SILVER: Single-loop variance reduction and application to federated learning

Kazusato Oko [1 2] **Shunta Akiyama** [1] **Denny Wu** [3 4] **Tomoya Murata** [2 5] **Taiji Suzuki** [1 2]

Abstract

Most variance reduction methods require multiple times of full gradient computation, which is time-consuming and hence a bottleneck in application to distributed optimization. We present a single-loop variance-reduced gradient estimator named SILVER (SIngle-Loop VariancE-Reduction) for the finite-sum non-convex optimization, which does not require multiple full gradients but nevertheless achieves the optimal gradient complexity. Notably, unlike existing methods, SILVER provably reaches second-order optimality, with exponential convergence in the Polyak-Łojasiewicz (PL) region, and achieves further speedup depending on the data heterogeneity. Owing to these advantages, SILVER serves as a new base method to design communication-efficient federated learning algorithms: we combine SILVER with local updates, which gives the best communication rounds and number of communicated gradients across all range of Hessian heterogeneity, and, at the same time, guarantees second-order optimality and exponential convergence in the PL region.

1. INTRODUCTION

1.1. Variance Reduced Finite-sum Optimization

We consider the finite-sum minimization problem, which is ubiquitous in ML optimization (Bottou et al., 2018):

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

Here each f_i is smooth and can be *nonconvex*. We aim to efficiently find a solution x that is an ε-first-order stationary point (i.e., $\|\nabla f(x)\|\le\varepsilon$) and furthermore an (ε,δ)-second-order stationary point (SOSP; i.e., $\|\nabla f(x)\|\le\varepsilon$ and $\lambda_{\min}(\nabla^2 f(x))\ge-\delta$).

To efficiently solve the problem (1), variance reduction is a technique in minibatch sampling to construct an estimator of the full batch gradient with a smaller variance than vanilla SGD by utilizing gradients at previous anchor points (Gower et al., 2020). One of the difficulties in variance reduction is that recursive updates of the gradient estimator easily accumulate the error and eventually buries the correct descent directions. Therefore, most variance reduction algorithms need to periodically compute the full batch (or large minibatch) gradient to refresh the estimator (Roux et al., 2012; Johnson and Zhang, 2013; Nguyen et al., 2017b; Fang et al., 2018; Zhou et al., 2020). However, such a gradient-refreshing step is time-consuming, complicates algorithms, and becomes a bottleneck in application to distributed optimization since it leads to periodic synchronization between the whole clients, which increases communication costs and is sometimes impractical due to too many clients.

In this context, recent studies have attempted to develop variance reduction algorithms that do not require multiple full gradient computations (convex: Nguyen et al. (2021); Beznosikov and Takáč (2021); non-convex: Cutkosky and Orabona (2019); Liu et al. (2020); Li et al. (2021b); Tran-Dinh et al. (2022)), which we refer to as *single-loop*.[1] Among them, Li et al. (2021b) introduced ZeroSARAH as a single-loop algorithm with optimal gradient complexity for nonconvex optimization.

However, these single-loop methods are *not as versatile as* to enjoy multiple advantages offered by popular variance reduction methods that use full gradients. For example, as a well-known full-gradient algorithm, SARAH (Nguyen et al., 2017a; 2022; Li, 2019) requires full gradients to achieve (i) the optimal gradient complexity in the nonconvex finite-sum optimization, (ii) second-order optimality, and (iii) exponential convergence under the strong convexity, each of which holds significant practical importance: (i) efficiently finding stationary points is the primary goal of nonconvex optimization. (ii) escaping from saddle points is necessary to guarantee the quality of the solution since first-order sta-

[1]The University of Tokyo [2]Center for Advanced Intelligence Project, RIKEN [3]New York University [4]Flatiron Institute [5]NTT DATA Mathematical Systems Inc. Correspondence to: Kazusato Oko <oko-kazusato@g.ecc.u-tokyo.ac.jp>.

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

[1]This definition of single-loop excludes PAGE (Li et al., 2021a) or Loopless SVRG / Loopless Katyusha (Kovalev et al., 2020), which compute full gradient stochastically.

tionary points can include a local maximum or a saddle point. (iii) strong convexity is often observed around local minima so that exploiting local strong convexity can accelerate the convergence. Moreover, (iv) the gradient complexity of SARAH can further be reduced when f_i are less heterogeneous (Beznosikov and Takáč, 2021; Murata and Suzuki, 2021), a necessary property in application to federated learning to yield efficient communication costs with local updates.

Existing single-loop methods only met a limited subset of the above desiderata (see Table 2 for details). For example, ZeroSARAH (Li et al., 2021b) satisfies (i) optimal gradient complexity the only single-loop algorithm, but not the others. No single-loop algorithm can provably find (ii) SOSPs, or (iii) achieves exponential convergence unless paying suboptimal gradient complexity for (i) such as SAGA (Defazio et al., 2014; Reddi et al., 2016b). Beznosikov and Takáč (2021) can yield (iv) speed-up with less heterogeneity but is only valid in strongly convex optimization.

These discussions motivate the following research question:

Can we develop a versatile single-loop variance reduction algorithm that met all the desiderata (i)-(iv)?

1.2. Federated Learning with Variance Reduction

Such a versatile algorithm will be especially useful as a base to design more versatile FL methods. Federated learning (FL) is a paradigm of distributed learning, where optimization is performed by exchanging model parameters updated by locally-stored data in clients, without sharing the data itself (Konečnỳ et al., 2016; Shokri and Shmatikov, 2015). We consider the finite P client setting:

$$\min_{x\in\mathbb{R}^d}\left\{f(x):=\frac{1}{P}\sum_{i=1}^{P}\mathbb{E}_j[f_{i,j}(x)]\right\}, \tag{2}$$

where $f_{i,j}$ is smooth, and clients and data are indexed by i and j. Because synchronization and communication between clients and the serve are bottlenecks, we aim to reduce *communication rounds* and, at the same time, the total number of communicated gradients, which we call *communication complexity*.

In FL, variance reduction is closely tied to reducing communication rounds. Variance reduction can correct the errors from local updates, where the parameters are updated locally between communications (McMahan et al., 2017; Li et al., 2020b; Liang et al., 2019). Thus, local updates combined with variance reduction can yield fewer communication rounds than centralized ones if (and only if) clients are less heterogeneous (Karimireddy et al., 2020; Woodworth et al., 2020a;b). In this paper, less heterogeneity specifically means that the Hessian of f_i are similar, i.e., $\|\nabla^2 f_i(x)-\nabla^2 f(x)\|\le\zeta(\ll 1)$ for all x, which is essential to surpass the $O(\varepsilon^{-2})$ communication rounds of centralized methods (Murata and Suzuki, 2021; Karimireddy et al., 2021; Patel et al., 2024).

On the other hand, to reduce communication complexity, client sampling, where only a subset of clients is used in each communication, has been used (McMahan et al., 2017; Karimireddy et al., 2020). This also reduces the risk of delay or fail at some client affecting the whole training (Li et al., 2020a). *Client sampling error*, due to this partial client participation, has also been addressed with the variance reduction technique.

However, existing FL algorithms still have limitations in their effectiveness and expandability, due to client sampling error. Look at Table 1, where p refers to client sample size out of the total P clients at each communication, K is the local update steps between communications, and ζ is the Hessian heterogeneity. While BVR-L-SGD (Murata and Suzuki, 2021) yields the best communication rounds of $O(\zeta\varepsilon^{-2})$, client sampling is not allowed. Similarly, while MimeMVR (Karimireddy et al., 2021) can also achieve efficient communication by utilizing less Hessian heterogeneity, increasingly larger sampling size is required to ignore the client sampling error as heterogeneity gets smaller. Moreover, the only algorithm with the second-order optimality guarantee also require full client participation and cannot handle client sampling error (BVR-L-PSGD (Murata and Suzuki, 2022)). Further, client sampling error prevents the algorithm from obtaining exponential convergence with local convexity, and the dependency on μ cannot be improved even with less heterogeneity (MimeSGD (Karimireddy et al., 2021)).

These problems clearly show limitation in controlling client sampling error. In other words, they are attributed to the fact that existing variance reduction methods used as base algorithms mostly requires multiple full gradient computations, while the lack of versatility prevents the use of single-loop methods. Therefore, developing a versatile single-loop method is expected to provide a fundamental solution to this problem, and to extend versatility of FL algorithms.

1.3. Our Contributions

Single-loop variance reduction. For the finite-sum problem (1), we propose a novel, completely-single-loop variance-reduction gradient estimator SILVER (SIngle-Loop VariancE-Reduction), offering all of these desired functionalities (i)-(iv).

(i) SILVER achieves the optimal gradient complexity for the smooth finite-sum optimization (1). (Theorem 1)

(ii) Just adding small noise, SILVER efficiently escapes from saddle points to provably find second-order sta-

Table 1. Comparison of communication rounds and complexity for (2).[2]

Algorithms	Communication rounds	Client sampling
FedAvg (NC) (Karimireddy et al., 2020)	$\frac{\sigma_c^2}{p\varepsilon^4} + \frac{\sigma_c}{\varepsilon^3} + \frac{1}{\varepsilon^2}$	✓
SCAFFOLD (NC) (Karimireddy et al., 2020)	$\frac{\sigma^2}{pK\varepsilon^4} + \frac{1}{\varepsilon^2}(\frac{P}{p})^{\frac{2}{3}}$	✓
SCAFFOLD (NC, quad) (Karimireddy et al., 2020)	$\frac{\sigma^2}{PK\varepsilon^4} + \frac{1}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2}$	×
MimeMVR (NC) (Karimireddy et al., 2021)	$\frac{\zeta'\sigma_c}{\sqrt{p}\varepsilon^3} + \frac{\sigma_c^2}{p\varepsilon^2} + \frac{1}{K\varepsilon^2} + \frac{\zeta'}{\varepsilon^2}$	✓
BVR-L-SGD (NC) (Murata and Suzuki, 2021)	$\frac{1}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2}$	×
DASHA-PP-PAGE (NC) (Tyurin and Richtárik, 2022a)	$\frac{1}{\varepsilon^2} + \frac{\omega\sqrt{P}}{p\varepsilon^2}$	✓
FL-SILVER (NC) (Theorem 4)	$\frac{1}{K\varepsilon^2} + \frac{\zeta\sqrt{P}}{p\varepsilon^2} + \frac{\zeta}{\varepsilon^2}$	✓
BVR-L-PSGD (SOSP) (Murata and Suzuki, 2022)	$(\frac{1}{K} + \zeta)(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4})$	×
FL-SILVER (SOSP) (Theorem 4)	$(\frac{1}{K} + \zeta)(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4})$	✓ $(p \gtrsim \sqrt{P} + \frac{\zeta^2}{\delta^2})$
MimeSGD (PL) (Karimireddy et al., 2021)	$\frac{\sigma_c^2}{\mu p\varepsilon^2} + \frac{L}{\mu}\log\varepsilon^{-1}$	✓
FL-SILVER (PL) (Theorem 4)	$(\frac{L}{\mu K} + \frac{\zeta\sqrt{P}}{\mu p} + \frac{\zeta}{\mu} + \frac{P}{p})\log\varepsilon^{-1}$	✓

※ NC: finding ε-first-order stationary points; SOSP: finding (ε, δ)-second-order stationary points; PL: finding ε-solutions under Polyak-Łojasiewicz (PL) condition with μ; Quad: only valid for quadratics.
P: total number of clients, p: client sample size (if allowed), K: local update steps between communications, ζ, ζ': Hessian heterogeneity ($\zeta \le \zeta'$), σ_c: variance of $\nabla f_i(x) - \nabla f(x)$, σ: variance of $\nabla f_{i,j}(x) - \nabla f_i(x)$, ω: compression rate of gradients.

tionary points. This is the first single-loop algorithm to yield second order optimality, except for Noisy SGD with a large minibatch (Jin et al., 2021). (Theorem 2)

(iii) When SILVER enters a locally convex region where the Polyak-Łojasiewicz (PL) condition (see Assumption 6) holds, it automatically switches to exponential convergence. (Theorem 3)

(iv) For application to FL, we analyze SILVER under Hessian-heterogeneity of ζ, and show the complexity of $O(n + \frac{\zeta\sqrt{n}}{\varepsilon^2})$ (Theorem 1). We also show the lower bound matching to this (Proposition 1).

Our algorithm is not merely a removal of the full-gradient from existing methods, but is based on a sophisticated combination of SARAH and SAGA. Please refer to Section 3.1 for the details of the construction.

Improved algorithm for FL. Furthermore, we demonstrate the usefulness of SILVER as a base algorithm for federated learning, leading to an efficient and versatile FL method named FL-SILVER.

- FL-SILVER improves BVR-L-SGD to the partial client participation setting. Specifically, by allowing client sampling, it can simultaneously achieve the best communication rounds of BVR-L-SGD and an improved communication complexity of $O\left(P + \frac{\zeta\sqrt{P}}{\varepsilon^2}\right)$. As the Hessian heterogeneity ζ gets smaller, FL-SILVER becomes increasingly more communication-efficient.
- Just adding small noise, FL-SILVER can find second-order stationary points without hurting communication rounds and complexity. In contrast to BVR-L-PSGD, FL-SILVER allows client sampling for $p \gtrsim \sqrt{P} + \frac{\zeta^2}{\delta^2}$.
- Under the PL condition, FL-SILVER yields exponential convergence. The exponential convergence rate gets faster as ζ reduces, whereas other algorithms do not exhibit this property even under strong convexity.

We remark that these arguments do not consider gradient compression, which requires additional assumptions but provides another way to reduce communication (e.g., DASHA-PP-PAGE (Tyurin and Richtárik, 2022a)). In Appendix A, we provide additional literature review including this.

We also mention about the storage cost. SILVER allocates a copy of x for each f_i. This is a common feature for single-loop algorithms (Defazio et al., 2014; Li et al., 2021b), while it is avoided by multiple full gradient computation (e.g., SARAH). Also, FL-SILVER requires for each client to store an auxiliary variable. However, it becomes increasingly common to hold local variables for variance reduction in FL, even when the base algorithms are not single-loop (Murata and Suzuki, 2021; Karimireddy et al., 2021), to reduce communication costs. Given these background, we believe that developing a versatile single-loop algorithm and applying it to FL have significant importance even by using storage cost, although some situations might prefer more storage efficient algorithms.

[2] After we uploaded the arXiv preprint, we noticed that a concurrent work (Patel et al., 2022) proposed an algorithm allowing for client sampling, and the round for NC is $O(\frac{1}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2} + \frac{\sigma_c^2}{p\varepsilon^2} + \frac{\zeta\sigma_c}{p\varepsilon^3})$. Thus they need to assume less heterogeneity of σ_c as well as ζ, which contrasts to our algorithm.

2. PRELIMINARIES

Here we formally describe the problem settings. We first introduce the gradient Lipschitzness. (a) the averaged Lipschitzness suffices to prove the first-order guarantee in expectation, while (b) component-wise Lipschitzness is required for the high-probability second-order guarantee.

Assumption 1.

(a) $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2 \le L^2\|x-y\|^2$ *and* $\mathbb{E}_j[\|\nabla f_{i,j}(x) - \nabla f_{i,j}(y)\|^2] \le L^2\|x-y\|^2$ *for all* x, y *and* i.

(b) $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|$ *and* $\|\nabla f_{i,j}(x) - \nabla f_{i,j}(y)\| \le L\|x-y\|$ *for all* x, y *and* i, j.

We also assume existence of the global infimum.

Assumption 2. *f has the global infimum:* $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ *and* $\Delta := f(x^0) - f^*$.

Sometimes gradient boundedness is required. (a) Inter-client gradient boundedness is for SILVER/FL-SILVER with Option I to completely remove full gradient, while Option II computes full gradient once at the initialization. The same assumption also appeared in ZeroSARAH (Li et al., 2021b). (b) Intra-client gradient boundedness is for FL-SILVER.

Assumption 3.

(a) $\|\nabla f_i(x^0) - \nabla f(x^0)\|^2 \le \sigma_c^2,\ \forall i.$

(b) $\|\nabla f_{i,j}(x) - \nabla f_i(x)\|^2 \le \sigma^2,\ \forall x, y$ *and* i, j.

For the second-order optimality, Hessian-Lipshitzness is commonly assumed (Ge et al., 2019; Li, 2019).

Assumption 4. *Each f_i is ρ-Hessian Lipschitz:* $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \le \rho\|x-y\|,\ \forall x, y$ *and* i.

For FL, we assume inter-client Hessian-heterogeneity. (a) is for the first-order guarantee, while (b) is for the second-order guarantee.

Assumption 5.

(a) $\frac{1}{n}\sum_{i=1}^n \|\nabla^2 f_i(x) - \nabla^2 f(x)\|^2 \le \zeta^2$ *for all* x.

(b) $\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \le \zeta$ *for all* x *and* i.

Remark 1. *Note that $\zeta \le \sqrt{2}L$ always holds. This assumption has been used in previous literature such as* BVR-L-SGD *(Murata and Suzuki, 2021), and* Mime *(Karimireddy et al., 2021). This is indeed necessary to show the superiority to the centralized methods. Moreover, we emphasize that changing ζ between $(0, \sqrt{2}L]$ can interpolate the i.i.d. data allocation and the completely heterogeneous case.*

Finally, we introduce the Polyak-Łojasiewicz (PL) condition (Polyak, 1963), a generalization of strong convexity. A μ-strongly convex function satisfies this with μ.

Assumption 6. *f satisfies the PL condition, i.e.,* $\|\nabla f(x)\|^2 \ge 2\mu\,(f(x) - f^*)$ *for any* $x \in \mathbb{R}^d$.

3. FINITE-SUM OPTIMIZATION WITH SILVER

Now we concretely describe the proposed algorithm SILVER and provide its convergence guarantees. In the pseudocode, $B(0,r)$ denotes the uniform distribution on the Euclidean ball in $\mathbb{R}^d$ with radius r.

Algorithm 1 SILVER(x^0, η, b, T, r)

Option I: $y_i^0 \leftarrow \nabla f_i(x^0)\ (i = 1, \cdots, n)$
Option II: Randomly sample minibatch I^0 with size b; $y_i^0 \leftarrow \frac{1}{b}\sum_{j\in I^0}\nabla f_j(x^0)\ (i = 1, \cdots, n)$
for $t = 1$ to T **do**
  Ramdomly sample minibatch $I^t \subseteq [n]$ with size b
  $x^t \leftarrow x^{t-1} - \frac{\eta}{n}\sum_{i=1}^n y_i^{t-1} + \xi^t \quad (\xi^t \sim B(0,r))$
  $y_i^t \leftarrow \begin{cases} \nabla f_i(x^t) & \text{for } i \in I^t \\ \frac{1}{b}\sum_{j\in I^t}(\nabla f_j(x^t) - \nabla f_j(x^{t-1})) + y_i^{t-1} & \text{for } i \notin I^t \end{cases}$

3.1. Algorithm Description

SILVER is carefully designed to combine SAGA (Defazio et al., 2014; Reddi et al., 2016b) and SARAH (Nguyen et al., 2017a;b), ensuring that it inherits the advantages of both. SAGA operates without multiple full gradients but offers suboptimal gradient complexity, while SARAH meets optimality and versatility (i)-(iv) but requires periodic full gradients. SILVER avoids the need for full gradients like SAGA while it constructs an accurate and versatile estimator like SARAH.

Recall that in SAGA's update, the discrepancy of the gradient estimator from true gradient at step t is decomposed as

$$\sum_{i\in I^t}\frac{\nabla f_i(x^t) - \nabla f_i(x^{T(t,i)})}{b} - \underbrace{\sum_{i=1}^n \frac{\nabla f_i(x^t) - \nabla f_i(x^{T(t,i)})}{n}}_{(\bigstar)},$$

where I^t is the randomly chosen minibatch with size b at the step t and $T(t,i)$ is the step when f_i is last sampled. SAGA stores $\nabla f_i(x^{T(t,i)})$ for each i. Thus, the second term $\bigstar$ is a change from the referable gradient of $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x^{T(t,i)})$, and the first term is an approximation of the second term $\bigstar$ using a minibatch with size b. Then, the variance of the gradient estimator is roughly bounded by $\frac{1}{b}\mathbb{E}_i[\|x^t - x^{T(t,i)}\|^2] \le \frac{t - \min T(t,i)}{b}\sum_{s=\min T(t,i)+1}^t L^2\|x^s - x^{s-1}\|^2$.

On the other hand, SARAH uses periodic full gradient computation, and recursively updates the reference gradient. The difference between SARAH's gradient estimator and the

true gradient can be written as

$$\sum_{s=T(t)+1}^{t}\Big(\nabla f(x^s)-\nabla f(x^{s-1})-\sum_{i\in I^s}\frac{\nabla f_i(x^s)-\nabla f_i(x^{s-1})}{b}\Big),$$

where $T(t)$ is the time of the last full gradient evaluation and I^s is the randomly chosen minibatch with size b at the step s. Then, the variance is bounded by $\frac{1}{b}\sum_{s=T(t)+1}^{t}L^2\|x^s-x^{s-1}\|^2$, and SARAH's estimator is better than that of SAGA by the $t-\min_i T(t,i)$ factor, which can be as large as $\frac{n}{b}$. Here, the key is the decomposition of the discrepancy into a sum of differences in gradients between adjacent steps, allowing for the utilization of the independence of I^s.

Based on the above discussion, we decompose SAGA's approximation target ★ into a sum of difference in gradients between adjacent steps $\nabla f_i(x^s)-\nabla f_i(x^{s-1})$, to utilize the independence of sampling at different time steps like SARAH. Then the target is decomposed as

$$\sum_{s=\min_i T(t,i)+1}^{t}\frac{1}{n}\sum_{i\in\tilde{I}_s^t}^{n}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})).$$

Here $\tilde{I}_s^t=[n]\setminus\bigcup_{\tau=s}^{t}I^t$, so that $\tilde{I}_s^t$ is the set of indexes not sampled between s and t.

However, the implementation and analysis do not go straightforward like SARAH, because $\tilde{I}_s^t$ depends not only on I^s but also on $I^{s+1},\dots,I^t$. For the implementation, we introduce an auxiliary variable y_i^t, and moreover, we explain how to efficiently compute the update in $O(b)$ time in the "efficient implementation" paragraph below. This yields Algorithm 1.

On the other hand, to evaluate the error, we show that the correlation between $\tilde{I}_s^t,\cdots,\tilde{I}_t^t$ is not too strong. This step makes the analysis more complicated, but in the end the optimal complexity is obtained as SARAH.

Efficient Implementation Note that we can update $\frac{1}{n}\sum_{i=1}^{n}y_i^t$ in $O(bd)$ time and using $O(nd)$ memory, where d is the parameter dimension. Indeed, first introduce an auxiliary variable v^t which is inductively defined by $v^t=\frac{1}{b}\sum_{i\in I^t}(\nabla f_i(x^t)-\nabla f_i(x^{t-1}))+v^{t-1}$ with $v^0=0$. For each i, define v_i^t with $v_i^0=0$ and update it as $v_i^t=v^t$ iff $i\in I^t$. We also define w_i^t with $w_i^0=y_i^0$ and update it as $w_i^t=y_i^t=\nabla f_i(x^t)$ iff $i\in I^t$. Now we can see that $\frac{1}{n}\sum_{i=1}^{n}y_i^t=\frac{1}{n}\sum_{i=1}^{n}y_i^{t-1}+\frac{1}{b}\sum_{i\in I^t}(\frac{n+b}{n}\nabla f_i(x^t)-\nabla f_i(x^{t-1}))-\frac{1}{n}\sum_{i\in I^t}(w_i^{t-1}+v_i^t-v_i^{t-1})$. Therefore, $\frac{1}{n}\sum_{i=1}^{n}y_i^t$ can be updated by only $O(bd)$ computation with $O(nd)$ memory.

3.2. Convergence Guarantees

Now we prove that SILVER satisfies the desired properties.

First-order Optimality

Theorem 1. *Under Assumptions 1-(a), 2, 5-(a), 3-(a) (only for Option II), if we choose* $\eta=\Theta(\frac{1}{L}\wedge\frac{b}{\zeta\sqrt{n}})$ *and* $r\le\frac{\eta\varepsilon}{2}$, *Algorithm 1 finds* ε*-first-order stationary points in expectation, i.e.,* $\min_{t\le T}\mathbb{E}[\|\nabla f(x^t)\|]\le\varepsilon$, *using*

$$n+Tb=O\left(n+\frac{\Delta\left(\zeta\sqrt{n}\vee Lb\right)}{\varepsilon^2}\right)\ \text{(Option I)},$$
$$(T+1)b=O\left(\frac{\Delta\left(\zeta\sqrt{n}\vee Lb\right)+\frac{n}{b}\sigma_c^2}{\varepsilon^2}\right)\ \text{(Option II)}$$

stochastic gradient queries.

SILVER with Option I removes the need of full gradient computation except for the initialization. At the same time, because $\zeta\le 2L$, the gradient complexity matches to the optimal rate (Li et al., 2021a) attained by existing algorithms that use multiple full gradients such as SARAH and SPIDER (Fang et al., 2018), proving (i). Option II completely removes full gradient, while dependency on σ_c is unavoidable in this case.

Moreover, the algorithm achieves speed-up when the heterogeneity ζ is small, which confirms (iv). SILVER is the first algorithm that achieves both (i) and (iv).

The proof in Appendix D.1 depends on the following lemma, which illustrates that the error from a certain step exponentially decays and never accumulates owing to our original combination of SAGA and SARAH, without full gradient computation.

Lemma 1. *Let* $g^t=\frac{1}{n}\sum_{i=1}^{n}y_i^t$ *be the gradient estimator and choose Option I. Then,*

$$\mathbb{E}[\left\|g^t-\nabla f(x^t)\right\|^2]\le\frac{30\zeta^2}{b}\sum_{s=1}^{t}(1-\frac{b}{4n})^{t-s}\mathbb{E}[\|x^s-x^{s-1}\|^2].$$

Lower Bound under Different Heterogeneity We complement the above result with the following lower bound, extending Fang et al. (2018); Li et al. (2021a) to include the ζ dependency.

Proposition 1. *Under Assumptions 1-(a), 2, and 5, any linear-span first-order algorithm requires*

$$\Omega\left(n+\frac{\Delta(\zeta\sqrt{n}+L)}{\varepsilon^2}\right)$$

stochastic gradients to find ε*-first-order stationary points of the problem* (1).

This matches to the upper bound in all parameters, so that SILVER is the optimal across all range of ζ. See Appendix F for the formal definition of the linear-span first-order algorithm and more details.

Second-order Optimality Here we state that SILVER can find second-order stationary points, to guarantee the second property. While it is usual to extend optimization methods to ensure second-order optimality (Ge et al., 2015; Jin et al., 2017; Vlatakis-Gkaragkounis et al., 2019), and variance reduction methods also have been applied to this (Ge et al., 2019; Fang et al., 2018; Li, 2019), all existing methods require a sub-routine for negative curvature extraction (e.g., SPIDER-SFO$^+$+Neon2 (Fang et al., 2018; Allen-Zhu and Li, 2018)) or periodic large minibatch as large as n or $O(\frac{\sigma_c^2}{\varepsilon^2})$ (e.g., SSRGD (Li, 2019); Stabilized SVRG (Ge et al., 2019)). On the other hand, only adding small noise, SILVER is the first completely-single-loop algorithm that can find SOSPs.

Theorem 2. *Under Assumptions 1-(b), 2, 4, and 5-(b), choose Option I, and let* $b = \tilde{\Omega}(\sqrt{n} + \frac{\zeta^2}{\delta^2})$, $\eta = \tilde{\Theta}(\frac{1}{L})$, $r = \tilde{O}\left(\frac{\varepsilon \wedge (\delta^2/\rho)}{L}\right)$. *Then, Algorithm 1 finds* (ε, δ)*-SOSPs using*

$$\tilde{O}\left(n + L\Delta\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right)b\right)$$

stochastic gradients, with high probability.

Since $\zeta \le 2L$, the complexity is comparable to best existing algorithms, such as SPIDER-SFO$^+$ and SSRGD. In addition to the main $\frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\delta^4}$ term, different algorithms have different additional terms, coming from respective analysis. Ours is $\frac{\zeta^2}{\varepsilon^2\delta^2} + \frac{\zeta^2}{\delta^6}$, which becomes smaller than that of SPIDER-SFO$^+$ when $n \gtrsim \delta^{-4}$ and that of SSRGD when $n \gtrsim \min\{\delta^{-3}, \delta\varepsilon^{-2}\}$. Furthermore, Remark 2 in Appendix D.2 introduces a small trick to turn this to $\frac{\zeta^2}{\varepsilon^2\delta} + \frac{\zeta^2}{\delta^5}$ or $\frac{n\delta}{\varepsilon^2} + \frac{n}{\delta^3}$.

We briefly explain the proof outline. As in Jin et al. (2017); Ge et al. (2019); Li (2019), we consider two sequences of $\{x^t\}$ with slightly different initial points while keeping all the other randomness the same. We show that, when negative curvature exists, these two sequence separate each other exponentially, and at least one of these sequences goes further from the initial point. This in turn means that if we perturb the algorithm a little around saddle points, then it can escape saddle points with high probability.

Although this high-level idea is classical, we confront several difficulties due to the single-loop structure. For example, while other algorithms refresh their gradient estimators around saddle points, our single-loop algorithm does not. Thus, we have to address the accumulated error, and it is not trivial whether our gradient estimator can find the right direction of the negative eigenvalue despite the accumulated error, which finally yields the $\frac{\zeta^2}{\varepsilon^2\delta^2} + \frac{\zeta^2}{\delta^6}$ term. For more details and complete proof, see Appendix D.2.

Note that our bound requires minibatch size of $b \gtrsim \sqrt{n} + \frac{\zeta^2}{\delta^2}$, and this minibatch size is common in many existing algorithms. In fact, SSRGD assumes $b \ge \sqrt{n}$ or $b \ge \frac{\sigma_c}{\varepsilon}$ and Stabilized SVRG assumes $b \ge n^{\frac{2}{3}}$. Considering that $\delta = O(\sqrt{\rho\varepsilon})$ is often assumed, our minibatch size is as moderate as those of existing algorithms. The necessity of this assumption comes from that if b is too small, the sampling error hides the right direction of negative curvature.

Exponential Convergence under PL Condition When SILVER enters locally convex regions, it automatically switches into exponential convergence phase.

Theorem 3. *Under Assumptions 1-(a), 2, 3-(a) (only for Option II), 5-(a), and 6, if we choose* η *as* $\eta = \Theta(\frac{1}{L} \wedge \frac{b}{\zeta\sqrt{n}})$, *and* $r \le \eta\sqrt{\frac{\varepsilon\mu}{6}} \wedge \sqrt{\frac{\eta\varepsilon b}{24n}}$, *Algorithm 1 finds an* ε*-solution in expectation, i.e.,* $\mathbb{E}[f(x^t) - f^*] \le \varepsilon$, *by using*

$$O\left(\left(\frac{Lb}{\mu} \vee \frac{\zeta\sqrt{n}}{\mu} \vee n\right)\log\frac{\Delta + \frac{n\sigma_c \mathbb{1}[\text{Option II}]}{bL}}{\varepsilon}\right)$$

stochastic gradients.

