate (4) for Scaffold. We perform grid search to find the best hyper-parameters for each algorithm including the number of local steps and the stepsizes. From Figure 2, we observe that DANE+-GD and FedRed-GD consistently achieve fast convergence due to the implicit similarities of the objective functions among the workers.

Practical choices of hyper-parameters. For FedRed, the main theories suggest that $p \sim \frac{\delta_A}{L}$ for convex problems and $p \sim \frac{\delta_B}{L}$ for non-convex instances. The stepsize η should be of order L. Hence, we can fix $\lambda = p\eta$ which has the same order as the similarity constant.

6. Conclusion

We propose a new federated optimization framework that simultaneously achieves both communication reduction and efficient local computations. Interesting future directions may include: theoretical analysis for client sampling and inexact control variates, extensions to the decentralized settings, accelerated version of FedRed, a more comprehensive study of FedRed and DANE+ for deep learning.

Acknowledgments

We acknowledge partial funding from a Google Scholar Research Award. The authors are thankful to the anonymous reviewers for their constructive comments.

Impact statement

This paper presents work that aims to advance the field of distributed Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Laurent Condat and Peter Richtárik. Randprox: Primal-dual optimization algorithms with randomized proximal updates. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cB4N3G5udUS.

Laurent Condat, Grigory Malinovsky, and Peter Richtárik. Tamuna: Accelerated federated learning with local training and partial participation. *arXiv preprint arXiv:2302.09832*, 2023.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10112–10121, 2022.

Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Can 5th generation local training methods support client sampling? yes! In *International Conference on Artificial Intelligence and Statistics*, pages 1055–1092. PMLR, 2023.

Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In *International conference on machine learning*, pages 4203–4227. PMLR, 2020.

Zhengmian Hu and Heng Huang. Tighter analysis for ProxSkip. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13469–13496. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hu23a.html.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. In *Advances in Neural Information Processing Systems*, 2021.

Ahmed Khaled and Chi Jin. Faster federated optimization under second-order similarity. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ElC6LYO4MfD.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. *Advances in Neural Information Processing Systems*, 35: 33494–33507, 2022.

Bo Li, Mikkel N. Schmidt, Tommy S. Alstrøm, and Sebastian U. Stich. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3964–3973, June 2023.

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.

Dachao Lin, Yuze Han, Haishan Ye, and Zhihua Zhang. Stochastic distributed optimization under average second-order similarity: Algorithms and analysis. *Advances in Neural Information Processing Systems*, 36, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

Konstantin Mishchenko, Rui Li, Hongxiang Fan, and Stylianos Venieris. Federated learning under second-order data heterogeneity, 2024. URL https://openreview.net/forum?id=jkhVrIllKg.

Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=h7FqQ6hCK18.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

Kumar Kshitij Patel, Lingxiao Wang, Blake Woodworth, Brian Bullins, and Nathan Srebro. Towards optimal communication complexity in distributed non-convex optimization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SNElc7QmMDe.

Kumar Kshitij Patel, Margalit Glasgow, Ali Zindari, Lingxiao Wang, Sebastian U Stich, Ziheng Cheng, Nirmit Joshi, and Nathan Srebro. The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. *arXiv preprint arXiv:2405.11667*, 2024.

Sashank J Reddi, Jakub Konečnỳ, Peter Richtárik, Barnabás Póczós, and Alex Smola. Aide: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

Abdurakhmon Sadiev, Dmitry Kovalev, and Peter Richtárik. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=W72rB0wwLVu.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1000–1008, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/shamir14.html.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1g2JnRcFX.

Ying Sun, Gesualdo Scutari, and Amir Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2):354–385, 2022.

Ye Tian, Gesualdo Scutari, Tianyu Cao, and Alexander Gasnikov. Acceleration in distributed optimization under similarity. In *International Conference on Artificial Intelligence and Statistics*, pages 5721–5756. PMLR, 2022.

Farshid Varno, Marzie Saghayi, Laya Rafiee Sevyeri, Sharut Gupta, Stan Matwin, and Mohammad Havaei.

Adabest: Minimizing client drift in federated learning via adaptive bias estimation. In *European Conference on Computer Vision*, pages 710–726. Springer, 2022.

Shusen Wang, Fred Roosta, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Xiao-Tong Yuan and Ping Li. On convergence of distributed approximate newton methods: Globalization, sharper bounds and beyond. *arXiv preprint arXiv:1908.02246*, 2019.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.

Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 362–370, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/zhangb15.html.

Ali Zindari, Ruichen Luo, and Sebastian U Stich. On the convergence of local SGD under third-order smoothness and hessian similarity. In *OPT 2023: Optimization for Machine Learning*, 2023. URL https://openreview.net/forum?id=OnJa9N4VwY.

Appendix

A. Technical Preliminaries

A.1. Basic Definitions

We use the following definitions throughout the paper.

Definition 3 (Convexity)**.** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-convex with $\mu \geq 0$ if $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} ||\mathbf{x} - \mathbf{y}||^2 \,. \tag{A.1}$$

Definition 4 (L-smooth)**.** *Let function $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. f is smooth if there exists $L \geq 0$ such that $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq L ||\mathbf{x} - \mathbf{y}|| \,. \tag{A.2}$$

A.2. Useful Lemmas

We frequently use the following helpful lemmas for the proofs.

Lemma 10. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. For any $\gamma > 0$, we have:*

$$-\frac{1}{2\gamma} ||\mathbf{x}||^2 - \frac{\gamma}{2} ||\mathbf{y}||^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle \leq \frac{1}{2\gamma} ||\mathbf{x}||^2 + \frac{\gamma}{2} ||\mathbf{y}||^2 \,, \tag{A.3}$$

$$||\mathbf{x} + \mathbf{y}||^2 \leq (1 + \gamma) ||\mathbf{x}||^2 + \Big(1 + \frac{1}{\gamma}\Big) ||\mathbf{y}||^2 \,. \tag{A.4}$$

Lemma 11. *Let $\mathbf{x} \in \mathbb{R}^d$ and $a > 0$. For any $\mathbf{y} \in \mathbb{R}^d$, we have:*

$$\langle \mathbf{x}, \mathbf{y} \rangle + \frac{a}{2} ||\mathbf{y}||^2 \geq -\frac{||\mathbf{x}||^2}{2a} \,. \tag{A.5}$$

Proof. The claim follows from the first-order optimality condition. □

Lemma 12. *Let $\{x_r\}_{r=0}^{+\infty}$ be a non-negative sequence such that $x_{r+1} - x_r \leq a_{r+1} \sqrt{x_{r+1}}$ for any $r \geq 0$, where $a_i \geq 0$ for any $i \geq 1$. Then for any $R \geq 1$, we have:*

$$x_R \leq \Big(\sqrt{x_0} + \sum_{r=1}^{R} a_r\Big)^2 \leq 2x_0 + 2\Big(\sum_{r=1}^{R} a_r\Big)^2 \,. \tag{A.6}$$

Proof. Indeed, for any $r \geq 0$, we have:

$$x_{r+1} - x_r = (\sqrt{x_{r+1}} - \sqrt{x_r})(\sqrt{x_{r+1}} + \sqrt{x_r}) \geq (\sqrt{x_{r+1}} - \sqrt{x_r}) \sqrt{x_{r+1}} \,. \tag{A.7}$$

It follows that:

$$\sqrt{x_{r+1}} \leq \sqrt{x_r} + a_{r+1} \,. \tag{A.8}$$

Summing up from $r = 0$ to $R - 1$, we get:

$$\sqrt{x_R} \leq \sqrt{x_0} + \sum_{r=1}^{R} a_r \,. \tag{A.9}$$

Taking the square on both sides, we have:

$$x_R \leq \Big(\sqrt{x_0} + \sum_{r=1}^{R} a_r\Big)^2 \leq 2x_0 + 2\Big(\sum_{r=1}^{R} a_r\Big)^2 \,. \tag{A.10}$$

□

Lemma 13. *Let $q > 0$ and let $(F_k)_{k=1}^{+\infty}$, $(D_k)_{k=1}^{+\infty}$ be two non-negative sequences such that: $F_{k+1} + D_{k+1} \leq q D_k$ for any $k \geq 0$. Then for any $K \geq 1$, it holds that:*

$$\frac{1}{S_k} \sum_{k=1}^{K} \frac{F_k}{q^k} + \frac{1 - q}{1 - q^K} D_K \leq \frac{1 - q}{\frac{1}{q^K} - 1} D_0 \,, \tag{A.11}$$

where $S_k := \sum_{k=1}^{K} \frac{1}{q^k}$.

Proof. Indeed, for any $k \geq 0$, we have:

$$\frac{F_{k+1}}{q^{k+1}} + \frac{D_{k+1}}{q^{k+1}} \leq \frac{D_k}{q^k} \ . \tag{A.12}$$

Summing up from $k = 0$ to $K - 1$, we get:

$$\sum_{k=1}^{K} \frac{F_k}{q^k} + \frac{D_K}{q^K} \leq D_0 \ . \tag{A.13}$$

Dividing both sides by $\sum_{k=1}^{K} \frac{1}{q^k}$ and using the fact that $\sum_{k=1}^{K} \frac{1}{q^k} = \frac{\frac{1}{q^K} - 1}{1-q}$, we obtain:

$$\frac{1-q}{\frac{1}{q^K} - 1} \sum_{k=1}^{K} \frac{F_k}{q^k} + \frac{1-q}{1-q^K} D_K \leq \frac{1-q}{\frac{1}{q^K} - 1} D_0 \ . \tag{A.14}$$

□

Lemma 14. *Let $\{x_i\}_{i=1}^n$ be a set of vectors in $\mathbb{R}^d$ and let $\mathbf{v} \in \mathbb{R}^d$ be an arbitrary vector. It holds that:*

$$\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{v}||^2 \ = ||\mathbf{x} - \mathbf{v}||^2 + \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x}||^2, \quad \textit{with} \quad \mathbf{x} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \ . \tag{A.15}$$

Proof. Indeed,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{v}||^2 &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x} + \mathbf{x} - \mathbf{v}||^2 & \text{(A.16)}\\
&= \frac{1}{n}\sum_{i=1}^{n} \Big[||\mathbf{x}_i - \mathbf{x}||^2 + 2\,\langle \mathbf{x}_i - \mathbf{x}, \mathbf{x} - \mathbf{v} \rangle + ||\mathbf{x} - \mathbf{v}||^2 \Big] & \text{(A.17)}\\
&= ||\mathbf{x} - \mathbf{v}||^2 + \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x}||^2 \ .
\end{aligned}$$

□

Lemma 15 (Nesterov (2018), Lemma 1.2.3)**.** *Smoothness* (A.2) *implies that there exists a quadratic upper bound on f:*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \mid + \frac{L}{2} ||\mathbf{y} - \mathbf{x}||^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \ . \tag{A.18}$$

Lemma 16 (Nesterov (2018), Theorem 2.1.12)**.** *Suppose a function f is L-smooth and μ-convex, then it holds that:*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu L}{\mu + L} ||\mathbf{x} - \mathbf{y}||^2 + \frac{1}{\mu + L} ||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})||^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{A.19}$$

Lemma 17 (Hessian Dissimilarity)**.** *Let $h_i : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable for any $i \in [n]$. Suppose for any $\mathbf{x} \in \mathbb{R}^d$ and any $p \geq 1$, it holds that:*

$$\frac{1}{n}\sum_{i=1}^{n} ||\nabla^2 h_i(\mathbf{x})||^p \leq \delta^p \ . \tag{A.20}$$

Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have:

$$\frac{1}{n}\sum_{i=1}^{n} ||\nabla h_i(\mathbf{x}) - \nabla h_i(\mathbf{y})||^p \leq \delta^p ||\mathbf{x} - \mathbf{y}||^p \ . \tag{A.21}$$

Proof. Since h_i is twice continuously differentiable, by Taylor's Theorem, we have:

$$\nabla h_i(\mathbf{y}) = \nabla h_i(\mathbf{x}) + \int_0^1 \nabla^2 h_i(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))[\mathbf{y} - \mathbf{x}]dt \ . \tag{A.22}$$

It follows that:

$$\begin{aligned}
||\nabla h_i(\mathbf{y}) - \nabla h_i(\mathbf{x})||^p &= \Big|\Big| \int_0^1 \nabla^2 h_i(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))[\mathbf{y} - \mathbf{x}]dt \Big|\Big|^p & \text{(A.23)}\\
&\leq \left[\int_0^1 ||\nabla^2 h_i(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))|| ||\mathbf{y} - \mathbf{x}|| dt \right]^p & \text{(A.24)}\\
&= ||\mathbf{y} - \mathbf{x}||^p \left[\int_0^1 ||\nabla^2 h_i(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))|| dt \right]^p & \text{(A.25)}\\
&\leq ||\mathbf{y} - \mathbf{x}||^p \int_0^1 ||\nabla^2 h_i(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))||^p dt \ . & \text{(A.26)}
\end{aligned}$$

where in the last inequality, we use the Jensen's inequality and the fact that $\tau \mapsto \tau^p$ is a convex function for $p \geq 1$.

By our assumption that $\frac{1}{n}\sum_{i=1}^n ||\nabla^2 h_i(\mathbf{z})||^p \leq \delta^p$ for any $\mathbf{z} \in \mathbb{R}^d$, we get:

$$\frac{1}{n}\sum_{i=1}^n ||\nabla h_i(\mathbf{y}) - \nabla h_i(\mathbf{x})||^p \leq ||\mathbf{y}-\mathbf{x}||^p \int_0^1 \frac{1}{n}\sum_{i=1}^n ||\nabla^2 h_i(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))||^p dt \tag{A.27}$$
$$\leq \delta^p ||\mathbf{y}-\mathbf{x}||^p \,.$$
□

Remark 18. *According to Lemma 17, if each f_i is twice continuously differentiable and satisfies: $\frac{1}{n}\sum_{i=1}^n ||\nabla^2 f_i(\mathbf{x}) - \nabla f^2(\mathbf{x})||^2 \leq \delta_A^2$, for any $\mathbf{x} \in \mathbb{R}^d$, then $\{f_i\}$ have δ_A-AHD. However the reverse does not hold. Let $n = 3$ and $h_i := f_i - f := \frac{1}{2}\mathbf{x}^T\mathbf{A}_i\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^2$ and*

$$\mathbf{A}_1 = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \quad \mathbf{A}_2 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{A}_3 = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix} \quad . \tag{A.28}$$

It holds that:

$$\frac{1}{n}\sum_{i=1}^n ||\nabla h_i(\mathbf{x}) - \nabla h_i(\mathbf{y})||^2 = \frac{9+1+4}{3}||\mathbf{x}-\mathbf{y}||^2 = \frac{14}{3}||\mathbf{x}-\mathbf{y}||^2 \,, \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^2 \,. \tag{A.29}$$

On the other hand,

$$\frac{1}{n}\sum_{i=1}^n ||\nabla^2 h_i(\mathbf{x})||^2 = \frac{9+1+9}{3} = \frac{19}{3}, \quad \forall \mathbf{x} \in \mathbb{R}^2 \,. \tag{A.30}$$

B. Proofs of main results

B.1. Convex results and proofs for Algorithm 1 with control variate (1)

Lemma 19. *Consider Algorithm 1. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be μ-convex with $\mu \geq 0$ for any $i \in [n]$. Then for any $r \geq 0$, and any $\mathbf{x} \in \mathbb{R}^d$, we have:*

$$f_i(\mathbf{x}) - \langle \mathbf{x}, \mathbf{h}_{i,r}\rangle + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2 \geq f_i(\mathbf{x}_{i,r+1}) - \langle \mathbf{x}_{i,r+1}, \mathbf{h}_{i,r}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2 \tag{B.1}$$
$$+ \langle \nabla F_{i,r}(\mathbf{x}_{i,r+1}), \mathbf{x}-\mathbf{x}_{i,r+1}\rangle + \frac{\mu+\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}||^2 \,. \tag{B.2}$$

Proof. Let $F_{i,r}(\mathbf{x}) := f_i(\mathbf{x}) - \langle \mathbf{x}, \mathbf{h}_{i,r}\rangle + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2$. Since $F_{i,r}$ is $(\mu+\lambda)$-convex, we have:

$$F_{i,r}(\mathbf{x}) \overset{(A.1)}{\geq} F_{i,r}(\mathbf{x}_{i,r+1}) + \langle \nabla F_{i,r}(\mathbf{x}_{i,r+1}), \mathbf{x}-\mathbf{x}_{i,r+1}\rangle + \frac{\mu+\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}||^2 \,. \tag{B.3}$$

Plugging in the definition of $F_{i,r}$, we get the claim. □

Lemma 20. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. Let $\lambda \geq \frac{\delta_A}{1-c}$ for some $c \in [0,1)$. Then for any $\mathbf{x} \in \mathbb{R}^d$ and any $r \geq 0$, we have:*

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}) + \frac{\mu+\lambda}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1}-\mathbf{x}||^2 + \frac{c\lambda}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2$$
$$\leq \frac{\lambda}{2}||\mathbf{x}^r-\mathbf{x}||^2 + \frac{1}{n}\sum_{i=1}^n \langle \nabla F_{i,r}(\mathbf{x}_{i,r+1}), \mathbf{x}_{i,r+1}-\mathbf{x}\rangle \,. \tag{B.4}$$

Proof. According to Lemma 19, we have:

$$f_i(\mathbf{x}) + \langle \mathbf{x}, \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r)\rangle + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2 \geq f_i(\mathbf{x}_{i,r+1}) + \langle \mathbf{x}_{i,r+1}, \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r)\rangle \tag{B.5}$$
$$+ \frac{\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2 + \langle \nabla F_{i,r}(\mathbf{x}_{i,r+1}), \mathbf{x}-\mathbf{x}_{i,r+1}\rangle + \frac{\mu+\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}||^2 \,. \tag{B.6}$$

By μ-convexity of f_i, we further get:

$$f_i(\mathbf{x}) + \langle \mathbf{x}, \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r)\rangle + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2 \overset{(A.1)}{\geq} f_i(\mathbf{x}^{r+1}) + \left\langle \nabla f_i(\mathbf{x}^{r+1}), \mathbf{x}_{i,r+1} - \mathbf{x}^{r+1}\right\rangle \tag{B.7}$$

$$+ \frac{\mu}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}||^2 + \langle \mathbf{x}_{i,r+1}, \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r)\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2 \tag{B.8}$$

$$+ \langle \nabla F_{i,r}(\mathbf{x}_{i,r+1}), \mathbf{x} - \mathbf{x}_{i,r+1}\rangle + \frac{\mu+\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}||^2 \; . \tag{B.9}$$

Taking the average on both sides over $i=1$ to n, we get:

$$\begin{aligned} f(\mathbf{x}) + \frac{\lambda}{2}||\mathbf{x}-\mathbf{x}^r||^2 &\geq f(\mathbf{x}^{r+1}) + \frac{1}{n}\sum_{i=1}^n \Big[\left\langle \nabla f_i(\mathbf{x}^{r+1}), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle + \langle \mathbf{x}_{i,r+1}, \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r)\rangle \Big] \\ &+ \frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}||^2 + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2 + \frac{\mu+\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}||^2 \\ &+ \frac{1}{n}\sum_{i=1}^n \langle \nabla F_{i,r}(\mathbf{x}_{i,r+1}), \mathbf{x}-\mathbf{x}_{i,r+1}\rangle \; . \end{aligned} \tag{B.10}$$

Let $h_i = f_i - f$. Note that $\frac{1}{n}\sum_{i=1}^n \left\langle \nabla f(\mathbf{x}^{r+1}), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle = \frac{1}{n}\sum_{i=1}^n \left\langle \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r), \mathbf{x}^{r+1}\right\rangle = 0$. It follows that:

$$\frac{1}{n}\sum_{i=1}^n \Big[\left\langle \nabla f_i(\mathbf{x}^{r+1}), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle + \langle \mathbf{x}_{i,r+1}, \nabla f(\mathbf{x}^r) - \nabla f_i(\mathbf{x}^r)\rangle \Big] \tag{B.11}$$

$$= \frac{1}{n}\sum_{i=1}^n \left\langle \nabla f_i(\mathbf{x}^{r+1}) - \nabla f(\mathbf{x}^{r+1}) - \nabla f_i(\mathbf{x}^r) + \nabla f(\mathbf{x}^r), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle \tag{B.12}$$

$$= \frac{1}{n}\sum_{i=1}^n \left\langle \nabla h_i(\mathbf{x}^{r+1}) - \nabla h_i(\mathbf{x}^r), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle \; . \tag{B.13}$$

We next split $\frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2$ into $\frac{c\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2 + \frac{(1-c)\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2$ with $0 \leq c \leq 1$. We combine the second part with $\frac{1}{n}\sum_{i=1}^n \left\langle \nabla h_i(\mathbf{x}^{r+1}) - \nabla h_i(\mathbf{x}^r), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle$ to get:

$$\begin{aligned} &\frac{(1-c)\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^r||^2 + \frac{1}{n}\sum_{i=1}^n \left\langle \nabla h_i(\mathbf{x}^{r+1}) - \nabla h_i(\mathbf{x}^r), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle \\ &\overset{(A.15)}{=} \frac{(1-c)\lambda}{2}||\mathbf{x}^{r+1}-\mathbf{x}^r||^2 + \frac{1}{n}\sum_{i=1}^n\left[\frac{(1-c)\lambda}{2}||\mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}||^2 + \left\langle \nabla h_i(\mathbf{x}^{r+1}) - \nabla h_i(\mathbf{x}^r), \mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}\right\rangle\right] \\ &\overset{(A.5)}{\geq} \frac{(1-c)\lambda}{2}||\mathbf{x}^{r+1}-\mathbf{x}^r||^2 - \frac{1}{2(1-c)\lambda}\frac{1}{n}\sum_{i=1}^n||\nabla h_i(\mathbf{x}^{r+1}) - \nabla h_i(\mathbf{x}^r)||^2 \\ &\overset{(3)}{\geq} \frac{(1-c)\lambda}{2}||\mathbf{x}^{r+1}-\mathbf{x}^r||^2 - \frac{\delta_A^2}{2(1-c)\lambda}||\mathbf{x}^{r+1}-\mathbf{x}^r||^2 \geq 0. \end{aligned}$$

where in the last inequality, we use the assumption that $\lambda \geq \frac{\delta_A}{1-c}$.

Plugging this inequality into (B.10), and dropping the non-negative $\frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,r+1}-\mathbf{x}^{r+1}||^2$, we get the claim. □

Corollary 21. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ-AHD. Let $\lambda \geq \delta_A$ and suppose that each local solver provides the exact solution. Then for any $r \geq 0$, we have:*

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \leq -\frac{\mu+\lambda}{2}||\mathbf{x}^{r+1}-\mathbf{x}^r||^2 \; , \tag{B.14}$$

and after R communication rounds, it holds that:

$$f(\mathbf{x}^R) - f(\mathbf{x}^\star) + \frac{\mu}{2}||\mathbf{x}^R-\mathbf{x}^\star||^2 \leq \frac{\mu}{2[(1+\frac{\mu}{\lambda})^R - 1]}||\mathbf{x}^0-\mathbf{x}^\star||^2 \leq \frac{\lambda}{2R}||\mathbf{x}^0-\mathbf{x}^\star||^2. \tag{B.15}$$

Proof. By the assumption that the subproblem is solved exactly, we have $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| = 0$ for any $i \in [n]$ and $r \geq 0$. According to Lemma 20 with $c = 1$, we have:

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}) + \frac{\mu+\lambda}{2}\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r+1} - \mathbf{x}||^2 \leq \frac{\lambda}{2}||\mathbf{x}^r - \mathbf{x}||^2 \,. \tag{B.16}$$

with $\lambda \geq \delta_A$. It follows that:

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}) \overset{\text{(A.15)}}{\leq} \frac{\lambda}{2}||\mathbf{x}^r - \mathbf{x}||^2 - \frac{\lambda+\mu}{2}||\mathbf{x}^{r+1} - \mathbf{x}||^2 \,. \tag{B.17}$$

Let $\mathbf{x} = \mathbf{x}_r$. The function value gap monotonically decreases as:

$$f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \leq -\frac{\mu+\lambda}{2}||\mathbf{x}^{r+1} - \mathbf{x}^r||^2 \,. \tag{B.18}$$

Let $\mathbf{x} = \mathbf{x}^\star$. We have:

$$\frac{2}{\mu+\lambda}\big(f(\mathbf{x}^{r+1}) - f(\mathbf{x}^\star)\big) \leq \underbrace{\Big(1 - \frac{\mu}{\mu+\lambda}\Big)}_{:=q}||\mathbf{x}^r - \mathbf{x}^\star||^2 - ||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 \,. \tag{B.19}$$

Dividing both sides by q^{r+1}, we get:

$$\frac{2}{\mu+\lambda}\frac{1}{q^{r+1}}\big(f(\mathbf{x}^{r+1}) - f(\mathbf{x}^\star)\big) \leq \frac{||\mathbf{x}^r - \mathbf{x}^\star||^2}{q^r} - \frac{||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2}{q^{r+1}} \,. \tag{B.20}$$

Summing up from $r = 0$ to $R - 1$, we have:

$$\Big[\frac{2}{\mu+\lambda}\sum_{r=1}^{R}\frac{1}{q^r}\Big]\big(f(\mathbf{x}^R) - f(\mathbf{x}^\star)\big) \overset{\text{(B.18)}}{\leq} \frac{2}{\mu+\lambda}\sum_{r=1}^{R}\frac{1}{q^r}\big(f(\mathbf{x}^r) - f(\mathbf{x}^\star)\big) \leq ||\mathbf{x}^0 - \mathbf{x}^\star||^2 - \frac{||\mathbf{x}^R - \mathbf{x}^\star||^2}{q^R} \,. \tag{B.21}$$

Using the fact that $\sum_{r=1}^{R}\frac{1}{q^r} = \frac{\frac{1}{q^R} - 1}{1-q}$ and rearranging, we have:

$$\begin{aligned} f(\mathbf{x}^R) - f^\star &\leq \frac{\mu+\lambda}{2}\frac{\frac{\mu}{\mu+\lambda}}{(1+\frac{\mu}{\lambda})^R - 1}||\mathbf{x}^0 - \mathbf{x}^\star||^2 - \frac{\mu+\lambda}{2}\frac{\frac{\mu}{\mu+\lambda}}{1 - q^R}||\mathbf{x}^R - \mathbf{x}^\star||^2 && \text{(B.22)} \\ &\leq \frac{\mu}{2[(1+\frac{\mu}{\lambda})^R - 1]}||\mathbf{x}^0 - \mathbf{x}^\star||^2 - \frac{\mu}{2}||\mathbf{x}^R - \mathbf{x}^\star||^2 \,. \end{aligned}$$

□

Lemma 22. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. After R communication rounds, we have:*

$$\frac{2}{\mu+\lambda}\sum_{r=1}^{R}\frac{f(\mathbf{x}^r) - f^\star}{q^r} + \frac{A_R}{q^R} + \frac{\lambda}{2(\mu+\lambda)}\sum_{r=1}^{R}\frac{B_{r-1}}{q^r} \leq 2||\mathbf{x}^0 - \mathbf{x}^\star||^2 + \frac{4}{(\mu+\lambda)^2}\Bigg(\sum_{r=1}^{R}\sqrt{\frac{C_r}{q^r}}\Bigg)^2 \,. \tag{B.23}$$

where $A_r := \frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r} - \mathbf{x}^\star||^2$, $B_r := \frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$, $C_r := \frac{1}{n}\sum_{i=1}^{n}||\nabla F_{i,r}(\mathbf{x}_{i,r})||^2$, and $q := \frac{\lambda}{\lambda+\mu}$.

Proof. According to Lemma 20 with $\mathbf{x} = \mathbf{x}^\star$, for any $r \geq 0$, we have:

$$\begin{aligned} &f(\mathbf{x}^{r+1}) - f(\mathbf{x}^\star) + \frac{\mu+\lambda}{2}\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 + \frac{\lambda}{4}\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2 && \text{(B.24)} \\ &\leq \frac{\lambda}{2}||\mathbf{x}^r - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^{n}||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||||\mathbf{x}_{i,r+1} - \mathbf{x}^\star|| && \text{(B.25)} \\ &\overset{\text{(A.15)}}{\leq} \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r} - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^{n}||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||||\mathbf{x}_{i,r+1} - \mathbf{x}^\star|| && \text{(B.26)} \\ &\leq \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r} - \mathbf{x}^\star||^2 + \sqrt{\frac{1}{n}\sum_{i=1}^{n}||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2} \,, && \text{(B.27)} \end{aligned}$$

where in the last inequality, we use the standard Cauchy-Schwarz inequality. We now use the simplified notations and divide both sides by $\frac{\mu+\lambda}{2}$ to get:

$$\frac{2}{\mu+\lambda}\big(f(\mathbf{x}^{r+1})-f^\star\big)+A_{r+1}+\frac{\lambda}{2(\mu+\lambda)}B_r \le \underbrace{(1-\frac{\mu}{\mu+\lambda})}_{:=q}A_r+\frac{2}{\mu+\lambda}\sqrt{C_{r+1}}\sqrt{A_{r+1}}\,. \tag{B.28}$$

Dividing both sides by q^{r+1} and summing up from $r=0$ to $R-1$, we obtain:

$$\frac{2}{\mu+\lambda}\sum_{r=1}^{R}\frac{f(\mathbf{x}^r)-f^\star}{q^r}+\frac{A_R}{q^R}+\frac{\lambda}{2(\mu+\lambda)}\sum_{r=1}^{R}\frac{B_{r-1}}{q^r}\le \underbrace{A_0+\frac{2}{\mu+\lambda}\sum_{r=1}^{R}\frac{\sqrt{C_r}\sqrt{A_r}}{q^r}}_{:=Q_R}\,, \tag{B.29}$$

from which we can deduce that $\frac{A_R}{q^R}\le Q_R$ for any $R\ge 1$.