SILVER with Option II completely removes the requirement of full gradient, even at the initial point, while obtaining the same complexity as SARAH, PAGE, and their variants (up to the $\log\sigma_c$ factor). The proof can be found in Appendix D.3.

It should be noted that the upper bound of r depends on μ, but it is not necessary to know the exact value of μ; it is sufficient to take r small enough. The disadvantage of taking r small is that the bound in Theorem 2 depends polylogarithmically on r^{-1}, but this polylogarithmic dependence would not become a practical issue.

4. APPLICATION TO FEDERATED LEARNING

The versatile single-loop method is useful as a base method for federated learning, because it allows simpler structure of the algorithm and brings various advantages. We demonstrate that SILVER serves as such, by developing its FL extension called FL-SILVER. The proposed algorithm combines SILVER with local updates, see Algorithm 2.

Specifically, FL-SILVER uses SILVER to bound the client sampling error by approximating the global gradient $\nabla f(x^t)$ based on the approximations of $\nabla f_i(x^t)$. Moreover, to bound the error from local updates, the approximations of $\nabla f_i(x^t)$ are constructed using a SARAH-type estimator within each client.

In the pseudocode, $\xi^{t,k}$ follows the uniform distribution on the Euclidean ball in $\mathbb{R}^d$ with radius r. Due to the space limitation, here we only present Option I. Although using one full participation only at x^0 is common as in MimeMVR

Algorithm 2 FL-SILVER$(x^0, \eta, p, b, T, K, r)$

for $i \in I^0 = [P]$ in parallel **do**
  Randomly select minibatch J_i^0 with size Kb
  $y_i^0 \leftarrow \frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)$
Send y_i^0 from $i \in I^0$ to the server
for $t = 1$ to T **do**
  Randomly sample one client i_t
  Send $\sum_{i=1}^P y_i^{t-1}$ and x^{t-1} from the server to i_t
  $x^{t,0} \leftarrow x^{t-1}$, $z^{t,0} \leftarrow 0$
  for $k = 1$ to K **do**
    $x^{t,k} \leftarrow x^{t,k-1} - \eta(\frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k-1}) + \xi^{t,k}$
    Randomly select minibatch $J_{i_t}^{t,k}$ with size b
    $z^{t,k} \leftarrow z^{t,k-1} + \frac{1}{b}\sum_{j\in J_{i_t}^{t,k}}(\nabla f_{i_t,j}(x^{t,k}) - \nabla f_{i_t,j}(x^{t,k-1}))$
  Send $x^{t,K}$ from i_t to the server; $x^t \leftarrow x^{t,K}$
  Randomly select p clients I^t; send x^t from i_t to I^t
  for $i \in I^t$ in parallel **do**
    Randomly select minibatch J_i^t with size Kb
    $y_i^t \leftarrow \frac{1}{Kb}\sum_{j\in J_i^t}\nabla f_{i,j}(x^t)$
    $\Delta y_i^t \leftarrow \frac{1}{Kb}\sum_{j\in J_i^t}(\nabla f_{i,j}(x^t) - \nabla f_{i,j}(x^{t-1}))$
  Send $\{(y_i^t, \Delta y_i^t)\}_{i\in I^t}$ from I^t to the server
  $y_i^t \leftarrow y_i^{t-1} + \frac{1}{p}\sum_{i\in I^t}\Delta y_i^t$ for $i \notin I^t$

Karimireddy et al. (2021), we can remove it with Option II. For the pseudocode and theorems for Option II can be found in Appendix D, as well as complete proofs.

4.1. Convergence Guarantees

Now, we present a summary of the various convergence guarantees of FL-SILVER in the following theorem.

Theorem 4. *Suppose Assumptions 1-(a), 2, 3-(b), and 5-(a) hold.*

- *If* $\eta = \Theta(\frac{1}{L} \wedge \frac{p}{\zeta\sqrt{P}} \wedge \frac{1}{\zeta K} \wedge \frac{p}{L\sqrt{P}})$, $r = O(\eta\varepsilon)$, $b = \Omega(\frac{\sigma^2}{PK\varepsilon^2})$, *and* $b \geq K$, *Algorithm 2 finds* ε*-first-order stationary points using*

$$O\left(1 + \left(\frac{L}{K} + \zeta\right)\left(1 \vee \frac{\sqrt{P}}{p}\right)\frac{\Delta}{\varepsilon^2}\right)$$

  communication rounds, in expectation.

- *If Assumption 4 and* $\delta < \zeta$ *additionally hold and we take* $p = \tilde{\Omega}(\sqrt{P} + \frac{\zeta^2}{\delta^2} + \frac{L^2}{Kb\delta^2})$, $\eta = \tilde{\Theta}(\frac{1}{L})$, $r = \tilde{O}(\frac{\varepsilon}{L})$, $b = \tilde{\Omega}(\frac{\sigma^2}{PK\varepsilon^2})$, *and* $b \geq K$, *Algorithm 2 finds* (ε, δ)*-second-order stationary points using*

$$\tilde{O}\left(1 + \Delta\left(\frac{L}{K} + \zeta\right)\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right)\right)$$

  communication rounds, with high probability.

- *If Assumption 6 additionally holds and we take* $\eta = \Theta\big(\frac{1}{L} \wedge \frac{\frac{p}{\sqrt{P}}\wedge 1}{L\vee(\zeta K)}\big)$ *and* $r = O(\eta\varepsilon)$, *Algorithm 2 finds* ε*-first-order stationary points using*

$$O\left(\left[\frac{L}{\mu K} \vee \frac{P}{p} \vee \frac{\zeta(\frac{\sqrt{P}}{p}\vee 1)}{\mu} \vee \frac{L(\frac{\sqrt{P}}{p}\vee 1)}{\mu\sqrt{bK}}\right]\log\frac{\Delta}{\varepsilon}\right)$$

  communication rounds, in expectation.

Before delving into specifics, let's make a few remarks.

- Multiplying by p and adding P yields the communication complexity. For the first-order convergence, as summarized in Table 1, FL-SILVER achieves the best communication rounds and communication complexity among the federated learning algorithm for the finite-client setting.
- All guarantees hold without specific modifications to each, and this versatility will be a significant benefit in practice. Specifically, we set $p = \tilde{\Theta}(\sqrt{P})$, $\eta = \tilde{\Theta}(\frac{1}{L})$, and $r = \tilde{O}(\frac{\varepsilon}{L})$, under a moderate local computational budget of $b \geq \tilde{\Omega}(\frac{\sigma}{PK\varepsilon^2})$ and $b \geq K$.

First-order Optimality First, we conduct a comparison with existing algorithms (see also Table 1). For the required communication rounds, FL-SILVER achieves the same best communication rounds as that of BVR-L-SGD (Murata and Suzuki, 2021), in the full client participation setting ($p = P$) or even allowing sampling of clients ($\sqrt{P} \leq p < P$). MimeMVR (Karimireddy et al., 2021) is another method that can achive communication rounds better than $O(\varepsilon^{-2})$, but they additionally require intra-client Hessian heterogeneity $\|\nabla^2 f_{i,j}(x) - \nabla^2 f_i(x)\|$ and variance σ_c to be small, which contrasts to our assumptions.

For the communication complexity of $\tilde{O}(P + \frac{\zeta\sqrt{P}}{\varepsilon^2})$, achieved with client sampling $p \leq \sqrt{P}$, is the best rate and strictly smaller than that of BVR-L-SGD (Murata and Suzuki, 2021). By recalling Proposition 1, the communication complexity is optimal in a sense that the server must receive information of $\Omega(P + \frac{\zeta\sqrt{P}}{\varepsilon^2})$ gradients to output ε-first-order solutions. Therefore, FL-SILVER simultaneously offers the best communication rounds and improved communication complexity as desired.

Both the number of communication rounds and communication complexity get smaller as as ζ get smaller. This heterogeneity is the necessary assumption for FL algorithms to become more communication efficient than centralized methods (Karimireddy et al., 2020; Murata and Suzuki, 2021), and our algorithm has successfully utilized this structural assumption. We also remark the local computational budget. Our requirement on b is moderate, because such an

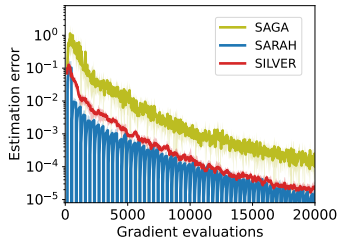


Figure 1. Accuracy of gradient estimator (finite-sum)

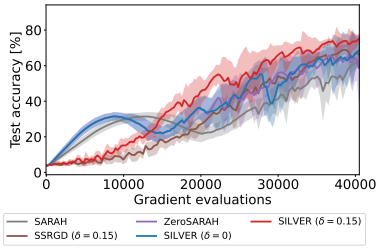


Figure 2. Test accuracy (finite-sum)

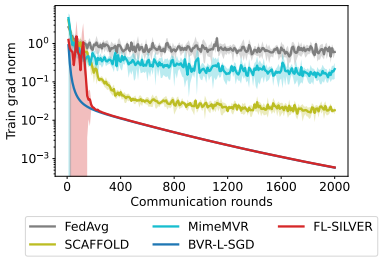


Figure 3. Gradient norm (FL)

assumption has also appeared in BVR-L-SGD (Murata and Suzuki, 2021) ($b = K$), MimeMVR (Karimireddy et al., 2021) (requiring the exact value of $\nabla f_i(x)$), and Patel et al. (2022) ($b = K$).

Second-order Optimality Furthermore, FL-SILVER can even find SOSPs, guaranteeing quality of the output. As the case of SILVER, all we need is just to add small noise. The proposed algorithm is the first FL algorithm that can find SOSPs while allowing sampling of the clients. It uses the same number of communication rounds as that of BVR-L-PSGD, but achieves the improved communication complexity when $\frac{\zeta^2}{\delta^2} \lesssim P$. The assumption of $\delta < \zeta$ can be removed with a small modification of the algorithm, see Appendix E.3.

Exponential Convergence under the PL Condition Also, FL-SILVER inherits exponential convergence under the PL condition, and gets more communication-efficient under less heterogeneity. In addition to the exponential convergence, the rate does not directly depend on the condition number $\frac{L}{\mu}$. Even if $\frac{L}{\mu}$ is large, less heterogeneity of clients ζ can ease the difficulty by variance reduction and local update, under a sufficiently large local computational budget. On the other hand, MimeSGD (Karimireddy et al., 2021) depends on condition number $\frac{L}{\mu}$ even if ζ is small, and other algorithms do so even with strong convexity. Therefore, in the strongly convex setting of FL, we are the first to prove that distributed optimization can achieve faster exponential convergence rate than centralized methods utilizing less heterogeneity of clients with local updates.

5. NUMERICAL EXPERIMENTS

Finally, we verify our theories by numerical experiments. Detailed explanation and additional experiments can be found in Appendix B. In all figures, the red line corresponds to our proposed method.

5.1. Accuracy of the Gradient Estimator of SILVER

We considered a classification of the capital letters using EMNIST (Cohen et al., 2017) with a two-layer neural network. We set $n = 130$, $b = 12 \coloneqq \sqrt{130}$, and the inner-loop length of SARAH to $\lfloor \frac{n}{b} \rfloor = 10$. We compared SAGA, SARAH, and SILVER (Option I) using a common step size $\eta = 0.01$, in terms of the squared error between true gradient $\nabla f(x^t)$ and the estimators. Remember that SAGA is a single-loop but has suboptimal complexity, while SARAH uses multiple full gradients to achieve optimal complexity.

According to Figure 1, the discrepancy of the SILVER estimator is clearly smaller than that of SAGA, and close to that of SARAH. Therefore, this shows that SILVER actually has as small variance yielding the optimal complexity as SARAH, while removing periodic full gradient computation, as desired.

5.2. Escaping Saddle Points with SILVER

We also compared SILVER (Option I) with SARAH, SSRGD (=SARAH+noise), and ZeroSARAH, in terms of the test accuracy. To increase the non-convexity, here a four-layer neural network was used and other than this construction of f_i was the same as previously. For SSRGD and SILVER, we added a small noise to see whether perturbation practically helps avoiding bad local minima.

In Figure 2, SILVER is faster than SARAH owing to avoidance of periodic full gradients. Moreover, perturbation avoids getting stuck in local minima and helps stable convergence; Contrary to noiseless SILVER, perturbed SILVER makes the accuracy increase almost monotonically. In summary, SILVER with small noise yielded the fastest convergence.

5.3. (Local) Exponential Convergence of FL-SILVER

For FL, we considered a little less heterogeneous classification of the capital letters, where each f_i consists of 90% data from one class and 10% from the rest, with a two-leyer neural network. We compared FL-SILVER (Option I) with FedAvg, SCAFFOLD, MimeMVR, and BVR-L-SGD, under $P = 104$, client minibatch size $p = 10 \coloneqq \sqrt{P}$ (except for BVR-L-SGD requiring $P = p = 104$), local minibatch size $b = 16$ and local update times $K = 10$.

In Figure 3, FL-SILVER achieves the smallest gradient

norm $\|\nabla f(x^t)\|$ and (locally) linear convergence. Moreover, it performs similarly to BVR-L-SGD, which is almost a special case of FL-SILVER with $P = p$. Thus, FL-SILVER can appropriately correct the errors from sampling of the clients and is about ten times more efficient than BVR-L-SGD in terms of communication complexity by allowing sampling of the clients.

6. CONCLUSION

This paper has developed a single-loop variance reduction method, and showed its first- and second-order optimality and linear convergence under the PL condition, with explicit dependency on the Hessian-heterogeneity. We have demonstrated that the proposed method serves as a useful base for a FL algorithm that allows client sampling and inherits these versatile benefits.

One of the interesting future research topics is the combination of the proposed FL method with communication compression (see also Appendix A.2). Since utilizing less heterogeneity and compressing gradients operate in orthogonal directions, combining these strategies will lead to a more communication-efficient FL algorithm.

Acknowledgements

The authors would like to thank Keigo Hara, Yuki Takezawa, and Yuki Yoshida for the discussions and helpful feedback. KO was partially supported by the IIW program of The University of Tokyo and JST ACT-X (JPMJAX23C4). SA was partially supported by JSPS KAKENHI (JP22J13388). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015).

Impact Statement

This paper addresses efficient non-convex optimization and its application to distributed optimization from a theoretical perspective, and is expected to lay a foundation for machine learning problems, particularly empirical loss minimization and privacy-protected training. Given the theoretical nature of this work, there would be no specific points to be discussed here.

References

Z. Allen-Zhu and Y. Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

A. Beznosikov and M. Takáč. Random-reshuffled SARAH does not need a full gradient computations. *arXiv preprint arXiv:2111.13322*, 2021.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60 (2):223–311, 2018.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

F. Chung and L. Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1): 79–127, 2006.

G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in neural information processing systems*, 32, 2019.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

R. Ge, Z. Li, W. Wang, and X. Wang. Stabilized SVRG: Simple variance reduction for nonconvex optimization. In *Conference on learning theory*, pages 1394–1448. PMLR, 2019.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

M. Grudzień, G. Malinovsky, and P. Richtárik. Improving accelerated federated learning with compression and importance sampling. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and*

Opportunities, 2023. URL https://openreview.net/forum?id=y0deDrXZaT.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021.

J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020.

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

Z. Li. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. *Advances in Neural Information Processing Systems*, 32, 2019.

Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021a.

Z. Li, S. Hanzely, and P. Richtárik. ZeroSARAH: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021b.

X. Liang, S. Shen, J. Liu, Z. Pan, E. Chen, and Y. Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

D. Liu, L. M. Nguyen, and Q. Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

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

T. Murata and T. Suzuki. Bias-variance reduced local SGD for less heterogeneous federated learning. In *International Conference on Machine Learning*, pages 7872–7881. PMLR, 2021.

T. Murata and T. Suzuki. Escaping saddle points with bias-variance reduced local perturbed sgd for communication efficient nonconvex distributed learning. *arXiv preprint arXiv:2202.06083*, 2022.

Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017a.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

L. M. Nguyen, K. Scheinberg, and M. Takáč. Inexact SARAH algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258, 2021.

L. M. Nguyen, M. van Dijk, D. T. Phan, P. H. Nguyen, T.-W. Weng, and J. R. Kalagnanam. Finite-sum smooth optimization with SARAH. *Computational Optimization and Applications*, pages 1–33, 2022.

K. K. Patel, L. Wang, B. Woodworth, B. Bullins, and N. Srebro. Towards optimal communication complexity in distributed non-convex optimization. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SNElc7QmMDe.

K. K. Patel, M. Glasgow, A. Zindari, L. Wang, S. U. Stich, Z. Cheng, N. Joshi, and N. Srebro. The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. *arXiv preprint arXiv:2405.11667*, 2024.

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

S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016a.

S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for smooth nonconvex optimization. In *2016 IEEE 55th conference on decision and control (CDC)*, pages 1971–1977. IEEE, 2016b.

N. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

R. Szlendak, A. Tyurin, and P. Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. *arXiv preprint arXiv:2110.03300*, 2021.

T. Tao and V. Vu. Random matrices: Universality of local spectral statistics of non-hermitian matrices. *The Annals of Probability*, 43(2):782–874, 2015.

Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071, 2022.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

A. Tyurin and P. Richtárik. A computation and communication efficient method for distributed nonconvex problems in the partial participation setting. *arXiv preprint arXiv:2205.15580*, 2022a.

A. Tyurin and P. Richtárik. Dasha: Distributed nonconvex optimization with communication compression, optimal oracle complexity, and no client synchronization. *arXiv preprint arXiv:2202.01268*, 2022b.

A. Tyurin, L. Sun, K. Burlachenko, and P. Richtárik. Sharper rates and flexible framework for nonconvex sgd with client and data sampling. *arXiv preprint arXiv:2206.02275*, 2022.

E.-V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in Neural Information Processing Systems*, 32, 2019.

B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020a.

B. E. Woodworth, K. K. Patel, and N. Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020b.

D. Zhou and Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pages 7574–7583. PMLR, 2019.

D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. *The Journal of Machine Learning Research*, 21(1):4130–4192, 2020.

Appendix

The appendix is organized as follows. Appendix A provides additional literature review. Appendix B explains the experimental settings presented in Section 5 and conducts additional experiments. Appendix C prepares concentration inequalites and a linear algebraic tool. Appendix D gives the proofs for SILVER. Appendix E gives the complete statements of the theoretical guarantees of FL-SILVER and their proofs. Finally, we prove the lower bound on the gradient complexity for the finite-sum smooth optimization in Appendix F.

A. ADDITIONAL LITERATURE REVIEW

A.1. (Single-loop) Variance Reduction Methods

Optimal Complexity for Finite-sum Smooth Optimization There are many variance reduction algorithms that utilize multiple full gradient computations during optimization (see Table 2). As we explained, SARAH (Nguyen et al., 2017a; 2022) is one of the most popular variance reduction algorithm, as well as its extension SSRGD (Li, 2019), and SARAH satisfies (i) the optimal gradient complexity for finite-sum smooth optimization, (ii) second-order optimality, (iii) exponential convergence with strong convexity, (iv) speed up with less heterogeneity. SARAH as appeared as the one of the three algorithms that achieves the optimal gradient complexity of $O(n + \frac{L\Delta\sqrt{n}}{\varepsilon^2})$ for finite-sum smooth optimization, as well as SPIDER-SFO$^+$ (Fang et al., 2018) and SNVRG (Zhou et al., 2020), developed upon SVRG (Reddi et al., 2016a). See the lower bound for Zhou and Gu (2019); Fang et al. (2018); Li et al. (2021a). We also mention PAGE (Li et al., 2021a). At each step, then stochastically determine whether to compute the full gradient to update the gradient estimator or to update the gradient estimator with minibatch gradient. Therefore, in expectation, then require $\Omega(\varepsilon^{-2})$ full gradients to obtain ε-first-order stationary points. Tyurin et al. (2022) provided the analysis of PAGE under the less heterogeneity assumption, but PAGE still requires multiple full gradients. Note that our definition of single-loop excludes any algorithm with multiple full gradient computations, because our goal is to develop a versatile variance reduction algorithm that avoids multiple full gradient computations of whatever form. Because regardless of whether it is deterministic or stochastic, full gradient computation slows down practical computational speed, and especially becomes a bottleneck in the application to distributed learning since this leads to periodic synchronization and communication between the whole client.

On the other hand, for single-loop variance reduction methods, SAGA (Reddi et al., 2016b) is a conventional single-loop algorithm and requires $O(n + \frac{Ln^{\frac{2}{3}}}{\varepsilon^2})$ gradient evaluations for solving (1), which is still sub-optimal from the optimal complexity of $O(n + \frac{L\sqrt{n}}{\varepsilon^2})$. After SAGA, several single-loop algorithms have been proposed. Hybrid SARAH Tran-Dinh et al. (2022) and iSARAH Nguyen et al. (2017a) considered single-loop variants of SARAH, but the complexity does not matches to $O(n + \frac{L\sqrt{n}}{\varepsilon^2})$ when ε is smaller than $1/\sqrt{n}$. STORM Cutkosky and Orabona (2019) achieves the complexity of $O(\sigma_c/\varepsilon^3)$, which is the optimal for stochastic optimization problem of $\min \mathbb{E}_i[f_i(x)]$, but not for the problem (1) when ε is small. Recently, ZeroSARAH achieved the optimal complexity for the problem (1) as the first single-loop algorithm. SILVER also achieves the same complexity, as well as other benefits.

Second-order optimality It is usual to extend an optimization algorithm to ensure second-order optimality (Ge et al., 2015; Jin et al., 2017; Vlatakis-Gkaragkounis et al., 2019; Allen-Zhu and Li, 2018), and variance reduction methods also have been applied to this (Stabilized SVRG (Ge et al., 2019), SPIDER-SFO$^+$ (Fang et al., 2018), and SSRGD (Li, 2019)). Since first-order stationary points can include a local maximum or a saddle point in nonconvex optimization, escaping them and finding SOSPs are necessary to guarantee the quality of the solution. However, no single-loop algorithm cannot find SOSPs. SILVER is the first single-loop methods that guarantees second-order optimality.

Exponential Convergence with Strong Convexity Many algorithms that use multiple full gradients, such as SVRG and SARAH, yield exponential convergence with strong convexity. For single-loop algorithms, SAGA (Reddi et al., 2016b) achieves the exponential convergence, but remember that its complexity for the problem (2) is suboptimal. Random-reshuffled SARAH (Beznosikov and Takáč, 2021) can yield the exponential convergence, but the guarantee was only shown for the strongly-convex case. In this context, SILVER is the first algorithm that achieves both optimal complexity in nonconvex settings and exponential convergence in strongly-convex settings.

Utilizing Less Heterogeneity The ability of utilizing less heterogeneity of $\{f_i\}_{i=1}^n$ is essential for variance reduction methods to serve as a base algorithm for communication-efficient federated learning. For example, SARAH has such an

Table 2. Stochastic gradient complexity for a nonconvex finite-sum problem (1).

Algorithms	Stochastic gradient complexity			Removal of multiple full gradients (single-loop?)
	Nonconvex	SOSP	PL condition	
(Noisy) SGD (Ghadimi and Lan, 2013; Ge et al., 2015; Karimi et al., 2016)	$\frac{\Delta\sigma_c^2}{\varepsilon^4}$	$\mathrm{poly}(\varepsilon^{-1},\delta^{-1},d,\sigma_c,\Delta)$	$\frac{\sigma_c^2}{\mu^2\varepsilon}\log\varepsilon^{-1}$	$\times$ ($b\gtrsim\sigma_c^2\varepsilon^{-2}$)
(Stabilized) SVRG (Reddi et al., 2016a; Ge et al., 2019; Johnson and Zhang, 2013)	$n+\frac{L\Delta n^{\frac{2}{3}}}{\varepsilon^2}$	$n+\Delta(\frac{n^{\frac{2}{3}}}{\varepsilon^2}+\frac{n^{\frac{2}{3}}}{\delta^4}+\frac{n}{\delta^3})$	$(n+\frac{L}{\mu})\log\varepsilon^{-1}$ (SC)	$\times$
SPIDER-SFO$^+$ (Fang et al., 2018)	$n+\frac{\zeta\Delta\sqrt{n}}{\varepsilon^2}$	$n+\Delta(\frac{\sqrt{n}}{\varepsilon^2}+\frac{1}{\varepsilon\delta^3}+\frac{1}{\delta^5})$	None	$\times$
SNVRG (Zhou et al., 2020)	$n+\frac{\zeta\Delta\sqrt{n}}{\varepsilon^2}$	None	$(n+\frac{\sqrt{n}}{\mu})\log^4\varepsilon^{-1}$	$\times$
SARAH (Nguyen et al., 2017b; 2022) and SSRGD (Li, 2019)	$n+\frac{\zeta\Delta\sqrt{n}}{\varepsilon^2}$	$n+\Delta(\frac{\sqrt{n}}{\varepsilon^2}+\frac{\sqrt{n}}{\delta^4}+\frac{n}{\delta^3})$	$(n+\frac{\zeta\sqrt{n}}{\mu})\log\varepsilon^{-1}$	$\times$
PAGE (Li et al., 2021a; Tyurin et al., 2022)	$n+\frac{\zeta\Delta\sqrt{n}}{\varepsilon^2}$	None	$(n+\frac{\zeta\sqrt{n}}{\mu})\log\varepsilon^{-1}$	$\times$
SAGA (Defazio et al., 2014; Reddi et al., 2016b)	$n+\frac{L\Delta n^{\frac{2}{3}}}{\varepsilon^2}$	None	$(n+\frac{L}{\mu})\log\varepsilon^{-1}$ (SC)	$\checkmark$
STORM (Cutkosky and Orabona, 2019)	$\frac{\zeta\sigma_c}{\varepsilon^3}+\frac{\sigma_c^2}{\varepsilon^2}$	None	None	$\checkmark$
Hybrid SARAH (Tran-Dinh et al., 2022)	$\frac{\Delta^{\frac{3}{2}}\sigma_c}{\varepsilon^3}+\frac{\Delta^{\frac{1}{2}}\sigma_c^3}{\varepsilon}$	None	None	$\checkmark$
iSARAH (Nguyen et al., 2021)	$n+\frac{\Delta^2+\sigma_c^4}{\varepsilon^4}$	None	None	$\checkmark$
Random-reshuffled SARAH (Beznosikov and Takáč, 2021)	None	None	$(\frac{L}{\mu}+\frac{\zeta n}{\mu})\log\varepsilon^{-1}$ (SC)	$\checkmark$
ZeroSARAH (Li et al., 2021b)	$n+\frac{\Delta\sqrt{n}}{\varepsilon^2}$	None	None	$\checkmark$ (only at x^0)
	$\frac{(\Delta+\sigma_c^2)\sqrt{n}}{\varepsilon^2}$	None	None	$\checkmark$
SILVER (Option I) (ours)	$n+\frac{\zeta\Delta\sqrt{n}}{\varepsilon^2}$	$n+\Delta(\frac{\sqrt{n}}{\varepsilon^2}+\frac{\sqrt{n}}{\delta^4}+\frac{\zeta^2}{\varepsilon^2\delta^2}+\frac{\zeta^2}{\delta^6})$	$(n+\frac{\zeta\sqrt{n}}{\mu})\log\varepsilon^{-1}$	$\checkmark$ (only at x^0)
SILVER (Option II) (ours)	$\frac{(\zeta\Delta+\sigma_c^2)\sqrt{n}}{\varepsilon^2}$	$(\Delta+\sigma_c^2)(\sqrt{n}+\frac{\zeta^2}{\delta^2})(\frac{1}{\varepsilon^2}+\frac{1}{\delta^4})$	$(n+\frac{\zeta\sqrt{n}}{\mu})\log\varepsilon^{-1}$	$\checkmark$

Note: Nonconvex: finding ε-first-order stationary points (polylogarithmic terms and the $\frac{1}{\varepsilon^2}$ term are omitted); SOSP: finding (ε,δ)-second-order stationary points (polylogarithmic terms are omitted); SC: only for the μ-strongly convex case; PL condition: finding ε-solution under μ-PL condition (polylogarithmic terms except for $\log\varepsilon^{-1}$ are omitted).

Here $\Delta=f(x^0)-\inf f(x)$, σ_c is the variance between $f_i(x)$, μ is the parameter for PL condition, and ζ is the Hessian-heterogeneity. Since $\zeta\le 2L$, $n+\frac{\zeta\Delta\sqrt{n}}{\varepsilon}$ (achieved by SILVER) is always better than or equal to $n+\frac{\Delta\sqrt{n}}{\varepsilon}$ (optimal complexity without less heterogeneity). Other parameters are assumed to be $O(1)$.

For SARAH, SPIDER-SFO$^+$, and STORM, although their original proofs did not consider the ζ dependency, the dependency on ζ is easily checked by following their proofs. Especially, in federated learning literature, such an ability is show SARAH's an ability of variance reduction methods to utilize less heterogeneity has widely been used. For SARAH and STORM, this ability is shown in Murata and Suzuki (2021) and Cutkosky and Orabona (2019), respectively. Also, the convergence rate of SARAH under the PL condition can be easily checked. SNVRG shows the condition when $n\ge\frac{\sigma^2}{\mu\varepsilon}$ (to find ε-solution $f(x)-f^*\le\varepsilon$); otherwise, replace n by $\frac{\sigma^2}{\mu\varepsilon}$.

ability, and have actually been utilized in federated learning as BVR-L-SGD (Murata and Suzuki, 2021). For many variance reduction algorithms that use multiple full gradients, including SARAH and SPIDER, we can easily check that they have such an ability. However, we cannot directly make ZeroSARAH satisfy this property. Here we explain the reason, which we think reveals the difficulty in simultaneously achieving the optimal complexity and speed-up with less heterogeneity.

Remember that SILVER (Option I) approximated

$$\sum_{s=\min_i T(t,i)+1}^{t}\frac{1}{n}\sum_{i\in\tilde{I}_s^t}^{n}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})).$$

with

$$\sum_{s=\min_i T(t,i)+1}^{t}\frac{|\tilde{I}_s^t|}{bn}\sum_{i\in I^s}^{n}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})).$$

Thus, the discrepancy between the true gradient and the gradient estimator is simply

$$\sum_{s=\min_i T(t,i)+1}^{n}\frac{1}{n}\sum_{i\in\tilde{I}_s^t}^{n}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)-\nabla f(x^{s-1})))$$
$$-\sum_{s=\min_i T(t,i)+1}^{n}\frac{|\tilde{I}_s^t|}{bn}\sum_{i\in I^s}^{n}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)-\nabla f(x^{s-1}))).$$

Note that each term consists of $\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))$, which is bounded by $\tilde{O}(\zeta\|x^s - x^{s-1}\|)$ with high probability.