We next upper bound Q_R. Let $Q_0:=A_0$. By definition, for any $R\ge 0$, it holds that:

$$Q_{R+1}-Q_R=\frac{2}{\mu+\lambda}\frac{\sqrt{C_{R+1}}\sqrt{A_{R+1}}}{q^{R+1}} \tag{B.30}$$

$$\le\frac{2}{\mu+\lambda}\frac{\sqrt{C_{R+1}}\sqrt{Q_{R+1}q^{R+1}}}{q^{R+1}} \tag{B.31}$$

$$=\frac{2}{\mu+\lambda}\frac{\sqrt{C_{R+1}}\sqrt{Q_{R+1}}}{\sqrt{q^{R+1}}}\,. \tag{B.32}$$

We now apply Lemma 12 with $a_{r+1}=\frac{2}{\mu+\lambda}\sqrt{\frac{C_{r+1}}{q^{r+1}}}$. For any $R\ge 1$, we get:

$$Q_R\le 2Q_0+\frac{4}{(\mu+\lambda)^2}\left(\sum_{r=1}^{R}\sqrt{\frac{C_r}{q^r}}\right)^2\,. \tag{B.33}$$

Plugging this upper bound into (B.29), we get the claim. □

Theorem 23. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i:\mathbb{R}^d\to\mathbb{R}$ be continuously differentiable and μ-convex with $\mu\ge 0$ for any $i\in[n]$. Assume that $\{f_i\}$ have δ_A-AHD. In general, suppose that the solutions returned by local solvers satisfy $\frac{1}{n}\sum_{i=1}^n||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2\le e_{r+1}^2$ for any $r\ge 0$ and $e_{r+1}\ge 0$. Let $\lambda\ge 2\delta_A$. After R communication rounds, we have:*

$$f(\bar{\mathbf{x}}^R)-f(\mathbf{x}^\star)+\frac{\mu}{2}\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_{i,R}-\mathbf{x}^\star||^2\le\frac{\mu}{(1+\frac{\mu}{\lambda})^R-1}||\mathbf{x}^0-\mathbf{x}^\star||^2+\frac{2}{\mu+\lambda}\sum_{r=1}^{R}e_r^2\le\frac{\lambda}{R}||\mathbf{x}^0-\mathbf{x}^\star||^2+\frac{2}{\mu+\lambda}\sum_{r=1}^{R}e_r^2\,. \tag{B.34}$$

where $\bar{\mathbf{x}}^R:=\arg\min_{\mathbf{x}\in\{\mathbf{x}^r\}_{r=1}^R}f(\mathbf{x})$.

Proof. Applying Lemma 22, dropping the non-negative $\frac{\lambda}{2(\mu+\lambda)}\sum_{r=1}^R\frac{B_{r-1}}{q^r}$, and plugging $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2\le e_{r+1}^2$ into the bound, we obtain:

$$\frac{2}{\mu+\lambda}\sum_{r=1}^{R}\frac{f(\mathbf{x}^r)-f^\star}{q^r}+\frac{A_R}{q^R}\le 2||\mathbf{x}^0-\mathbf{x}^\star||^2+\frac{4}{(\mu+\lambda)^2}\left(\sum_{r=1}^{R}\frac{e_r}{(\sqrt{q})^r}\right)^2 \tag{B.35}$$

$$\le 2||\mathbf{x}^0-\mathbf{x}^\star||^2+\frac{4}{(\mu+\lambda)^2}\sum_{r=1}^{R}e_r^2\sum_{r=1}^{R}\frac{1}{q^r}\,. \tag{B.36}$$

Define $\bar{\mathbf{x}}^R:=\arg\min_{\mathbf{x}\in\{\mathbf{x}^r\}_{r=1}^R}f(\mathbf{x})$. Dividing both sides by $\sum_{r=1}^R\frac{1}{q^r}=\frac{\frac{1}{q^R}-1}{1-q}$, we get:

$$\frac{2}{\mu+\lambda}\big(f(\bar{\mathbf{x}}^R)-f^\star\big)+(1-q)A_R\le\frac{2(1-q)}{\frac{1}{q^R}-1}||\mathbf{x}^0-\mathbf{x}^\star||^2+\frac{4}{(\mu+\lambda)^2}\sum_{r=1}^{R}e_r^2\,. \tag{B.37}$$

Plugging in the definition of q, we get the claim. □

Theorem 24. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. Suppose that the solutions returned by local solvers satisfy $\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq e_r^2 \sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$ for any $r \geq 0$ with $e_r \geq 0$. Let $\lambda \geq 2\delta_A$ and let $\sum_{r=0}^{+\infty} e_r^2 \leq \frac{\lambda(\mu+\lambda)}{8}$. After R communication rounds, we have:*

$$f(\bar{\mathbf{x}}^R) - f(\mathbf{x}^\star) + \frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,R} - \mathbf{x}^\star||^2 \leq \frac{\mu}{[(1+\frac{\mu}{\lambda})^R - 1]}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \leq \frac{\lambda}{R}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \ . \tag{B.38}$$

where $\bar{\mathbf{x}}^R := \arg\min_{\mathbf{x} \in \{\mathbf{x}^r\}_{r=1}^R} f(\mathbf{x})$.

Proof. Applying Lemma 22, and plugging $\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq e_r^2 \sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$ into the bound, we obtain:

$$\frac{2}{\mu+\lambda}\sum_{r=1}^R \frac{f(\mathbf{x}^r) - f^\star}{q^r} + \frac{A_R}{q^R} + \frac{\lambda}{2(\mu+\lambda)}\sum_{r=1}^R \frac{B_{r-1}}{q^r} \leq 2||\mathbf{x}^0 - \mathbf{x}^\star||^2 + \frac{4}{(\mu+\lambda)^2}\left(\sum_{r=1}^R e_{r-1}\sqrt{\frac{B_{r-1}}{q^r}}\right)^2 \tag{B.39}$$

$$\leq 2||\mathbf{x}^0 - \mathbf{x}^\star||^2 + \frac{4}{(\mu+\lambda)^2}\sum_{r=1}^R e_{r-1}^2 \sum_{r=1}^R \frac{B_{r-1}}{q^r} \ . \tag{B.40}$$

Let $\frac{4}{(\mu+\lambda)^2}\sum_{r=1}^R e_{r-1}^2 \leq \frac{\lambda}{2(\mu+\lambda)}$. We get:

$$\frac{2}{\mu+\lambda}\sum_{r=1}^R \frac{f(\mathbf{x}^r) - f^\star}{q^r} + \frac{A_R}{q^R} \leq 2||\mathbf{x}^0 - \mathbf{x}^\star||^2 \ . \tag{B.41}$$

Dividing both sides by $\sum_{r=1}^R \frac{1}{q^r} = \frac{\frac{1}{q^R} - 1}{1-q}$, we get:

$$\frac{2}{\mu+\lambda}\big(f(\bar{\mathbf{x}}^R) - f^\star\big) + (1-q)A_R \leq \frac{2(1-q)}{\frac{1}{q^R} - 1}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \ . \tag{B.42}$$

Plugging in the definition of q, we get the claim. □

Corollary 25. *Consider Algorithm 1 with control variate (1) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$ and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. Suppose that each local solver returns a solution such that $\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq \frac{\lambda(\mu+\lambda)}{8(r+1)(r+2)}\sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$ for any $r \geq 0$. Let $\lambda = 2\delta_A$. Then after R communication rounds, we have:*

$$f(\bar{\mathbf{x}}^R) - f(\mathbf{x}^\star) + \frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,R} - \mathbf{x}^\star||^2 \leq \frac{\mu}{[(1+\frac{\mu}{2\delta_A})^R - 1]}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \leq \frac{2\delta_A}{R}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \ . \tag{B.43}$$

where $\bar{\mathbf{x}}^R := \arg\min_{\mathbf{x} \in \{\mathbf{x}^r\}_{r=1}^R} f(\mathbf{x})$.

Suppose each device uses the standard gradient descent initialized at $\mathbf{x}^r$, Then the total number of local steps required at each round r is no more than:

$$K_r = \Theta\left(\frac{L}{\mu+\lambda}\ln\left(\frac{L}{\sqrt{\lambda(\mu+\lambda)}}(r+2)\right)\right) \qquad (\textit{DANE+-GD}) \ . \tag{B.44}$$

Suppose each device uses the fast gradient descent initialized at $\mathbf{x}^r$, Then the total number of local steps required at each round r is no more than:

$$K_r = \Theta\left(\sqrt{\frac{L}{\mu+\lambda}}\ln\left(\frac{L}{\sqrt{\lambda(\mu+\lambda)}}(r+2)\right)\right) \qquad (\textit{DANE+-FGD}) \ . \tag{B.45}$$

Proof. According to Theorem (24), to achieve the corresponding convergence rate, the solutions returned by local solvers should satisfy $\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq e_r^2 \sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$ for any $r \geq 0$, with $\sum_{r=0}^{+\infty} e_r^2 \leq \frac{\lambda(\mu+\lambda)}{8}$.

Let $\mathbf{x}_{i,r}^\star := \arg\min_{\mathbf{x}}\{F_{i,r}(\mathbf{x})\}$. Note that $F_{i,r}$ is $(\mu+\lambda)$-strongly convex. This gives:

$$||\mathbf{x}_{i,r+1} - \mathbf{x}^r|| \geq ||\mathbf{x}^r - \mathbf{x}_{i,r}^\star|| - ||\mathbf{x}_{i,r+1} - \mathbf{x}_{i,r}^\star|| \tag{B.46}$$

$$\geq ||\mathbf{x}^r - \mathbf{x}_{i,r}^\star|| - \frac{1}{\mu+\lambda}||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \ . \tag{B.47}$$

Therefore, suppose we want to have $\sum_{i=1}^n ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \le e_r^2 \sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$, it is sufficient to have:

$$||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \le e_r \Big[||\mathbf{x}^r - \mathbf{x}_{i,r}^\star|| - \frac{1}{\mu+\lambda}||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||\Big], \qquad \forall i \in [n]\ , \tag{B.48}$$

or equivalently:

$$||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \le \frac{e_r}{1+\frac{e_r}{\lambda+\mu}}||\mathbf{x}^r - \mathbf{x}_{i,r}^\star||, \qquad \forall i \in [n]\ . \tag{B.49}$$

By our choice of $e_r = \frac{\lambda(\mu+\lambda)}{8(r+1)(r+2)}$, we have:

$$e_r \ge \frac{\sqrt{\lambda(\mu+\lambda)}}{4(r+2)}, \quad \text{and} \quad 1 + \frac{e_r}{\lambda+\mu} \le 1 + \frac{\mu+\lambda}{(r+1)(\mu+\lambda)} \le 2\ . \tag{B.50}$$

Thus, the accuracy condition is sufficient to be:

$$||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \le \frac{\lambda(\mu+\lambda)}{64(r+2)^2}||\mathbf{x}^r - \mathbf{x}_{i,r}^\star||^2\ . \tag{B.51}$$

DANE+-GD: Local Gradient Descent. Recall that for any ν_1-convex and ν_2-smooth function g, initial point $\mathbf{y}_0$ and any number of iteration steps k, the standard gradient method has the following convergence guarantee: $||\nabla g(\mathbf{y}_k)||^2 \le \mathcal{O}\Big(\nu_2^2 \exp\big(-(\nu_1/\nu_2)k\big)||\mathbf{y}_0 - \mathbf{y}^\star||^2\Big)$ where $\mathbf{y}^\star$ is the minimizer of g (See Theorem 3.12 (Bubeck et al., 2015) and use the fact that $g(\mathbf{y}_k) - g^\star \ge \frac{1}{2\nu_2}||\nabla g(\mathbf{y}_k)||^2$). Now we use this result to compute the required local steps K_r to satisfy (B.51). Recall that $F_{i,r}$ is $(\mu+\lambda)$-convex and $(L+\lambda)$-smooth. For any $r \ge 0$, we have that:

$$\Theta\Big((L+\lambda)^2 \exp(-\frac{\mu+\lambda}{L+\lambda}K_r)\Big) \le \Theta\bigg(\frac{\lambda(\mu+\lambda)}{(r+2)^2}\bigg)\ . \tag{B.52}$$

This gives:

$$K_r \ge \Theta\bigg(\frac{L}{\mu+\lambda}\ln\Big(\frac{L}{\sqrt{\lambda(\mu+\lambda)}}(r+2)\Big)\bigg)\ . \tag{B.53}$$

where we use the fact that $\lambda \lesssim L$.

DANE+-FGD: Local Fast Gradient Descent. Recall that for any ν_1-convex and ν_2-smooth function g, initial point $\mathbf{y}_0$ and any number of iteration steps k, the fast gradient method has the following convergence guarantee: $||\nabla g(\mathbf{y}_k)||^2 \le \mathcal{O}\Big(\nu_2^2 \exp\big(-\sqrt{\nu_1/\nu_2}k\big)||\mathbf{y}_0 - \mathbf{y}^\star||^2\Big)$ where $\mathbf{y}^\star$ is the minimizer of g (See Theorem 3.18 (Bubeck et al., 2015) and use the fact that $g(\mathbf{y}_k) - g^\star \ge \frac{1}{2\nu_2}||\nabla g(\mathbf{y}_k)||^2$) . Now we use this result to compute the required local steps K_r to satisfy (B.51). Recall that $F_{i,r}$ is $(\mu+\lambda)$-convex and $(L+\lambda)$-smooth. For any $r \ge 0$, we have that:

$$\Theta\Big((L+\lambda)^2 \exp(-\sqrt{\frac{\mu+\lambda}{L+\lambda}}K_r)\Big) \le \Theta\bigg(\frac{\lambda(\mu+\lambda)}{(r+2)^2}\bigg)\ . \tag{B.54}$$

This gives:

$$K_r \ge \Theta\bigg(\sqrt{\frac{L}{\mu+\lambda}}\ln\Big(\frac{L}{\sqrt{\lambda(\mu+\lambda)}}(r+2)\Big)\bigg)\ . \tag{B.55}$$

where we use the fact that $\lambda \lesssim L$.

□

B.2. Non-convex results and proofs for Algorithm 1 with control variate (1)

Lemma 26. *Consider Algorithm 1 with control variate (1). Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. For any $r \ge 0$, $i \in [n]$, and any $\mathbf{x} \in \mathbb{R}^d$, it holds that:*

$$||\nabla F_{i,r}(\mathbf{x}) - \nabla f(\mathbf{x})|| \le (\delta_B + \lambda)||\mathbf{x}^r - \mathbf{x}||\ . \tag{B.56}$$

Proof. Let $h_i := f - f_i$. Using the definition of $\nabla F_{i,r}(\mathbf{x})$, we get:

$$\begin{aligned} ||\nabla F_{i,r}(\mathbf{x}) - \nabla f(\mathbf{x})|| &= ||\nabla h_i(\mathbf{x}^r) - \nabla h_i(\mathbf{x}) + \lambda(\mathbf{x} - \mathbf{x}^r)|| && \text{(B.57)} \\ &\le ||\nabla h_i(\mathbf{x}) - \nabla h_i(\mathbf{x}^r)|| + \lambda||\mathbf{x} - \mathbf{x}^r|| && \text{(B.58)} \\ &\overset{(2)}{\le} \delta_B||\mathbf{x} - \mathbf{x}^r|| + \lambda||\mathbf{x} - \mathbf{x}^r||\ . \end{aligned}$$

□

Lemma 27. *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. For any $i \in [n]$ and any $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{x}^r \in \mathbb{R}^d$, we have:*

$$f_i(\mathbf{x}^r) + \langle \nabla h_i(\mathbf{x}^r), \mathbf{x}^r - \mathbf{y}\rangle - f_i(\mathbf{y}) - \frac{\lambda}{2}||\mathbf{y} - \mathbf{x}^r||^2 \leq f(\mathbf{x}^r) - f(\mathbf{y}) - \frac{\lambda - \delta_B}{2}||\mathbf{y} - \mathbf{x}^r||^2 \,. \tag{B.59}$$

For Algorithm 1 with control variate 1, the left-hand side is equal to $F_{i,r}(\mathbf{x}^r) - F_{i,r}(\mathbf{y})$.

Proof. Let $h_i := f - f_i$. Using the definition of $\nabla F_{i,r}(\mathbf{x})$, we get:

$$F_{i,r}(\mathbf{x}^r) - F_{i,r}(\mathbf{y}) = f_i(\mathbf{x}^r) + \langle \nabla h_i(\mathbf{x}^r), \mathbf{x}^r - \mathbf{y}\rangle - f_i(\mathbf{y}) - \frac{\lambda}{2}||\mathbf{y} - \mathbf{x}^r||^2 \tag{B.60}$$

$$= f(\mathbf{x}^r) - f(\mathbf{y}) + h_i(\mathbf{y}) - h_i(\mathbf{x}^r) - \langle \nabla h_i(\mathbf{x}^r), \mathbf{y} - \mathbf{x}^r\rangle - \frac{\lambda}{2}||\mathbf{y} - \mathbf{x}^r||^2 \tag{B.61}$$

$$\overset{(2)}{\leq} f(\mathbf{x}^r) - f(\mathbf{y}) - \frac{\lambda - \delta_B}{2}||\mathbf{y} - \mathbf{x}^r||^2 \,. \tag{B.62}$$

□

Lemma 28. *Consider Algorithm 1 with control variate (1). Let the global model be updated by choosing an arbitrary local model with an index set $(i_r)_{r=0}^{+\infty}$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Suppose that the solutions returned by local solvers satisfy $F_{i,r}(\mathbf{x}_{i,r+1}) \leq F_{i,r}(\mathbf{x}^r)$ for any $i \in [n]$. Then for any $r \geq 0$, it holds that:*

$$f(\mathbf{x}_{i,r+1}) + \frac{\lambda - \delta_B}{2}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2 \leq f(\mathbf{x}^r) \,. \tag{B.63}$$

and thus:

$$f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r) \,. \tag{B.64}$$

Proof. According to Lemma 27 with $\mathbf{y} = \mathbf{x}_{i,r+1}$, we get:

$$F_{i,r}(\mathbf{x}^r) - F_{i,r}(\mathbf{x}_{i,r+1}) \leq f(\mathbf{x}^r) - f(\mathbf{x}_{i,r+1}) - \frac{\lambda - \delta_B}{2}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2 \,. \tag{B.65}$$

By the assumption that the function value decreases locally, we have that:

$$f(\mathbf{x}_{i,r+1}) + \frac{\lambda - \delta_B}{2}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2 \leq f(\mathbf{x}^r) \,. \tag{B.66}$$

□

Since the previous display holds for any $i \in [n]$ and $\mathbf{x}^{r+1} = \mathbf{x}_{i_r,r+1}$, we have: $f(\mathbf{x}^{r+1}) \leq f(\mathbf{x}^r)$.

Theorem 29. *Consider Algorithm 1 with control variate (1). Let the global model be updated by choosing an arbitrary local model with an index set $(i_r)_{r=0}^{+\infty}$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Suppose that the solutions returned by local solvers satisfy $F_{i,r}(\mathbf{x}_{i,r+1}) \leq F_{i,r}(\mathbf{x}^r)$ and $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \leq e_{r+1}$ for any $r \geq 0$ and any $i \in [n]$. Let $\lambda = a\delta_B$ with $a > 1$. Then after R communication rounds, we have:*

$$||\nabla f(\bar{\mathbf{x}}^R)||^2 \leq \frac{4(a+1)^2}{(a-1)} \frac{\delta_B(f(\mathbf{x}^0) - f^\star)}{R} + 2\frac{1}{R}\sum_{r=1}^{R} e_r^2 \,. \tag{B.67}$$

where $\bar{\mathbf{x}}^R = \arg\min_{\mathbf{x} \in \{\mathbf{x}^r\}_{r=0}^R} \{||\nabla f(\mathbf{x})||\}$.

Proof. According to Lemma 28, for any $r \geq 0$ and any $i \in [n]$, it holds that

$$f(\mathbf{x}_{i,r+1}) + \frac{\lambda - \delta_B}{2}||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2 \leq f(\mathbf{x}^r) \,. \tag{B.68}$$

It remains to lower bound $||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2$. Recall that the solution returned by the local solver satisfies $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \leq e_{r+1}$. Using Lemma 26 with $\mathbf{x} = \mathbf{x}_{i,r+1}$, we get:

$$e_{r+1} \geq ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})|| \geq ||\nabla f(\mathbf{x}_{i,r+1})|| - \delta_B||\mathbf{x}^r - \mathbf{x}_{i,r+1}|| - \lambda||\mathbf{x}_{i,r+1} - \mathbf{x}^r|| \,. \tag{B.69}$$

It follows that:

$$||\mathbf{x}_{i,r+1} - \mathbf{x}^r||^2 \geq \left(\frac{||\nabla f(\mathbf{x}_{i,r+1})||}{\delta_B + \lambda} - \frac{e_{r+1}}{\delta_B + \lambda}\right)^2 \overset{(A.3)}{\geq} \frac{||\nabla f(\mathbf{x}_{i,r+1})||^2}{2(\lambda + \delta_B)^2} - \frac{e_{r+1}^2}{(\lambda + \delta_B)^2} \,. \tag{B.70}$$

Plugging this bound into the previous display, we get, for any $i \in [n]$:

$$\frac{\lambda - \delta_B}{4(\lambda + \delta_B)^2} ||\nabla f(\mathbf{x}_{i,r+1})||^2 \leq f(\mathbf{x}^r) - f(\mathbf{x}_{i,r+1}) + \frac{\lambda - \delta_B}{2(\lambda + \delta_B)^2} e_{r+1}^2 \ . \tag{B.71}$$

Substituting $\lambda = a\delta_B$, using the fact that $\mathbf{x}^{r+1} = \mathbf{x}_{i_r,r+1}$ and the fact that the previous display holds for any $i \in [n]$, we obtain:

$$\frac{a-1}{4(a+1)^2\delta_B} ||\nabla f(\mathbf{x}^{r+1})||^2 \leq f(\mathbf{x}^r) - f(\mathbf{x}^{r+1}) + \frac{a-1}{2(a+1)^2\delta_B} e_{r+1}^2 \ . \tag{B.72}$$

Summing up from $r = 0$ to $r = R - 1$, and dividing both sides by R, we get the claim. □

Corollary 30. *Consider Algorithm 1 with control variate (1). Let the global model be updated by choosing an arbitrary local model with an index set $(i_r)_{r=0}^{+\infty}$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Suppose for any $r \geq 0$, each local solver runs the standard gradient descent starting from $\mathbf{x}^r$ until $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq \frac{18\delta_B(f(\mathbf{x}^0)-f^\star)}{R}$ is satisfied. Let $\lambda_r = 2\delta_B$. After R communication rounds, we have:*

$$||\nabla f(\bar{\mathbf{x}}^R)||^2 \leq \frac{72\delta_B(f(\mathbf{x}^0) - f^\star)}{R} \ . \tag{B.73}$$

where $\bar{\mathbf{x}}^R = \arg\min_{\mathbf{x} \in \{\mathbf{x}^r\}_{r=0}^R} \{||\nabla f(\mathbf{x})||\}$. For any $r \geq 0$, the required number of local steps K_r at round r is no more than $\Theta\Big(\frac{L}{\delta_B} R\Big)$.

Proof. According to Theorem 29, for any $R \geq 1$, we have:

$$||\nabla f(\bar{\mathbf{x}}^R)||^2 \leq \frac{36\delta_B(f(\mathbf{x}^0) - f^\star)}{R} + 2\frac{1}{R}\sum_{r=1}^{R} e_r^2 \ . \tag{B.74}$$

For any $1 \leq r \leq R$, let $e_r^2 \leq \frac{18\delta_B(f(\mathbf{x}^0)-f^\star)}{R}$. We further get:

$$||\nabla f(\bar{\mathbf{x}}^R)||^2 \leq \frac{72\delta_B(f(\mathbf{x}^0) - f^\star)}{R} \ . \tag{B.75}$$

We next estimate the number of local steps required to have $||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq \frac{18\delta_B(f(\mathbf{x}^0)-f^\star)}{R}$. By running standard gradient descent on $F_{i,r}$ at each round r for $K_{i,r}$ steps and returning the point with the minimum gradient norm, we get:

$$||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq \frac{(\lambda + L)(F_{i,r}(\mathbf{x}^r) - F_{i,r}^\star)}{K_{i,r}} \ . \tag{B.76}$$

where $F_{i,r}^\star = \inf_x\{F_{i,r}(\mathbf{x})\}$. According to Lemma 27, for any $\mathbf{y} \in \mathbb{R}^d$, we have:

$$F_{i,r}(\mathbf{x}^r) - F_{i,r}(\mathbf{y}) \leq f(\mathbf{x}^r) - f(\mathbf{y}) - \frac{\lambda - \delta_B}{2} ||\mathbf{y} - \mathbf{x}^r||^2 \tag{B.77}$$

$$\leq f(\mathbf{x}^r) - f(\mathbf{y}) \leq f(\mathbf{x}^r) - f^\star \ . \tag{B.78}$$

It follows that:

$$F_{i,r}(\mathbf{x}^r) - F_{i,r}^\star \leq f(\mathbf{x}^r) - f^\star \ . \tag{B.79}$$

According to Lemma 28, the function value monotonically decreases. Hence, we get:

$$F_{i,r}(\mathbf{x}^r) - F_{i,r}^\star \leq f(\mathbf{x}^0) - f^\star, \qquad \text{and} \qquad ||\nabla F_{i,r}(\mathbf{x}_{i,r+1})||^2 \leq \frac{(\lambda + L)(f(\mathbf{x}^0) - f^\star)}{K_{i,r}} \ . \tag{B.80}$$

Let $\frac{(\lambda+L)(f(\mathbf{x}^0)-f^\star)}{K_{i,r}} \leq \frac{18\delta_B(f(\mathbf{x}^0)-f^\star)}{R}$. We obtain, for any $i \in [n]$:

$$K_{i,r} \geq \frac{\lambda + L}{18\delta_B} R = \Theta\Big(\frac{L}{\delta_B} R\Big) \ . \tag{B.81}$$

This concludes the proof. □

B.3. Algorithm 1 with a special control variate under strong convexity

In this section, we consider the following choice of control variate:

$$\mathbf{h}_{i,r+1} := m(\mathbf{x}^{r+1} - \mathbf{x}_{i,r+1}) + \mathbf{h}_{i,r} \ . \tag{B.82}$$

with $m \geq 0$ and $\frac{1}{n}\sum_{i=1}^n \mathbf{h}_{i,0} = 0$.

Lemma 31. *Consider Algorithm 1 with the standard averaging and control variate (B.82) with $m = \lambda$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-convex with $\mu > 0$. Suppose that $n \geq 2$ and that each local solver provides the exact solution, then for any $\mathbf{x} \in \mathbb{R}^d$, we have:*

$$\begin{aligned} &||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 + \frac{2\mu L}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 + \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r+1} - \nabla f_i(\mathbf{x}^\star)||^2 \leq ||\mathbf{x}^r - \mathbf{x}^\star||^2 \\ &+ \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 - \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2 - \frac{2}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \ . \end{aligned} \tag{B.83}$$

Proof. Recall that $\mathbf{h}_{i,r+1} := \lambda(\mathbf{x}^{r+1} - \mathbf{x}_{i,r+1}) + \mathbf{h}_{i,r}$ and thus we have:

$$||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 + \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r+1} - \nabla f_i(\mathbf{x}^\star)||^2 \overset{(B.82)}{=} ||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}^{r+1} - \mathbf{x}_{i,r+1} + \frac{1}{\lambda}(\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star))||^2 \tag{B.84}$$

$$\overset{(A.15)}{=} \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^\star - \frac{1}{\lambda}(\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star))||^2 \ . \tag{B.85}$$

where we use the fact that $\frac{1}{n}\sum_{i=1}^n [\mathbf{x}_{i.r+1} - \frac{1}{\lambda}(\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star))] = \mathbf{x}^{r+1}$.