On the other hand, for ZeroSARAH, the discrepancy between the true gradient $\nabla f(x^t)$ and the gradient estimator is written as

$$\sum_{s=1}^{t} \frac{(1-\lambda)^{t-s}}{b}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))$$
$$+ \lambda \sum_{s=1}^{t-1} \frac{(1-\lambda)^{t-s}}{b} \sum_{i \in I^{s+1}} \left[\nabla f_i(x^s) - \nabla f_i(x^{T(s,i)}) - \frac{1}{n}\sum_{i'=1}^{n}(\nabla f_{i'}(x^s) - \nabla f_{i'}(x^{T(s,i')}))\right].$$

Thus, for the second term, each component $\left\|\nabla f_i(x^s) - \nabla f_i(x^{T(s,i)}) - \frac{1}{n}\sum_{i'=1}^{n}(\nabla f_{i'}(x^s) - \nabla f_{i'}(x^{T(s,i')}))\right\|$ cannot be bounded by $\tilde{O}(\zeta\|x^s - x^{T(s,i)}\|)$ when ζ is small. This is understood as follows. There might exist i' such that $T(s,i') < T(s,i)$. If there are so many i', $\left\|\nabla f_i(x^s) - \nabla f_i(x^{T(s,i)}) - \frac{1}{n}\sum_{i'=1}^{n}(\nabla f_{i'}(x^s) - \nabla f_{i'}(x^{T(s,i')}))\right\| \gtrsim L\|x^{T(s,i)+1} - x^{T(s,i)}\|$ might hold.

This demonstrate that simultaneously achieving the optimal complexity for the finite-sum smooth optimization and speed-up with less heterogeneity with a single-loop algorithm is far from trivial, and our construction of the estimator is the key to enabling this.

A.2. Communication-efficient Federated Learning Utilizing Properties of the Problem

When f_i are just L-smooth, we cannot obtain better communication rounds and communication complexity than those of centralized methods. In order to show the superiority of local methods in terms of communication rounds and communication complexity, we typically assume less inter-client heterogeneity on f_i. Also, when x is a high ($d \gg 1$) dimensional vector, compressability of the vectors is sometimes assumed and utilized.

Less Inter-client Heterogeneity with Local Updates When $f_{i,j}$ are L-smooth but f_i are a little less heterogeneous, i.e., $\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \zeta$, it has been proven that local updates can achieve better communication rounds and complexity, and therefore this less heterogeneity assumption has become popular in the analysis. SCAFFOLD (Karimireddy et al., 2020) proved such a result for quadratic functions, without client sampling. BVR-L-SGD (Murata and Suzuki, 2021) and MimeMVR (Karimireddy et al., 2021) can achieve $O(\frac{\zeta}{\varepsilon^2})$ communication rounds to find ε-first-order stationary points, which is superior to the lower bound for centralized methods of $O(\frac{L}{\varepsilon^2})$ when $\zeta \leq L$. However, $p = P$ (Murata and Suzuki, 2021) or $p = O(\sigma_c^2\varepsilon^{-2})$ (Karimireddy et al., 2021) are required, and thus these methods do not deal with client sampling error efficiently. Also, Murata and Suzuki (2022) proved that $O(\frac{\zeta}{\varepsilon^2} + \frac{\zeta}{\delta^4})$ communicatoun rounds are sufficient to find (ε, δ)-second-order stationary points, but this also requires full client sample at each communication round. We resolved this issue by using SILVER, a novel single-loop variance reduction method, to control the client sampling error efficiently, hence improving communication complexity of them (by allowing client sampling) while maintaining the best communication rounds of $O(\frac{\zeta}{\varepsilon^2})$ (for first-order optimality) and $O(\frac{\zeta}{\varepsilon^2} + \frac{\zeta}{\delta^4})$ (for second-order optimality).

We also mention federated learning under the PL condition or the strong convexity. Under the PL condition, MimeSGD (Karimireddy et al., 2021) yields the communication rounds of $O(\frac{\sigma_c^2}{\mu p \varepsilon^2} + \frac{L}{\mu}\log\varepsilon^{-1})$. FL-SILVER can achieve the communication rounds of $O(\frac{L}{\mu K} + \frac{\zeta(\frac{\sqrt{P}}{p} \wedge 1)}{\mu} + \frac{P}{p})$. Contrary to MimeSGD, FL-SILVER achieves the exponential convergence while allowing client sampling, and notably, FL-SILVER can mitigate the dependency on μ when clients are less heterogeneous ($\zeta \leq L$), which is the first such result to the best of our knowledge. For the strongly convex case, Mishchenko et al. (2022); Grudzień et al. (2023) proved that $O((\frac{P}{p} + \sqrt{\frac{PL}{p\mu}})\log\varepsilon^{-1})$ communication rounds, which is the accelerated rate. Note that they do not show any benefit of local updates under less heterogeneity assumption. Thus when $\zeta \leq \sqrt{\mu}$, our algorithm is superior than theirs. Combining our algorithm and theirs would be an interesting future work.

Compression of Gradients In addition to less heterogeneity, compression of gradients is also proposed as a means to reduce communication cost. Especially, in Gorbunov et al. (2021), the unbiased quantization operator Q is assumed to satisfy $\mathbb{E}[\|Q(x) - x\|^2] \leq \omega\|x\|^2$, and $\sup_x \mathbb{E}[\|Q(x)\|] \leq \zeta_Q \leq d$. Gorbunov et al. (2021) achieved the communication

rounds multiplied by $1/d$ (so that this matches to the usual definition of communication rounds when $\zeta_Q = d$) of $O(\frac{1}{\varepsilon^2}(\frac{\zeta_Q}{d} + \sqrt{\frac{\zeta_Q w}{dP}}))$, but this requires full client participation at each communication. Based on this, Szlendak et al. (2021) invented a new compressor and achieved the adjusted communication round of $O(\frac{1/P+\zeta/\sqrt{P}}{\varepsilon^2})$, but again, this requires full client participation. Moreover, they consider the problem (1) when f_i are distributed, which is a bit different from (2). By multiplying P to $O(\frac{1/P+\zeta/\sqrt{P}}{\varepsilon^2})$ of Szlendak et al. (2021), we obtain $O(\frac{1+\zeta\sqrt{P}}{\varepsilon^2})$, which is the same as the communication complexity of SILVER, but the actual number of communication rounds of theirs is $O(\frac{1}{\varepsilon^2})$, while ours is $O(\frac{\zeta}{\varepsilon^2})$. Tyurin and Richtárik (2022b;a) extended Gorbunov et al. (2021) to local updates. Especially, Tyurin and Richtárik (2022a) yielded the communication complexity of $O(\frac{1}{\varepsilon^2} + \frac{\omega\sqrt{P}}{p\varepsilon^2})$ while allowing partial client participation. However, the benefit of local updates was not shown.

For these reasons, analyses under less heterogeneity and compression-based methods are essentially from different spirits; assumptions are different, and evaluation criteria are sometimes different as well. We leave it for future work to extend Szlendak et al. (2021) to the problem (2) and show the benefit of local updates.

B. EXPERIMENTAL DETAILS AND ADDITIONAL EXPERIMENT

B.1. Experimental Details

B.1.1. Accuracy of the Gradient Estimator of SILVER

We considered a classification of the capital letters using EMNIST By_Class dataset (Cohen et al., 2017). The original dataset consists of $814,255$ images of handwritten uppercase and lowercase letters and numbers 0-9. Because the number of data points in each class is not balanced and the number of images of lowercase letters is relatively small, we only used the images of uppercase letters for the experiment. To balance the number of data points between each class, we took the following procedure. We repeatedly sampled 100 data points five times per each uppercase letter, which yields $26 \times 5 = 130$ groups of sampled data. For each group i, we defined f_i as the average of the cross-entropy loss between the output of the model and the true class, over the 100 data points belonging to the group. As a model, we adopted a two-layer fully-connected neural network. We added L_2-regularizer with a regularization parameter of $\lambda = 0.01$ to the empirical risk.

We compared SILVER with SARAH (Nguyen et al., 2017a;b) and SAGA (Defazio et al., 2014; Reddi et al., 2016b). SAGA is single-loop, while SARAH requires full gradient computation periodically. Also, the theoretical gradient complexity of SAGA is $O(n + \frac{n^{\frac{2}{3}}}{\varepsilon^2})$ and that of SARAH is $O(n + \frac{\sqrt{n}}{\varepsilon^2})$. Remind that SILVER was designed to have as small variance as that of SARAH, while removing the need of periodic full gradient computation; SILVER is single-loop but achieves the gradient complexity of $O(n + \frac{\sqrt{n}}{\varepsilon^2})$.

We took the minibatch size as $b = 12 \fallingdotseq \sqrt{n} = \sqrt{130}$ and the inner-loop length of SARAH to $m = \lfloor \frac{n}{b} \rfloor = 10$. (Note that SARAH refreshes its gradient estimator at every $m = 10$ steps.) We set the learning rate to $\eta = 0.01$ for all algorithms, since the larger step size tend to increase the discrepancy, meaning that it is not fair to compare algorithms with different step sizes to discuss the discrepancy. We plotted the mean of the five trials with different random seeds and the sample variance is also shown in the corresponding (lighter) color for each algorithm.

According to Figure 1, the discrepancy of the SILVER estimator was clearly smaller than that of SAGA, and close to that of SARAH. Therefore, this result validated that our strategy actually worked well.

B.1.2. Escaping Saddle Points with SILVER

We prepared f_i with the same data as the above experiment and employing a four-layer fully-connected neural network. We implemented SARAH (Nguyen et al., 2017a;b), SSRGD (Li, 2019), and ZeroSARAH (Li et al., 2021b). SARAH is a popular variance reduction algorithm with periodic full gradient computations, SSRGD adds noise to SARAH, and ZeroSARAH is a recent single-loop variance reduction method with no theoretical second-order optimality guarantee. We set the minibatch size to $b = 12$ for all algorithms, the inner-loop length of SARAH and SSRGD to $m = \lfloor \frac{n}{b} \rfloor = 10$, and $\lambda = \frac{b}{n} \fallingdotseq 0.092$ for ZeroSARAH. Note that (Li et al., 2021b) adopted $\lambda = \frac{b}{2n}$, but we found that $\lambda = \frac{b}{n}$ was more stable in this setting. The learning rate for each method was tuned individually, from $\eta \in \{1.0, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001\}$, so that the test accuracy after 2000 iterations is the highest. For SSRGD and noisy SILVER, we added small noise of $r = 0.15$. We plotted the mean of the ten trials with different random seeds and the sample variance is also shown in the corresponding

(lighter) color for each algorithm.

B.1.3. Faster and (Locally) Exponential Convergence of FL-SILVER

We verified the performance of FL-SILVER for nonconvex federated learning. For the federated learning problem (2), we again considered the classification of the capital letters (Cohen et al., 2017). This time each f_i consists of $100 \times q\%$ of the data from one class and $100 \times (1-q)\%$ of the data from the other classes, following Murata and Suzuki (2021; 2022). Specifically, we prepared $P = 104$ clients, and for each class of alphabets, we distributed $q \times 100\%$ of the images into four clients, and the rest into the remaining 100 clients. We call this grouping as a dataset with the heterogeneity parameter of q. This makes each f_i a little less heterogeneous, and this time we chose $q = 0.9$. Then, we constructed $f_{i,j}$ with the cross-entropy loss and a two-leyer neural network with width of the hidden layer 100, following Murata and Suzuki (2021). L_2-regularizer with a scale of $\lambda = 0.01$ is added to the empirical risk.

We compared FL-SILVER with FedAvg (McMahan et al., 2017), SCAFFOLD (Karimireddy et al., 2020), MimeMVR (Karimireddy et al., 2021), and BVR-L-SGD (Murata and Suzuki, 2021). Especially BVR-L-SGD (Murata and Suzuki, 2021) can be seen as an almost special case of FL-SILVER when $p = P$ (full client sampling at every communication round). For each algorithm, we set $p = 10 \coloneqq \sqrt{P} = \sqrt{104}$ as the number of the clients used at each communication (except for BVR-L-SGD, which requires $p = P = 104$). Note that, according to Theorems 5 and 6, setting $p \simeq \sqrt{P}$ in FL-SILVER theoretically guarantees that the required number of communication rounds of FL-SILVER are not affected by the client sampling and is the same order as that of BVR-L-(P)SGD, which requires $p = P$ to obtain the theoretical guarantee. Then, we set the number of local update to $K = 10$ and the local minibatch size as $b = 16$. We tuned the learning rate for each algorithm individually from $\{1.0, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001\}$, so that the test accuracy after 2000 outer-loop iterations is the highest. For a fair comparison, the global learning rate of SCAFFOLD was set to $\eta = 1$, as is done in the original paper (Karimireddy et al., 2020), and MimeMVR adopted a momentum parameter of $a = 0.1$ as the authors of the paper reported as the best. We plotted the mean of the five trials with different random seeds and the sample variance is also shown in the corresponding (lighter) color for each algorithm.

In Figure 3, FL-SILVER achieved the smallest gradient norm $\|\nabla f(x^t)\|$ and the linear convergence at the neighborhood of solutions. This can be seen as an empirical verification of the exponential convergence utilizing strong convexity around a local minima, which contrasts FL-SILVER to FedAvg, SCAFFOLD, and MimeMVR. Figure 3 shows that FL-SILVER achieved the higher test accuracy with fewer communication, compared to FedAvg, SCAFFOLD, and MimeMVR. In addition, we observe that FL-SILVER performed similarly to BVR-L-SGD the best method in terms of communication rounds (and again, BVR-L-SGD can be seen as a special case of FL-SILVER without client sampling). FL-SILVER appropriately corrected the error from sampling of the clients while achieving the same performance as that of BVR-L-SGD, and FL-SILVER was about ten times more efficient than BVR-L-SGD in terms of communication complexity by allowing sampling of the clients.

B.2. Additional Experiments

B.2.1. Performance under Changing Heterogeneity

To exhibit how accurately FL-SILVER can control the variance between clients, we measured the performance of FL-SILVER under changing heterogeneity. We changed heterogeneity parameter in the range of $q \in \{0.04 \text{ (i.i.d.)}, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0 \text{ (completely heterogeneous)}\}$, and compared FL-SILVER with FedAvg, in terms of both train and test accuracy. All other settings were the same as those of the experiment for Figure 3. Figure 4 shows the average of five trials with different random seeds.

According to Figure 4, while FedAvg decreased the train and test accuracy as the heterogeneity increases, the performance of FL-SILVER with $q = 1.0$ (totally heterogeneous) was only slightly worse than that with $q = 0.04$. The fact that FL-SILVER, using sampling of clients, was little affected by the strong heterogeneity shows that the client sampling error and the error from local update were successfully corrected.

B.3. Escaping Saddle Points with FL-SILVER

Theorem 4 guarantees second-order optimality of FL-SILVER. To validate this theoretical result, we compared noisy FL-SILVER with FedAvg (McMahan et al., 2017), SCAFFOLD (Karimireddy et al., 2020), MimeMVR (Karimireddy et al., 2021), BVR-L-SGD (Murata and Suzuki, 2021), BVR-L-PSGD (Murata and Suzuki, 2022), and noiseless FL-SILVER.

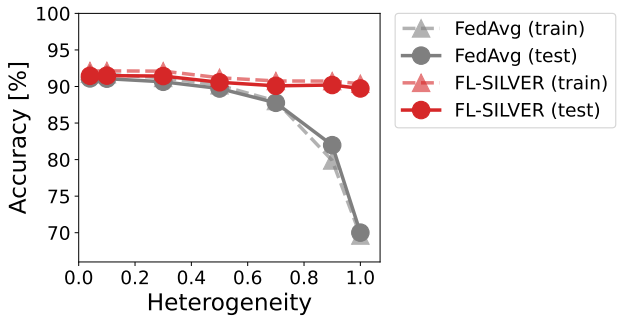


Figure 4. Performance under changing heterogeneity

For noisy FL-SILVER and BVR-L-PSGD, we added small noise of $r = 0.015$. Note that FL-SILVER with small noise and BVR-L-PSGD only have theoretical guarantee of the second-order optimality. Then, we constructed $f_{i,j}$ with the cross-entropy loss with $q = 0.7$ (see Appendix B.1.3) and a three-layer fully-connedted neural network, following Murata and Suzuki (2022). L_2-regularizer with a scale of $\lambda = 0.01$ is added to the empirical risk. We plotted the mean of the five trials with different random seeds. We set $P = 104$, $p = 10$, $K = 10$, and $b = 16$. This time we omitted the sample variance for clearer presentation.

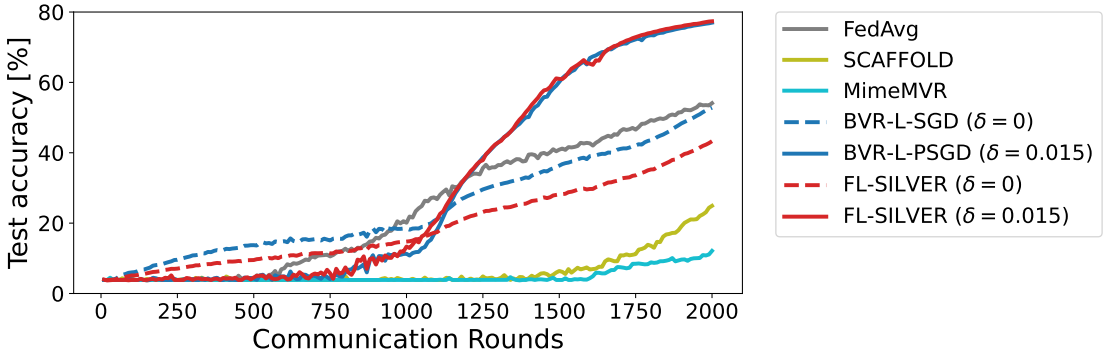


Figure 5. Small perturbation helps faster convergence

The result is shown in Figure 5. We can clearly observe that FL-SILVER with small noise and BVR-L-PSGD achieved the highest test accuracy. Also note that, while SILVER used client sampling, the performance of FL-SILVER was almost the same as BVR-L-PSGD. This shows that FL-SILVER allows sampling of clients without hurting the required number of communication rounds by setting $p = \sqrt{P}$, which is consistent with the theory; according to Theorem 4, setting $p \coloneqq \sqrt{P}$ theoretically guarantees that the convergence rate of FL-SILVER is not affected by the client sampling and achieves the same number of communication complexity as that of BVR-L-PSGD to find SOSPs. As a result, while achieving the most stable and fastest training, FL-SILVER was as ten times efficient as BVR-L-PSGD, in terms of communication complexity (the total number of gradients communicated between the clients and the server).

B.3.1. Comparison with SARAH by changing the learning rate

Here we provide comparison of SILVER with SARAH (Nguyen et al., 2017a;b), which is one of the most prevailing variance reduction algorithm with theoretical optimal gradient complexity of $O\left(n + \frac{\sqrt{n}}{\varepsilon^2}\right)$, which is the same as SILVER, and which uses periodic full gradients.

As is done in the experiment for Figure 1, we prepared f_i in the following way. We repeatedly sampled 100 data points five times per each uppercase letter, which yields $26 \times 5 = 130$ groups of sampled data. For each group i, we define f_i as the average of the cross-entropy loss between the output of the model and the true class over the 100 data points belonging to the group. As a model, we adopted a two-layer fully-connected neural network. We set the minibatch size to $b = 12 \coloneqq \sqrt{n} = \sqrt{130}$ for both algorithms, and the inner-loop length of SARAH to $m = \lfloor \frac{n}{b} \rfloor = 10$. We

added L_2-regularizer to the empirical risk with a fixed regularization parameter of $\lambda = 0.01$. We compared SILVER with SARAH in terms of the training loss, the norm of the gradient computed by the whole training data, the test loss, and the test accuracy, under the same number of stochastic gradient accesses. We changed the learning rate η between $\{0.1, 0.03, 0.01, 0.003, 0.001\}$. We plotted the mean of the five trials with different random seeds and the sample variance is also shown in the corresponding (lighter) color for each algorithm.

Figure 6 shows the result. We clearly observe that the proposed algorithm SILVER slightly faster than SARAH in all range of learning rate η, owing to the removal of multiple full gradient evaluations of SILVER. The trajectories of SILVER are as stable as SARAH in all settings. This result shows that we can remove the requirement of periodic full gradient evaluation without hurting the stability during optimization with SILVER.

B.4. Computing infrastructures

- OS: Ubuntu 16.04.5
- CPU: Intel(R) Xeon(R) CPU E5-2680 v4 2.40GHz
- CPU Memory: 512GB
- GPU: Nvidia Tesla V100 (32GB)
- Programming language: Python 3.6.13
- Deep learning framework: PyTorch 1.7.1

C. TOOLS

We prepare some concentration inequalities and linear algebraic tool for later use.

C.1. Concentration Inequalities

In the following, multiple times we use concentration inequalities that we introduce below. When we say with high probability, it means that the event happens with a failure probability less than an inverse of a sufficient large polynomial in all relevant parameters, i.e., $n, d, \varepsilon, \delta, P, L, \rho, \sigma, \sigma_c$. High-probability bounds come together with logarithmic constants on these parameters. To simplify the analysis, with a slight abuse of notation, we use the same notation C_1 for different polylogarithmic constants that come from concentraion inequalities, and C_1 may vary from line to line.

Proposition 2 (Vector Bernstein inequality (Tropp, 2012))**.** *Let $x_1, \cdots, x_k$ be a finite sequence of independent, random, d-dimensional vectors and $\nu \in (0, 1)$. Assume that each vector satisfies*

$$\|x_i - \mathbb{E}[x_i]\| \leq R \quad \textit{almost surely.}$$

Define

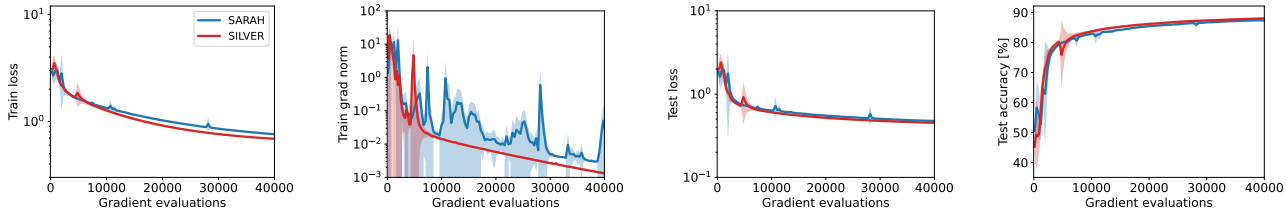
$$\sigma^2 = \sum_{i=1}^{k} \mathbb{E}[\|x_i - \mathbb{E}[x_i]\|^2]$$

Then, with high probability,

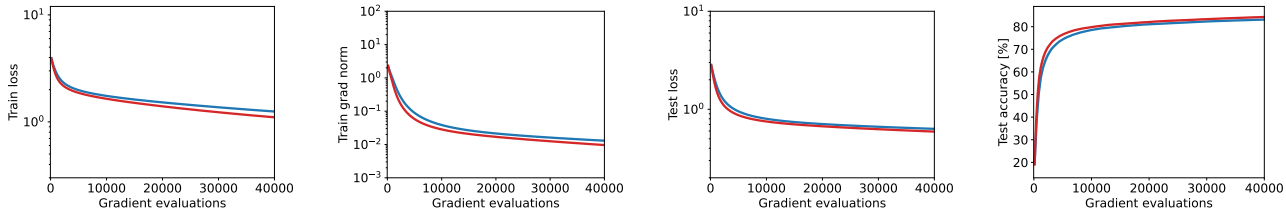
$$\left\| \sum_{i=1}^{k} (x_i - \mathbb{E}[x_i]) \right\|^2 \leq C_1 \cdot (\sigma^2 + R^2).$$

Proposition 3 (Vector Bernstein inequality without replacement)**.** *Let $A = (a_1, a_2, \cdots, a_k)$ be d-dimensional fixed vectors, $X = (x_1, \cdots, x_l)$ $(l \leq k)$ be a random sample without replacement from A. Assume that $\sum_{i=1}^{k} a_i = 0$ and that each vector satisfies*

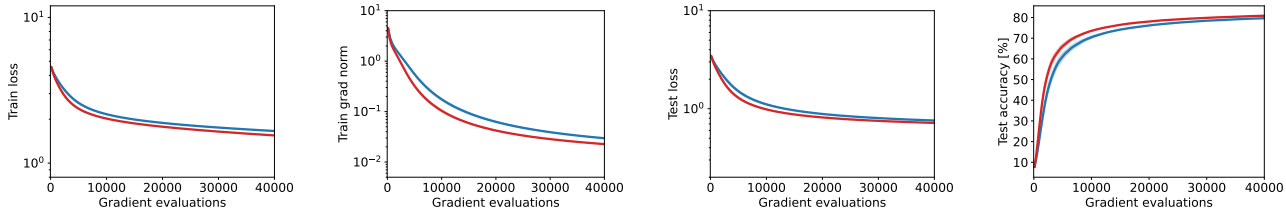
$$\|a_i\| \leq R.$$



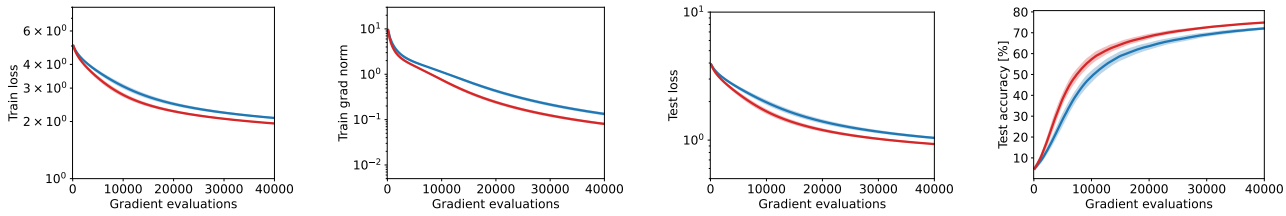
(a) Train loss ($\eta = 0.1$) (b) Gradient norm ($\eta = 0.1$) (c) Test loss ($\eta = 0.1$) (d) Test accuracy ($\eta = 0.1$)



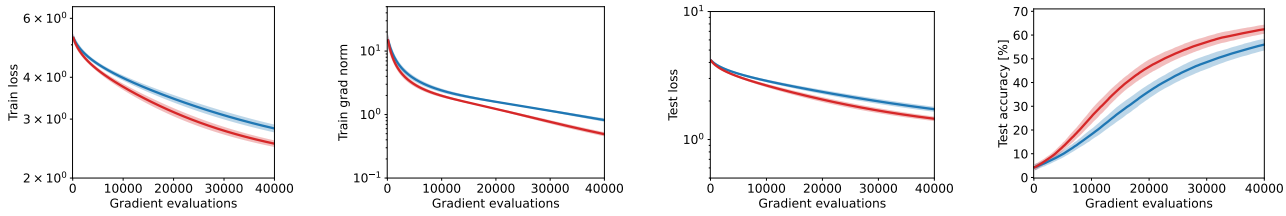
(e) Train loss ($\eta = 0.03$) (f) Gradient norm ($\eta = 0.03$) (g) Test loss ($\eta = 0.03$) (h) Test accuracy ($\eta = 0.03$)



(i) Train loss ($\eta = 0.01$) (j) Gradient norm ($\eta = 0.01$) (k) Test loss ($\eta = 0.01$) (l) Test accuracy ($\eta = 0.01$)



(m) Train loss ($\eta = 0.003$) (n) Gradient norm ($\eta = 0.003$) (o) Test loss ($\eta = 0.003$) (p) Test accuracy ($\eta = 0.003$)



(q) Train loss ($\eta = 0.001$) (r) Gradient norm ($\eta = 0.001$) (s) Test loss ($\eta = 0.001$) (t) Test accuracy ($\eta = 0.001$)

Figure 6. Comparison with SARAH by changing the learning rate

Define

$$\sigma^2 = \frac{1}{k}\sum_{i=1}^{k} \|a_i\|^2.$$

Then, for each $l < k$, with high probability,

$$\left\|\sum_{i=1}^{l} x_i\right\|^2 \leq C_1 \cdot (l\sigma^2 + R^2).$$

Because we are not aware of a rigorous proof of such a result, we attach its complete proof at the end of this subsection.

Proposition 4 (Azuma-Hoeffding inequality with high probability (Chung and Lu, 2006; Tao and Vu, 2015))**.** *Let $\{x_i\}$ be a d-dimensional vector sequence and martingale with respect to a filtration $\{\mathcal{F}_i\}$. Assume that each x_i satisfies $\mathbb{E}[x_i|\mathcal{F}_{i-1}] = 0$ and*

$$\|x_i\| \leq R_i \quad \textit{with probability } 1 - \nu_i$$

for $\nu_i \in (0,1)$ $(i = 1, \ldots, k)$. Then, with high probability,

$$\left\|\sum_{i=1}^{k} x_i\right\|^2 \leq C_1 \sum_{i=1}^{k} R_i^2.$$

Proof of Proposition 3. First, we consider the case $l \leq \frac{k}{2}$. Let $y_i = \sum_{j=1}^{i} x_j$ and consider a filtration $\mathcal{F}_i = \sigma(x_1, \cdots, x_i)$. Then, we have

$$\mathbb{E}\left[y_{i+1} \,|\mathcal{F}_i\right] = y_i + \frac{1}{k-i}\left(\sum_{j=1}^{n} a_j - \sum_{j=1}^{i} x_j\right) = \frac{k-i-1}{k-i} y_i.$$

This means that $\left\{\frac{1}{k-i}y_i\right\}_{i=0}^{l}$ is martingale with respect to $\{\mathcal{F}_i\}$. We have that this martingale satisfies the assumptions of Proposition 5 (see below) with $R'^2 = \frac{R^2}{(k-l)^2}$ and $\sigma'^2 = \frac{2\sigma^2}{(k-l)^2}$. In fact, we have

$$\begin{aligned}\left\|\frac{1}{k-i-1}y_{i+1} - \mathbb{E}\left[\frac{1}{k-i-1}y_{i+1}\middle|\mathcal{F}_i\right]\right\|^2 &= \left\|\frac{1}{k-i-1}x_{i+1} - \mathbb{E}\left[\frac{1}{k-i-1}x_{i+1}\middle|\mathcal{F}_i\right]\right\|^2 \\ &\leq \left\|\frac{1}{k-i-1}x_{i+1}\right\|^2 \leq \frac{R^2}{(k-i-1)^2} \leq \frac{R^2}{(k-l)^2},\end{aligned}$$

where the equality follows since $x_1 \ldots, x_i$ are $\mathcal{F}_i$-measurable, and

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{k-i-1}y_{i+1} - \mathbb{E}\left[\frac{1}{k-i-1}y_{i+1}\middle|\mathcal{F}_i\right]\right\|^2\middle|\mathcal{F}_i\right] &\leq \mathbb{E}\left[\left\|\frac{1}{k-i-1}x_{i+1}\right\|^2\middle|\mathcal{F}_i\right] \\ &= \frac{1}{(k-i-1)^2}\cdot\frac{1}{k-i}\left(\sum_{j=1}^{k}\|a_j\|^2 - \sum_{j=1}^{i}\|x_j\|^2\right) \\ &\leq \frac{1}{(k-l)^2}\cdot\frac{2}{k}\sum_{i=1}^{k}\|a_i\|^2 \quad \left(\because k-i \geq \frac{k}{2}\right) \\ &= \frac{2\sigma^2}{(k-l)^2}.\end{aligned}$$

Thus, we use Proposition 5 (see below) to obtain

$$\mathbb{P}\left[\|y_l\| \geq t\right] \leq (d+1) \cdot \exp\left(\frac{-t^2}{2l\sigma^2 + Rt/3}\right).$$

What remains is the case of $l \geq \frac{k}{2}$. Since $\sum_{i=1}^{l} x_i = -\sum_{i=l+1}^{k} x_i$ holds, we can apply the above bound for $\sum_{i=l+1}^{k} x_i$. Thus, we have the first assertion for all $l < k$. The second assertion follows by setting $t = C_1 \cdot (l\sigma^2 + R)$.