Note that $\nabla F_{i,r}(\mathbf{x}_{i,r+1}) = \mathbf{0} = \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r} + \lambda(\mathbf{x}_{i,r+1} - \mathbf{x}^r)$, it follows that:

$$||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 + \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r+1} - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.86}$$

$$= \frac{1}{n}\sum_{i=1}^n \Big|\Big|\mathbf{x}^r - \mathbf{x}^\star - \frac{1}{\lambda}\big(\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\big)\Big|\Big|^2 \tag{B.87}$$

We now upper bound the last display. Unrolling it gives:

$$\frac{1}{n}\sum_{i=1}^n \Big|\Big|\mathbf{x}^r - \mathbf{x}^\star - \frac{1}{\lambda}\big(\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\big)\Big|\Big|^2 \tag{B.88}$$

$$= ||\mathbf{x}^r - \mathbf{x}^\star||^2 - \frac{2}{\lambda}\frac{1}{n}\sum_{i=1}^n \langle \mathbf{x}^r - \mathbf{x}^\star, \nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\rangle + \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.89}$$

Note that the inner product can be lower bounded by:

$$\langle \mathbf{x}^r - \mathbf{x}^\star, \nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\rangle \tag{B.90}$$

$$= \langle \mathbf{x}^r - \mathbf{x}_{i,r+1}, \nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\rangle + \langle \mathbf{x}_{i,r+1} - \mathbf{x}^\star, \nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\rangle \tag{B.91}$$

$$\overset{(A.19)}{\geq} \frac{1}{\lambda}\langle \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}, \nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\rangle + \frac{\mu L}{\mu+L}||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 + \frac{1}{\mu+L}||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.92}$$

$$= \frac{1}{\lambda}||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 + \frac{1}{\lambda}\langle \nabla f_i(\mathbf{x}^\star) - \mathbf{h}_{i,r}, \nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\rangle + \frac{\mu L}{\mu+L}||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 \tag{B.93}$$

$$+ \frac{1}{\mu+L}||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.94}$$

$$= \frac{1}{\lambda}||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 - \frac{1}{\lambda}||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 + \frac{1}{\lambda}\langle \nabla f_i(\mathbf{x}^\star) - \mathbf{h}_{i,r}, \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}\rangle \tag{B.95}$$

$$+ \frac{\mu L}{\mu+L}||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 + \frac{1}{\mu+L}||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \ . \tag{B.96}$$

Plugging (B.96) into (B.89), we obtain:

$$\frac{1}{n}\sum_{i=1}^n \Big|\Big|\mathbf{x}^r - \mathbf{x}^\star - \frac{1}{\lambda}\big(\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\big)\Big|\Big|^2 \tag{B.97}$$

$$\leq ||\mathbf{x}^r - \mathbf{x}^\star||^2 - \frac{2\mu L}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 - \left(\frac{1}{\lambda^2} + \frac{2}{\lambda(\mu+L)}\right)\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.98}$$

$$+ \frac{2}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 + \frac{2}{\lambda^2}\frac{1}{n}\sum_{i=1}^n \langle \mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star), \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}\rangle\ . \tag{B.99}$$

Further note that:

$$- ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 + 2\langle \mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star), \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}\rangle \tag{B.100}$$

$$= -\Big(||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 + \langle \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}, \nabla f_i(\mathbf{x}_{i,r+1}) + \mathbf{h}_{i,r} - 2\nabla f_i(\mathbf{x}^\star)\rangle\Big) \tag{B.101}$$

$$+ 2\langle \mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star), \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}\rangle \tag{B.102}$$

$$= -||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 - ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2\ . \tag{B.103}$$

Hence (B.99) can be simplified to:

$$\frac{1}{n}\sum_{i=1}^n \Big|\Big|\mathbf{x}^r - \mathbf{x}^\star - \frac{1}{\lambda}\big(\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)\big)\Big|\Big|^2 \tag{B.104}$$

$$\leq ||\mathbf{x}^r - \mathbf{x}^\star||^2 - \frac{2\mu L}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 + \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 - \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2 \tag{B.105}$$

$$- \frac{2}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2\ . \tag{B.106}$$

Plugging this upper bound into (B.87) and rearranging give the claim. □

Corollary 32. *Consider Algorithm 1 with the standard averaging and control variate* (B.82) *with* $m = \lambda = \sqrt{\mu L}$*. Let* $f_i : \mathbb{R}^d \to \mathbb{R}$ *be* L*-smooth and* μ*-strongly convex with* $\mu > 0$*. Suppose that* $n \geq 2$ *and that each local solver provides the exact solution, then after* R *communication rounds, we have:*

$$\Phi^R \leq \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \frac{\mu}{\sqrt{L}} + 2\sqrt{\mu}}\right)^R \Phi^0\ , \tag{B.107}$$

where $\Phi^r := ||\mathbf{x}^r - \mathbf{x}^\star||^2 + \frac{(\mu+L)}{\mu L(\mu+L) + 2\mu L\sqrt{\mu L}}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2$.

Proof. According to Lemma 31, for any $r \geq 0$, we have:

$$\begin{aligned} &||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 + \frac{2\mu L}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,r+1} - \mathbf{x}^\star||^2 + \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r+1} - \nabla f_i(\mathbf{x}^\star)||^2 \leq ||\mathbf{x}^r - \mathbf{x}^\star||^2 \\ &+ \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 - \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2 - \frac{2}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2\ . \end{aligned} \tag{B.108}$$

It remains to upper bound the last two terms. Note that:

$$||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.109}$$

$$= ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 + 2\langle \mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star), \nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}\rangle + ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2 \tag{B.110}$$

$$\overset{(A.3)}{\geq} (1 - 2\frac{1}{2\beta_r})||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 + (1 - 2\frac{\beta_r}{2})||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2\ . \tag{B.111}$$

It follows that:

$$- \frac{1}{\lambda^2}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2 - \frac{2}{\lambda(\mu+L)}\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \nabla f_i(\mathbf{x}^\star)||^2 \tag{B.112}$$

$$\leq -\frac{2}{\lambda(\mu+L)}(1 - \frac{1}{\beta_r})\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2 + \left[\frac{2}{\lambda(\mu+L)}(\beta_r - 1) - \frac{1}{\lambda^2}\right]\frac{1}{n}\sum_{i=1}^n ||\nabla f_i(\mathbf{x}_{i,r+1}) - \mathbf{h}_{i,r}||^2\ . \tag{B.113}$$

Let the second coefficient be zero. We get the largest possible $\beta_r = \frac{\mu+L}{2\lambda} + 1$. We obtain the main recurrence:

$$||\mathbf{x}^{r+1} - \mathbf{x}^\star||^2 + \frac{(\mu+L)}{\lambda^2(\mu+L)+2\mu L\lambda}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r+1} - \nabla f_i(\mathbf{x}^\star)||^2 \leq \tag{B.114}$$

$$\max\Big\{\frac{\lambda(\mu+L)}{\lambda(\mu+L)+2\mu L}, 1 - \frac{2\lambda}{2\lambda+\mu+L}\Big\}\Big(||\mathbf{x}^r - \mathbf{x}^\star||^2 + \frac{(\mu+L)}{\lambda^2(\mu+L)+2\mu L\lambda}\frac{1}{n}\sum_{i=1}^n ||\mathbf{h}_{i,r} - \nabla f_i(\mathbf{x}^\star)||^2\Big) . \tag{B.115}$$

The left component of the contraction factor is increasing in λ while the right component is decreasing in λ. Therefore, it is clear that the best λ is such that the left and the right components are equal. This gives us exactly $\lambda^\star = \sqrt{\mu L}$ and $\frac{\lambda^\star(\mu+L)}{\lambda^\star(\mu+L)+2\mu L} = 1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\frac{\mu}{\sqrt{L}}+2\sqrt{\mu}}$. □

In the context of Algorithm 3, control Variate (B.82) becomes:

$$\mathbf{h}_{i,k+1} := m_k(\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k+1}) + \mathbf{h}_{i,k} . \tag{B.116}$$

with $m_k \geq 0$ and $\frac{1}{n}\sum_{i=1}^n \mathbf{h}_{i,0} = 0$.

B.4. Convex result and proof for Algorithm 2 with control variate (4)

The proofs in this section are similar to the ones for Algorithm 3.

Theorem 33. *Consider Algorithm 2 with control variate (4) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$ for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD. By choosing $p = \frac{\lambda+\mu/2}{\eta+\mu/2}$ and $\eta \geq \lambda \geq \delta_A$, for any $K \geq 1$, it holds that:*

$$\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] + \frac{\mu}{4}\mathbb{E}\Big[||\tilde{\mathbf{x}}_K - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,K} - \mathbf{x}^\star||^2\Big] \leq \frac{\mu}{2}\frac{||\mathbf{x}_0 - \mathbf{x}^\star||^2}{(1+\frac{\mu}{2\eta})^K - 1} \leq \frac{\eta||\mathbf{x}_0 - \mathbf{x}^\star||^2}{K} . \tag{B.117}$$

where $\bar{\bar{\mathbf{x}}}_K := \sum_{k=1}^K \frac{1}{q^k}\bar{\mathbf{x}}_k / \sum_{k=1}^K \frac{1}{q^k}$, $\bar{\mathbf{x}}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,k}$, and $q := 1 - \frac{\mu}{2\eta+\mu}$.

Proof. Recall that the updates of Algorithm 3 satisfies:

$$\mathbf{x}_{i,k+1} = \arg\min\{f_i(\mathbf{x}) - \langle \mathbf{x}, \mathbf{h}_{i,k}\rangle + \frac{\eta}{2}||\mathbf{x} - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2}||\mathbf{x} - \tilde{\mathbf{x}}_k||^2\} . \tag{B.118}$$

Using strong convexity, we get:

$$f_i(\mathbf{x}^\star) - \langle \mathbf{x}^\star, \mathbf{h}_{i,k}\rangle + \frac{\eta}{2}||\mathbf{x}^\star - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2}||\mathbf{x}^\star - \tilde{\mathbf{x}}_k||^2 \overset{(A.1)}{\geq} f_i(\mathbf{x}_{i,k+1}) - \langle \mathbf{x}_{i,k+1}, \mathbf{h}_{i,k}\rangle + \frac{\eta}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 \tag{B.119}$$

$$+ \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\eta+\lambda+\mu}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 . \tag{B.120}$$

By convexity of f_i, we further get:

$$f_i(\mathbf{x}^\star) - \langle \mathbf{x}^\star, \mathbf{h}_{i,k}\rangle + \frac{\eta}{2}||\mathbf{x}^\star - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2}||\mathbf{x}^\star - \tilde{\mathbf{x}}_k||^2 \overset{(A.1)}{\geq} f_i(\bar{\mathbf{x}}_{k+1}) + \langle \nabla f_i(\bar{\mathbf{x}}_{k+1}), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1})\rangle - \langle \mathbf{x}_{i,k+1}, \mathbf{h}_{i,k}\rangle \tag{B.121}$$

$$+ \frac{\eta}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\eta+\lambda+\mu}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 . \tag{B.122}$$

Taking the average on both sides over $i = 1$ to n, we get:

$$f(\mathbf{x}^\star) + \frac{\eta}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 + \frac{\lambda}{2}||\bar{\mathbf{x}}_k - \mathbf{x}^\star||^2 \geq f(\bar{\mathbf{x}}_{k+1}) + \frac{1}{n}\sum_{i=1}^n\Big[\langle \nabla f_i(\bar{\mathbf{x}}_{k+1}), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle - \langle \mathbf{x}_{i,k+1}, \mathbf{h}_{i,k}\rangle\Big]$$
$$+ \frac{1}{n}\sum_{i=1}^n \frac{\eta}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 + \frac{1}{n}\sum_{i=1}^n \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{1}{n}\sum_{i=1}^n \frac{\eta+\lambda+\mu}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 . \tag{B.123}$$

Note that: $\frac{1}{n}\sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_{k+1}), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle = \frac{1}{n}\sum_{i=1}^n \langle \bar{\mathbf{x}}_{k+1}, \mathbf{h}_{i,k}\rangle$. Let $h_i := f_i - f$. It follows that:

$$\frac{1}{n}\sum_{i=1}^n\Big[\langle \nabla f_i(\bar{\mathbf{x}}_{k+1}), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle - \langle \mathbf{x}_{i,k+1}, \mathbf{h}_{i,k}\rangle\Big] + \frac{1}{n}\sum_{i=1}^n \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 \tag{B.124}$$

$$\overset{(A.15)}{=} \frac{1}{n}\sum_{i=1}^n\Big[\langle \nabla h_i(\bar{\mathbf{x}}_{k+1}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}||^2 + \frac{\lambda}{2}||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2\Big] \tag{B.125}$$

$$\overset{(A.5)}{\geq} \frac{1}{n}\sum_{i=1}^n\Big[-\frac{1}{2\lambda}||\nabla h_i(\bar{\mathbf{x}}_{k+1}) - \nabla h_i(\tilde{\mathbf{x}}_k)||^2 + \frac{\lambda}{2}||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2\Big] \tag{B.126}$$

$$\overset{(3)}{\geq} \frac{1}{n}\sum_{i=1}^n\Big[-\frac{\delta_A^2}{2\lambda}||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\lambda}{2}||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2\Big] \geq 0\ . \tag{B.127}$$

where in the last inequality, we use the fact that $\lambda \geq \delta_A$.

The main recurrence is then simplified as (after dropping the non-negative $\frac{\eta}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2$):

$$\begin{aligned} f(\mathbf{x}^\star) + \frac{\eta}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 + \frac{\lambda}{2}||\bar{\mathbf{x}}_k - \mathbf{x}^\star||^2 &\geq f(\bar{\mathbf{x}}_{k+1}) + \frac{1}{n}\sum_{i=1}^n\frac{\eta}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 + \frac{1}{n}\sum_{i=1}^n\frac{\eta+\lambda+\mu}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 \\ &\overset{(A.15)}{\geq} f(\bar{\mathbf{x}}_{k+1}) + \frac{\eta+\mu/2}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 + \frac{\lambda+\mu/2}{2}||\bar{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2\ . \end{aligned} \tag{B.128}$$

Taking expectation w.r.t θ_k, we have:

$$\frac{\lambda+\mu/2}{2p}\,\mathbb{E}_{\theta_k}[||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2] = \frac{\lambda+\mu/2}{2}||\bar{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2 + \frac{\lambda+\mu/2}{2}(\frac{1}{p} - 1)||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2\ , \tag{B.129}$$

Adding both sides by $\frac{\lambda+\mu/2}{2}(\frac{1}{p}-1)||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2$ and taking the expectation w.r.t θ_k, we get:

$$\begin{aligned} f(\mathbf{x}^\star) + \frac{\eta}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 + \Big(\frac{\lambda+\mu/2}{2p} - \frac{\mu}{4}\Big)||\bar{\mathbf{x}}_k - \mathbf{x}^\star||^2 \geq \mathbb{E}_{\theta_k}\Big[f(\bar{\mathbf{x}}_{k+1}) + \frac{\eta+\mu/2}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 \\ + \frac{\lambda+\mu/2}{2p}||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2\Big]\ . \end{aligned} \tag{B.130}$$

By our choice of $p = \frac{\lambda+\mu/2}{\eta+\mu/2}$ with $\eta \geq \lambda$, we get:

$$\frac{\eta+\mu/2}{2} = \frac{\lambda+\mu/2}{2p}, \quad \frac{\lambda+\mu/2}{2p} - \frac{\mu}{4} = \frac{\eta}{2}\ . \tag{B.131}$$

Taking the full expectation of the main recurrence, dividing both sides by $\frac{\eta+\mu/2}{2}$, applying Lemma 13, using the convexity of f, and multiplying both sides by $\frac{\eta+\mu/2}{2}$, we get:

$$\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] + \frac{\mu}{4}\,\mathbb{E}\Big[||\tilde{\mathbf{x}}_K - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,K} - \mathbf{x}^\star||^2\Big] \leq \frac{\mu}{2}\frac{||\mathbf{x}_0 - \mathbf{x}^\star||^2}{(1+\frac{\mu}{2\eta})^K - 1}\ . \tag{B.132}$$

□

B.5. Non-convex result and proof for Algorithm 2 with control variate (4)

Theorem 34. *Consider Algorithm 2 with control variate (4) and randomized averaging with random index set $\{i_k\}_{k=0}^{+\infty}$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD. Let $\lambda = \delta_B$, $p = \frac{\lambda}{\eta}$ and $\eta \geq 4\delta_B$. For any $K \geq 1$, it holds that:*

$$\mathbb{E}\Big[||\nabla f(\bar{\mathbf{x}}_K)||^2\Big] \leq \frac{150\eta(f(\mathbf{x}^0) - f^\star)}{K}\ , \tag{B.133}$$

where $\bar{\mathbf{x}}_K$ is uniformly sampled from $(\mathbf{x}_{i_k,k})_{k=0}^{K-1}$.