□

Proposition 5 (Freedman's inequality for matrix martingales)**.** *Consider a matrix martingale $\{Y_i \mid i = 0, 1, \cdots\}$ with respect to a filtration $\{\mathcal{F}_i\}$, whose values are matrices with dimension $d_1 \times d_2$, and let $\{X_i \mid i = 1, 2, \cdots\}$ be the difference sequence. Assume that each of the difference sequence is uniformly bounded:*

$$\|X_i\|^2 \leq R'^2 \quad \textit{almost surely.}$$

Also, assume that each i satisfies

$$\max\left\{\left\|\mathbb{E}[X_i X_i^\top \mid \mathcal{F}_{i-1}]\right\|, \left\|\mathbb{E}[X_i^\top X_i \mid \mathcal{F}_{i-1}]\right\|\right\} \leq \sigma'^2 \quad \textit{almost surely.}$$

Then, for all $t \geq 0$ and for each l,

$$\mathbb{P}\left[\|Y_l\| \geq t\right] \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{l\sigma'^2 + R't/3}\right).$$

C.2. Linear Algebraic Tool

The following lemma is due to Murata and Suzuki (2022). We provide its proof below for completeness.

Lemma 2 (Murata and Suzuki (2022))**.** *Let A be a $d \times d$ symmetric matrix with the smallest and largest eigenvalues $\lambda_{\min} < 0$ and $\lambda_{\max} < 1$, respectively. Then, for $k = 0, 1, \cdots$, it holds that*

$$\|A(I - A)^k\| \leq -\lambda_{\min}(1 - \lambda_{\min})^k + \frac{1}{k+1}.$$

Proof. Since A is diagonalizable, we write $A = \sum_{i=1}^{d} \lambda_i e_i e_i^\top$, where $e_1, \ldots, e_d$ are normalized eigenvectors and $\lambda_{\min} = \lambda_1 \leq \cdots \leq \lambda_d = \lambda_{\max}$ are the corresponding eigenvalues. Then, it holds that

$$A(I - A)^k = \sum_{i=1}^{d} \lambda_i (1 - \lambda_i)^k e_i e_i^\top.$$

Thus, the remaining is to evaluate $\max_i |\lambda_i(1 - \lambda_i)^k|$. After some algebra, we get

$$\begin{aligned} 0 < \lambda(1-\lambda)^k &\leq \begin{cases} -\lambda(1-\lambda)^k & (\text{if } \lambda \leq 0) \\ \frac{1}{k+1}\left(\frac{k}{k+1}\right)^k & \left(\text{if } \lambda > 0; \text{ the equality holds with } \lambda = \frac{1}{1+k}\right) \end{cases} \\ &\leq -\lambda_{\min}(1 - \lambda_{\min})^k + \frac{1}{k+1}, \end{aligned}$$

which concludes the proof. □

D. MISSING PROOFS FOR SILVER

D.1. First-order Optimality (Proof of Theorem 1)

This section proves Theorem 1, the optimal gradient complexity for finding first-order stationary points. The proof crucially depends on the following lemma, which shows that the error from previous steps exponentially decays and never accumulates owing to our update SILVER, without full gradient computation.

Lemma 1 (formal). *Let Assumptions 5-(a) and 3-(a) (only for Option II) hold. Let* $g^t = \frac{1}{n}\sum_{i=1}^n y_i^t$ *and all the other variables be as stated in Algorithm 1. Then,*

$$\mathbb{E}\|g^t - \nabla f(x^t)\|^2 \leq \frac{30\zeta^2}{b}\sum_{s=1}^{t}(1-\frac{b}{4n})^{t-s+1}\mathbb{E}\left[\|x^s - x^{s-1}\|^2\right] + \frac{9\sigma_c^2\mathbb{1}[\text{Option II}]}{b}(1-\frac{b}{n})^t.$$

Here $\mathbb{1}[\text{Option II}] = 1$ *for Option II and* 0 *otherwise.*

For the proof of this lemma, we decompose the error as follows:

$$\begin{aligned}
&\|g^t - \nabla f(x^t)\|^2 \\
&= \left\|\frac{1}{n}\sum_{s=1}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I_s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) - \sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}))\right) + \frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0 - \nabla f_i(x^0))\right\|^2 \\
&\leq 3\underbrace{\left\|\frac{1}{n}\sum_{s=1}^{t}\sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))\right\|^2}_{\text{(a)}} \\
&\quad + 3\underbrace{\left\|\frac{1}{n}\sum_{s=1}^{t}\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))\right\|^2}_{\text{(b)}} + 3\underbrace{\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0 - \nabla f_i(x^0))\right\|^2}_{\text{(c)}}.
\end{aligned} \tag{3}$$

For each of (a), (b), and (c), we apply one of the following lemmas, which yields the assertion.

Lemma 3. *Under Assumption 5-(a), the term (a) is bounded as follows:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{s=1}^{t}\sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))\right\|^2\right] \leq \frac{4\zeta^2}{b}\sum_{s=1}^{t}(1-\frac{b}{2n})^{t-s+1}\mathbb{E}[\|x^s - x^{s-1}\|^2].$$

Lemma 4. *Under Assumption 5-(a), the term (b) is bounded as follows:*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{s=1}^{t}\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))\right\|^2\right] \leq \frac{6\zeta^2}{b}\sum_{s=1}^{t}(1-\frac{b}{4n})^{t-s+1}\mathbb{E}\left[\|x^s - x^{s-1}\|^2\right].$$

Lemma 5. *For Option I,* $(\mathrm{c}) = 0$*. For Option II, under 3-(a), we have*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0 - \nabla f_i(x^0))\right\|^2\right] \leq \frac{3\sigma_c^2}{b}(1-\frac{b}{n})^t.$$

We prove these auxiliary lemmas as follows.

Proof of Lemma 3. First, we have that

$$\begin{aligned}
\text{(a)} &= \left\| \frac{1}{n} \sum_{s=1}^{t} \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\
&\leq \frac{\sum_{s=1}^{t} E[|\tilde{I}_s^t|](1-\frac{b}{2n})^{-t+s-1}}{n^2} \sum_{s=1}^{t} \frac{(1-\frac{b}{2n})^{t-s+1}}{\mathbb{E}[|\tilde{I}_s^t|]} \left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\
&\leq \frac{\sum_{s=1}^{t} n(1-\frac{b}{4n})^{-t+s-1}}{n^2} \sum_{s=1}^{t} \frac{(1-\frac{b}{2n})^{t-s+1}}{\mathbb{E}[|\tilde{I}_s^t|]} \left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2, \qquad (4)
\end{aligned}$$

where we used the Cauthy-Schwarts inequality for the second line, and $\mathbb{E}[|\tilde{I}_s^t|] = n(1-\frac{b}{n})^{t-s+1}$ and $(1-\frac{b}{n})(1-\frac{b}{2n})^{-1} \leq 1-\frac{b}{4n}$ $(b \leq n)$ for the third line. Note that $\tilde{I}_s^t$ depends only on $I_s, \cdots, I_t$, while x^s and x^{s-1} depend only on $I_0, \cdots, I^{s-1}$. Therefore, by conditioning on $I_0, \cdots, I^{s-1}$ and $|\tilde{I}_s^t|$, each term of (4) is bounded by

$$\mathbb{E}\left[\left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \middle| I_0, \cdots, I^{s-1}, |\tilde{I}_s^t| \right] = |\tilde{I}_s^t| \zeta^2 \|x^s - x^{s-1}\|^2.$$

We apply this to (4) and obtain that

$$\begin{aligned}
\mathbb{E}[\text{(a)}] &\leq \frac{\sum_{s=1}^{t} n(1-\frac{b}{4n})^{t-s+1}}{n^2} \sum_{s=1}^{t} \frac{(1-\frac{b}{2n})^{t-s+1}}{\mathbb{E}[|\tilde{I}_s^t|]} \mathbb{E}[|\tilde{I}_s^t| \zeta^2 \|x^s - x^{s-1}\|^2 | I_0, \cdots, I^{s-1}, |\tilde{I}_s^t|] \\
&\leq \frac{\sum_{s=1}^{t} n(1-\frac{b}{4n})^{t-s+1} \zeta^2}{n^2} \sum_{s=1}^{t} (1-\frac{b}{2n})^{t-s+1} \mathbb{E}\left[\|x^s - x^{s-1}\|^2 \frac{\mathbb{E}[\tilde{I}_s^t | I_0, \cdots, I^{s-1}]}{\mathbb{E}[|\tilde{I}_s^t|]} \right] \\
&\leq \frac{4\zeta^2}{b} \sum_{s=1}^{t} (1-\frac{b}{2n})^{t-s+1} \mathbb{E}[\|x^s - x^{s-1}\|^2],
\end{aligned}$$

which concludes the proof. □

Proof of Lemma 4. We decompose the target as follows:

$$\begin{aligned}
\text{(b)} &= \left\| \frac{1}{n} \sum_{s=1}^{t} \frac{|\tilde{I}_s^t|}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\
&\leq 2 \left\| \frac{1}{n} \sum_{s=1}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2 \\
&\quad + 2 \left\| \frac{1}{n} \sum_{s=1}^{t} \frac{|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2. \qquad (5)
\end{aligned}$$

For the first term, because $\mathbb{E}[|\tilde{I}_s^t|] = n(1-\frac{b}{n})^{t-s+1}$ and independentness of $I^1, \cdots, I^t$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\| \frac{1}{n} \sum_{s=1}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2 \right] \\
&\leq \frac{\zeta^2}{b} \sum_{s=1}^{t} (1-\frac{b}{n})^{2(t-s+1)} \mathbb{E}[\|x^s - x^{s-1}\|^2]. \qquad (6)
\end{aligned}$$

On the other hand, for the second term, we further decompose it as

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{s=1}^{t}\frac{|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|]}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\\
&\leq\frac{\sum_{s=1}^{t}(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{2n})^{t-s+1}}{n^2b^2}\\
&\qquad\times\sum_{s=1}^{t}\frac{(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|])^2}{(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{2n})^{t-s+1}}\left\|\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\\
&\leq\frac{2}{b^3n}\sum_{s=1}^{t}\frac{(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|])^2}{(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{2n})^{t-s+1}}\left\|\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2,
\end{aligned}$$

where we used the Cauthy-Schwartz for the first inequality. Taking the expectation of the last line yields

$$\begin{aligned}
&\mathbb{E}\left[\frac{2}{b^3n}\sum_{s=1}^{t}\frac{(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|])^2}{(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{2n})^{t-s+1}}\left\|\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\right]\\
&=\mathbb{E}\left[\frac{2}{b^3n}\sum_{s=1}^{t}\frac{\mathbb{E}[(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|])^2|I^s]}{(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{2n})^{t-s+1}}\left\|\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\right]\\
&=\mathbb{E}\left[\frac{2}{b^3n}\sum_{s=1}^{t}\frac{n(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{n})^{t-s+1}}{(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{2n})^{t-s+1}}\left\|\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\right]\\
&\leq\mathbb{E}\left[\frac{1}{b^4}\sum_{s=1}^{t}(1-\frac{b}{4n})^{t-s+1}b\zeta^2\|x^s-x^{s-1}\|^2\right]\\
&=\frac{2\zeta^2}{b^3}\sum_{s=1}^{t}(1-\frac{b}{4n})^{t-s+1}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right],
\end{aligned}\tag{7}$$

where we used $\mathbb{E}[(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|])^2|I^s]=n(1-(1-\frac{b}{n})^{t-s+1})(1-\frac{b}{n})^{t-s+1}$ for the second equality.

Applying (6) and (7) to (5), we have

$$\mathbb{E}[(\mathrm{b})]\leq\frac{6\zeta^2}{b}\sum_{s=1}^{t}(1-\frac{b}{4n})^{t-s+1}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right],$$

which concludes the proof. □

Proof of Lemma 5. As for Option I, the assertion directly follows from the definition $y_i^0=\nabla f_i(x^0)$ $(i=1,\cdots,n)$. Henceforth, we prove the bound for the Option II. We first decompose

$$\begin{aligned}
(\mathrm{c})&=\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0-\nabla f_i(x^0))\right\|^2\\
&\leq\left\|\frac{|\tilde{I}_1^t|}{nb}\sum_{i\in I^0}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2+\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2.
\end{aligned}$$

By taking expectation, we have

$$\begin{aligned}
(\mathrm{c}) &= \mathbb{E}\left[\frac{\mathbb{E}[|\tilde{I}_1^t|^2|I^0]}{b^2n^2}\left|\sum_{i\in I^0}(\nabla f_i(x^0)-\nabla f(x^0))\right|^2\right]+\mathbb{E}\left[\frac{|\tilde{I}_1^t|}{n^2}\sigma_c^2\right]\\
&\le \mathbb{E}\left[\frac{n(1-(1-\frac{b}{n})^t)(1-\frac{b}{n})^t+n^2(1-\frac{b}{n})^{2t}}{b^2n^2}\cdot\left|\sum_{i\in I^0}(\nabla f_i(x^0)-\nabla f(x^0))\right|^2\right]+\frac{n(1-\frac{b}{n})^t}{n^2}\sigma_c^2\\
&\le \frac{2\sigma_c^2}{b}(1-\frac{b}{n})^t+\frac{\sigma_c^2}{n}(1-\frac{b}{n})^t\le\frac{3\sigma_c^2}{b}(1-\frac{b}{n})^t,
\end{aligned}$$

which concludes the proof. □

Now we have Lemma 1, which bounds the variance of our gradient estimator. We combine this with the following descent lemma, which ensures decrease of the function values. Note that Assumption 1-(a) implies L-gradient Lipschitzness of f.

Lemma 6. *Let f be an L-gradient Lipschitz function and $x^t := x^{t-1}-\eta g^{t-1}+\xi^{t-1}$ with $\|\xi^{t-1}\|\le r$. Then,*

$$f(x^t)\le f(x^{t-1})+\eta\|\nabla f(x^{t-1})-g^{t-1}\|^2-\frac{\eta}{2}\|\nabla f(x^{t-1})\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^t-x^{t-1}\|^2+\frac{r^2}{\eta}$$

holds.

Proof. Starting from the direct result from L-gradient Lipschitzness, we have

$$\begin{aligned}
f(x^t)&\le f(x^{t-1})+\langle\nabla f(x^{t-1}),x^t-x^{t-1}\rangle+\frac{L}{2}\|x^t-x^{t-1}\|^2\\
&= f(x^{t-1})+\left\langle\nabla f(x^{t-1})-g^{t-1}+\frac{\xi^{t-1}}{\eta},x^t-x^{t-1}\right\rangle+\left\langle g^{t-1}-\frac{\xi^{t-1}}{\eta},x^t-x^{t-1}\right\rangle+\frac{L}{2}\|x^t-x^{t-1}\|^2\\
&= f(x^{t-1})+\left\langle\nabla f(x^{t-1})-g^{t-1}+\frac{\xi^{t-1}}{\eta},x^t-x^{t-1}\right\rangle-\left(\frac{1}{\eta}-\frac{L}{2}\right)\|x^t-x^{t-1}\|^2\\
&= f(x^{t-1})+\frac{\eta}{2}\left\|\nabla f(x^{t-1})-g^{t-1}+\frac{\xi^{t-1}}{\eta}\right\|^2-\frac{\eta}{2}\|\nabla f(x^{t-1})\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^t-x^{t-1}\|^2 \qquad (8)\\
&\le f(x^{t-1})+\eta\|\nabla f(x^{t-1})-g^{t-1}\|^2-\frac{\eta}{2}\|\nabla f(x^{t-1})\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^t-x^{t-1}\|^2+\eta\left\|\frac{\xi^{t-1}}{\eta}\right\|^2 \qquad (9)\\
&\le f(x^{t-1})+\eta\|\nabla f(x^{t-1})-g^{t-1}\|^2-\frac{\eta}{2}\|\nabla f(x^{t-1})\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^t-x^{t-1}\|^2+\frac{r^2}{\eta}, \qquad (10)
\end{aligned}$$

where we used $x^t-x^{t-1}=\eta v^{t-1}+\xi^{t-1}$ and $\langle a-b,b\rangle=\frac{1}{2}(\|a-b\|^2-\|a\|^2+\|b\|^2)$ for (8), $\|a+b\|^2\le 2(\|a\|^2+\|b\|^2)$ for (9), and $\|\xi^{t-1}\|\le r$ for (10). □

By combining Lemmas 1 and 6, we obtain the desired first-order convergence guarantee.

Proof of Theorem 1. We sum up Lemma 6 over all $t=1,2,\cdots,T$ to get

$$\begin{aligned}
&\sum_{t=1}^T\|\nabla f(x^{t-1})\|^2\le\frac{2}{\eta}\left[(f(x^0)-f(x^t))-\sum_{t=1}^T\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^t-x^{t-1}\|^2+\eta\sum_{t=1}^T\|\nabla f(x^{t-1})-g^{t-1}\|^2\right]+\frac{2Tr^2}{\eta^2},\\
&\therefore\left(\min_{1\le t\le T-1}\mathbb{E}[\|\nabla f(x^{t-1})\|]\right)^2\le\frac{2\Delta}{\eta T}-\frac{1}{\eta T}\sum_{t=1}^T\left(\frac{1}{\eta}-L\right)\mathbb{E}[\|x^t-x^{t-1}\|^2]+\frac{2}{T}\sum_{t=1}^T\mathbb{E}[\|\nabla f(x^{t-1})-g^{t-1}\|^2]+\frac{\varepsilon^2}{2},
\end{aligned} \qquad (11)$$

where we used $r\le\frac{\eta\varepsilon}{2}$.

We let $g^t = \frac{1}{n}\sum_{i=1}^n y_i^t$. Applying Lemma 1 yields that

$$\sum_{t=1}^T \mathbb{E}[\|\nabla f(x^{t-1}) - g^{t-1}\|^2] \leq \sum_{t=1}^T \frac{30\zeta^2}{b}\sum_{s=1}^{t-1}(1-\frac{b}{4n})^{t-s}\mathbb{E}\left[\|x^s - x^{s-1}\|^2\right] + \sum_{t=1}^T \frac{9\sigma_c^2 \mathbb{1}[\text{Option II}]}{b}(1-\frac{b}{n})^{t-1}$$
$$\leq \frac{120n\zeta^2}{b^2}\sum_{t=1}^T \mathbb{E}\left[\|x^t - x^{t-1}\|^2\right] + \frac{9n\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2}. \tag{12}$$

Applying (12) to (11), we have

$$\left(\min_{1\leq t\leq T-1}\mathbb{E}[\|\nabla f(x^{t-1})\|]\right)^2 \leq \frac{2\Delta}{\eta T} - \frac{1}{\eta T}\sum_{t=1}^T\left(\frac{1}{\eta} - L - \frac{240\eta n\zeta^2}{b^2}\right)\mathbb{E}[\|x^t - x^{t-1}\|^2] + \frac{\varepsilon^2}{2} + \frac{9n\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2 T}.$$

By taking $\eta \leq \frac{1}{2L} \wedge \frac{b}{\zeta\sqrt{480n}}$, we obtain that

$$\left(\min_{1\leq t\leq T-1}\mathbb{E}[\|\nabla f(x^{t-1})\|]\right)^2 \leq \frac{2\Delta}{\eta T} + \frac{\varepsilon^2}{2} + \frac{9n\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2 T}.$$

By taking $T \geq \frac{8\Delta}{\eta\varepsilon^2} + \frac{36n\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2\varepsilon^2}$, we obtain Theorem 1. Especially, when $\eta \leq \frac{1}{2L}\wedge\frac{b}{\zeta\sqrt{240n}}$, $T \geq \frac{8\Delta}{\varepsilon^2}(2L + \zeta\sqrt{480n}/b) + \frac{36n\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2\varepsilon^2}$. □

D.2. Second-order Optimality (Proof of Theorem 2)

The goal of this subsection is to prove that SILVER can find second-order stationary points, as the first single-loop algorithm.

Theorem 2 (full version). *Under Assumptions 1-(b), 2, 3-(a) (only for Option II), 4, and 5-(b), and let* $b = \tilde{\Omega}(\sqrt{n} + \frac{\zeta^2}{\delta^2})$, $\eta = \tilde{\Theta}(\frac{1}{L})$, $r = \tilde{O}\left(\frac{\varepsilon\vee(\delta^2/\rho)}{L}\right)$. *Then, Algorithm 1 finds* (ε,δ)*-SOSPs using*

$$\tilde{O}\left(n + \left(L\Delta + \frac{\mathbb{1}[\text{Option II}]n\sigma_c^2}{b^2}\right)\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right)b\right)$$

stochastic gradients, with high probability.

The proof follows that of (Jin et al., 2017; Ge et al., 2019; Li, 2019); Let x^{τ_0} be a point such as $\lambda_{\min}(\nabla f(x^{\tau_0})) \leq -\delta$. Around that point, we consider two points x_1 and x_2 such that $\langle x_1, \mathsf{e}\rangle \approx \langle x_2, \mathsf{e}\rangle$, where e is the eigenvector of $\lambda_{\min}(\nabla f(x^{\tau_0}))$. Then, two coupled sequences that SILVER generates from the two initial points (x_1 and x_2) will be separated exponentially, as long as they are in a small region around the initial points. This means that if we add some noise to the sequence around a saddle point, then with a certain probability, the algorithm can move away from the saddle point.

We again emphasize that, although this high-level proof outline is classical, we face the difficulties arising from the single-loop structure of the algorithm.

First, we need to prove the high-probability bound on the gradient estimator. Our estimator is more correlated than existing ones due to the $|\tilde{I}_s^t|$ term, thus requiring more delicate analysis than those for StabilizedSVRG (Ge et al., 2019) and SSRGD (Li, 2019).

Also, Many existing algorithms compute periodic full gradient and can refresh their gradient estimators around saddle points. In contrast, our single-loop algorithm does not use full gradient, meaning that we have to deal with the error accumulated before that point, and it is not trivial whether such errors can be sufficiently small so that the direction of the negative eigenvalue can be found by the gradient estimator. Regarding this point, we found that taking minibatch size as large as $b \gtrsim \sqrt{n} + \frac{\zeta^2}{\delta^2}$ is sufficient. We note that a classical choice of δ is $\delta = O(\sqrt{\varepsilon})$ (Nesterov and Polyak, 2006; Jin et al., 2017; Li, 2019), and in this case b should be taken as $O(\sqrt{n} + \frac{1}{\varepsilon})$, which as moderate as existing literature (Ge et al., 2019; Li, 2019).

The exponential separation of two sequences is formalized in the following lemma.

Lemma 7 (Small stuck region)**.** *Let Assumptions 1-(b), 4, 5-(b) hold. Let $\{x^t\}$ be a sequence generated by* SILVER *and suppose that there exists a step τ_0 such that $-\gamma := \lambda_{\min}(\nabla^2 f(x^{\tau_0})) \leq -\delta$ holds. We denote the smallest eigenvector direction of $\lambda_{\min}(\nabla^2 f(x^{\tau_0}))$ by e. Moreover, we define a coupled sequence $\{\tilde{x}^t\}$ by running* SILVER *with $\tilde{x}^0 = x^0$ and share the same choice of randomness, i.e., minibatches and noises with $\{x^t\}$, except for the noise at some step $\tau(> \tau_0)$: $\tilde{\xi}^\tau = \xi^\tau - r_e \mathsf{e}$ with $r_e \geq \frac{r\nu}{\sqrt{d}}$ $(\nu < 1)$. Let $w^t = x^t - \tilde{x}^t$, $g^t = \frac{1}{n}\sum_{i=1}^n y_i^t$, $\tilde{g}^t = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i^t$, and $h^t = g^t - \nabla f(x^t) - (\tilde{g}^t - \nabla f(\tilde{x}^t))$. Here $\tilde{y}_i^t$ is the counterpart of y_i^t and corresponds to $\{\tilde{x}^t\}$.*

Then, there exists a constant $C_2 = \tilde{\Theta}(1)$ such that if we take $b = \tilde{\Omega}(\sqrt{n} + \frac{\zeta^2}{\delta^2})$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_2 = O\left(\frac{\log \frac{\gamma}{C_2 \rho r_e}}{\eta\delta}\right) = \tilde{O}\left(\frac{L}{\delta}\right)$, it holds that

$$\max_{\tau_0 \leq t < \tau + T_2} \{\|x^t - x^\tau\|, \|\tilde{x}^t - x^\tau\|\} \geq \frac{\delta}{C_2 \rho}$$

with high probability.

In the following, we first prove the high probability bound on the error in g^t (high-probability version of Lemma 1), and then prove Lemma 7.

Lemma 1 (High probability bound). *Let Assumptions 5-(b), and 3-(a) (only for Option II) hold. Let $g^t = \frac{1}{n}\sum_{i=1}^n y_i^t$ and all the other variables be as stated in Algorithm 1. Then, with high probability,*

$$\|g^t - \nabla f(x^t)\|^2 \leq \frac{C_1 \zeta^2}{b} \sum_{s=\max\{1, t-T_1\}}^{t} \|x^s - x^{s-1}\|^2 + \frac{C_1 \sigma_c^2 \mathbb{1}[\text{Option II}] \mathbb{1}[t \leq T_1]}{b},$$

where $T_1 = \tilde{\Theta}(\frac{n}{b})$ and $C_1 = \tilde{O}(1)$. Here $\mathbb{1}[\text{Option II}] = 1$ for Option II and 0 otherwise, and $\mathbb{1}[t \leq T_1] = 1$ when $t \leq T_1$ and 0 otherwise.

Proof. With high probability, $\tilde{I}_s^t = \emptyset$ if $t - s = \tilde{\Omega}(\frac{n}{b})$. Thus we assume this event happens below. Thus, by taking $T_1 = \tilde{\Theta}(\frac{n}{b})$ and following (3), we have

$$\begin{aligned}
&\|g^t - \nabla f(x^t)\|^2 \\
&\leq 3 \underbrace{\left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1\}}^{t} \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2}_{\text{(a)}} \\
&+ 3 \underbrace{\left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1\}}^{t} \frac{|\tilde{I}_s^t|}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2}_{\text{(b)}} + 3 \underbrace{\left\| \frac{\mathbb{1}[t \leq T_1]}{n} \sum_{i \in \tilde{I}_1^t} (y_i^0 - \nabla f_i(x^0)) \right\|^2}_{\text{(c)}}.
\end{aligned}$$

We bound (a), (b), and (c) separately by showing the high-probability bounds of Lemmas 3 to 5. For (a),

$$\begin{aligned}
\text{(a)} &\leq \frac{T_1 + 1}{n^2} \sum_{s=\max\{1,t-T_1\}}^{t} \left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\
&\leq \frac{T_1 + 1}{n^2} \sum_{s=\max\{1,t-T_1\}}^{t} C_1 |\tilde{I}_s^t| \zeta^2 \|x^s - x^{s-1}\|^2 \leq \frac{C_1 \zeta^2}{b} \sum_{s=\max\{1,t-T_1\}}^{t} \|x^s - x^{s-1}\|^2,
\end{aligned}$$

where we used the vector Bernstein inequality without replacement (Proposition 3) for the second inequality, under the condition that $|\tilde{I}_s^t|$ is fixed.

For (b),

$$\text{(b)} \leq 2\left\|\frac{1}{n}\sum_{s=\max\{1,t-T_1\}}^{t}\frac{\mathbb{E}[|\tilde{I}_s^t|]}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2 \tag{13}$$

$$+\frac{2T_1}{n^2}\sum_{s=\max\{1,t-T_1\}}^{t}\left(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|]\right)^2\left\|\frac{1}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2. \tag{14}$$

For the first term (13),

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{s=\max\{1,t-T_1\}}^{t}\frac{\mathbb{E}[|\tilde{I}_s^t|]}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\\
&=\sum_{s=\max\{1,t-T_1\}}^{t}\left\|\frac{\mathbb{E}[|\tilde{I}_s^t|]}{bn^2}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\\
&\leq\sum_{s=\max\{1,t-T_1\}}^{t}\left\|\frac{\mathbb{E}[|\tilde{I}_s^t|]}{bn^2}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)-\nabla f(x^{s-1})))\right\|^2\\
&\leq\frac{C_1\mathbb{E}[|\tilde{I}_s^t|]^2\zeta^2}{bn^2}\sum_{s=\max\{1,t-T_1\}}^{t}\|x^s-x^{s-1}\|^2\leq\frac{C_1\zeta^2}{b}\sum_{s=\max\{1,t-T_1\}}^{t}\|x^s-x^{s-1}\|^2,
\end{aligned}\tag{15}$$

where we used the independentness of I^s in the first equality and the vector Bernstein inequality (Proposition 3) for the second last inequality.

On the other hand, for the second term (14), we have $\left(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|]\right)^2\leq C_1 n$. This can be checked as follows. We prepare a "reverse" filtration $\tilde{\mathcal{F}}=\{\tilde{\mathcal{F}}_s^t\}_{s=t}^{\max\{1,t-T_1\}}$ with $\tilde{\mathcal{F}}_s^t=\sigma(I_t,I_{t-1},\cdots,I_s)$. Because we have $\mathbb{E}_s\left[|\tilde{I}_{s+1}^t|-|\tilde{I}_s^t|\mid\tilde{\mathcal{F}}_{s+1}^t\right]=\frac{b}{n}|\tilde{I}_{s+1}^t|$, the following relation holds: $\mathbb{E}_s\left[|\tilde{I}_s^t|\,\middle|\,\tilde{\mathcal{F}}_{s+1}^t\right]=\left(1-\frac{b}{n}\right)|\tilde{I}_{s+1}^t|$. Hence, the process $\{u_s^t:=|\tilde{I}_s^t|-\left(1-\frac{b}{n}\right)|\tilde{I}_{s+1}^t|\mid t>s\geq t-T_1\}$ is a martingale with respect to $\tilde{\mathcal{F}}$ and satisfies $\mathbb{E}_s\left[u_s^t\,\middle|\,\tilde{\mathcal{F}}_{s+1}^t\right]=0$. In addition, let $A=\{\underbrace{1,\cdots,1}_{|\tilde{I}_{s+1}^t|},\underbrace{0,\cdots,0}_{n-|\tilde{I}_{s+1}^t|}\}$ and $\tilde{A}=(\tilde{a}_1,\cdots,\tilde{a}_b)$ be a random sample without replacement from A, Then, u_s^t conditioned on $\tilde{\mathcal{F}}_{s+1}^t$ follows the same distribution as that of $\sum_{l=1}^b\tilde{a}_i-\mathbb{E}\left[\sum_{l=1}^b\tilde{a}_i\right]$. This means that, using Proposition 3, we have $\|u_s^t\|^2\leq C_1 b$ with high probability. Finally, we apply Proposition 4 to bound $|\tilde{I}_s^t|=\sum_{\tau=t}^s\left(1-\frac{b}{n}\right)^{(\tau-s)}u_\tau^t+n\left(1-\frac{b}{n}\right)^{(t-s+1)}$, which yields that $\left(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|]\right)^2\leq C_1\sum_{\tau=t}^s\left(1-\frac{b}{n}\right)^{(\tau-s)}b\leq C_1 b\cdot\frac{n}{b}=C_1 n$ with high probability. Therefore,

$$\begin{aligned}
&\frac{T_1}{n^2}\sum_{s=\max\{1,t-T_1\}}^{t}\left(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|]\right)^2\left\|\frac{1}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2\\
&\leq\frac{C_1\zeta^2}{b^2}\sum_{s=\max\{1,t-T_1\}}^{t}\|x^s-x^{s-1}\|^2.
\end{aligned}\tag{16}$$

Combining (15) and (16) yields $\text{(b)}\leq\frac{C_1\zeta^2}{b}\sum_{s=\max\{1,t-T_1\}}^{t}\|x^s-x^{s-1}\|^2$ with high probability.

Finally, we consider (c). As for Option I, the assertion directly follows from the definition of y_i^0. We prove the bound for

Option II. We have

$$\begin{aligned}(\mathrm{c}) &\leq 2\left\|\frac{1}{n}\frac{|\tilde{I}_1^t|}{b}\sum_{i\in I^1}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2+2\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2,\\ &\leq \frac{C_1|\tilde{I}_1^t|^2}{n^2b}\sigma_c^2+\frac{C_1|\tilde{I}_1^t|}{n^2}\sigma_c^2\leq\frac{C_1}{b}\sigma_c^2.\end{aligned}$$

Putting everything all together, we obtain the assertion with high probability. □

Now we prove Lemma 7. We need the following auxiliary lemma.

Lemma 8. *Under Assumptions 4 and 5-(b), we take* $T_1=\tilde{\Theta}(\frac{n}{b})$, $T_2=\Theta(\frac{\log\frac{\delta}{C_2\rho r_e}}{\eta\gamma})$, *and assume* $\max_{\tau_0\leq t\leq\tau+T_2}\{\|x^t-x^\tau\|,\|\tilde{x}^t-x^\tau\|\}<\frac{\delta}{C_2\rho}$. *Then, the following holds uniformly for all* $t\leq\tau+T_2$ *with high probability:*

$$\|h^t\|\leq\begin{cases}0 & (t<\tau)\\ \dfrac{C_1\zeta r_e}{\sqrt{b}} & (t=\tau)\\ \dfrac{C_1\zeta r_e}{\sqrt{b}}+\dfrac{C_1\zeta}{\sqrt{b}}\sqrt{\displaystyle\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\|w^s-w^{s-1}\|^2}+\dfrac{C_1\delta}{C_2\sqrt{b}}\sqrt{\displaystyle\sum_{s=\max\{\tau,t-T_1\}}^{t}\|w^s\|^2} & (\textit{otherwise}).\end{cases}$$

Proof. As for the case $t<\tau$, the assertion directly follows from the definition of $\{\tilde{x}^t\}$. For the proof of the rest cases, we use notations as follows:

$$\begin{aligned}H&=\nabla^2 f(x^{\tau_0}),\\ H_i&=\nabla^2 f_i(x^{\tau_0}),\\ \mathrm{d}H^t&=\int_0^1(\nabla^2 f(\tilde{x}^t+\theta(x^t-\tilde{x}^t))-H)\mathrm{d}\theta,\\ \mathrm{d}H_i^t&=\int_0^1(\nabla^2 f_i(\tilde{x}^t+\theta(x^t-\tilde{x}^t))-H_i)\mathrm{d}\theta.\end{aligned}$$

Moreover, to simplify the notation, we denote

$$u_i^s:=(\nabla f_i(x^s)-\nabla f_i(\tilde{x}^s))-(\nabla f_i(x^{s-1})-\nabla f_i(\tilde{x}^{s-1}))-(\nabla f(x^s)-\nabla f(\tilde{x}^s))+(\nabla f(x^{s-1})-\nabla f(\tilde{x}^{s-1})).$$

We have that $\mathbb{E}_i[u_i^s]=0$, where the expectation is taken over the choice of i. Furthermore, for $s\geq\tau+1$, by using the Hessian-heterogenity (Assumption 5) and the Hessian Lipschitzness (Assumption 4), we have that

$$\begin{aligned}\|u_i^s\|&=\|(\nabla f_i(x^s)-\nabla f_i(\tilde{x}^s))-(\nabla f_i(x^{s-1})-\nabla f_i(\tilde{x}^{s-1}))-(\nabla f(x^s)-\nabla f(\tilde{x}^s))+(\nabla f(x^{s-1})-\nabla f(\tilde{x}^{s-1}))\|\\ &=\Big\|\int_0^1\nabla^2 f_i(\tilde{x}^s-\theta(x^s-\tilde{x}^s))(x^s-\tilde{x}^s)\mathrm{d}\theta-\int_0^1\nabla^2 f_i(\tilde{x}^{s-1}-\theta(x^{s-1}-\tilde{x}^{s-1}))(x^{s-1}-\tilde{x}^{s-1})\mathrm{d}\theta\\ &\quad-\int_0^1\nabla^2 f(\tilde{x}^s-\theta(x^s-\tilde{x}^s))(x^s-\tilde{x}^s)\mathrm{d}\theta+\int_0^1\nabla^2 f(\tilde{x}^{s-1}-\theta(x^{s-1}-\tilde{x}^{s-1}))(x^{s-1}-\tilde{x}^{s-1})\mathrm{d}\theta\Big\|\\ &=\|(H_i+\mathrm{d}H_i^s)w^s-(H_i+\mathrm{d}H_i^{s-1})w^{s-1}-(H+\mathrm{d}H^s)w^s+(H+\mathrm{d}H^{s-1})w^{s-1}\|\\ &\leq\|H_i-H\|\|w^s-w^{s-1}\|+(\|\mathrm{d}H_i^s\|+\|\mathrm{d}H^s\|)\|w^s\|+(\|\mathrm{d}H_i^{s-1}\|+\|\mathrm{d}H^{s-1}\|)\|w^{s-1}\|\\ &\leq\zeta\|w^s-w^{s-1}\|+2\rho\max_{0\leq\theta\leq1}\{\|\tilde{x}^s-\theta(x^s-\tilde{x}^s)-x^\tau\|\}\|w^s\|+2\rho\max_{0\leq\theta\leq1}\{\|\tilde{x}^{s-1}-\theta(x^{s-1}-\tilde{x}^{s-1})-x^\tau\|\}\|w^{s-1}\|\\ &=\zeta\|w^s-w^{s-1}\|+2\rho\max\{\|x^s-x^\tau\|,\|\tilde{x}^s-x^\tau\|\}\|w^s\|+2\rho\max\{\|x^{s-1}-x^\tau\|,\|\tilde{x}^{s-1}-x^\tau\|\}\|w^{s-1}\|\\ &<\zeta\|w^s-w^{s-1}\|+\frac{2\delta}{C_2}\|w^s\|+\frac{2\delta}{C_2}\|w^{s-1}\|,\end{aligned}\tag{17}$$

where we use $\max_{\tau_0\le t\le \tau+T_2}\{\|x^t-x^\tau\|,\|\tilde{x}^t-x^\tau\|\}<\frac{\delta}{C_2\rho}$ for the last inequality. For $s=\tau$, by Assumption 5, we have $\|u_i^\tau\|=\|(\nabla f_i(x^\tau)-\nabla f_i(\tilde{x}^\tau))-(\nabla f(x^\tau)-\nabla f(\tilde{x}^\tau))\|\le 2\zeta\|x^\tau-\tilde{x}^\tau\|=2\zeta r_e$.

Recall the discussion in Lemma 1, we have

$$
\begin{aligned}
h^t &= g^t-\nabla f(x^t)-\tilde{g}^t+\nabla f(\tilde{x}^t)\\
&= \frac{1}{n}\sum_{s=\max\{\tau,t-T_1\}}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))-\sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))\right)\\
&\quad-\frac{1}{n}\sum_{s=\max\{\tau,t-T_1\}}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}(\nabla f_i(\tilde{x}^s)-\nabla f_i(\tilde{x}^{s-1}))-\sum_{i\in\tilde{I}_s^t}(\nabla f_i(\tilde{x}^s)-\nabla f_i(\tilde{x}^{s-1}))\right)\\
&= \frac{1}{n}\sum_{s=\max\{\tau,t-T_1\}}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}u_i^s-\sum_{i\in\tilde{I}_s^t}u_i^s\right)\\
&= \begin{cases}\frac{1}{n}\left(\frac{|\tilde{I}_\tau^\tau|}{b}\sum_{i\in I^\tau}u_i^\tau-\sum_{i\in\tilde{I}_\tau^\tau}u_i^\tau\right) & (t=\tau)\\ \frac{1}{n}\left(\frac{|\tilde{I}_\tau^t|}{b}\sum_{i\in I^\tau}u_i^\tau-\sum_{i\in\tilde{I}_\tau^t}u_i^\tau\right)+\frac{1}{n}\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}u_i^s\right)-\frac{1}{n}\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\sum_{i\in\tilde{I}_s^t}u_i^s & (t\ge\tau+1).\end{cases}
\end{aligned}
$$

As for the first term in both cases, we have

$$
\left\|\frac{1}{n}\left(\frac{|\tilde{I}_\tau^t|}{b}\sum_{i\in I^\tau}u_i^\tau-\sum_{i\in\tilde{I}_\tau^t}u_i^\tau\right)\right\|\le\left\|\frac{1}{n}\frac{|\tilde{I}_\tau^t|}{b}\sum_{i\in I^\tau}u_i^\tau\right\|+\left\|\frac{1}{n}\sum_{i\in\tilde{I}_\tau^t}u_i^\tau\right\|\le\frac{|\tilde{I}_\tau^t|}{n}\frac{2C_1\zeta r_e}{\sqrt{b}}\le\frac{2C_1\zeta r_e}{\sqrt{b}},\tag{18}
$$

by using Proposition 3 and $\|u_i^\tau\|\le 2\zeta r_e$, with high probability.

For the second term in the case $t\ge\tau+1$, we follow how we bounded (b) in the proof of Lemma 1 (High probability bound). We just replace $\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)-\nabla f(x^{s-1}))$ by u_i^s and use (17) to obtain that

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}u_i^s\right\| &\le \frac{2C_1}{\sqrt{b}}\sqrt{\sum_{s=\max\{1,t-T_1\}}^{t}\left(\zeta\|w^s-w^{s-1}\|+\frac{2\delta}{C_2}\|w^s\|+\frac{2\delta}{C_2}\|w^{s-1}\|\right)^2}\\
&\le \frac{2C_1}{\sqrt{b}}\sqrt{\sum_{s=\max\{1,t-T_1\}}^{t}\left(\zeta\|w^s-w^{s-1}\|+\frac{2\delta}{C_2}\|w^s\|+\frac{2\delta}{C_2}\|w^{s-1}\|\right)^2}.
\end{aligned}\tag{19}
$$

with high probability.

Finally, we bound the last term in the case $t\ge\tau+1$. By using Proposition 3, we obtain

$$
\begin{aligned}
\left\|\frac{1}{n}\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\sum_{i\in\tilde{I}_s^t}u_i^s\right\| &\le \frac{\sqrt{T_1+1}}{n}\sqrt{\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\left\|\sum_{i\in\tilde{I}_s^t}u_i^s\right\|^2}\\
&\le \frac{C_1}{\sqrt{nb}}\sqrt{\sum_{s=\max\{\tau+1,t-T_1\}}^{t}C_1^2b\left(\zeta\|w^s-w^{s-1}\|+\frac{2\delta}{C_2}\|w^s\|+\frac{2\delta}{C_2}\|w^{s-1}\|\right)^2}\\
&\le \frac{C_1}{\sqrt{b}}\sqrt{\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\left(\zeta\|w^s-w^{s-1}\|+\frac{2\delta}{C_2}\|w^s\|+\frac{2\delta}{C_2}\|w^{s-1}\|\right)^2}
\end{aligned}\tag{20}
$$

with high probability.

Combining (18), (19), and (20), we have

$$\begin{aligned}\|h^t\| &\le \frac{2C_1\zeta r_e}{\sqrt{b}} + \frac{2C_1+C_1}{\sqrt{b}}\sqrt{\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\left(\zeta\|w^s-w^{s-1}\| + \frac{2\delta}{C_2}\|w^s\| + \frac{2\delta}{C_2}\|w^{s-1}\|\right)^2} \\ &\le \frac{C_1\zeta r_e}{\sqrt{b}} + \frac{C_1\zeta}{\sqrt{b}}\sqrt{\sum_{s=\max\{\tau+1,t-T_1\}}^{t}\|w^s-w^{s-1}\|^2} + \frac{C_1\delta}{C_2\sqrt{b}}\sqrt{\sum_{s=\max\{\tau,t-T_1\}}^{t}\|w^s\|^2}\end{aligned}$$

with high probability for all $t > \tau$. For $t = \tau$, (18) directly implies the desired bound. □

Now, we are ready to prove Lemma 7.

Proof of Lemma 7. We assume the contrary, i.e., $\max_{\tau_0\le t\le\tau+T_2}\{\|x^t-x^\tau\|, \|\tilde{x}^t-x^\tau\|\} < \frac{\delta}{C_2\rho}$, and show the following by induction: for $\tau \le t \le \tau+T_2$,

$$\begin{aligned}&\text{(a)}\quad \frac{1}{2}(1+\eta\gamma)^{t-\tau}r_e \le \|w^t\| \le 2(1+\eta\gamma)^{t-\tau}r_e \\ &\text{(b)}\quad \|w^t-w^{t-1}\| \le \begin{cases} r_e & (\text{for } t=\tau) \\ 3\eta\gamma(1+\eta\gamma)^{t-\tau}r_e & (\text{for } t\ge\tau+1)\end{cases} \\ &\text{(c)}\quad \|h^t\| \le \frac{C_1\gamma}{C_2}(1+\eta\gamma)^{t-\tau}r_e.\end{aligned}$$

Then, (a) yields contradiction by taking $t-\tau = T_2 = \Theta\left(\frac{\log\frac{\delta}{C_2\rho r_e}}{\eta\gamma}\right)$ since it holds that

$$\max_{\tau_0\le t<\tau+T_2}\{\|x^t-x^\tau\|, \|\tilde{x}^t-x^\tau\|\} \ge \frac{1}{2}\|x^t-\tilde{x}^t\| = \frac{1}{2}\|w^t\| \ge \frac{\delta}{C_2\rho}.$$

It is easy to check (a) and (b) for $t=\tau$. As for (c), by taking $b \ge \frac{\zeta^2C_2^2}{\delta^2}$, $\|h^\tau\| \le \frac{C_1}{C_2}\delta r_e \le \frac{C_1}{C_2}\gamma r_e$ holds with high probability by Lemma 8.

Now, we derive that (a), (b), and (c) are true for $t+1$ if they are true for $t=\tau,\tau+1,\cdots,t$. For $t\ge\tau+1$, we can decompose w^t as

$$\begin{aligned}w^t &= w^{t-1} - \eta\left(g^{t-1}-\tilde{g}^{t-1}\right) \\ &= w^{t-1} - \eta\left(\nabla f(x^{t-1}) - \nabla f(\tilde{x}^{t-1}) + g^{t-1} - \nabla f(x^{t-1}) - \tilde{g}^{t-1} + \nabla f(\tilde{x}^{t-1})\right) \\ &= w^{t-1} - \eta\left(\int_0^1 \nabla^2 f(\tilde{x}^{t-1}+\theta(x^{t-1}-\tilde{x}^{t-1}))(x^{t-1}-\tilde{x}^{t-1})\mathrm{d}\theta + g^{t-1} - \nabla f(x^{t-1}) - \tilde{g}^{t-1} + \nabla f(\tilde{x}^{t-1})\right) \\ &= w^{t-1} - \eta\left((\mathrm{d}H^{t-1}+H)w^{t-1} + g^{t-1} - \nabla f(x^{t-1}) - \tilde{g}^{t-1} + \nabla f(\tilde{x}^{t-1})\right) \\ &= (I-\eta H)w^{t-1} - \eta(\mathrm{d}H^{t-1}w^{t-1} + h^{t-1}) \\ &= (I-\eta H)^{t-\tau}w^\tau - \eta\sum_{s=\tau}^{t-1}(I-\eta H)^{t-1-s}(\mathrm{d}H^s w^s + h^s) \\ &= (1+\eta\gamma)^{t-\tau}r_e\mathsf{e} - \eta\sum_{s=\tau}^{t-1}(I-\eta H)^{t-1-s}(\mathrm{d}H^s w^s + h^s). \qquad (21)\end{aligned}$$

According to this decomposition, we verify (a), (b), and (c).

Verifying (a) The first term of (21) satisfies

$$\|(1+\eta\gamma)^{t+1-\tau} r_e \mathsf{e}\| = (1+\eta\gamma)^{t+1-\tau} r_e.$$

Thus, it suffices to bound the norm of $\eta \sum_{s=\tau}^{t-1} (I-\eta H)^{t-1-s}(\mathrm{d}H^s w^s + h^s)$ by $\frac{1}{2}(1+\eta\gamma)^{t-\tau} r_e$. We have

$$\left\| \eta \sum_{s=\tau}^{t} (I-\eta H)^{t-s} \mathrm{d}H_s w_s \right\| \le \eta \sum_{s=\tau}^{t} \|I-\eta H\|^{t-s} \, \|\mathrm{d}H^s\| \, \|w^s\| \tag{22}$$

$$\le \eta (1+\eta\gamma)^{t-\tau} r_e \sum_{s=\tau}^{t} \|\mathrm{d}H^s\| \tag{23}$$

$$\le \eta (1+\eta\gamma)^{t-\tau} r_e T_2 \frac{\delta}{C_2} \tag{24}$$

$$\le \frac{1}{4}(1+\eta\gamma)^{t-\tau} r_e. \tag{25}$$

For (22), we used the facts that the maximum eigenvalue of ηH is at most $\eta L \le 1$ when $\eta \le \frac{1}{L}$ and that the minimum eigenvalue is $-\eta\gamma$, which imply $\|I-\eta H\| \le 1+\eta\gamma$. (23) follows from the assumptions on $\|w_s\|$. For (24), we used $t \le \tau + T_2$ and

$$\begin{aligned}
\|\mathrm{d}H^s\| &= \left\| \int_0^1 (\nabla^2 f(\tilde{x}^s + \theta(x^s - \tilde{x}^s)) - H) \mathrm{d}\theta \right\| \\
&\le \max_{0\le\theta\le 1} \rho \|\tilde{x}^s + \theta(x^s - \tilde{x}^s) - x^{\tau_0}\| \\
&= \max_{0\le\theta\le 1} \rho \max\{\|x^s - x^{\tau_0}\|, \|\tilde{x}^s - x^{\tau_0}\|\} < \rho \frac{\delta}{C_2 \rho} = \frac{\delta}{C_2},
\end{aligned}$$

where the first inequality follows from the hessian Lipschitzness (Assumption 4). The final inequality (25) holds when we take C_2 as $C_2 \ge 4\delta\eta T_2 = \tilde{\Theta}(1)$.

In addition, we have

$$\left\| \eta \sum_{s=\tau}^{t} (I-\eta H)^{t-s} h^s \right\| \le \eta \sum_{s=\tau}^{t} \|I-\eta H\|^{t-s} \, \|h^s\|$$

$$\le \eta \sum_{s=\tau}^{t-1} (1+\eta\gamma)^{t-s} \frac{3C_1\gamma}{C_2} (1+\eta\gamma)^{s-\tau} r_e \tag{26}$$

$$= \eta T_2 \frac{C_1\gamma}{C_2} (1+\eta\gamma)^{t-\tau} r_e$$

$$\le \frac{1}{4}(1+\eta\gamma)^{t-\tau} r_e. \tag{27}$$

Note that (26) can be checked by the same argument as (22) and the inductive hypothesis. (27) holds when we take $C_2 \ge 4\eta\gamma T_2 C_1 = \tilde{\Theta}(1)$.

Combining (25) and (27), we can bound the second term of (21) as desired, which concludes (a) holds for $t \ge \tau + 1$.

Verifying (b) For $t \ge \tau + 1$, we have

$$\begin{aligned}
&w_{t+1} - w_t \\
&= (1+\eta\gamma)^{t-\tau+1} r_e \mathsf{e} - \eta \sum_{s=\tau}^{t} (I-\eta H)^{t-s} (\mathrm{d}H^s w^s + h^s) - \left((1+\eta\gamma)^{t-\tau} r_e \mathsf{e} - \eta \sum_{s=\tau}^{t-1} (I-\eta H)^{t-1-s} (\mathrm{d}H^s w^s + h^s) \right) \\
&= \eta\gamma (1+\eta\gamma)^{t-\tau} r_e \mathsf{e} - \eta \sum_{s=\tau}^{t-1} \eta H (I-\eta H)^{t-1-s} (\mathrm{d}H^s w^s + h^s) - \eta (\mathrm{d}H^t w^t + h^t).
\end{aligned}$$

As for the first term, we can bound its norm as

$$\|\eta\gamma(1+\eta\gamma)^{t-\tau}r_e\mathsf{e}\| \le \eta\gamma(1+\eta\gamma)^{t-\tau}r_e.$$

The norm of the second term can be bounded by using (a) and (b) for $\tau+1,\cdots,t-1$ and Lemma 2 as follows:

$$\begin{aligned}
&\left\|\eta\sum_{s=\tau}^{t-1}\eta H(I-\eta H)^{t-1-s}(\mathrm{d}H^s w^s + h^s)\right\| \\
&\le \sum_{s=\tau}^{t-1}\eta\left\|\eta H(I-\eta H)^{t-1-s}\right\|\left(\|\mathrm{d}H^s\|\|w^s\| + \|h^s\|\right) \\
&\le \sum_{s=\tau}^{t-1}\eta\left\|\eta H(I-\eta H)^{t-1-s}\right\|\left(\frac{\delta}{C_2}(1+\eta\gamma)^{s-\tau}r_e + \frac{C_1\gamma}{C_2}(1+\eta\gamma)^{s-\tau}r_e\right) \\
&\le \sum_{s=\tau}^{t-1}\eta\left\|\eta H(I-\eta H)^{t-1-s}\right\|\left(\frac{\delta}{C_2}+\frac{C_1\gamma}{C_2}\right)(1+\eta\gamma)^{s-\tau}r_e \\
&\le \sum_{s=\tau}^{t-1}\eta\left(\eta\gamma(1+\eta\gamma)^{t-1-s}+\frac{1}{t-s}\right)\left(\frac{\delta}{C_2}+\frac{C_1\gamma}{C_2}\right)(1+\eta\gamma)^{s-\tau}r_e \\
&\le \eta\left(\eta\gamma T_2+\log T_2\right)\left(\frac{\delta}{C_2}+\frac{C_1\gamma}{C_2}\right)(1+\eta\gamma)^{t-\tau}r_e.
\end{aligned}$$

Since $T_2 = \tilde{\Theta}\left(\frac{1}{\eta\gamma}\right)$ and $\gamma \ge \delta$, setting $C_2 = \tilde{\Theta}(1)$ and $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$ with sufficiently large hidden constants yields $(\eta\gamma T_2+\log T_2)\left(\frac{\delta}{C_2}+\frac{C_1\gamma}{C_2}\right) \le \gamma$. Thus, the second term is bounded by $\eta\gamma(1+\eta\gamma)^{t-\tau}r_e$.

Finally, we consider the third term. We have $\|\mathrm{d}H^t w^t\| \le \frac{\delta}{C_2}r_e(1+\eta\gamma)^{t-\tau}r_e$ and $\|h^t\| \le \frac{C_1\gamma}{C_2}(1+\eta\gamma)^{t-\tau}r_e$ by the inductive hypothesis. Thus, taking C_2 sufficiently large, the third term is bounded by $\eta\gamma(1+\eta\gamma)^{t-\tau}r_e$.

Combining these bounds, we get (b) for $t+1$.

Verifying (c) By using Lemma 8 and the inductive hypothesis, we have

$$\begin{aligned}
\|h^{t+1}\| &\le \frac{C_1\zeta r_e}{\sqrt{b}} + \frac{C_1\zeta}{\sqrt{b}}\sqrt{\sum_{s=\max\{\tau,t-T_1\}}^{t}\|w^s-w^{s-1}\|^2} + \frac{C_1\delta}{C_2\sqrt{b}}\sqrt{\sum_{s=\max\{\tau,t-T_1\}}^{t}\|w^s\|^2} \\
&\le \frac{C_1\zeta}{\sqrt{b}}r_e + \frac{C_1\zeta\sqrt{n}\eta\gamma}{b}(1+\eta\gamma)^{t-\tau}r_e + \frac{C_1\sqrt{n}\delta}{C_2 b}(1+\eta\gamma)^{t-\tau}r_e \\
&\le \left(\frac{C_1\zeta}{\sqrt{b}} + \frac{C_1\zeta\sqrt{n}\eta\gamma}{b} + \frac{C_1\delta\sqrt{n}}{C_2 b}\right)(1+\eta\gamma)^{t-\tau}r_e
\end{aligned}$$

with high probability for all t. Taking $b \ge \Theta(\frac{C_2^2\zeta^2}{\delta^2}\wedge C_2^2\sqrt{n})$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $C_2 = O(C_1) = \tilde{O}(1)$ gives $\frac{C_1\zeta}{\sqrt{b}} + \frac{C_1\zeta\sqrt{n}\eta\gamma}{b} + \frac{C_1\delta\sqrt{n}}{C_2 b} \le \frac{C_1}{C_2}\gamma$. Thus, we obtain that (c) holds for $t+1$.

Thus, we complete the induction step, and hence, the assertion follows. □

From Lemma 7, we can ensure that SILVER escapes saddle points with high probability.

Lemma 9. *Let Assumptions 1-(b), 4, 5-(b) hold. Let $\{x^t\}$ be a sequence generated by SILVER and $\tau_0(\ge 0)$ be a step where $-\gamma := \lambda_{\min}(\nabla^2 f(x^{\tau_0})) \le -\delta$ holds. We denote the eigenvector with the eigenvalue $\lambda_{\min}(\nabla^2 f(x^{\tau_0}))$ by e. We take $b = \tilde{\Omega}(\sqrt{n}+\frac{\zeta^2}{\delta^2})$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_2 = \tilde{O}\left(\frac{L}{\delta}\right)$. Then, for arbitrary $\tau > \tau_0$, it holds that*

$$\mathbb{P}\left[\max_{\tau_0\le t\le\tau+T_2}\|x^t - x^{\tau_0}\| \ge \frac{\delta}{C_2\rho} \mid I^0,\cdots,I^\tau,\xi^1,\cdots,\xi^\tau\right] \ge 1-2\nu,$$

Proof. Let A be a subset of $B(0,r)$ such that each $a \in A$ satisfies

$$\mathbb{P}\left[\max_{\tau_0 \le t \le \tau+T_2} \|x^t - x^{\tau_0}\| > \frac{\delta}{C_2\rho} \mid I^0, \cdots, I^\tau, \xi^1, \cdots, \xi^\tau, \xi^{\tau+1} = a\right] \le 1-\nu.$$

Then, no two elements, $\xi_{\tau+1}$ and $\tilde{\xi}_{\tau+1}$ such that $\xi_{\tau+1} - \tilde{\xi}_{\tau+1} = r_e \mathsf{e}$ with $r_e \ge \frac{\nu r}{\sqrt{d}}$, can be elements of A at the same time since by Lemma 7, it holds that

$$\max_{\tau_0 \le t \le \tau+T_2} \{\|x^t - x^{\tau_0}\|, \|\tilde{x}^t - x^{\tau_0}\|\} \ge \frac{\delta}{C_2\rho}$$

with high probability. Let $V_d(r)$ be the volume of Euclidean ball with radius r in $\mathbb{R}^d$. Then, we have

$$\frac{\mathrm{Vol}(A)}{V_d(r)} \le \frac{r_e V_{d-1}(r)}{V_d(r)} = \frac{r_e \Gamma(\frac{d}{2}+1)}{\sqrt{\pi} r \Gamma(\frac{d}{2}+\frac{1}{2})} \le \frac{r_e}{\pi r}\left(\frac{d}{2}+1\right)^{\frac{1}{2}} \le \frac{r_e\sqrt{d}}{r} \le \nu.$$

This means that A occupies at least $1-\frac{1}{T}$ of the volumes of $B(0,r)$. From this fact and the definition of A, we have

$$\mathbb{P}\left[\max_{\tau_0 \le t \le \tau+T_2} \|x^t - x^{\tau_0}\| \ge \frac{\delta}{C_2\rho} \mid I^0, \cdots, I^\tau, \xi^1, \cdots, \xi^\tau\right] \ge 1-\nu-\nu = 1-2\nu,$$

which gives the conclusion. □

We are now ready to prove the main theorem of this subsection, which guarantees that the algorithm finds (ε, δ)-second-order stationary point with high probability.

Proof of Theorem 2. Since $T_2 = \frac{C_3 \log \frac{\delta}{C_2\rho r_e}}{\eta\gamma}$ depends on x^{τ_0} (since γ depends on $\nabla^2 f(x^{\tau_0})$), we take $T_2 = \frac{C_3 \log \frac{\delta}{C_2\rho r_e}}{\eta\delta}$ instead from now. Note that this replacement does not affect whether Lemma 9 holds.

We devide $\{t = 0, 1, \cdots, T-1\}$ into $\lceil \frac{T}{2T_2} \rceil$ *phases*: $P^s = \{2sT_2 \le t < 2(s+1)T_2\}$ $\left(s = 0, \cdots, \lceil \frac{T}{2T_2} \rceil - 1\right)$. For each phase, we define a^s as a random variable defined by

$$a^s = \begin{cases} 1 & \left(\text{if } \sum_{t\in P^s} \mathbb{1}[\|\nabla f(x^t)\| > \varepsilon] > T_2\right), \\ 2 & \left(\text{if there exists } t \text{ such that } 2sT_2 \le t < (2s+1)T_2, \|\nabla f(x^t)\| \le \varepsilon \text{ and } \lambda_{\min}(\nabla^2 f(x^t)) \le -\delta\right), \\ 3 & \left(\text{if there exists } t \text{ such that } 2sT_2 \le t < (2s+1)T_2, \|\nabla f(x^t)\| \le \varepsilon \text{ and } \lambda_{\min}(\nabla^2 f(x^t)) > -\delta\right). \end{cases}$$

Note that $\mathbb{P}[a^s \in \{1,2,3\}] = 1$ for each τ. This is because if there does not exist t between $2sT_2 \le t < (2s+1)T_2$ such that $\|\nabla f(x^t)\| \le \varepsilon$ (i.e., neither $a^\tau = 2$ nor 3), then we have $\sum_{t\in P^s} \mathbb{1}[\|\nabla f(x^t)\| > \varepsilon] \ge \sum_{t=2sT_2}^{(2s+1)T_2-1} \mathbb{1}[\|\nabla f(x^t)\| > \varepsilon] = T_2$, meaning $a^s = 1$. We denote $N_1 = \sum_{s=0}^{\frac{T}{2T_2}-1} \mathbb{1}[a^s = 1]$, $N_2 = \sum_{s=0}^{\frac{T}{2T_2}-1} \mathbb{1}[a^s = 2]$, and $N_3 = \sum_{s=0}^{\frac{T}{2T_2}-1} \mathbb{1}[a^s = 3]$.