Proof. Let $h_i := f - f_i$. Recall that for any $i \in [n]$, the update of Algorithm 2 satisfies: $F_{i,k}(\mathbf{x}_{i,k+1}) \leq F_{i,k}(\mathbf{x}_{i,k})$ where $F_{i,k}(\mathbf{x}) := f_i(\mathbf{x}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}\rangle + \frac{\eta}{2}||\mathbf{x} - \mathbf{x}_{i,k}||^2 + \frac{\lambda}{2}||\mathbf{x} - \tilde{\mathbf{x}}_k||^2$. This gives:

$$f_i(\mathbf{x}_{i,k}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 \geq f_i(\mathbf{x}_{i,k+1}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\eta}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2\ . \tag{B.134}$$

According to Lemma 27 with $\mathbf{x}^r = \mathbf{x}_{i,k}$ and $\mathbf{y} = \mathbf{x}_{i,k+1}$, we get:

$$f_i(\mathbf{x}_{i,k}) + \langle \nabla h_i(\mathbf{x}_{i,k}), \mathbf{x}_{i,k} - \mathbf{x}_{i,k+1}\rangle - f_i(\mathbf{x}_{i,k+1}) \le f(\mathbf{x}_{i,k}) - f(\mathbf{x}_{i,k+1}) + \frac{\delta_B}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 \ . \tag{B.135}$$

Substituting this inequality into the previous display, we get:

$$\begin{aligned} f(\mathbf{x}_{i,k}) + \frac{\lambda}{2}||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 \ge f(\mathbf{x}_{i,k+1}) &- \langle \nabla h_i(\mathbf{x}_{i,k}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}\rangle \\ &+ \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\eta - \delta_B}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 \ . \end{aligned} \tag{B.136}$$

For any $\alpha > 0$, we have that:

$$- \langle \nabla h_i(\mathbf{x}_{i,k}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}\rangle = \langle \nabla h_i(\tilde{\mathbf{x}}_k) - \nabla h_i(\mathbf{x}_{i,k}), \mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}\rangle \tag{B.137}$$

$$\overset{\text{(A.3)}}{\ge} -\frac{||\nabla h_i(\tilde{\mathbf{x}}_k) - \nabla h_i(\mathbf{x}_{i,k})||^2}{2\alpha} - \frac{\alpha||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2}{2} \tag{B.138}$$

$$\ge -\frac{\delta_B^2||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k}||^2}{2\alpha} - \frac{\alpha||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2}{2} \ . \tag{B.139}$$

It follows that:

$$f(\mathbf{x}_{i,k}) + \Big(\frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha}\Big)||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 \ge f(\mathbf{x}_{i,k+1}) + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\eta - \delta_B - \alpha}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 \ . \tag{B.140}$$

Suppose that $\eta - \delta_B - \alpha > 0$. We lower bound the last term. Recall that $\mathbf{x}_{i,k+1}$ is a stationary point of $F_{i,k}$. We have:

$$\nabla f_i(\mathbf{x}_{i,k+1}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \eta(\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}) + \lambda(\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k) = 0 \ , \tag{B.141}$$

which implies:

$$\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k} = -\frac{1}{\eta + \lambda}\big(\nabla f_i(\mathbf{x}_{i,k+1}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)\big) \ . \tag{B.142}$$

It follows that:

$$(\eta + \lambda)||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}|| = ||\nabla f_i(\mathbf{x}_{i,k+1}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)|| \tag{B.143}$$

$$\ge ||\nabla f(\mathbf{x}_{i,k+1})|| - ||\nabla h_i(\tilde{\mathbf{x}}_k) - \nabla h_i(\mathbf{x}_{i,k+1})|| - \lambda||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k|| \tag{B.144}$$

$$\overset{(2)}{\ge} ||\nabla f(\mathbf{x}_{i,k+1})|| - \delta_B||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k+1}|| - \lambda||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k|| \ . \tag{B.145}$$

This gives:

$$(\eta + \lambda)^2||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 \overset{\text{(A.3)}}{\ge} \frac{||\nabla f(\mathbf{x}_{i,k+1})||^2}{4} - \frac{\delta_B^2||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k+1}||^2}{2} - \lambda^2||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 \ . \tag{B.146}$$

Plugging this inequality into (B.140), we get, for any $i \in [n]$:

$$f(\mathbf{x}_{i,k}) + A||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 \ge f(\mathbf{x}_{i,k+1}) + B||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + C||\nabla f(\mathbf{x}_{i,k+1})||^2 \ . \tag{B.147}$$

where $A := \frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha} + \frac{\lambda^2(\eta - \delta_B - \alpha)}{2(\lambda+\eta)^2}$, $B := \frac{\lambda}{2} - \frac{\delta_B^2(\eta - \delta_B - \alpha)}{4(\eta+\lambda)^2}$ and $C := \frac{\eta - \delta_B - \alpha}{8(\eta+\lambda)^2}$.

Since the previous display holds for any $i \in [n]$, we can take the expectation w.r.t θ_k and i_k to get:

$$\mathbb{E}_{i_k,\theta_k}[f(\mathbf{x}_{i_k,k}) + A||\mathbf{x}_{i_k,k} - \tilde{\mathbf{x}}_k||^2] \ge \mathbb{E}_{i_k,\theta_k}[f(\mathbf{x}_{i_k,k+1}) + B||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_k||^2 + C||\nabla f(\mathbf{x}_{i_k,k+1})||^2] \ . \tag{B.148}$$

Recall that with probability p, $\tilde{\mathbf{x}}_{k+1} = \mathbf{x}_{i_k,k+1}$. It follows that:

$$\mathbb{E}_{i_k,\theta_k}[||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_{k+1}||^2] = (1-p)\,\mathbb{E}_{i_k,\theta_k}[||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_k||^2] \ . \tag{B.149}$$

Substituting this identity into the previous display, we get:

$$\mathbb{E}_{i_k,\theta_k}[f(\mathbf{x}_{i_k,k}) + A||\mathbf{x}_{i_k,k} - \tilde{\mathbf{x}}_k||^2] \ge \mathbb{E}_{i_k,\theta_k}\left[f(\mathbf{x}_{i_k,k+1}) + \frac{B}{1-p}||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_{k+1}||^2 + C||\nabla f(\mathbf{x}_{i_k,k+1})||^2\right] \ . \tag{B.150}$$

Note that i_{k+1} and i_k follow the same distribution and are independent of $(\theta_k)_{k=0}^{+\infty}$. It follows that: $\mathbb{E}_{i_k,\theta_k}[f(\mathbf{x}_{i_k,k+1})] = \mathbb{E}_{i_{k+1},i_k,\theta_k}[f(\mathbf{x}_{i_{k+1},k+1})]$, $\mathbb{E}_{i_k,\theta_k}[||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_{k+1}||^2] = \mathbb{E}_{i_{k+1},i_k,\theta_k}[||\mathbf{x}_{i_{k+1},k+1} - \tilde{\mathbf{x}}_{k+1}||^2]$, and $\mathbb{E}_{i_k,\theta_k}[||\nabla f(\mathbf{x}_{i_k,k+1})||^2] = \mathbb{E}_{i_{k+1},i_k,\theta_k}[||\nabla f(\mathbf{x}_{i_{k+1},k+1})||^2]$, Taking expectation w.r.t i_{k+1} on both sides of the previous display, substituting these three identities, and then taking the full expectation, we obtain our main recurrence:

$$\mathbb{E}[f(\mathbf{x}_{i_k,k}) + A||\mathbf{x}_{i_k,k} - \tilde{\mathbf{x}}_k||^2] \geq \mathbb{E}\left[f(\mathbf{x}_{i_{k+1},k+1}) + \frac{B}{1-p}||\mathbf{x}_{i_{k+1},k+1} - \tilde{\mathbf{x}}_{k+1}||^2 + C||\nabla f(\mathbf{x}_{i_{k+1},k+1})||^2\right] . \tag{B.151}$$

Suppose that $A \leq \frac{B}{1-p}$ and that $\eta - \delta_B - \alpha > 0$. Summing up from $k = 0$ to $K - 1$ and dividing both sides by K, we get:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[||\nabla f(\mathbf{x}_{i_k,k})||^2\Big] \leq \frac{f(\mathbf{x}^0) - f^\star}{CK} , \tag{B.152}$$

where we use the fact that $\mathbb{E}[||\mathbf{x}_{i_0,0} - \tilde{\mathbf{x}}_0||^2] = 0$.

We next choose the parameters to satisfy $A \leq \frac{B}{1-p}$ and $\eta - \delta_B - \alpha > 0$. Plugging in the definitions of A and B, we get:

$$\frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha} + \frac{\lambda^2(\eta - \delta_B - \alpha)}{2(\lambda+\eta)^2} + \frac{\delta_B^2(\eta - \delta_B - \alpha)}{4(\eta+\lambda)^2(1-p)} \leq \frac{\lambda}{2(1-p)} . \tag{B.153}$$

This is equivalent to:

$$(1-p)\frac{\delta_B^2}{2\alpha} + \frac{\big(\delta_B^2 + 2(1-p)\lambda^2\big)(\eta - \delta_B - \alpha)}{4(\eta+\lambda)^2} \leq \lambda p . \tag{B.154}$$

Note that:

$$(1-p)\frac{\delta_B^2}{2\alpha} + \frac{\big(\delta_B^2 + 2(1-p)\lambda^2\big)(\eta - \delta_B - \alpha)}{4(\eta+\lambda)^2} \leq \frac{\delta_B^2}{2\alpha} + \frac{\big(\delta_B^2 + 2\lambda^2\big)(\eta - \delta_B - \alpha)}{4(\eta+\lambda)^2} \tag{B.155}$$

$$= \frac{\delta_B^2}{2\alpha} - \frac{\delta_B^2 + 2\lambda^2}{4(\eta+\lambda)^2}\alpha + \frac{\big(\delta_B^2 + 2\lambda^2\big)(\eta - \delta_B)}{4(\eta+\lambda)^2} . \tag{B.156}$$

Let $\lambda = \delta_B$, $p = \frac{\lambda}{\eta}$, $\eta \geq \lambda$ and $\alpha = \frac{2}{3}\eta$. We have:

$$\frac{\delta_B^2}{2\alpha} - \frac{\delta_B^2 + 2\lambda^2}{4(\eta+\lambda)^2}\alpha + \frac{\big(\delta_B^2 + 2\lambda^2\big)(\eta - \delta_B)}{4(\eta+\lambda)^2} \leq \lambda^2\left[\frac{1}{2\alpha} + \frac{3(\eta - \alpha)}{4(\eta+\lambda)^2}\right] \leq \lambda^2\Big[\frac{3}{4\eta} + \frac{1}{4\eta}\Big] = \lambda p . \tag{B.157}$$

At the same time, η should satisfy $\eta > 3\delta_B$ such that $\eta - \delta_B - \alpha > 0$. Let $\eta \geq 4\delta_B$, we have:

$$C := \frac{\eta - \delta_B - \alpha}{8(\eta+\lambda)^2} = \frac{\frac{\eta}{3} - \delta_B}{8(\eta + \delta_B)^2} \geq \frac{\frac{\eta}{3} - \frac{\eta}{4}}{8(\eta + \frac{1}{4}\eta)^2} = \frac{\eta}{150} . \tag{B.158}$$

□

B.6. Convex result and proof for Algorithm 3 with control variate (4)

Lemma 35. *Consider Algorithm 3 with control variate (4) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$ and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD with $\delta_A \leq L$. Assume that $\mathbb{E}[g_i(\mathbf{x})] = \nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ and that $\mathbb{E}[||g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})||^2] \leq \sigma^2$. Let $\lambda \geq \delta_A$ and $\eta \geq L$. Then for any $k \geq 0$, it holds that:*

$$\mathbb{E}\left[f(\mathbf{x}^\star) + \left[\frac{\lambda + \mu/2}{2p} - \frac{\mu}{4}\right]||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 + \frac{\eta-\mu}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2\right] \geq \mathbb{E}\left[f(\bar{\mathbf{x}}_{k+1}) + \frac{\lambda + \mu/2}{2p}||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2 + \frac{\eta - \mu/2}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2\right] - \frac{\sigma^2}{2(\eta - L)} . \tag{B.159}$$

Proof. Let $G_{i,k}(\mathbf{x}) := f_i(\mathbf{x}_{i,k}) + \langle g_i(\mathbf{x}_{i,k}), \mathbf{x} - \mathbf{x}_{i,k}\rangle + \frac{\eta}{2}||\mathbf{x} - \mathbf{x}_{i,k}||^2$. Recall that the update of Algorithm 3 satisfies:

$$\mathbf{x}_{i,k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d}\{G_{i,k}(\mathbf{x}) - \langle \mathbf{h}_{i,k}, \mathbf{x}\rangle + \frac{\lambda}{2}||\mathbf{x} - \tilde{\mathbf{x}}_k||^2\} . \tag{B.160}$$

Using strong convexity, we get:

$$G_{i,k}(\mathbf{x}^\star) - \langle \mathbf{h}_{i,k}, \mathbf{x}^\star\rangle + \frac{\lambda}{2}||\mathbf{x}^\star - \tilde{\mathbf{x}}_k||^2 \overset{(A.1)}{\geq} G_{i,k}(\mathbf{x}_{i,k+1}) - \langle \mathbf{h}_{i,k}, \mathbf{x}_{i,k+1}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\lambda+\eta}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 . \tag{B.161}$$

Denote the randomness coming from $g_i(\mathbf{x}_{i,k})$ by ξ_i and denote all the randomness $\{\xi_i\}_{i=1}^n$ by ξ.