According to Lemma 9, with probability $1-2\nu$ (here ν can be arbitrary small), it holds that if $a^s = 2$ then that phase successes escaping saddle points. Specifically, by taking $\tau = (2s+1)T_2$, we have

$$\max_{\tau \le t' < \tau+T_2} \|x^\tau - x^{t'}\| > \frac{\delta}{C_2\rho} \tag{28}$$

holds. (28) further leads to

$$T_2 \sum_{t=2\tau T_2}^{2(\tau+1)T_2-1} \|x^{t+1} - x^t\|^2 > \left(\frac{\delta}{C_2\rho}\right)^2 \left(\iff \sum_{t=2\tau T_2}^{2(\tau+1)T_2-1} \|x^{t+1} - x^t\|^2 > \frac{\delta^2}{T_2 C_2^2 \rho^2}\right). \tag{29}$$

On the other hand, by combining Lemma 1 (High probability bound) and Lemma 6, we have

$$\begin{aligned} &\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 \\ &\le \frac{2}{\eta}\left[\left(f(x^0) - f(x^t)\right) - \left(\frac{1}{2\eta} - \frac{L}{2} - \frac{C_1\eta\zeta^2 n}{b^2}\right)\sum_{t=1}^{T} \|x^t - x^{t-1}\|^2 + \frac{\eta C_1 T_1 \sigma_c^2 \mathbb{1}[\text{Option II}]}{b}\right] + \frac{2Tr^2}{\eta^2} \end{aligned}$$

with high probability. By taking $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, applying $b \geq \zeta\sqrt{n}$ and $f(x^0) - f(x^t) \leq \Delta$, and rearranging terms, we obtain

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T}\|x^t - x^{t-1}\|^2 \leq \frac{2\Delta}{\eta} + \frac{2C_1T_1\sigma_c^2\mathbb{1}[\text{Option II}]}{b} + \frac{2Tr^2}{\eta^2}$$

From the definition of $a^\tau = 1$ and (29), that the left-hand side is bounded as

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T}\|x^t - x^{t-1}\|^2 \geq N_1T_2\varepsilon^2 + \frac{\delta^2 N_2}{2\eta^2 T_2 C_2^2\rho^2}.$$

Thus, it holds that

$$\max\left\{N_1T_2\varepsilon^2, N_2T_2 \cdot \frac{\delta^2}{2\eta^2T_2^2C_2^2\rho^2}\right\} \leq \frac{2\Delta}{\eta} + \frac{2C_1T_1\sigma_c^2\mathbb{1}[\text{Option II}]}{b} + \frac{2Tr^2}{\eta^2} \tag{30}$$

By the parameter settings, we have $\frac{2\eta^2T_2^2C_2^2\rho^2}{\delta^2} = \tilde{O}\left(\frac{\rho^2}{\delta^4}\right)$. From this, $(N_1 + N_2)T_2 \leq \tilde{O}\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right) \times$ (the right-hand side of (30)). Taking $T \geq \tilde{\Omega}\left(\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right)\left(\frac{\Delta}{\eta} + \frac{n\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2}\right)\right) = \tilde{\Omega}\left(\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right)\left(L\Delta + \frac{\mathbb{1}[\text{Option II}]n\sigma_c^2}{b^2}\right)\right)$ and $r = \tilde{O}(\eta(\varepsilon \vee \delta^2/\rho)) = \tilde{O}((\varepsilon \vee \delta^2/\rho)/L)$, there exists s such that $a^s = 3$, which concludes the proof. □

Remark 2. *Although our main interest in this paper is to develop a simple algorithm with convergence to second-order stationary points, it can be easily shown that adaptive selection of minibatch size can reduce the gradient complexity. In Lemma 7, if we carefully check the proof, we can see that the condition $b = \tilde{\Theta}(\sqrt{n} + \frac{\zeta^2}{\delta^2})$ is needed only for the step τ. On the other hand, for all $\tau_0 \leq t \leq \tau + T_2$ except for $t = \tau$, $b = \tilde{\Theta}(\sqrt{n})$ is sufficient. Because we consider $\tau = (2\tau + 1)T_2$ $\left(\tau = 0, \cdots, \frac{T}{2T_2} - 1\right)$, if we take $b = \tilde{\Theta}(\sqrt{n} + \frac{\zeta^2}{\delta^2})$ only at $t = (2\tau + 1)T_2$ $\left(\tau = 0, \cdots, \frac{T}{2T_2} - 1\right)$ and $b = \tilde{\Theta}(\sqrt{n})$ at the other steps, the above argument still holds with a slight modification. Then, the gradient complexity is reduced to*

$$\tilde{O}\left(n + L\Delta\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\rho^2\sqrt{n}}{\delta^4} + \frac{\zeta^2}{L\varepsilon^2\delta} + \frac{\zeta^2\rho^2}{L\delta^5}\right)\right) \quad \text{(Option I)},$$

$$.\tilde{O}\left(\left(L\Delta + \sigma_c^2\right)\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\rho^2\sqrt{n}}{\delta^4} + \frac{\zeta^2}{L\varepsilon^2\delta} + \frac{\zeta^2\rho^2}{L\delta^5}\right)\right) \quad \text{(Option II)}.$$

In the classical setting $\delta = O(\sqrt{\rho\varepsilon})$, This bound is better than SSRGD *(Li, 2019) when $n \geq \varepsilon^{-1/2}$, than* Stabilized *when $n \geq \varepsilon^{-1/3}$, and than* SPIDER-SFO$^+$(+Neon2) *(Fang et al., 2018; Allen-Zhu and Li, 2018) no matter what n and δ are. We also note that, by carefully looking the proof of SSRGD (Li, 2019), we find that they implicitly limits their analysis to the case of $\frac{L^2}{\delta^2} \lesssim n$.*

D.3. Exponential Convergence under PL Condition (proof of Theorem 3)

In this subsection, we prove that SILVER automatically switches to the exponential convergence when the PL condition (Assumption 6) holds.

Proof of Theorem 3. According to the descent lemma (Lemma 6) and PL condition (Assumption 6), we have that

$$\begin{aligned} f(x^t) &\leq f(x^{t-1}) + \eta\|\nabla f(x^{t-1}) - g^{t-1}\|^2 - \frac{\eta}{2}\|\nabla f(x^{t-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 + \frac{r^2}{\eta} \\ &\leq f(x^{t-1}) + \eta\|\nabla f(x^{t-1}) - g^{t-1}\|^2 - \frac{\eta\mu}{2}(f(x^{t-1}) - f(x^*)) - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 + \frac{r^2}{\eta}. \end{aligned}$$

Rearranging the terms yields

$$f(x^t) - f^* \leq (1 - \frac{\eta\mu}{2})(f(x^{t-1}) - f^*) + \eta\|\nabla f(x^{t-1}) - g^{t-1}\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 + \frac{r^2}{\eta}. \tag{31}$$

By applying Lemma 1 to this, we obtain that

$$\mathbb{E}[f(x^t)-f^*] \le (1-\frac{\eta\mu}{2})\mathbb{E}[(f(x^{t-1})-f^*)] - \left(\frac{1}{2\eta}-\frac{L}{2}\right)\mathbb{E}[\|x^t-x^{t-1}\|^2] + \frac{30\eta\zeta^2}{b}\sum_{s=1}^{t-1}(1-\frac{b}{4n})^{t-s}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right] + \frac{9\eta\sigma_c^2\mathbb{1}[\text{Option II}]}{b}(1-\frac{b}{4n})^{t-1} + \frac{r^2}{\eta}. \tag{32}$$

Multiplying both sides of (32) by $(1-\alpha)^{T-t}$ with some $0<\alpha<1$ and summing up over all $t=1,2,\cdots,T$, we get

$$\begin{aligned}
&\mathbb{E}[f(x^T)-f^*] + \underline{\sum_{t=1}^{T-1}(1-\alpha)^{T-t}\mathbb{E}[f(x^t)-f^*]} \\
&\le (1-\frac{\eta\mu}{2})(1-\alpha)^{T-1}(f(x^0)-f^*) + \underline{\sum_{t=1}^{T-1}(1-\frac{\eta\mu}{2})(1-\alpha)^{T-t-1}\mathbb{E}[f(x^t)-f^*]} \\
&\quad - \sum_{s=1}^{T}(1-\alpha)^{T-s}\left(\frac{1}{2\eta}-\frac{L}{2}-\frac{30\eta\zeta^2}{b}\sum_{t=s+1}^{T}(1-\frac{b}{4n})^{t-s}(1-\alpha)^{s-t}\right)\mathbb{E}[\|x^s-x^{s-1}\|^2] \\
&\quad + \sum_{t=1}^{T}(1-\alpha)^{T-t}(1-\frac{b}{4n})^{t-1}\frac{9\eta\sigma_c^2\mathbb{1}[\text{Option II}]}{b} + \frac{r^2}{\eta}\sum_{t=1}^{T}(1-\alpha)^{T-t}.
\end{aligned} \tag{33}$$

We let $\eta = \frac{1}{2L}\wedge\frac{b}{4\zeta\sqrt{30n}}$ and $\alpha = \frac{\mu}{4L}\wedge\frac{\mu b}{8\zeta\sqrt{30n}}\wedge\frac{b}{8n} = \frac{\mu\eta}{2}\wedge\frac{b}{8n}$. The choice of η ensures that (33) ≤ 0. All of $(1-\frac{\eta\mu}{2})$ can be replaced by $(1-\alpha)$, and especially the underlined parts cancel out each other. Also, $\sum_{t=1}^{T}(1-\alpha)^{T-t}(1-\frac{b}{4n})^{t-1} \le \sum_{t=1}^{T}(1-\alpha)^{T-t}(1-\frac{b}{8n})^{2(t-1)} \le 2\sum_{t=1}^{T}(1-\alpha)^{T}(1-\frac{b}{8n})^{t} \le \frac{16n}{b}(1-\alpha)^T$. Now, we obtain that

$$\mathbb{E}[f(x^T)-f^*] \le (1-\alpha)^T(f(x^0)-f^*) + (1-\alpha)^T\frac{144n\eta\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2} + \frac{r^2}{\eta\alpha}.$$

We take $T \ge \frac{1}{\alpha}\log\frac{3(\Delta+\frac{144n\eta\sigma_c^2\mathbb{1}[\text{Option II}]}{b^2})}{\varepsilon} = \Theta((\frac{L}{\mu}\vee\frac{\zeta\sqrt{n}}{\mu b}\vee\frac{n}{b})\log\frac{\Delta+\frac{n\sigma_c\mathbb{1}[\text{Option II}]}{bL}}{\varepsilon})$, so that the first and second terms are bounded by $\frac{\varepsilon}{3}$, respectively. Here Tb matches the desired gradient complexity. Also, by taking $r \le \sqrt{\frac{\eta\alpha}{3}} = \eta\sqrt{\frac{\varepsilon\mu}{6}}\wedge\sqrt{\frac{\eta\varepsilon b}{24n}}$, the last term is also bounded by $\frac{\varepsilon}{3}$. Therefore, the assertion follows. □

E. MISSING STATEMENTS AND PROOFS FOR FL-SILVER

This section applies SILVER to communication-efficient federated learning. The full version of the algorithm is provided as Algorithm 3. Theorem 4 is divided into three subsections, each of which corresponds to the first-order optimality, the second-order optimality, and exponential convergence under the PL condition.

E.1. First-order Optimality

We first consider the first-order optimality. We give the full statement of our theorem below.

Theorem 5. *Let Assumptions 1-(a), 2, 3-(a) (Only for Option II), 3-(b), and 5-(a) hold. Let $r \le \frac{\eta\varepsilon}{2}$ and $PKb \ge \frac{192\sigma^2}{\varepsilon^2}$. Set η as*

$$\eta = \Theta\left(\frac{1}{L}\vee\frac{p}{\zeta K\sqrt{P}}\wedge\frac{p\sqrt{b}}{L\sqrt{PK}}\wedge\frac{\sqrt{b}}{L\sqrt{K}}\wedge\frac{1}{\zeta K}\right).$$

Then, Algorithm 2 finds an ε-first-order stationary points $x^{t,k}$ for problem (2) in expectation, in

$$T = O\left(\left[\left[\frac{L}{K}\vee\zeta\left(\frac{\sqrt{P}}{p}\vee 1\right)\vee\frac{L}{\sqrt{bK}}\left(\frac{\sqrt{P}}{p}\vee 1\right)\right]\Delta + \mathbb{1}[\text{Option II}]\left(\frac{\sigma P}{p^2Kb}+\frac{\sigma_c^2 P}{p^2}\right)\right]\frac{1}{\varepsilon^2}\right)$$

Algorithm 3 FL-SILVER$(x^0, \eta, p, b, T, K, r)$ (full version)

Option I: $I^0 \leftarrow [P]$
Option II: Randomly select p clients I^0
for $i \in I^0$ in parallel **do**
  Randomly select minibatch J_i^0 with size Kb
  $y_i^0 \leftarrow \frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)$
Send y_i^0 from $i \in I^0$ to the server
if Option II: $y_i^0 \leftarrow \frac{1}{p}\sum_{i\in I^0} y_i^0$ $(i = 1, \cdots, P)$ // Not send y_i^0 to clients; Store y_i^0 in server until i sampled.
for $t = 1$ to T **do**
  Randomly sample one client i_t and send $\sum_{i=1}^P y_i^{t-1}$ and x^{t-1} from the server to i_t
  $x^{t,0} \leftarrow x^{t-1},\ z^{t,0} \leftarrow 0$
  for $k = 1$ to K **do**
    $x^{t,k} \leftarrow x^{t,k-1} - \eta(\frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k-1}) + \xi^{t,k}$ $(\xi^{t,k} \sim B(0, r))$
    Randomly select minibatch $J_{i_t}^{t,k}$ with size b
    $z^{t,k} \leftarrow z^{t,k-1} + \frac{1}{b}\sum_{j\in J_{i_t}^{t,k}}(\nabla f_{i_t,j}(x^{t,k}) - \nabla f_{i_t,j}(x^{t,k-1}))$
  Send $x^{t,K}$ from i_t to the server; $x^t \leftarrow x^{t,K}$
  Randomly select p clients I^t and send x^t from i_t to I^t
  for $i \in I^t$ in parallel **do**
    Randomly select minibatch J_i^t with size Kb
    $y_i^t \leftarrow \frac{1}{Kb}\sum_{j\in J_i^t}\nabla f_{i,j}(x^t)$
    $\Delta y_i^t \leftarrow \frac{1}{Kb}\sum_{j\in J_i^t}(\nabla f_{i,j}(x^t) - \nabla f_{i,j}(x^{t-1}))$
  Send $\{(y_i^t, \Delta y_i^t)\}_{i\in I^t}$ from I^t to the server
  $y_i^t \leftarrow y_i^{t-1} + \frac{1}{p}\sum_{i\in I^t}\Delta y_i^t$ for $i \notin I^t$
  // Practically, we update $\sum_{i=1}^P y_i^t$ in $O(pd)$ time in the server with efficient update of SILVER.

communication rounds and

$$\mathbb{1}[\text{Option I}]P + Tp = \mathbb{1}[\text{Option I}]P + O\left(\left[\left[\frac{Lp}{K}\vee\zeta\left(\sqrt{P}\vee p\right)\vee\frac{L}{\sqrt{bK}}\left(\sqrt{P}\vee p\right)\right]\Delta + \mathbb{1}[\text{Option II}]\left(\frac{\sigma P}{pKb} + \frac{\sigma_c^2 P}{p}\right)\right]\frac{1}{\varepsilon^2}\right)$$

communication complexity (total number of communicated gradients).

Remark 3. *By letting $b = K$ achieves the optimal communication rounds (and communication complexity) under the fixed local computational budget bK:*

$$T = O\left(\left[\left[\frac{L}{K}\vee\zeta\left(\frac{\sqrt{P}}{p}\vee 1\right)\right]\Delta + \mathbb{1}[\text{Option II}]\left(\frac{\sigma P}{p^2K^2} + \frac{\sigma_c^2 P}{p^2}\right)\right]\frac{1}{\varepsilon^2}\right).$$

One can see that our local minibatch size is a moderate choice compared to the literature; Murata and Suzuki (2021) uses $b \geq K$ at every round and local full batch gradient periodically, and Karimireddy et al. (2021) uses local full batch gradient every round. Also, we remark that $bKP \geq \frac{192\sigma^2}{\varepsilon^2}$ is satisfied when P is large even under the local budget bK is small.

From now, we prove Theorem 5. For convenience, define $\tilde{y}_i^0$ as

$$\tilde{y}_i^0 := \begin{cases}\nabla f_i(x^0) & \text{(Option I)},\\ \frac{1}{p}\sum_{i\in I^0}\nabla f_i(x^0) & \text{(Option II)}.\end{cases}$$

For each $t \geq 1$ and i, define $T(t, i)$ as the last step t is sampled:

$$T(t,i) := \begin{cases}\max\{s|\ 1 \leq s \leq t,\ i \in I^s\} & (\text{if } \{s|\ 1 \leq s \leq t,\ i \in I^s\} \neq \emptyset),\\ 0 & (\text{otherwise}).\end{cases}$$

In the following, a slight abuse of notation, we identify $(t,k)=(t',k')$ if $t+Kk=t'+Kk'$. We say $(t,k)\le(t',k')$ if $t+Kk<t'+Kk'$.

Similarly to Lemma 1, we bound the difference between $\frac{1}{P}\sum_{i=1}^P y_i^{t-1}+z^{t,k}$ and $\nabla f(x^{t,k})$.

Lemma 10. *Suppose that Assumptions 1-(a), 3-(a) (Only for Option II), 3-(b), and 5-(a) hold. Then, regarding Algorithm 2, we have*

$$\begin{aligned}
&\mathbb{E}[\|\frac{1}{P}\sum_{i=1}^P y_i^{t-1}+z^{t,k}-\nabla f(x^{t,k})\|^2]\\
&\le\left[\frac{180\zeta^2}{p}+\frac{12L^2}{pKb}\right]\sum_{s=1}^t(1-\frac{p}{4P})^{t-s+1}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right]+\left[\frac{6L^2}{b}+6\zeta^2K\right]\sum_{l=1}^k\mathbb{E}[\|x^{t,l}-x^{t,l-1}\|^2]\\
&\quad+\frac{54\sigma_c^2\mathbb{1}[\text{Option II}]}{p}(1-\frac{p}{P})^t+\frac{12\sigma^2}{PKb}+\frac{6(1-\frac{p}{P})^t\sigma^2\mathbb{1}[\text{Option II}]}{pKb}.
\end{aligned}$$

Proof. We decompose the error into several parts. First, we observe that

$$\frac{1}{P}\sum_{i=1}^P y_i^{t-1}+z^{t,k}-\nabla f(x^{t,k})=\frac{1}{P}\sum_{i=1}^P y_i^{t-1}-\nabla f(x^{t-1})+z^{t,k}-(\nabla f(x^{t,k})-\nabla f(x^{t,0})). \tag{34}$$

Similarly to the proof of Lemma 1 (formal), we can expand $\frac{1}{P}\sum_{i=1}^P y_i^t-\nabla f(x^t)$ as

$$\begin{aligned}
&\frac{1}{P}\sum_{i=1}^P y_i^t-\nabla f(x^t)\\
&=\underbrace{\frac{1}{P}\sum_{s=1}^t\left(\frac{|\tilde I_s^t|}{p}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))-\sum_{i\in\tilde I_s^t}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))\right)+\frac{1}{P}\sum_{i\in\tilde I_1^t}(\tilde y_i^0-\nabla f_i(x^0))}_{\text{(a)}}\\
&\quad+\underbrace{\sum_{s=1}^t\frac{|\tilde I_s^t|}{PpKb}\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s)-\nabla f_{i,j}(x^{s-1})-(\nabla f_i(x^s)-\nabla f_i(x^{s-1})))}_{\text{(b)}}\\
&\quad+\underbrace{\frac{1}{PKb}\sum_{i=1}^P\mathbb{1}[T(t,i)\ge1]\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)})-\nabla f_i(x^{T(t,i)}))}_{\text{(c)}}+\underbrace{\frac{1}{P}\sum_{i=1}^P\mathbb{1}[T(t,i)=0](y_i^0-\tilde y_i^0)}_{\text{(d)}}.
\end{aligned}\tag{35}$$

Also, we have

$$\begin{aligned}
z^{t,k}-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))&=\sum_{l=1}^k\frac{1}{b}\sum_{j\in J^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))\\
&=\underbrace{\sum_{l=1}^k\left(\frac{1}{b}\sum_{j\in J^{t,l}}((\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))-(\nabla f_{i_t}(x^{t,l})-\nabla f_{i_t}(x^{t,l-1})))\right)}_{\text{(e)}}\\
&\quad+\underbrace{\nabla f_{i_t}(x^{t,k})-\nabla f_{i_t}(x^{t,0})-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))}_{\text{(f)}}.
\end{aligned}\tag{36}$$

Therefore, by (34), (35), and (36), we have

$$\left\|\frac{1}{P}\sum_{i=1}^P y_i^{t-1}+z^{t,k}-\nabla f(x^{t,k})\right\|^2\le6\|\text{(a)}\|^2+6\|\text{(b)}\|^2+6\|\text{(c)}\|^2+6\|\text{(d)}\|^2+6\|\text{(e)}\|^2+6\|\text{(f)}\|^2.$$

For each term, we apply one of the following auxiliary lemmas, which directly yields the assertion. □

Lemma 11. *Under Assumptions 5-(a) and 3-(a) (only for Option II), the term* (a) *is bounded as follows:*

$$\mathbb{E}\left[\left\|\frac{1}{P}\sum_{s=1}^{t}\left(\frac{|\tilde{I}_s^t|}{p}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))-\sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))\right)+\frac{1}{P}\sum_{i\in\tilde{I}_1^t}(\tilde{y}_i^0-\nabla f_i(x^0))\right\|^2\right]$$
$$\leq \frac{30\zeta^2}{p}\sum_{s=1}^{t}(1-\frac{p}{4P})^{t-s+1}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right]+\frac{9\sigma_c^2\mathbb{1}[\text{Option II}]}{p}(1-\frac{p}{P})^t.$$

Lemma 12. *Under Assumptions 1-(a), the term* (b) *is bounded as follows:*

$$\mathbb{E}\left[\left\|\underbrace{\sum_{s=1}^{t}\frac{|\tilde{I}_s^t|}{PpKb}\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s)-\nabla f_{i,j}(x^{s-1})-(\nabla f_i(x^s)-\nabla f_i(x^{s-1})))}_{(*)}\right\|^2\right]$$
$$\leq \frac{2L^2(1-\frac{p}{P})^{t-s+1}}{pKb}\sum_{s=1}^{t}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right].$$

Lemma 13. *Under Assumption 3-(b), the term* (c) *is bounded as follows:*

$$\mathbb{E}\left[\left\|\frac{1}{PKb}\sum_{i=1}^{P}\mathbb{1}[T(t,i)\geq 1]\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)})-\nabla f_i(x^{T(t,i)}))\right\|^2\right]\leq\frac{\sigma^2}{PKb}.$$

Lemma 14. *Under Assumption 3-(b), the term* (d) *is bounded as follows:*

$$\mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0](y_i^0-\tilde{y}_i^0)\right\|^2\right]\leq\begin{cases}\dfrac{(1-\frac{p}{P})^t\sigma^2}{PKb} & \text{(Option I)},\\ \dfrac{(1-\frac{p}{P})^t\sigma^2}{pKb} & \text{(Option II)}.\end{cases}$$

Lemma 15. *Under Assumption 1-(a), the term* (e) *is bounded as follows:*

$$\mathbb{E}\left[\left\|\sum_{l=1}^{k}\left(\frac{1}{b}\sum_{j\in J^{t,l}}((\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))-(\nabla f_{i_t}(x^{t,l})-\nabla f_{i_t}(x^{t,l-1})))\right)\right\|^2\right]\leq\frac{L^2}{b}\sum_{l=1}^{k}\mathbb{E}[\|x^{t,l}-x^{t,l-1}\|^2].$$

Lemma 16. *Under Assumption 5-(a), the term* (f) *is bounded as follows:*

$$\mathbb{E}\left[\left\|\nabla f_{i_t}(x^{t,k})-\nabla f_{i_t}(x^{t,0})-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))\right\|^2\right]\leq\zeta^2K\sum_{l=1}^{k}\mathbb{E}[\|x^{t,l}-x^{t,l-1}\|^2].$$

Proof of Lemma 11. By replacing p and P by b and n, the term (a) is exactly the same as (3) in the proof of Lemma 1 (formal). Thus we simply follow the proof of Lemma 1 (formal) to obtain the assertion. □

Proof of Lemma 12. Consider the conditional expectation on $|\tilde{I}_s^t|$ for all s. each cross term of $(*)$ is still mean-zero on this

conditioning, and therefore

$$\begin{aligned}
\mathbb{E}[(\mathrm{b})] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{s=1}^{t}\frac{|\tilde{I}_s^t|}{PpKb}\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s)-\nabla f_{i,j}(x^{s-1})-(\nabla f_i(x^s)-\nabla f_i(x^{s-1})))\right\|^2\middle| |\tilde{I}_1^t|,\cdots,|\tilde{I}_t^t|\right]\right] \\
&\le \sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{|\tilde{I}_s^t|}{PpKb}\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s)-\nabla f_{i,j}(x^{s-1})-(\nabla f_i(x^s)-\nabla f_i(x^{s-1})))\right\|^2\middle| |\tilde{I}_1^t|,\cdots,|\tilde{I}_t^t|\right]\right] \\
&\le \sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left[\frac{|\tilde{I}_s^t|^2L^2}{PpK^2b}\|x^s-x^{s-1}\|^2\middle| |\tilde{I}_1^t|,\cdots,|\tilde{I}_t^t|\right]\right]. \\
&= \frac{L^2}{P^2pKb}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{E}\left[|\tilde{I}_s^t|^2|I^s,\cdots,I^t\right]\|x^s-x^{s-1}\|^2\right] \\
&\le \frac{L^2}{P^2pKb}\sum_{s=1}^{t}\mathbb{E}\left[[P(1-(1-\frac{p}{P})^{t-s+1})(1-\frac{p}{P})^{t-s+1}+P^2(1-\frac{p}{P})^{2(t-s+1)}]\|x^s-x^{s-1}\|^2\right] \\
&\le \frac{2L^2(1-\frac{p}{P})^{t-s+1}}{pKb}\sum_{s=1}^{t}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right],
\end{aligned}$$

which concludes the proof. □

Proof of Lemma 13. By conditioning (c) on $T(t,i)$, the cross terms of $\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)})-\nabla f_i(x^{T(t,i)}))$ for two different i are mean-zero. Thus,

$$\begin{aligned}
\mathbb{E}[(\mathrm{c})] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{PKb}\sum_{i=1}^{P}\mathbb{1}[T(t,i)\ge 1]\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)})-\nabla f_i(x^{T(t,i)}))\right\|^2\middle| T(t,i)\right]\right] \\
&= \frac{1}{P^2K^2b^2}\sum_{i=1}^{P}\mathbb{E}\left[\mathbb{E}\left[\left\|\mathbb{1}[T(t,i)\ge 1]\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)})-\nabla f_i(x^{T(t,i)}))\right\|^2\middle| T(t,i)\right]\right] \\
&\le \frac{1}{P^2K^2b^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)})-\nabla f_i(x^{T(t,i)}))\right\|^2\right]\le\frac{\sigma^2}{PKb},
\end{aligned}$$

which concludes the proof. □

Proof of Lemma 14. We first consider Option I:

$$\begin{aligned}
\mathbb{E}[(\mathrm{d})] &= \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0]\left(\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_i(x^0)-\nabla f_{i,j}(x^0)\right)\right\|^2\right] \\
&\le \frac{1}{P^2}\sum_{i=1}^{P}\mathbb{E}\left[\mathbb{1}[T(t,i)=0]\left\|\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)-\nabla f_i(x^0)\right\|^2\right]. \qquad (37)
\end{aligned}$$

For the second equality, we considered conditional expectation on $T(t,i)$ similarly to the proof of Lemma 13. Because $T(t,i)$ and $\left\|\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)-\nabla f_{i,j}(x^0)\right\|^2$ are independent and $\mathbb{E}[\mathbb{1}[T(t,i)=0]]=(1-\frac{p}{P})^t$, we have

$$(37)\le\frac{(1-\frac{p}{P})^t\sigma^2}{PKb},$$

which proves the claim for Option I.

For Option II, in the same way as we obtained (37), we have

$$\mathbb{E}[(\mathrm{d})] = \mathbb{E}\left[\left\|\frac{1}{P}\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0]\left(\frac{1}{p}\sum_{i\in I^0}\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)-\nabla f_i(x^0)\right)\right\|^2\right]$$
$$\leq \frac{1}{P^2}\sum_{i=1}^{P}\mathbb{E}\left[\left\|\mathbb{1}[T(t,i)=0]\left(\frac{1}{p}\sum_{i\in I^0}\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)-\nabla f_i(x^0)\right)\right\|^2\right], \tag{38}$$

which is further bounded, because $T(t,i)$ is independent of I^0 and J_i^0 and $\mathbb{E}[\mathbb{1}[T(t,i)=0]] = (1-\frac{p}{P})^t$, as

$$(38) \leq \frac{1}{P}\mathbb{E}\left[\left\|\mathbb{1}[T(t,i)=0]\left(\frac{1}{p}\sum_{i\in I^0}\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)-\nabla f_i(x^0)\right)\right\|^2\right] \leq \frac{(1-\frac{p}{P})^t\sigma^2}{pKb},$$

which gives the assertion for Option II. □

Proof of Lemma 15. Because $\nabla f_{i_t,j}(x^{t,l}) - \nabla f_{i_t,j}(x^{t,l-1})) - \nabla f_{i_t}(x^{t,l}) + \nabla f_{i_t}(x^{t,l-1})$ is mean-zero and independent from $J^{t,1},\cdots,J^{t,l-1}$, we have

$$\mathbb{E}\left[\left\|\sum_{l=1}^{k}\left(\frac{1}{b}\sum_{j\in J^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))-(\nabla f_{i_t}(x^{t,l})-\nabla f_{i_t}(x^{t,l-1}))\right)\right\|^2\right] \leq \frac{L^2}{b}\sum_{l=1}^{k}\mathbb{E}[\|x^{t,l}-x^{t,l-1}\|^2]$$

□

Proof of Lemma 16. Because of Assumption 5, we have

$$\mathbb{E}\left[\left\|\nabla f_{i_t}(x^{t,k})-\nabla f_{i_t}(x^{t,0})-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))\right\|^2\right]$$
$$\leq \zeta^2\mathbb{E}[\|x^{t,k}-x^{t,0}\|^2] \leq \zeta^2 K\sum_{l=1}^{k}\mathbb{E}[\|x^{t,l}-x^{t,k-1}\|^2].$$

□

Proof of Theorem 5. According to Lemma 6, for all $t\geq 1$ and $1\leq k\leq K$, we have

$$f(x^{t,k}) \leq f(x^{t,k-1}) + \eta\|\nabla f(x^{t,k-1}) - (\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k-1})\|^2 - \frac{\eta}{2}\|\nabla f(x^{t,k-1})\|^2 - \left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,k}-x^{t,k-1}\|^2 + \frac{r^2}{\eta}.$$

We sum up this over all $(t',k') \leq (T,K)$ to get

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2$$
$$\leq \frac{2}{\eta}\left[(f(x^0)-f(x^T)) - \sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,k}-x^{t,k-1}\|^2 + \eta\sum_{t=1}^{T}\|\nabla f(x^{t,k-1}) - (\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k-1})\|^2\right] + \frac{2TKr^2}{\eta^2},$$
$$\therefore \left(\min_{\substack{1\leq t\leq T\\ 1\leq k\leq K}}\mathbb{E}[\|\nabla f(x^{t,k-1})\|]\right)^2 \leq \frac{2\Delta}{\eta TK} - \frac{1}{\eta TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{\eta}-L\right)\mathbb{E}[\|x^{t,k}-x^{t,k-1}\|^2]$$
$$+\frac{2}{TK}\sum_{t=1}^{T}\sum_{k=1}^{T}\mathbb{E}[\|\nabla f(x^{t,k-1}) - (\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k-1})\|^2] + \frac{\varepsilon^2}{2}. \tag{39}$$

Here we used $f(x^0) - f(x^T) \leq \Delta$ and $r \leq \frac{\eta\varepsilon}{2}$.

According to Lemma 10, we have

$$\begin{aligned}
&\frac{2}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k}-\nabla f(x^{t,k})\|^2] \\
&\leq \left[\frac{360PK^2\zeta^2}{p^2}+\frac{24L^2PK}{p^2b}+\frac{12L^2K}{b}+12\zeta^2K^2\right]\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}[\|x^{t,k}-x^{t,k-1}\|^2] \\
&\quad+\frac{108\sigma_c^2P\mathbb{1}[\text{Option II}]}{p^2T}+\frac{24\sigma^2}{PKb}+\frac{12\sigma^2P\mathbb{1}[\text{Option II}]}{p^2bKT}. \qquad (40)
\end{aligned}$$

We take η as

$$\eta \leq \frac{1}{2L}\vee\frac{p}{\zeta K\sqrt{2880P}}\wedge\frac{p\sqrt{b}}{L\sqrt{192PK}}\wedge\frac{\sqrt{b}}{L\sqrt{96K}}\wedge\frac{1}{\sqrt{96}\zeta K} \qquad (41)$$

so that $\frac{360PK\zeta^2}{p^2}+\frac{24L^2P}{p^2b}+\frac{12L^2K^2}{b}+12\zeta^2K\leq\frac{1}{2\eta^2}$ and $L\leq\frac{1}{2\eta}$ hold. Moreover, we apply the assumption $\frac{24\sigma^2}{PKb}\leq\frac{\varepsilon^2}{8}$ (or just compute the local full gradient instead of sampling J_i^t), and let

$$T \geq \left[\frac{16\Delta}{K\eta}+\mathbb{1}[\text{Option II}](\frac{96\sigma P}{p^2Kb}+\frac{864\sigma_c^2P}{p^2})\right]\frac{1}{\varepsilon^2}.$$

so that $\frac{2\Delta}{\eta TK}, \frac{108\sigma_c^2P\mathbb{1}[\text{Option II}]}{p^2T}, \frac{12\sigma^2P\mathbb{1}[\text{Option II}]}{p^2bKT}\leq\frac{\varepsilon^2}{8}$ hold. Especially, when we take η that satisfy the equality in (41), T should satisfy

$$T \geq \left[\left[\frac{32L}{K}\vee\zeta\left(\frac{384\sqrt{5P}}{p}\vee 64\sqrt{6}\right)\vee\frac{L}{\sqrt{bK}}\left(\frac{128\sqrt{3P}}{p}\vee 64\sqrt{6}\right)\right]\Delta+\mathbb{1}[\text{Option II}]\left(\frac{96\sigma P}{p^2Kb}+\frac{864\sigma_c^2P}{p^2}\right)\right]\frac{1}{\varepsilon^2}.$$

Finally, applying (40) to (39) yields that

$$\begin{aligned}
\left(\min_{\substack{1\leq t\leq T\\1\leq k\leq K}}\mathbb{E}[\|\nabla f(x^{t,k-1})\|]\right)^2 &\leq \frac{2\Delta}{\eta TK}-\frac{1}{\eta TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{\eta}-L\right)\mathbb{E}[\|x^{t,k}-x^{t,k-1}\|^2] \\
&\quad+\underbrace{\frac{1}{2\eta^2TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}[\|x^{t,k}-x^{t,k-1}\|^2]+\frac{3\varepsilon^2}{8}}_{(40)}+\frac{\varepsilon^2}{2}\leq\varepsilon^2,
\end{aligned}$$

which concludes the proof. □

E.2. Second-order Optimality

We show that FL-SILVER can efficiently find second-order stationary points. The full theorem is provided as follows.

Theorem 6. *We assume Assumptions 1-(b), 2, 3-(b), 4, and 5-(b). For Option II, we additionally assume 3-(a). Set $\delta<\zeta$, $p=\tilde{\Theta}(\sqrt{P}+\frac{\zeta^2}{\delta^2}+\frac{L^2}{Kb\delta^2}), b\geq K$, $b=\tilde{\Omega}(\frac{\sigma^2}{PK\varepsilon^2})$, $\eta=\tilde{\Theta}(\frac{1}{L})$, $r=\tilde{O}(\frac{\varepsilon}{L})$, and $\nu\in(0,1)$. Then, Algorithm 2 finds (ε,δ)-second-order stationary points using*

$$\tilde{O}\left(1+\left[\Delta\left(\frac{L}{K}+\zeta\right)+\mathbb{1}[\text{Option II}]\left(\frac{\sigma P}{p^2Kb}+\frac{\sigma_c^2P}{p^2}\right)\right]\left(\frac{1}{\varepsilon^2}+\frac{\rho^2}{\delta^4}\right)\right)$$

communication rounds and

$$\tilde{O}\left(P+\left(\sqrt{P}+\frac{\zeta^2}{\delta^2}+\frac{L^2}{Kb\delta^2}\right)\left[\Delta\left(\frac{L}{K}+\zeta\right)+\mathbb{1}[\text{Option II}]\left(\frac{\sigma P}{p^2Kb}+\frac{\sigma_c^2P}{p^2}\right)\right]\left(\frac{1}{\varepsilon^2}+\frac{\rho^2}{\delta^4}\right)\right)$$

communication complexity, with probability at least $1-\nu$.

We first prepare a high-probability version of Theorem 5.

Lemma 17. *Let Assumptions 1-(b), 2, 3-(a) (Only for Option II), 3-(b), and 5-(b) hold. Set η as*

$$\eta = \tilde{\Theta}\left(\frac{1}{L} \vee \frac{p}{\zeta K\sqrt{P}} \wedge \frac{p\sqrt{b}}{L\sqrt{PK}} \wedge \frac{\sqrt{b}}{L\sqrt{K}} \wedge \frac{1}{\zeta K}\right).$$

Then Algorithm 2 satisfies

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\|x^{t,k}-x^{t,k-1}\|^2 \leq \frac{2\Delta}{\eta} + \frac{C_1\sigma^2 T}{Pb} + C_1\mathbb{1}[\text{Option II}]K\left[\frac{\sigma_c^2 P}{p^2} + \frac{\sigma^2 P}{p^2 Kb}\right] + \frac{2r^2}{\eta^2}$$

with high probability.

Proof. Similarly to Lemma 1 (high probability bound), $\tilde{I}_s^t = \emptyset$ with high probability if $t - s = \tilde{\Omega}(\frac{P}{p})$. We set $T_3 = \tilde{\Theta}(\frac{P}{p})$ so that with high probability, by following the proof of Lemma 10, we have

$$\begin{aligned}
&\frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} + z^{t,k} - \nabla f(x^{t,k}) \\
&= \underbrace{\frac{1}{P}\sum_{s=\max\{1,t-T_3\}}^{t}\left(\frac{|\tilde{I}_s^t|}{p}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})) - \sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s)-\nabla f_i(x^{s-1}))\right) + \frac{1}{P}\sum_{i\in\tilde{I}_1^t}(\tilde{y}_i^0 - \nabla f_i(x^0))}_{\text{(a)}'} \\
&\quad + \underbrace{\sum_{s=\max\{1,t-T_3\}}^{t}\frac{|\tilde{I}_s^t|}{PpKb}\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1})))}_{\text{(b)}'} \\
&\quad + \underbrace{\frac{1}{PKb}\sum_{i=1}^{P}\mathbb{1}[T(t,i)\geq 1]\sum_{j\in J_i^{T(t,i)}}(\nabla f_{i,j}(x^{T(t,i)}) - \nabla f_i(x^{T(t,i)}))}_{\text{(c)}'} + \underbrace{\frac{1}{P}\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0](y_i^0 - \tilde{y}_i^0)}_{\text{(d)}'}. \\
&\quad + \underbrace{\sum_{l=1}^{k}\left(\frac{1}{b}\sum_{j\in J^{t,l}}((\nabla f_{i_t,j}(x^{t,l}) - \nabla f_{i_t,j}(x^{t,l-1})) - (\nabla f_{i_t}(x^{t,l}) - \nabla f_{i_t}(x^{t,l-1})))\right)}_{\text{(e)}'} \\
&\quad - \underbrace{\nabla f_{i_t}(x^{t,k}) - \nabla f_{i_t}(x^{t,0}) - (\nabla f(x^{t,k}) - \nabla f(x^{t,0}))}_{\text{(f)}'}.
\end{aligned}$$

Because $\left\|\frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} + z^{t,k} - \nabla f(x^{t,k})\right\|^2 \leq 6\|\text{(a)}'\|^2 + 6\|\text{(b)}'\|^2 + 6\|\text{(c)}'\|^2 + 6\|\text{(d)}'\|^2 + 6\|\text{(e)}'\|^2 + 6\|\text{(f)}'\|^2$, we bound each by referring to Lemmas 11 to 16.

For (a)', just following Lemma 1 (High probablity bound) gives

$$\|\text{(a)}'\|^2 \leq \frac{C_1\zeta^2}{b}\sum_{s=\max\{1,t-T_3\}}^{t}\|x^s - x^{s-1}\|^2 + \frac{C_1\sigma_c^2\mathbb{1}[\text{Option II}]\mathbb{1}[t\leq T_3]}{b}.$$

For (b)', we first bound $\left|\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1})))\right|^2$ by $C_1L^2pKb\|x^s - x^{s-1}\|^2$ using Proposition 2 and fixing I^s, and then apply Proposition 4 by conditioning on $|\tilde{I}_s^t|$ to obtain

$$\|\text{(b)}'\|^2 \leq \sum_{s=\max\{1,t-T_3\}}^{t}\frac{|\tilde{I}_s^t|^2 C_1L^2}{P^2pKb}\|x^s - x^{s-1}\|^2 \leq \frac{C_1L^2}{pKb}\sum_{s=\max\{1,t-T_3\}}^{t}\|x^s - x^{s-1}\|^2.$$

For (c)', by conditioning on $T(t,i)$, Proposition 2 yields

$$\|(\mathrm{c})'\|^2 \leq \frac{1}{P^2K^2b^2}C_1\left(\sum_{i=1}^{P}\mathbb{1}[T(t,i)\geq 1]\right)\sigma^2 \leq \frac{C_1\sigma^2}{PKb}.$$

For (d)', if $t > T_3$, it is equal to zero with high probability because $\mathbb{1}[T(t,i)=0]$ for all i. If $t < T_3$, for Option I, by conditioning on $I^0, \cdots, I^T$, Proposition 2 yields

$$\|(\mathrm{d})'\|^2 = \left\|\frac{1}{P}\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0]\left(\frac{1}{Kb}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)-\nabla f_i(x^0)\right)\right\|^2 \leq C_1\frac{1}{P^2}\left(\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0]\right)\frac{C_1\sigma^2}{Kb}\leq\frac{C_1\sigma^2}{PKb}$$

and for Option II, Proposition 2 yields

$$\begin{aligned}\|(\mathrm{d})'\|^2 &= \left\|\frac{1}{P}\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0]\left(\frac{1}{p}\sum_{i'\in I^0}\frac{1}{Kb}\sum_{j\in J_{i'}^0}\nabla f_{i',j}(x^0)-\nabla f_{i'}(x^0)\right)\right\|^2\\ &\leq \frac{1}{P^2}\left(\sum_{i=1}^{P}\mathbb{1}[T(t,i)=0]\right)^2\frac{C_1\sigma^2}{pKb}\leq\frac{C_1\sigma^2}{pKb}.\end{aligned}$$

For (e)', first apply Proposition 2 $\frac{1}{b}\sum_{j\in J^{t,l}}((\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))-(\nabla f_{i_t}(x^{t,l})-\nabla f_{i_t}(x^{t,l-1})))$ and then apply Proposition 4 to obtain

$$\|(\mathrm{e})'\|^2 \leq \frac{C_1L^2}{b}\sum_{l=1}^{k}\|x^{t,l}-x^{t,l-1}\|^2.$$

Finally, for (f)', just consider high probability bound with respect to the randomness of i_t yields

$$\|(\mathrm{e})'\|^2 \leq C_1\zeta^2K\sum_{l=1}^{k}\|x^{t,l}-x^{t,l-1}\|^2.$$

Putting everything all together, we obtain the bound on the gradient estimator:

$$\begin{aligned}\|\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k}-\nabla f(x^{t,k})\|^2 &\leq C_1\left[\frac{\zeta^2}{p}+\frac{L^2}{pKb}\right]\sum_{s=\max\{1,t-T_3\}}^{t}\mathbb{E}\left[\|x^s-x^{s-1}\|^2\right]\\ &+C_1\left[\frac{L^2}{b}+\zeta^2K\right]\sum_{l=1}^{k}\mathbb{E}[\|x^{t,l}-x^{t,l-1}\|^2]+\frac{C_1\sigma^2}{PKb}+C_1\mathbb{1}[\text{Option II}]\mathbb{1}[t\leq T_3]\left[\frac{\sigma_c^2}{p}+\frac{\sigma^2}{pKb}\right],\end{aligned}$$

which implies

$$\begin{aligned}\frac{2}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\|\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k}-\nabla f(x^{t,k})\|^2 &\leq C_1\left[\frac{PK^2\zeta^2}{p^2}+\frac{L^2PK}{p^2b}+\frac{L^2K}{b}+\zeta^2K^2\right]\sum_{t=1}^{T}\sum_{k=1}^{K}\|x^{t,k}-x^{t,k-1}\|^2\\ &+\frac{C_1\sigma^2}{PKb}+\frac{C_1\mathbb{1}[\text{Option II}]}{T}\left[\frac{\sigma_c^2P}{p^2}+\frac{\sigma^2P}{p^2Kb}\right],\end{aligned}\tag{42}$$

where we used $T_3=\tilde{O}(\frac{p}{P})$.

Similarly to (39), when $r \leq \frac{\eta\varepsilon}{2}$, we have

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2+\frac{1}{2\eta^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\|x^{t,k}-x^{t,k-1}\|^2\leq\frac{2\Delta}{\eta}-\frac{1}{\eta}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{2\eta}-L\right)\|x^{t,k}-x^{t,k-1}\|^2 \\ +2\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(x^{t,k-1})-(\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k-1})\|^2]+\frac{2r^2}{\eta^2}. \tag{43}$$

We take η as

$$\eta=\tilde{\Theta}\left(\frac{1}{2L}\vee\frac{p}{\zeta K\sqrt{P}}\wedge\frac{p\sqrt{b}}{L\sqrt{PK}}\wedge\frac{\sqrt{b}}{L\sqrt{K}}\wedge\frac{1}{\zeta K}\right)$$

so that $\|x^{t,k}-x^{t,k-1}\|^2$ terms cancel out in (42) and RHS of (43). Then we obtain that

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2+\frac{1}{2\eta^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\|x^{t,k}-x^{t,k-1}\|^2\leq\frac{2\Delta}{\eta}+\frac{C_1\sigma^2T}{Pb}+C_1\mathbb{1}[\text{Option II}]K\left[\frac{\sigma_c^2P}{p^2}+\frac{\sigma^2P}{p^2Kb}\right]+\frac{2r^2}{\eta^2}$$

with high probability. □

Similarly to the proof of SILVER, the key argument is the exponential separation of two coupled trajectories with different initial values.

Lemma 18 (Small Stuck Region)**.** *Let Assumptions 1-(b), 4, and 5-(b) hold. Assume $\delta<\frac{1}{\zeta}$. Let $\{x^{t,k}\}$ be a sequence generated by* FL-SILVER *and (τ_0,κ_0) $(0\leq\kappa_0<K)$ be a step where $-\gamma:=\lambda_{\min}(\nabla^2 f(x^{\tau_0,\kappa_0}))\leq-\delta$ holds. We denote the eigenvector with the eigenvalue $\lambda_{\min}(\nabla^2 f(x^{\tau_0,\kappa_0}))$ by* e*. Moreover, let $\{\tilde{x}^{t,k}\}$ by a coupled sequence that is generated by* FL-SILVER *with $\tilde{x}^0=x^0$ and shares the same choice of randomness with $\{x_t\}$ i.e., client samplings, minibatches and noises, except for the noise at a step $(\tau_0,K)>(\tau_0,\kappa_0)$: $\tilde{\xi}^{\tau_0,K}=\xi^{\tau_0,K}-r_e\mathsf{e}$ with $r_e\geq\frac{r\nu}{\sqrt{d}}$ $(0<\nu<1)$. Let $w^{t,k}=x^{t,k}-\tilde{x}^{t,k}$, $w^t=x^t-\tilde{x}^t$, $g^{t,k}=\frac{1}{P}\sum_{i=1}^{P}y_i^{t-1}+z^{t,k}$, $\tilde{g}^{t,k}=\frac{1}{P}\sum_{i=1}^{P}\tilde{y}_i^{t-1}+z^{t,k}$, $h^t=\frac{1}{P}\sum_{i=1}^{P}y_i^t-\nabla f(x^t)-\left(\frac{1}{P}\sum_{i=1}^{P}\tilde{y}_i^t-\nabla f(\tilde{x}^t)\right)$, and $h^{t,k}=(z^{t,k}-(\nabla f(x^{t,k})-\nabla f(x^{t,0})))-(\tilde{z}^{t,k}-(\nabla f(\tilde{x}^{t,k})-\nabla f(\tilde{x}^{t,0})))$. Using these notations, $g^{t,k}-\nabla f(x^{t,k})-(\tilde{g}^{t,k}-\nabla f(\tilde{x}^{t,k}))=h^{t-1}+h^{t,k}$ holds.*

Then, there exists a sufficiently large constants $C_3=\tilde{O}(1)$ with which the following holds: If we take $p=\tilde{\Omega}(\sqrt{P}+\frac{\zeta^2}{\delta^2}+\frac{L^2}{Kb\delta^2})$, $b\geq K$, $K=O(\frac{L}{\zeta})$, $\eta=\tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_4=O(\frac{C_4\log\frac{\delta}{\rho r_e}}{\eta\gamma})=\tilde{O}\left(\frac{L}{\delta}\right)$, with high probability, we have

$$\max_{(\tau_0,\kappa_0)\leq(t,k)<(\tau_0+1,T_4)}\{\|x^{\tau,k}-x^{\tau_0,\kappa_0}\|,\|\tilde{x}^{\tau,\kappa}-x^{\tau_0,\kappa_0}\|\}\geq\frac{\delta}{C_3\rho}.$$

In order to show Lemma 18, we prepare the two following lemmas, which bound the difference between gradient estimation errors of the two sequences.

Lemma 19. *Under the same assumption as that of Lemma 18, we assume $\max_{(\tau_0,\kappa_0)\leq(t,k)<(\tau_0+1,T_5)}\{\|x^{\tau,k}-x^{\tau_0,\kappa_0}\|,\|\tilde{x}^{\tau,\kappa}-x^{\tau_0,\kappa_0}\|\}<\frac{\delta}{C_3\rho}$. Then, the following holds uniformly for all $(\tau_0,\kappa_0)\leq(t,k)\leq(\tau_0+1,T_4)$ with high probability:*

$$\|h^t\|\leq\begin{cases}0 & (t<\tau_0),\\ \left(\frac{\zeta}{\sqrt{p}}+\frac{L}{\sqrt{pKb}}\right)C_1r_e & (t=\tau_0),\\ \left(\frac{\zeta}{\sqrt{p}}+\frac{L}{\sqrt{pKb}}\right)C_1r_e+\left(\frac{\zeta\sqrt{K}}{\sqrt{p}}+\frac{L}{\sqrt{pb}}\right)C_1\sqrt{\sum_{s=\max\{\tau_0+1,t-T_3\}}^{t}\sum_{k=1}^{K}\|w^{s,k}-w^{s,k-1}\|^2} & \\ \quad+\frac{C_1\delta}{C_3\sqrt{p}}\sqrt{\sum_{s=\max\{\tau_0,t-T_3\}}^{t}\|w^s\|^2} & (t\geq\tau_0+1),\end{cases}$$

where $T_3=\tilde{\Theta}(\frac{P}{p})$, and $C_1=\tilde{O}(1)$ is a sufficiently large constant.

Lemma 20. *Under the same assumption as that of Lemma 18, the following holds uniformly for all $t \geq \tau_0 + 1$ and $k \geq 0$ with high probability:*

$$\left\|h^{t,k}\right\| \leq \zeta \sum_{l=1}^{k} \|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_3}\|w^{t,k}\| + \frac{2\delta}{C_3}\|w^{t,0}\| + \frac{C_1}{\sqrt{b}} \sqrt{\sum_{l=1}^{k} \left(L\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_3}\|w^{t,l}\| + \frac{2\delta}{C_3}\|w^{t,l-1}\| \right)^2}.$$

For $t \leq \tau_0$, we have $\left\|h^{t,k}\right\| = 0$.

Proof of Lemma 19. As for the case $t < \tau_0$, the assertion directly follows from the definition of $\{\tilde{x}^{t,k}\}$. For the proof of the rest cases, we use notations as follows:

$$\begin{aligned}
H &= \nabla^2 f(x^{\tau_0,\kappa_0}), \\
H_i &= \nabla^2 f_i(x^{\tau_0,\kappa_0}) \\
H_{i,j} &= \nabla^2 f_{i,j}(x^{\tau_0,\kappa_0}), \\
\mathrm{d}H^{t,k} &= \int_0^1 (\nabla^2 f(\tilde{x}^{t,k} + \theta(x^{t,k} - \tilde{x}^{t,k})) - H)\mathrm{d}\theta, \\
\mathrm{d}H_i^{t,k} &= \int_0^1 (\nabla^2 f_i(\tilde{x}^{t,k} + \theta(x^{t,k} - \tilde{x}^{t,k})) - H_i)\mathrm{d}\theta, \\
\mathrm{d}H_{i,j}^{t,k} &= \int_0^1 (\nabla^2 f_{i,j}(\tilde{x}^{t,k} + \theta(x^{t,k} - \tilde{x}^{t,k})) - H_{i,j})\mathrm{d}\theta.
\end{aligned}$$

Moreover, we denote

$$u_i^s \coloneqq (\nabla f_i(x^s) - \nabla f_i(\tilde{x}^s)) - (\nabla f_i(x^{s-1}) - \nabla f_i(\tilde{x}^{s-1})) - (\nabla f(x^s) - \nabla f(\tilde{x}^s)) + (\nabla f(x^{s-1}) - \nabla f(\tilde{x}^{s-1}))$$

and

$$u_{i,j}^s \coloneqq (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(\tilde{x}^s)) - (\nabla f_{i,j}(x^{s-1}) - \nabla f_{i,j}(\tilde{x}^{s-1})) - (\nabla f_i(x^s) - \nabla f_i(\tilde{x}^s)) + (\nabla f_i(x^{s-1}) - \nabla f_i(\tilde{x}^{s-1})).$$

Note that $\mathbb{E}_i[u_i^s] = 0$ (expectation with respect to the choice of i) and $\mathbb{E}_j[u_{i,j}^s] = 0$ (expectation with respect to the choice of j) hold. Using 1-(b), 4, 5 and $\max_{(\tau_0,\kappa_0) \leq (t,k) < (\tau_0+1, T_5)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} < \frac{\delta}{C_3\rho}$, we can derive that

$$\|u_i^s\| \leq \zeta\|w^s - w^{s-1}\| + \frac{2\delta}{C_3}\|w^s\| + \frac{2\delta}{C_2}\|w^{s-1}\| \quad \text{and} \quad \|u_{i,j}^s\| \leq L\|w^s - w^{s-1}\| + \frac{2\delta}{C_3}\|w^s\| + \frac{2\delta}{C_2}\|w^{s-1}\|$$

for $s \geq \tau_0 + 1$, by similar argument to the proof of Lemma 8. For $t = \tau_0$, we have $\|u_i^{\tau_0}\| = \|(\nabla f_i(x^{\tau_0}) - \nabla f_i(\tilde{x}^{\tau_0})) - (\nabla f(x^{\tau_0}) - \nabla f(\tilde{x}^{\tau_0}))\| \leq \zeta\|x^{\tau_0} - \tilde{x}^{\tau_0}\| = \zeta r_e$ and $\|u_{i,j}^{\tau_0}\| = \|(\nabla f_{i,j}(x^{\tau_0}) - \nabla f_{i,j}(\tilde{x}^{\tau_0})) - (\nabla f(x^{\tau_0}) - \nabla f(\tilde{x}^{\tau_0}))\| \leq L\|x^{\tau_0} - \tilde{x}^{\tau_0}\| = Lr_e$.

As we did in Lemma 8, for $t \geq \tau_0 + 1$, we have

$$\begin{aligned}
h^t = &\underbrace{\frac{1}{P}\left(\frac{|\tilde{I}_{\tau_0}^t|}{p}\sum_{i\in I^{\tau_0}} u_i^{\tau_0} - \sum_{i\in\tilde{I}_{\tau_0}^t} u_i^{\tau_0}\right)}_{(a)} + \underbrace{\frac{1}{PKb}\left(\frac{|\tilde{I}_{\tau_0}^{\tau_0}|}{p}\sum_{i\in I^{\tau_0}}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0} - \sum_{i\in\tilde{I}_{\tau_0}^{\tau_0}}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0}\right)}_{(b)} \\
&+ \underbrace{\frac{1}{P}\sum_{s=\max\{\tau_0+1,t-T_3\}}^{t}\left(\frac{|\tilde{I}_s^t|}{p}\sum_{i\in I^s} u_i^s\right) - \frac{1}{P}\sum_{s=\max\{\tau_0+1,t-T_3\}}^{t}\sum_{i\in\tilde{I}_s^t} u_i^s}_{(c)} \\
&+ \underbrace{\frac{1}{PKb}\sum_{s=\max\{\tau_0+1,t-T_3\}}^{t}\left(\frac{|\tilde{I}_s^t|}{p}\sum_{i\in I^s}\sum_{j\in J_i^s} u_{i,j}^s\right) - \frac{1}{PKb}\sum_{s=\max\{\tau_0+1,t-T_3\}}^{t}\sum_{i\in\tilde{I}_s^t}\sum_{j\in J_i^s} u_{i,j}^s}_{(d)}
\end{aligned}$$

with high probability uniformly for all t. For $t = \tau_0$, $h^{\tau_0} = (\mathrm{a}) + (\mathrm{b})$ holds.

Recall the argument in Lemma 8. We have that

$$\|(\mathrm{a})\| \leq \frac{C_1 \zeta r_e}{\sqrt{p}} \quad \text{and} \quad \|(\mathrm{c})\| \leq \frac{C_1}{\sqrt{p}} \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3\}} \left(\zeta \|w^s - w^{s-1}\| + \frac{2\delta}{C_3}\|w^s\| + \frac{2\delta}{C_3}\|w^{s-1}\| \right)^2}$$

hold with high probability.

Moreover, by conditioning on I^{τ_0}, Proposition 2 yields that

$$\|(\mathrm{b})\| \leq \left\| \frac{1}{P} \frac{|\tilde{I}^t_{\tau_0}|}{p} \sum_{i \in I^{\tau_0}} \sum_{j \in J_i^{\tau_0}} u_{i,j}^{\tau_0} \right\| + \left\| \frac{1}{P} \sum_{i \in \tilde{I}^t_{\tau_0}} \sum_{j \in J_i^{\tau_0}} u_{i,j}^{\tau_0} \right\| \leq \frac{|\tilde{I}^t_{\tau_0}|}{P} \frac{C_1 L r_e}{\sqrt{pKb}} + \frac{\sqrt{|\tilde{I}^t_{\tau_0}|} C_1 L r_e}{P\sqrt{Kb}} \leq \frac{C_1 L r_e}{\sqrt{pKb}},$$

with high probability.