Recall that f_i is μ-strongly convex and L-smooth. Suppose that $\eta \geq L$, we get:

$$\mathbb{E}_\xi[G_{i,k}(\mathbf{x}^\star)] \overset{(A.1)}{\leq} f_i(\mathbf{x}^\star) + \frac{\eta-\mu}{2}||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 \; , \tag{B.162}$$

and

$$\mathbb{E}_\xi[G_{i,k}(\mathbf{x}_{i,k+1})] \overset{(A.18)}{\geq} \mathbb{E}_\xi\Big[f_i(\mathbf{x}_{i,k+1}) + \langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1} - \mathbf{x}_{i,k}\rangle + \frac{\eta - L}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2\Big] \tag{B.163}$$

$$\overset{(A.5)}{\geq} \mathbb{E}_\xi\Big[f_i(\mathbf{x}_{i,k+1}) - \frac{||g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k})||^2}{2(\eta - L)}\Big] \tag{B.164}$$

$$\geq \mathbb{E}_\xi[f_i(\mathbf{x}_{i,k+1})] - \frac{\sigma^2}{2(\eta-L)} \; . \tag{B.165}$$

where in the last inequality, we use the assumption that $\mathbb{E}_{\xi_i}[||g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})||^2] \leq \sigma^2$ for any $\mathbf{x} \in \mathbb{R}^d$.

Taking the expectation w.r.t ξ on both sides of (B.161), plugging in these bounds and taking the average on both sides over $i = 1$ to n, we get:

$$f(\mathbf{x}^\star) + \frac{\lambda}{2}||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 + \frac{\eta-\mu}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 \tag{B.166}$$

$$\geq \mathbb{E}_\xi\Big[\frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}_{i,k+1}) + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\lambda+\eta}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 - \frac{1}{n}\sum_{i=1}^n\langle \mathbf{h}_{i,k}, \mathbf{x}_{i,k+1}\rangle\Big] - \frac{\sigma^2}{2(\eta-L)} \tag{B.167}$$

$$\overset{(A.1)}{\geq} \mathbb{E}_\xi\Big[f(\bar{\mathbf{x}}_{k+1}) + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\lambda+\eta}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 + \frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}||^2 \tag{B.168}$$

$$+ \frac{1}{n}\sum_{i=1}^n\langle \nabla f_i(\bar{\mathbf{x}}_{k+1}) - \nabla f(\bar{\mathbf{x}}_{k+1}) - \nabla f_i(\tilde{\mathbf{x}}_k) + \nabla f(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle\Big] - \frac{\sigma^2}{2(\eta-L)} \; , \tag{B.169}$$

where in the last inequality, we use the strong convexity of f_i and the fact that:

$$\frac{1}{n}\sum_{i=1}^n\langle \mathbf{h}_{i,k}, \bar{\mathbf{x}}_{k+1}\rangle = \frac{1}{n}\sum_{i=1}^n\langle \nabla f(\bar{\mathbf{x}}_{k+1}), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle = 0 \; . \tag{B.170}$$

Let $h_i := f_i - f$. By the assumption that $\{f_i\}$ have δ_A-AHD, it follows that:

$$\frac{1}{n}\sum_{i=1}^n\langle \nabla h_i(\bar{\mathbf{x}}_{k+1}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 \tag{B.171}$$

$$\overset{(A.15)}{=} \frac{1}{n}\sum_{i=1}^n\langle \nabla h_i(\bar{\mathbf{x}}_{k+1}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}\rangle + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}||^2 + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2 \tag{B.172}$$

$$\overset{(A.5)}{\geq} -\frac{1}{2\lambda}\frac{1}{n}\sum_{i=1}^n||\nabla h_i(\bar{\mathbf{x}}_{k+1}) - \nabla h_i(\tilde{\mathbf{x}}_k)||^2 + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2 \tag{B.173}$$

$$\overset{(3)}{\geq} -\frac{\delta_A^2}{2\lambda}||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^n||\bar{\mathbf{x}}_{k+1} - \tilde{\mathbf{x}}_k||^2 \geq 0 \; . \tag{B.174}$$

where we use the assumption that $\lambda \geq \delta_A$ in the last inequality. We can then simplify the main recurrence as:

$$\begin{aligned} f(\mathbf{x}^\star) + \frac{\lambda}{2}||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 + \frac{\eta-\mu}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 &\geq \mathbb{E}_\xi\Big[f(\bar{\mathbf{x}}_{k+1}) + \frac{\lambda+\eta}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2\Big] - \frac{\sigma^2}{2(\eta-L)} \\ &\overset{(A.15)}{\geq} \mathbb{E}_\xi\Big[f(\bar{\mathbf{x}}_{k+1}) + \frac{\eta-\mu/2}{2}\frac{1}{n}\sum_{i=1}^n||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2 \\ &\quad + \frac{\lambda+\mu/2}{2}||\bar{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2\Big] - \frac{\sigma^2}{2(\eta-L)} \; . \end{aligned} \tag{B.175}$$

where we drop the non-negative $\frac{\mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k+1} - \bar{\mathbf{x}}_{k+1}||^2$.

Taking expectation w.r.t θ_k, we have:

$$\frac{\lambda + \mu/2}{2p}\mathbb{E}_{\theta_k,\xi}[||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2] = \frac{\lambda + \mu/2}{2}\mathbb{E}_{\xi}[||\bar{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2] + \frac{\lambda + \mu/2}{2}(\frac{1}{p} - 1)||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 \,, \tag{B.176}$$

Adding both sides by $\frac{\lambda+\mu/2}{2}(\frac{1}{p} - 1)||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2$ and taking the expectation over θ_k on both sides of the main recurrence, we get:

$$\begin{aligned} f(\mathbf{x}^\star) + \left[\frac{\lambda + \mu/2}{2p} - \frac{\mu}{4}\right]||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 + \frac{\eta - \mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2 \geq \mathbb{E}_{\theta_k,\xi}\Bigg[f(\bar{\mathbf{x}}_{k+1}) + \frac{\lambda + \mu/2}{2p}||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2 \\ + \frac{\eta - \mu/2}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2\Bigg] - \frac{\sigma^2}{2(\eta - L)} \,. \end{aligned} \tag{B.177}$$

Taking the full expectation on both sides, we get the claim. □

Theorem 36. *Consider Algorithm 3 with control variate (4) and the standard averaging. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, μ-convex with $\mu \geq 0$ and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_A-AHD with $\delta_A + \mu \leq L$. Assume that $\mathbb{E}[g_i(\mathbf{x})] = \nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ and that $\mathbb{E}[||g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})||^2] \leq \sigma^2$. By choosing $\lambda = \delta_A$, $p = \frac{\lambda+\mu/2}{\eta-\mu/2}$, and $\eta > L$, for any $K \geq 1$, it holds that:*

$$\begin{aligned} \mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] + \frac{\mu}{4}\mathbb{E}\left[||\tilde{\mathbf{x}}_K - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,K} - \mathbf{x}^\star||^2\right] &\leq \frac{\mu}{2}\frac{1}{(1 + \frac{\mu}{2\eta - 2\mu})^K - 1}||\mathbf{x}_0 - \mathbf{x}^\star||^2 + \frac{\sigma^2}{2(\eta - L)} \\ &\leq \frac{\eta}{K}||\mathbf{x}^0 - \mathbf{x}^\star||^2 + \frac{\sigma^2}{2(\eta - L)} \,. \end{aligned} \tag{B.178}$$

where $\bar{\bar{\mathbf{x}}}_K := \sum_{k=1}^K \frac{1}{q^k}\bar{\mathbf{x}}_k / \sum_{k=1}^K \frac{1}{q^k}$, $\bar{\mathbf{x}}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,k}$, and $q := 1 - \frac{\mu}{2\eta - \mu}$.

To reach ε-accuracy, i.e. $\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] \leq \varepsilon$, by choosing $\eta = \frac{\sigma^2}{\varepsilon} + L$, we get:

$$K \leq \left\lceil \left(\frac{2L}{\mu} + \frac{2\sigma^2}{\mu\varepsilon}\right)\ln\left(1 + \frac{\mu||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon}\right)\right\rceil \leq \left\lceil \frac{2L||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon} + \frac{2\sigma^2||\mathbf{x}^0 - \mathbf{x}^\star||^2}{\varepsilon^2}\right\rceil \,. \tag{B.179}$$

Proof. According to Lemma 35, we have:

$$\begin{aligned} \mathbb{E}\Bigg[f(\mathbf{x}^\star) + \left[\frac{\lambda + \mu/2}{2p} - \frac{\mu}{4}\right]||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 + \frac{\eta - \mu}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2\Bigg] \geq \mathbb{E}\Bigg[f(\bar{\mathbf{x}}_{k+1}) + \frac{\lambda + \mu/2}{2p}||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2 \\ + \frac{\eta - \mu/2}{2}\frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2\Bigg] - \frac{\sigma^2}{2(\eta - L)} \,. \end{aligned} \tag{B.180}$$

By our choices of parameters: $\lambda = \delta_A$, and $p = \frac{\lambda+\mu/2}{\eta-\mu/2}$, we have:

$$\frac{\lambda + \mu/2}{2p} = \frac{\eta - \mu/2}{2}, \quad \frac{\eta - \mu}{\eta - \mu/2} = \frac{\frac{\lambda+\mu/2}{2p} - \frac{\mu}{4}}{\frac{\lambda+\mu/2}{2p}} = 1 - \frac{\mu}{2\eta - \mu} \,. \tag{B.181}$$

Plugging these parameters into the main recurrence and dividing both sides by $\frac{\eta-\mu/2}{2}$ we get:

$$\begin{aligned} &\mathbb{E}\left[\frac{2}{\eta - \mu/2}[f(\bar{\mathbf{x}}_{k+1}) - f^\star] + \left[||\tilde{\mathbf{x}}_{k+1} - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k+1} - \mathbf{x}^\star||^2\right]\right] \\ &\leq \mathbb{E}\Bigg[\underbrace{\left(1 - \frac{\mu}{2\eta - \mu}\right)}_{:=q}\left[||\tilde{\mathbf{x}}_k - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,k} - \mathbf{x}^\star||^2\right]\Bigg] + \frac{2\sigma^2}{(\eta - L)(2\eta - \mu)} \,. \end{aligned} \tag{B.182}$$

Applying Lemma 13, using the convexity of f and multiplying both sides by $\frac{\eta-\mu/2}{2}$, we get:

$$\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star] + \frac{\mu}{4}\mathbb{E}\left[||\tilde{\mathbf{x}}_K - \mathbf{x}^\star||^2 + \frac{1}{n}\sum_{i=1}^n ||\mathbf{x}_{i,K} - \mathbf{x}^\star||^2\right] \leq \frac{\mu}{2}\frac{1}{(1 + \frac{\mu}{2\eta - 2\mu})^K - 1}||\mathbf{x}_0 - \mathbf{x}^\star||^2 + \frac{\sigma^2}{2(\eta - L)} \,. \tag{B.183}$$

This concludes the proof for the first claim.

We next prove the second claim. To achieve $\mathbb{E}[f(\bar{\bar{\mathbf{x}}}_K) - f^\star]$, it is sufficient to let

$$\begin{cases} \frac{\mu}{2}\frac{1}{(1+\frac{\mu}{2\eta-2\mu})^{K}-1}||\mathbf{x}_0-\mathbf{x}^\star||^2 \le \frac{\varepsilon}{2} \\ \frac{\sigma^2}{2(\eta-L)} = \frac{\varepsilon}{2} \end{cases} \Rightarrow \begin{cases} K \ge \frac{2\eta-\mu}{\mu}\ln\Big(1+\frac{\mu||\mathbf{x}^0-\mathbf{x}^\star||^2}{\varepsilon}\Big) \\ \eta = \frac{\sigma^2}{\varepsilon}+L \end{cases} . \tag{B.184}$$

Plugging $\eta = \frac{\sigma^2}{\varepsilon} + L$ into the expression for K, we get:

$$K \ge \Big(\frac{2L}{\mu} + \frac{2\sigma^2}{\mu\varepsilon}\Big)\ln\Big(1+\frac{\mu||\mathbf{x}^0-\mathbf{x}^\star||^2}{\varepsilon}\Big) . \tag{B.185}$$

□

B.7. Non-convex result and proof for Algorithm 3 with control variate (4)

Theorem 37. *Consider Algorithm 3 with control variate (4) and randomized averaging with random index set $\{i_k\}_{k=0}^{+\infty}$. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and L-smooth for any $i \in [n]$. Assume that $\{f_i\}$ have δ_B-BHD with $\delta_B \le L$. Assume that $\mathbb{E}[g_i(\mathbf{x})] = \nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ and that $\mathbb{E}[||g_i(\mathbf{x}) - \nabla f_i(\mathbf{x})||^2] \le \sigma^2$. By choosing $\eta = 3L + \sqrt{9L^2 + \frac{L\sigma^2 K}{f(\mathbf{x}^0)-f^\star}}$, $\lambda = \delta_B$ and $p = \frac{\delta_B}{L}$, for any $K \ge 1$, it holds that:*

$$\mathbb{E}\big[||\nabla f(\bar{\mathbf{x}}_K)||^2\big] \le \frac{96L(f(\mathbf{x}^0)-f^\star)}{K} + 24\sqrt{\frac{L(f(\mathbf{x}^0)-f^\star)}{K}}\sigma . \tag{B.186}$$

where $\bar{\mathbf{x}}_K$ is uniformly sampled from $(\mathbf{x}_{i_k,k})_{k=0}^{K-1}$.

Proof. Let $h_i := f - f_i$. Let $G_{i,k}(\mathbf{x}) := f_i(\mathbf{x}_{i,k}) + \langle g_i(\mathbf{x}_{i,k}), \mathbf{x}-\mathbf{x}_{i,k}\rangle + \frac{\eta}{2}||\mathbf{x}-\mathbf{x}_{i,k}||^2$. Recall that for any $i \in [n]$, the update of Algorithm 3 satisfies:

$$\mathbf{x}_{i,k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d}\{G_{i,k}(\mathbf{x}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}\rangle + \frac{\lambda}{2}||\mathbf{x}-\tilde{\mathbf{x}}_k||^2\} . \tag{B.187}$$

Using strong convexity, we have:

$$\begin{aligned} f_i(\mathbf{x}_{i,k}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k}-\tilde{\mathbf{x}}_k||^2 \overset{(A.1)}{\ge} & G_{i,k}(\mathbf{x}_{i,k+1}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k+1}-\tilde{\mathbf{x}}_k||^2 \\ & + \frac{\lambda+\eta}{2}||\mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}||^2 . \end{aligned} \tag{B.188}$$

Denote the randomness coming from $g_i(\mathbf{x}_{i,k})$ by ξ_i and denote all the randomness $\{\xi_i\}_{i=1}^n$ by ξ. Denote the randomness from the random selection of the index at iteration k by i_k.

Let $\eta \ge L$. We get:

$$\mathbb{E}_\xi[G_{i,k}(\mathbf{x}_{i,k+1})] \overset{(A.18)}{\ge} \mathbb{E}_\xi\Big[f_i(\mathbf{x}_{i,k+1}) + \langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1}-\mathbf{x}_{i,k}\rangle + \frac{\eta-L}{2}||\mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}||^2\Big] . \tag{B.189}$$

It follows that:

$$\begin{aligned} f_i(\mathbf{x}_{i,k}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k}-\tilde{\mathbf{x}}_k||^2 \ge \mathbb{E}_\xi\Big[& f_i(\mathbf{x}_{i,k+1}) + \langle \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k+1}-\tilde{\mathbf{x}}_k||^2 \\ & + \langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1}-\mathbf{x}_{i,k}\rangle + \frac{\lambda+2\eta-L}{2}||\mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}||^2\Big] . \end{aligned} \tag{B.190}$$

According to Lemma 27 with $\mathbf{x}^r = \mathbf{x}_{i,k}$ and $\mathbf{y} = \mathbf{x}_{i,k+1}$, we get:

$$f_i(\mathbf{x}_{i,k}) + \langle \nabla h_i(\mathbf{x}_{i,k}), \mathbf{x}_{i,k}-\mathbf{x}_{i,k+1}\rangle - f_i(\mathbf{x}_{i,k+1}) \le f(\mathbf{x}_{i,k}) - f(\mathbf{x}_{i,k+1}) + \frac{\delta_B}{2}||\mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}||^2 . \tag{B.191}$$

Substituting this inequality into the previous display, we get:

$$\begin{aligned} f(\mathbf{x}_{i,k}) + \frac{\lambda}{2}||\mathbf{x}_{i,k}-\tilde{\mathbf{x}}_k||^2 \ge \mathbb{E}_\xi\Big[& f(\mathbf{x}_{i,k+1}) - \langle \nabla h_i(\mathbf{x}_{i,k}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}\rangle - \frac{\delta_B}{2}||\mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}||^2 \\ & + \langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1}-\mathbf{x}_{i,k}\rangle + \frac{\lambda}{2}||\mathbf{x}_{i,k+1}-\tilde{\mathbf{x}}_k||^2 + \frac{\lambda+2\eta-L}{2}||\mathbf{x}_{i,k+1}-\mathbf{x}_{i,k}||^2\Big] . \end{aligned} \tag{B.192}$$

For any $\alpha > 0$, we have that:

$$
\begin{aligned}
-\langle \nabla h_i(\mathbf{x}_{i,k}) - \nabla h_i(\tilde{\mathbf{x}}_k), \mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}\rangle &= \langle \nabla h_i(\tilde{\mathbf{x}}_k) - \nabla h_i(\mathbf{x}_{i,k}), \mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}\rangle && \text{(B.193)}\\
&\overset{\text{(A.3)}}{\geq} -\frac{||\nabla h_i(\tilde{\mathbf{x}}_k) - \nabla h_i(\mathbf{x}_{i,k})||^2}{2\alpha} - \frac{\alpha||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2}{2} && \text{(B.194)}\\
&\geq -\frac{\delta_B^2||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k}||^2}{2\alpha} - \frac{\alpha||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2}{2}\ . && \text{(B.195)}
\end{aligned}
$$

Plugging this inequality into (B.192), we obtain, for any $i \in [n]$:

$$
\begin{aligned}
f(\mathbf{x}_{i,k}) + \Big(\frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha}\Big)||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 \geq \mathbb{E}_\xi\Big[f(\mathbf{x}_{i,k+1}) + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{\lambda + 2\eta - L - \delta_B - \alpha}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 \\
+ \langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1} - \mathbf{x}_{i,k}\rangle \Big]\ . \qquad \text{(B.196)}
\end{aligned}
$$

We now lower-bound the last two terms. Recall that $\mathbf{x}_{i,k+1}$ satisfies:

$$
g_i(\mathbf{x}_{i,k}) + \eta(\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k) = 0\ . \qquad \text{(B.197)}
$$

which implies:

$$
\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k} = -\frac{1}{\eta + \lambda}\big(g_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)\big)\ . \qquad \text{(B.198)}
$$

It follows that:

$$
\begin{aligned}
&\mathbb{E}_\xi\Big[\langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1} - \mathbf{x}_{i,k}\rangle\Big] && \text{(B.199)}\\
&= \mathbb{E}_\xi\left[\left\langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), -\frac{1}{\eta+\lambda}\big(g_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)\big)\right\rangle\right] && \text{(B.200)}\\
&= -\frac{1}{\eta+\lambda}\,\mathbb{E}_\xi\Big[\langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}) + \nabla f_i(\mathbf{x}_{i,k})\rangle\Big] && \text{(B.201)}\\
&= -\frac{1}{\eta+\lambda}\,\mathbb{E}_\xi\big[||g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k})||^2\big]\ , && \text{(B.202)}
\end{aligned}
$$

and that:

$$
\begin{aligned}
\mathbb{E}_\xi[||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2] &= \frac{1}{(\eta+\lambda)^2}\,\mathbb{E}_\xi\Big[||g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}) + \nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)||^2\Big] && \text{(B.203)}\\
&= \frac{1}{(\eta+\lambda)^2}\left[\mathbb{E}_\xi\Big[||g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k})||^2\Big] + \mathbb{E}_\xi\Big[||\nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)||^2\Big]\right]\ . && \text{(B.204)}
\end{aligned}
$$

where we use the assumption that $g_i(\mathbf{x}_{i,k})$ is an unbiased estimator of $\nabla f_i(\mathbf{x}_{i,k})$. The last two terms of (B.196) can thus be lower bounded by:

$$
\begin{aligned}
&\mathbb{E}_\xi\left[\frac{\lambda + 2\eta - L - \delta_B - \alpha}{2}||\mathbf{x}_{i,k+1} - \mathbf{x}_{i,k}||^2 + \langle g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k}), \mathbf{x}_{i.k+1} - \mathbf{x}_{i,k}\rangle\right] && \text{(B.205)}\\
&\geq A\,\mathbb{E}_\xi\Big[||\nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)||^2\Big] - B\,\mathbb{E}_\xi\Big[||g_i(\mathbf{x}_{i,k}) - \nabla f_i(\mathbf{x}_{i,k})||^2\Big] && \text{(B.206)}\\
&\geq A\,\mathbb{E}_\xi\Big[||\nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)||^2\Big] - B\sigma^2\ . && \text{(B.207)}
\end{aligned}
$$

where $A := \frac{\lambda + 2\eta - L - \delta_B - \alpha}{2(\eta+\lambda)^2}$ and $B := \frac{\lambda + L + \delta_B + \alpha}{2(\eta+\lambda)^2}$.

Further note that:

$$
\begin{aligned}
||\nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)|| &\geq ||\nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k)|| - \lambda||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k}|| && \text{(B.208)}\\
&\geq ||\nabla f(\mathbf{x}_{i,k})|| - ||\nabla h_i(\tilde{\mathbf{x}}_k) - \nabla h_i(\mathbf{x}_{i,k})|| - \lambda||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k}|| && \text{(B.209)}\\
&\overset{(2)}{\geq} ||\nabla f(\mathbf{x}_{i,k})|| - \delta_B||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k|| - \lambda||\tilde{\mathbf{x}}_k - \mathbf{x}_{i,k}|| && \text{(B.210)}\\
&= ||\nabla f(\mathbf{x}_{i,k})|| - (\delta_B + \lambda)||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||\ . && \text{(B.211)}
\end{aligned}
$$

This gives:

$$||\nabla f_i(\mathbf{x}_{i,k}) + \nabla h_i(\tilde{\mathbf{x}}_k) + \lambda(\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k)||^2 \overset{(A.3)}{\geq} \frac{||\nabla f(\mathbf{x}_{i,k})||^2}{2} - (\delta_B + \lambda)^2 ||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 . \tag{B.212}$$

Substituting all the previous displays into (B.196), we get:

$$f(\mathbf{x}_{i,k}) + C||\mathbf{x}_{i,k} - \tilde{\mathbf{x}}_k||^2 + B\sigma^2 \geq \mathbb{E}_\xi \left[f(\mathbf{x}_{i,k+1}) + \frac{\lambda}{2}||\mathbf{x}_{i,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{A}{2}||\nabla f(\mathbf{x}_{i,k})||^2 \right] . \tag{B.213}$$

where $C := \frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha} + A(\delta_B + \lambda)^2$.

Since the previous display holds for any $i \in [n]$, we can take the expectation w.r.t. θ_k and i_k and get:

$$\mathbb{E}_{i_k,\theta_k} \left[f(\mathbf{x}_{i_k,k}) + C||\mathbf{x}_{i_k,k} - \tilde{\mathbf{x}}_k||^2 + B\sigma^2 \right] \geq \mathbb{E}_{i_k,\theta_k,\xi} \left[f(\mathbf{x}_{i_k,k+1}) + \frac{\lambda}{2}||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_k||^2 + \frac{A}{2}||\nabla f(\mathbf{x}_{i_k,k})||^2 \right] , \tag{B.214}$$

Recall that with probability p, $\tilde{\mathbf{x}}_{k+1} = \mathbf{x}_{i_k,k+1}$. It follows that:

$$\mathbb{E}_{i_k,\theta_k,\xi}[||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_{k+1}||^2] = (1-p)\,\mathbb{E}_{i_k,\theta_k,\xi}[||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_k||^2] , \tag{B.215}$$

Substituting this identity into the previous display, we get:

$$\mathbb{E}_{i_k,\theta_k} \left[f(\mathbf{x}_{i_k,k}) + C||\mathbf{x}_{i_k,k} - \tilde{\mathbf{x}}_k||^2 + B\sigma^2 \right] \geq \mathbb{E}_{i_k,\theta_k,\xi} \left[f(\mathbf{x}_{i_k,k+1}) + \frac{\lambda}{2(1-p)}||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_{k+1}||^2 + \frac{A}{2}||\nabla f(\mathbf{x}_{i_k,k})||^2 \right] . \tag{B.216}$$

Note that i_k and i_{k+1} follow the same distribution and are independent of $\{\theta_k\}_{k=0}^{+\infty}$. It follows that:

$$\mathbb{E}_{i_k,\theta_k,\xi}\left[f(\mathbf{x}_{i_k,k+1}) \right] = \mathbb{E}_{i_{k+1},i_k,\theta_k,\xi}\left[f(\mathbf{x}_{i_{k+1},k+1}) \right], \text{ and } \mathbb{E}_{i_k,\theta_k,\xi}\left[||\mathbf{x}_{i_k,k+1} - \tilde{\mathbf{x}}_{k+1}||^2 \right] = \mathbb{E}_{i_{k+1},i_k,\theta_k,\xi}\left[||\mathbf{x}_{i_{k+1},k+1} - \tilde{\mathbf{x}}_{k+1}||^2 \right] . \tag{B.217}$$

Taking expectation w.r.t i_{k+1} on both sides of the previous display, substituting these two identities, and then taking the full expectation, we obtain our main recurrence:

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_{i_k,k})] + C\,\mathbb{E}[||\mathbf{x}_{i_k,k} - \tilde{\mathbf{x}}_k||^2] + B\sigma^2 \geq \mathbb{E}[f(\mathbf{x}_{i_{k+1},k+1})] &+ \frac{\lambda}{2(1-p)}\,\mathbb{E}[||\mathbf{x}_{i_{k+1},k+1} - \tilde{\mathbf{x}}_{k+1}||^2] \\ &+ \frac{A}{2}\,\mathbb{E}[||\nabla f(\mathbf{x}_{i_k,k})||^2] . \end{aligned} \tag{B.218}$$

It is clear now we need to choose parameters such that $C \leq \frac{\lambda}{2(1-p)}$, which is equivalent to:

$$\frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha} + \frac{\lambda + 2\eta - L - \delta_B - \alpha}{2(\eta+\lambda)^2}(\delta_B + \lambda)^2 \leq \frac{\lambda}{2(1-p)} , \tag{B.219}$$

while at the same time keeping A sufficiently large.

By choosing $p = \frac{\delta_B}{L}$, $\lambda = \delta_B$, $\eta \geq 3L$ and $\alpha = 5L$, we obtain:

$$\frac{\lambda}{2} + \frac{\delta_B^2}{2\alpha} + \frac{\lambda + 2\eta - L - \delta_B - \alpha}{2}\frac{(\delta_B+\lambda)^2}{(\eta+\lambda)^2} \leq \frac{\delta_B}{2} + \frac{\delta_B^2}{10L} + (\eta - 3L)\frac{4\delta_B^2}{\eta^2} \leq \frac{\delta_B}{2} + \frac{\delta_B^2}{2L} \leq \frac{\lambda}{2(1-p)} , \tag{B.220}$$

where in the last second inequality, we use the fact that $\frac{\eta - 3L}{\eta^2} \leq \frac{1}{12L}$ for any $\eta \geq 3L$.

Plugging in these parameters into (B.218) , summing up from $k = 0$ to $K-1$ and dividing both sides by K, we obtain:

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\Big[||\nabla f(\mathbf{x}_{i_k,k})||^2 \Big] \leq \frac{2(f(\mathbf{x}^0) - f^\star)}{AK} + \frac{2B\sigma^2}{A} , \tag{B.221}$$

where we use the fact that $\mathbb{E}[||\mathbf{x}_{i_0,0} - \tilde{\mathbf{x}}_0||] = 0$. Plugging the definition of A and B, we obtain:

$$\frac{2(f(\mathbf{x}^0) - f^\star)}{AK} + \frac{2B\sigma^2}{A} = \frac{2(\eta + \delta_A)^2}{\eta - 3L}\frac{(f(\mathbf{x}^0) - f^\star)}{K} + \frac{2(\delta_B + 3L)}{\eta - 3L}\sigma^2 \tag{B.222}$$

$$\leq \frac{8\eta^2}{\eta - 3L}\frac{(f(\mathbf{x}^0) - f^\star)}{K} + \frac{8L}{\eta - 3L}\sigma^2 . \tag{B.223}$$

Let $v_1 := \frac{f(\mathbf{x}^0)-f^\star}{K}$ and $v_2 := L\sigma^2$. The upper bound can be written as: $8\big[v_1(\eta-3L)+\frac{9v_1L^2+v_2}{\eta-3L}+6v_1L\big]$. Minimizing the bound w.r.t. η over $\eta \geq 3L$, we get $\eta^\star = 3L+\sqrt{9L^2+\frac{v_2}{v_1}}$. Plugging this choice into the upper bound and using the fact that $\eta^\star \geq 6L$, $\eta^\star \geq 3L+\sqrt{\frac{v_2}{v_1}}$, and $(\eta^\star)^2 \leq 18L^2+18L^2+2\frac{v_2}{v_1}$, we get:

$$\frac{8\eta^2}{\eta-3L}\frac{(f(\mathbf{x}^0)-f^\star)}{K}+\frac{8L}{\eta-3L}\sigma^2 \leq \frac{96L(f(\mathbf{x}^0)-f^\star)}{K}+16\sqrt{\frac{v_2}{v_1}}\frac{f(\mathbf{x}^0)-f^\star}{K}+8L\sigma^2\sqrt{\frac{v_1}{v_2}} \tag{B.224}$$

$$= \frac{96L(f(\mathbf{x}^0)-f^\star)}{K}+24\sqrt{\frac{L(f(\mathbf{x}^0)-f^\star)}{K}}\sigma\,. \tag{B.225}$$

□