For (d), we first bound

$$\left\| \frac{|\tilde{I}^t_s|}{p} \sum_{i \in I^s} \sum_{j \in J_i^s} u_{i,j}^s \right\| \leq \frac{C_1 |\tilde{I}^t_s| \sqrt{Kb}}{\sqrt{p}} \left(L\|w^s - w^{s-1}\| + \frac{2\delta}{C_3}\|w^s\| + \frac{2\delta}{C_3}\|w^{s-1}\| \right),$$

$$\left\| \sum_{i \in \tilde{I}^t_s} \sum_{j \in J_i^s} u_{i,j}^s \right\| \leq C_1 \sqrt{|\tilde{I}^t_s| Kb} \left(L\|w^s - w^{s-1}\| + \frac{2\delta}{C_3}\|w^s\| + \frac{2\delta}{C_3}\|w^{s-1}\| \right)$$

with high probability, using Proposition 2. Then, by Proposition 4, we have

$$\|(\mathrm{d})\| \leq \frac{C_1}{\sqrt{pKb}} \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3\}}^{t} \left(L\|w^s - w^{s-1}\| + \frac{2\delta}{C_3}\|w^s\| + \frac{2\delta}{C_3}\|w^{s-1}\| \right)^2}.$$

By combining all these, we have

$$\begin{aligned}
\|g^t\| &\leq \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}} \right) C_1 \zeta r_e + \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}} \right) C_1 \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3\}}^{t} \|w^s - w^{s-1}\|^2} \\
&\quad + \frac{C_1 \delta}{C_3} \sqrt{\sum_{s=\max\{\tau_0, t-T_3\}}^{t} \|w^s\|^2} \\
&\leq \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}} \right) C_1 \zeta r_e + \left(\frac{\zeta\sqrt{K}}{\sqrt{p}} + \frac{L}{\sqrt{pb}} \right) C_1 \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3\}}^{t} \sum_{k=1}^{K} \|w^{s,k} - w^{s,k-1}\|^2} \\
&\quad + \frac{C_1 \delta}{C_3 \sqrt{p}} \sqrt{\sum_{s=\max\{\tau_0, t-T_3\}}^{t} \|w^s\|^2}
\end{aligned}$$

with high probability. Thus, we get the assertion for $t \geq \tau_0 + 1$. For $t = \tau_0$, the bounds on (a) and (b) imply the desired bound. □

Proof of Lemma 20. Let

$$u_i^{t,l} \coloneqq (\nabla f_i(x^{t,l}) - \nabla f_i(\tilde{x}^{t,l})) - (\nabla f_i(x^{t,0}) - \nabla f_i(\tilde{x}^{t,0})) - (\nabla f(x^{t,l}) - \nabla f(\tilde{x}^{t,l})) + (\nabla f(x^{t,0}) - \nabla f(\tilde{x}^{t,0}))$$

and

$$\begin{aligned}u_{i,j}^{t,l} :=&(\nabla f_{i,j}(x^{t,l}) - \nabla f_{i,j}(\tilde{x}^{t,l})) - (\nabla f_{i,j}(x^{t,l-1}) - \nabla f_{i,j}(\tilde{x}^{t,l-1}))\\ &- (\nabla f_i(x^{t,l}) - \nabla f_i(\tilde{x}^{t,l})) + (\nabla f_u(x^{t,l-1}) - \nabla f_i(\tilde{x}^{t,l-1}))\end{aligned}$$

By their definitions, $h^{t,k} = u_{i_t}^{t,l} + \frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}^{t,l}} u_{i_t,j}^{t,l}$ holds. We can bound the norm of them as

$$\|u_{i_t}^{t,k}\| \le \zeta\|w^{t,k} - w^{t,0}\| + \frac{2\delta}{C_3}\|w^{t,k}\| + \frac{2\delta}{C_3}\|w^{t,0}\| \le \zeta\sum_{l=1}^{k}\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_3}\|w^{t,k}\| + \frac{2\delta}{C_3}\|w^{t,0}\| \tag{44}$$

and

$$\|u_{i_t,j}^{t,l}\| \le L\|w^{t,l} - w^{t,l}\| + \frac{2\delta}{C_3}\|w^{t,l}\| + \frac{2\delta}{C_3}\|w^{t,l-1}\|.$$

Thus, applying Proposition 2 and Proposition 4 to $\frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}^{t,l}} u_{i_t,j}^{t,l}$, we get

$$\left\|\frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}^{t,l}} u_{i_t,j}^{t,l}\right\| \le \frac{C_1^2}{\sqrt{b}}\sqrt{\sum_{l=1}^{k}\left(L\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_3}\|w^{t,l}\| + \frac{2\delta}{C_3}\|w^{t,l-1}\|\right)^2} \tag{45}$$

with high probability for all t and K.

Substituting (44) and (45) to $h^{t,k} = u_{i_t}^{t,l} + \frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}^{t,l}} u_{i_t,j}^{t,l}$, we get the desired bound. □

Now, we are ready to prove Lemma 18.

Proof of Lemma 18. We assume the contrary and show the following by induction, for $(\tau_0+1, 0) \le (t,k) \le (\tau_0+1, T_4)$:

$$\begin{aligned}&\text{(a)}\quad \frac{1}{2}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \le \|w^{t,k}\| \le 2(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e\\ &\text{(b)}\quad \|w^{t,k} - w^{t,k-1}\| \le \begin{cases} r_e & (\text{for } (t,k) = (\tau_0+1, 0))\\ 3\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e & (\text{for } (t,k) > (\tau_0+1, 0))\end{cases}\\ &\text{(c)}\quad \|h^{t-1} + h^{t,k}\| \le \frac{C_1\gamma}{C_3}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e.\end{aligned}$$

Then, (a) yields contradiction by taking $(t,k) - (\tau_0+1, 0) = T_4 = \Theta\left(1 + \frac{\log\frac{\delta}{C_2\rho r_e}}{\eta\gamma K}\right)$ to break the assumption.

It is easy to check (a) and (b) for and $t = \tau_0 + 1$ and $k = 0$. As for (c), checking the initial condition at $(t,k) = (\tau_0+1, 0)$ requires assumption on the size of p. According to Lemma 19, taking $p \ge C_1 C_3^2(\frac{\zeta^2}{\delta^2} + \frac{L^2}{\delta^2 K b})$, $\|g^{\tau_0}\| \le \frac{C_1}{C_3}\delta r_e \le C_1\gamma r_e$ holds.

Now, we derive that (a), (b) and (c) are true for $(t, k+1)$, assuming that they are true for all $(\tau_0+1, 0), \cdots, (t,k)$. To this end, we consider the decomposition of $w^{t,k}$ as follows:

$$\begin{aligned}w^{t,k+1} &= w^{t,k} - \eta\left(g^{t,k} - \tilde{g}^{t,k}\right)\\ &= (1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e \mathsf{e} - \eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} (I - \eta H)^{(t-s)K+(k-l)}(\mathrm{d}H^{s,l} w^{s,l} + h^{s-1} + h^{s,l}),\end{aligned} \tag{46}$$

for $(t, k+1) \ge (\tau_0+1, 1)$.

Verifying (a) The first term $(1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e \mathsf{e}$ of (46) satisfies

$$\|(1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e \mathsf{e}\| = (1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e.$$

Then, focus on bounding $\eta \sum_{(s,l)=(\tau_0+1,0)}^{t,k} (I-\eta H)^{(t-s)K+(k-l)} (\mathrm{d}H^{s,l} w^{s,l} + h^{s-1} + h^{s,l})$ by $\frac{1}{2}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e$. We have

$$\begin{aligned}
\left\| \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} (I-\eta H)^{(t-s)K+(k-l)} \mathrm{d}H^{s,l} w^{s,l} \right\| &\le \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} \|I-\eta H\|^{(t-s)K+(k-l)} \left\|\mathrm{d}H^{s,l}\right\| \left\|w^{s,l}\right\| \\
&\le 2\eta(1+\eta\gamma)^{(t-s)K+(k-l)+(s-\tau_0-1)K+l} r_e \sum_{(s,l)=(\tau_0+1,0)}^{t,k} \left\|\mathrm{d}H^{s,l}\right\| \\
&\le 2\eta(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e T_4 K \frac{\delta}{C_3} \\
&\le \frac{2\eta\delta T_4}{C_3}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \\
&\le \frac{1}{4}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e. \qquad (47)
\end{aligned}$$

The last inequality follows from the definition of $T_4 = \Theta(\frac{\log \frac{\delta}{C_3 \rho r_e}}{\eta\gamma}) \le \frac{C_3}{8\eta\delta}$. when we take C_3 sufficiently large.

In addition, we have

$$\begin{aligned}
\left\| \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} (I-\eta H)^{(t-s)K+(k-l)} (h^{s-1} + h^{s,l}) \right\| &\le \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} \|I-\eta H\|^{(t-s)K+(k-l)} \left\|g^{s-1} + h^{s,l}\right\| \\
&\le \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} (1+\eta\gamma)^{(t-s)K+(k-l)} \frac{C_1\gamma}{C_3} (1+\eta\gamma)^{(s-\tau_0-1)K+l} \\
&\le \frac{\eta\gamma T_4}{C_3}(1+\eta\gamma)^{(t-\tau_0-1)K+k} \\
&\le \frac{1}{4}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e. \qquad (48)
\end{aligned}$$

For the final inequality, we again use $T_4 = \Theta(\frac{\log \frac{\delta}{C_3 \rho r_e}}{\eta\gamma}) \le \frac{C_3}{4\eta\delta}$ with sufficiently large C_3.

Combining (47) and (48), we get (a) for $(t, k+1)$ as desired.

Verifying (b) For $(t,k) \ge (\tau_0+1, 0)$, we have

$$\begin{aligned}
& w^{t,k+1} - w^{t,k} \\
&= (1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e \mathsf{e} - \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} (I-\eta H)^{(t-s)K+(k-l)} (\mathrm{d}H^{s,l} w^{s,l} + h^{s-1} + h^{s,l}) \\
&\quad - (1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \mathsf{e} - \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)} (I-\eta H)^{(t-s)K+(k-l)} (\mathrm{d}H^{s,l} w^{s,l} + h^{s-1} + h^{s,l}) \\
&= \eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \mathsf{e} \\
&\quad - \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)} \eta H (I-\eta H)^{(t-s)K+(k-l)} (\mathrm{d}H^{s,l} w^{s,l} + h^{s-1} + h^{s,l}) - \eta(\mathrm{d}H_t w_t + h^{t-1} + h^{t,k}).
\end{aligned}$$

As for the first term, we can bound it as

$$\|\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \mathsf{e}\| \le \eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e.$$

Evaluating the second term requires (a) and (b) for $(\tau_0+1,0),\cdots,(t,k-1)$ and Lemma 2:

$$\begin{aligned}
&\left\|\eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)}\eta H(I-\eta H)^{(t-s)K+(k-l)}(\mathrm{d}H^{s,l}w^{s,l}+h^{s-1}+h^{s,l})\right\|\\
&\le\sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)}\eta\left\|\eta H(I-\eta H)^{(t-s)K+(k-l)}\right\|\left(\|\mathrm{d}H^{s,l}\|\|w^{s,l}\|+\|h^{s-1}+h^{s,l}\|\right)\\
&\le\sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)}\eta\left\|\eta H(I-\eta H)^{(t-s)K+(k-l)}\right\|\left(\frac{\delta}{C_3}(1+\eta\gamma)^{(s-\tau_0-1)K+l}r_e+\frac{C_1\gamma}{C_3}(1+\eta\gamma)^{(s-\tau_0-1)K+l}r_e\right)\\
&\le\sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)}\eta\left(\eta\gamma(1+\eta\gamma)^{(t-s)K+(k-l)}+\frac{1}{(t-s)K+(k-l)}\right)\left(\frac{\delta}{C_3}+\frac{C_1\gamma}{C_3}\right)(1+\eta\gamma)^{(s-\tau_0-1)K+l}r_e\\
&\le\eta\left(\eta\gamma T_5+\log T_5\right)\left(\frac{\delta}{C_3}+\frac{C_1\gamma}{C_3}\right)(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e.
\end{aligned}$$

Since $T_4=\tilde{O}\left(\frac{1}{\eta\gamma}\right)$ and $\gamma\ge\delta$, setting $C_3=\tilde{O}(1)$ with sufficiently large C_3 yields $(\eta\gamma T_5+\log T_5)\left(\frac{\delta}{C_3}+\frac{C_1\gamma}{C_3}\right)\le\gamma$. Thus, the second term is bounded by $\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e$.

Finally, we consider the third term. We have $\|\mathrm{d}H^{t,k}w^{t,k}\|\le\frac{\delta}{C_3}r_e(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e$ and $\|g^{t-1}+h^{t,k}\|\le\frac{2C_1\gamma}{C_3}(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e$. Thus, by taking C_3 sufficiently large, the third term is bounded by $\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e$.

By combining these bounds, we get (b) for $(t,k+1)$.

Verifying (c) Using Lemma 19 and assumptions, we have

$$\begin{aligned}
&\|h^{t+1}\|\\
&\le\left(\frac{\zeta}{\sqrt{p}}+\frac{L}{\sqrt{pKb}}\right)C_1r_e+\left(\frac{\zeta\sqrt{K}}{\sqrt{p}}+\frac{L}{\sqrt{pb}}\right)C_1\sqrt{\sum_{s=\max\{\tau_0+1,t-T_3\}}^{t}\sum_{k=1}^{K}\|w^{s,k}-w^{s,k-1}\|^2}\\
&\qquad+\frac{C_1\delta}{C_3\sqrt{p}}\sqrt{\sum_{s=\max\{\tau_0,t-T_3\}}^{t}\|w^s\|^2}\\
&\le\left[\frac{\zeta C_1}{\sqrt{p}}+\frac{LC_1}{\sqrt{pKb}}+\left(\frac{C_1\zeta KT_3^{\frac{1}{2}}}{\sqrt{p}}+\frac{C_1LK^{\frac{1}{2}}T_3^{\frac{1}{2}}}{\sqrt{pb}}\right)3\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+K}+\frac{C_1T_3^{\frac{1}{2}}\delta}{C_3\sqrt{pK}}(1+\eta\gamma)^{(t-\tau_0-1)K+K}\right]r_e\\
&=\left[\frac{\zeta C_1}{\sqrt{p}}+\frac{LC_1}{\sqrt{pKb}}+\left(\frac{C_1\zeta P^{\frac{1}{2}}K}{p}+\frac{C_1L\sqrt{PK}}{p\sqrt{b}}\right)3\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+K}+\frac{C_1\sqrt{P}\delta}{C_3p}(1+\eta\gamma)^{(t-\tau_0-1)K+K}\right]r_e.
\end{aligned}$$

Taking $p\ge C_1C_3^2(\sqrt{P}+\frac{C_3^2\zeta^2}{\delta^2}+\frac{C_3^2L^2}{\delta^2Kb})$, $b\ge K$, and $K=O\left(\frac{L}{\zeta}\right)$, we have $\|h^{t+1}\|\le\frac{C_1\gamma}{C_3}(1+\eta\gamma)^{(t-\tau_0)K}$.

Moreover, Lemma 20 states that, for $k<K$,

$$\begin{aligned}
&\left\|h^{t,k+1}\right\|\\
&\le\zeta\sum_{l=1}^{k+1}\|w^{t,l}-w^{t,l-1}\|+\frac{2\delta}{C_3}\|w^{t,k+1}\|+\frac{2\delta}{C_3}\|w^{t,0}\|+\frac{C_1}{\sqrt{b}}\sqrt{\sum_{l=1}^{k+1}\left(L\|w^{t,l}-w^{t,l-1}\|+\frac{2\delta}{C_3}\|w^{l,k}\|+\frac{2\delta}{C_3}\|w^{t,l-1}\|\right)^2}
\end{aligned}$$

holds with high probability. If (a) and (b) hold for all $(s,l)\le(t,k+1)$, then we have

$$\begin{aligned}
\left\|h^{t,k+1}\right\|\le{}&3\zeta K\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}+\frac{8\delta}{C_3}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}\\
&+\frac{C_1\sqrt{K}}{\sqrt{b}}L\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}+\frac{C_1\sqrt{K}\delta}{\sqrt{b}}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}.
\end{aligned}$$

Taking $b \geq K$, $\eta = \tilde{\Theta}(\frac{1}{L})$, and $K = O\left(\frac{L}{\zeta}\right)$, we have $\|h^{t,k+1}\| \leq \frac{C_1\gamma}{C_3}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}$. Thus, we obtain that (c) holds for $(t, k+1)$.

Therefore, we have completed the induction step and have $\frac{1}{2}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \leq \|w^t\|$ for all $(\tau_0+1, 0) \leq (t,k) < (\tau_0+1, T_5)$ with $T_4 = \Theta(\frac{\log \frac{\delta}{C_3\rho r_e}}{\eta\gamma})$. Taking C_3 sufficiently large, we have $\frac{1}{2}(1+\eta\gamma)^{(\tau_0+1-\tau_0-1)K+T_5} r_e \geq \frac{\delta}{C_3\rho}$. This yields contradiction against the assumption and the desired assertion follows. □

From Lemma 18, we can show that FL-SILVER escapes saddle points with high probability. We have the following lemma, and the proof is the same as that of Lemma 9.

Lemma 21. *Let $\{x^{t,k}\}$ be a sequence generated by* FL-SILVER *and (τ_0, κ_0) $(0 \leq \kappa_0 < K)$ be a step where $-\gamma := \lambda_{\min}(\nabla^2 f(x^{\tau_0,\kappa_0})) \leq -\delta$ holds. We take $\nu \in (0,1)$, $p \geq \tilde{\Omega}(\sqrt{P} + \frac{\zeta^2}{\delta^2} + \frac{L^2}{Kb\delta^2}), b \geq \sqrt{K}$ and, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_4 = \Theta(\frac{\log \frac{\delta}{C_3\rho r_e}}{\eta\gamma}) = \tilde{\Theta}\left(\frac{L}{\delta}\right)$, with sufficiently large $C_3 = \tilde{O}(1)$. Then,*

$$\mathbb{P}\left[\max_{(\tau_0,\kappa_0)\leq(t,k)<(\tau_0+1,T_5)} \|x^{t,k} - x^{\tau_0,\kappa_0+1}\| \geq \frac{\delta}{C_3\rho} \mid I^0, \cdots, I^\tau, i_0, \cdots, i_{\tau_0}, \xi^{0,0}, \cdots, \xi^{\tau_0,\kappa_0}\right] \geq 1-\nu.$$

Finally, we show the main theorem of this subsection, which guarantees that the algorithm finds (ε, δ)-second-order stationary point with high probability.

Proof of Theorem 2. Since $T_4 = \Theta(\frac{\log \frac{\delta}{C_3\rho r_e}}{\eta\gamma})$ depends on x^{τ_0,κ_0}, we take $T_4 = \Theta(\frac{\log \frac{\delta}{C_3\rho r_e}}{\eta\delta})$ from now instead. This change does not affect whether Lemma 21 holds. Also, we let $T_5 = \lceil 1 + \frac{T_4}{K} \rceil$.

We divide $\{t = 0, 1, \cdots, T-1\}$ into the following $\lfloor \frac{T}{2T_5} \rfloor$ *phases*: $P^s = \{2sT_5 \leq t < 2(s+1)T_5\}$ $\left(\tau = 0, \cdots, \lfloor \frac{T}{2T_5} \rfloor - 1\right)$. For each phase, we define a^s as a random variable taking values

$$a^s = \begin{cases} 1 & \left(\text{if } \sum_{t\in P^s}\sum_{k=0}^K \mathbb{1}[\|\nabla f(x^{t,k})\| > \varepsilon] > KT_5\right) \\ 2 & \left(\text{if there exists } t \text{ such that } (2sT_5,0) \leq (t,k) < ((2s+1)T_5, 0), \|\nabla f(x^{t,k})\| \leq \varepsilon \text{ and } \lambda_{\min}(\nabla^2 f(x^{t,k})) \leq -\delta\right) \\ 3 & \left(\text{if there exists } t \text{ such that } (2sT_5,0) \leq (t,k) < ((2s+1)T_5, 0), \|\nabla f(x^{t,k})\| \leq \varepsilon \text{ and } \lambda_{\min}(\nabla^2 f(x^{t,k})) > -\delta\right). \end{cases}$$

Note that $\mathbb{P}[a^s = 1,2,3] = 1$ for each s. This is because if there does not exist t between $(2\tau T_5, 0) \leq (t,k) < ((2s+1)T_5, 0)$ such that $\|\nabla f(x^{t,k})\| \leq \varepsilon$ (i.e., neither $a^s = 2$ nor 3), then we have $\sum_{t\in P^s}\sum_{k=0}^K \mathbb{1}[\|\nabla f(x^{t,k})\| > \varepsilon] \geq \sum_{t=2sT_5}^{(2\tau+1)T_5-1}\sum_{k=0}^K \mathbb{1}[\|\nabla f(x^{t,k})\| > \varepsilon] = T_5K$, meaning $a^s = 1$. We denote $N_1 = \sum_{s=0}^{\lfloor \frac{T}{2T_5} \rfloor} \mathbb{1}[a^s = 1]$, $N_2 = \sum_{s=0}^{\lfloor \frac{T}{2T_5} \rfloor} \mathbb{1}[a^s = 2]$, and $N_3 = \sum_{s=0}^{\lfloor \frac{T}{2T_5} \rfloor} \mathbb{1}[a^s = 3]$.

According to Lemma 21, with high probability over all s, it holds that if $a^s = 2$ then that phase successes escaping saddle points; i.e., there exists $(2\tau T_5, 0) \leq (t,k) < ((2\tau+1)T_5, 0)$ and

$$\max_{(t,k)\leq(t',k')<((2s+2)T_5,0)} \|x^{t',k'} - x^{t,k}\| > \frac{\delta}{C_3\rho} \tag{49}$$

holds. Eq. (49) further leads to

$$T_5K \sum_{t=2sT_5}^{2(s+1)T_5-1}\sum_{k=1}^K \|x^{t,k} - x^{t,k-1}\|^2 > \left(\frac{\delta}{C_3\rho}\right)^2 \Leftrightarrow \sum_{t=2sT_5}^{2(s+1)T_5-1}\sum_{k=1}^K \|x^{t,k} - x^{t,k-1}\|^2 > \frac{\delta^2}{T_5KC_3^2\rho^2}. \tag{50}$$

On the other hand, in Lemma 17 (high probability bound), we derived that

$$\sum_{t=1}^T\sum_{k=1}^K \|\nabla f(x^{t,k-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^T\sum_{k=1}^K \|x^{t,k} - x^{t,k-1}\|^2 \tag{51}$$

$$\leq \frac{2\Delta}{\eta} + \frac{C_1\sigma^2 T}{Pb} + C_1\mathbb{1}[\text{Option II}]K\left[\frac{\sigma_c^2 P}{p^2} + \frac{\sigma^2 P}{p^2 Kb}\right] + \frac{2r^2}{\eta^2}, \tag{52}$$

when $\eta = \tilde{\Theta}\left(\frac{1}{L} \vee \frac{p}{\zeta K\sqrt{P}} \wedge \frac{p\sqrt{b}}{L\sqrt{PK}} \wedge \frac{\sqrt{b}}{L\sqrt{K}} \wedge \frac{1}{\zeta K}\right)$.

From the definition of $a^s = 1$ and (50), We know that (51) is bounded as

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \|\nabla f(x^{t,k-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T}\sum_{k=1}^{K} \|x^{t,k} - x^{t,k-1}\|^2 \geq N_1 T_5 K\varepsilon^2 + \frac{\delta^2 N_2}{2\eta^2 T_5 K C_3^2 \rho^2}.$$

Thus, $N_1 T_5 K \leq \frac{1}{\varepsilon^2} \times$ (the right-hand side of (52)) and $N_2 T_5 K \leq \frac{2\eta^2 C_3^2 \rho^2 K^2 T_5^2}{\delta^2} \times$ (the right-hand side of (52)) holds. Here, $\frac{2\eta^2 T_5^2 K^2 C_3^2 \rho^2}{\delta^2} = \tilde{O}\left(\frac{\rho^2}{\delta^4} + \frac{\eta^2 K^2}{\delta^2}\right) = \tilde{O}\left(\frac{\rho^2}{\delta^4}\right)$, when $K = O\left(\frac{L}{\zeta}\right) \leq O\left(\frac{L}{\delta}\right)$. From this, $(N_1 + N_2)T_5 \leq \tilde{O}\left(\frac{1}{K\varepsilon^2} + \frac{\rho^2}{K\delta^4}\right) \times$ (the right-hand side of (52)). Taking $T \geq \tilde{O}\left(\frac{1}{K\varepsilon^2} + \frac{\rho^2}{K\delta^4}\right) \times$ (the right-hand side of (52)) $\geq 2(N_1 + N_2 + 1)T_5$, there exists s such that $a^s = 3$, which concludes the proof. □

E.3. Finding Second-Order Stationary Points When Clients are Homogeneous ($\zeta \ll \frac{1}{\delta}$)

In the previous subsection, we assumed that $\zeta \geq \frac{1}{\delta}$. Here, we introduce a simple trick to remove this assumption and give its convergence analysis.

Let $T_6 = \tilde{\Theta}\left(\frac{L}{\delta}\right)$ with a sufficiently large hidden constant. In line 15-16 of FL-SILVER, when $k \equiv T_6$, we randomly select sample $J_{i_t}^{t,k}$ of size $\frac{C_1 C_3^2 L^2}{\delta^2} + b > b$, and update $z^{t,k}$ as $z^{t,k} \leftarrow z^{t,k-1} + \frac{1}{|J_{i_t}^{t,k}|}\sum_{j \in J_{i_t}^{t,k}} (\nabla f_{i_t,j}(x^{t,k}) - \nabla f_{i_t,j}(x^{t,k-1}))$. This increases the number of gradient evaluations in each inner-loop by $\tilde{O}(K/(L/\delta)) \times \tilde{O}(L^2/\delta^2) = \tilde{O}(KL/\delta) \lesssim \tilde{O}(K^2) \lesssim \tilde{O}(Kb)$. Hence, this does not affect the inner-loop complexity more than by constant factors.

Then, the following lemma holds, which stands as generalization of Lemma 18.

Lemma 22 (Small stuck region)**.** *Let $\{x^{t,k}\}$ be a sequence generated by* FL-SILVER *and (τ_0, κ_0) be a step where $-\gamma := \lambda_{\min}(\nabla^2 f(x^{\tau_0,\kappa_0})) \leq -\delta$ holds. We denote the smallest eigenvector direction of $\lambda_{\min}(\nabla^2 f(x^{\tau_0,\kappa_0}))$ as* e*. Moreover, we define a coupled sequence $\{\tilde{x}^{t,k}\}$ by running* FL-SILVER *with $\tilde{x}^0 = x^0$ and the same choice of all randomness i.e., client samplings, minibatches and noises, but the noise at some step $(\tau, \kappa) > (\tau_0, \kappa_0)$, satisfying $\kappa \equiv T_6$;* We let $\tilde{\xi}^{\tau,\kappa} = \xi^{\tau,\kappa} - r_e$e with $r_e \geq \frac{r\nu}{\sqrt{d}}$. Let $w^{t,k} = x^{t,k} - \tilde{x}^{t,k}$, $w^t = x^t - \tilde{x}^t$, $g^{t,k} = \frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k}$, $\tilde{g}^t = \frac{1}{P}\sum_{i=1}^P \tilde{y}_i^{t-1} + z^{t,k}$, $h^t = \frac{1}{P}\sum_{i=1}^P y_i^t - \nabla f(x^t) - \left(\frac{1}{P}\sum_{i=1}^P \tilde{y}_i^t - \nabla f(\tilde{x}^t)\right)$, and $h^{t,k} = (z^{t,k} - (\nabla f(x^{t,k}) - \nabla f(x^{t,0}))) - (\tilde{z}^{t,k} - (\nabla f(\tilde{x}^{t,k}) - \nabla f(\tilde{x}^{t,0})))$. Then, $g^{t,k} - \nabla f(x^{t,k}) - (\tilde{g}^{t,k} - \nabla f(\tilde{x}^{t,k})) = h^{t-1} + h^{t,k}$.

There exists a sufficiently large constants $C_3 = \tilde{O}(1)$,, with which the following holds: If we take $p \geq C_1\sqrt{P} + \frac{C_1 C_3^2 \zeta^2}{\delta^2} + \frac{C_1 C_3^2 L^2}{Kb\delta^2}$, $b \geq \sqrt{K}$ and, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, with high probability, we have

$$\max_{(\tau_0,\kappa_0) \leq (t,k) < (\tau_0,\kappa_0+3T_6)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} \geq \frac{\delta}{C_3\rho}.$$

Proof of Lemma 22. We assume K is at least as large as $3T_6$. When $K - 2T_6 \leq \kappa_0 < K - 1$, taking $T_6 \geq T_4$ yields the assertion, considering the two coupled sequence initialized at (κ_0, K), according to a slight modification of Lemma 18.

Otherwise, we let (τ, κ) as the first step after (τ_0, κ_0) with $\kappa \equiv T_6$. Then, it suffice to show that, with high probability,

$$\max_{(\tau,\kappa) \leq (t,k) < (\tau,\kappa+T_6)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} \geq \frac{\delta}{C_3\rho}. \tag{53}$$

Since $K \geq 3T_6$ and $\kappa_0 < K - 2T_6$ imply $h^{t-1} = 0$ for all $(\tau, \kappa) \leq (t, k) < (\tau, \kappa + T_6)$, $h^{t-1} + h^{t,k} = h^{t,k}$ holds. Then, $\|h^{\tau,\kappa}\| = \left\|u_{i_t}^{\tau,\kappa} + \frac{1}{|J_{i_t}^{\tau,\kappa}|}\sum_{j \in J_{i_t}^{\tau,\kappa}} u_{i_t,j}^{\tau,\kappa}\right\| \leq \zeta r_e + \frac{L}{\sqrt{|J_{i_t}^{\tau,\kappa}|}} r_e \leq 2\delta r_e$, using Proposition 2. Moreover, for $(\tau, k) > (\tau, \kappa)$,

when we assume $\max_{(\tau,\kappa)\le(t,k)<(\tau,\kappa+T_6)}\{\|x^{\tau,k}-x^{\tau_0,\kappa_0}\|,\|\tilde{x}^{\tau,\kappa}-x^{\tau_0,\kappa_0}\|\}<\frac{\delta}{C_3\rho}$,

$$\begin{aligned}
&\|h^{\tau,k}\|\\
&=\left\|u_{i_\tau}^{\tau,k}+\sum_{l=\kappa}^{k}\frac{1}{|J_{i_t}^{\tau,l}|}\sum_{j\in J_{i_t}^{\tau,\kappa}}u_{i_t,j}^{\tau,l}\right\|\\
&\le\zeta\sum_{l=\tau}^{k}\|w^{\tau,l}-w^{\tau,l-1}\|+\frac{2\delta}{C_3}\|w^{\tau,k}\|+\frac{2\delta}{C_3}\|w^{\tau,0}\|+\delta r_e+\frac{C_1^2}{\sqrt{b}}\sqrt{\sum_{l=1}^{k}\left(L\|w^{\tau,k}-w^{\tau,k-1}\|+\frac{2\delta}{C_3}\|w^{\tau,k}\|+\frac{2\delta}{C_3}\|w^{\tau,k-1}\|\right)^2}.
\end{aligned}$$

Assuming that (a) $\frac{1}{2}(1+\eta\gamma)^{k-\kappa}r_e\le\|w^{t,k}\|\le 2(1+\eta\gamma)^{k-\kappa}r_e$ and (b) $\|w^{t,k}-w^{t,k-1}\|\le 3\eta\gamma(1+\eta\gamma)^{k-\kappa}r_e$ for $(\tau,\kappa)<(t,k)<(\tau,\kappa+T_6)$, we get $\|h^{t,k}\|\le\frac{2C_1\gamma}{C_3}(1+\eta\gamma)^{k-\kappa}$. Thus, following the discussion in Lemma 18 and taking T_6 as large as T_5, we have (53). □

Previously, we only focused on the noise at the last local step (κ_0,K). Thus, if the number of steps required to escape saddle points $T_5=\tilde{O}(\frac{L}{\delta})$ is smaller than the local steps $K=\tilde{O}(\frac{L}{\zeta})$, the algorithm sometimes have to wait more than T_4 steps for the last local step. Therefore, taking $K\ge T_4$ was useless to reduce the number of communication rounds (at least for the theoretical proof). On the other hand, based on Lemma 22, when FL-SILVER approaches a saddle point, FL-SILVER does not need to wait next communication, and can escape the stack region within $2T_6$ local steps, even if $T_6\ll K$. This allows to us to take K larger than $O(\frac{L}{\delta})$, and leads to removal of the assumption $\delta<\frac{1}{\zeta}$ from Theorem 6.

E.4. Convergence under PL condition

The full statement is as follows:

Theorem 7. *Suppose that Assumptions 1-(a), 2, 3-(a) (only for Option II), 3-(b), 5-(a), and 6 hold. Choose*

$$\eta=\Theta\left(\frac{1}{L}\wedge\frac{(\frac{p}{\sqrt{P}}\wedge 1)}{\zeta K}\wedge\frac{(\frac{p}{\sqrt{P}}\wedge 1)\sqrt{b}}{L\sqrt{K}}\right),$$

and assume that $r\le\eta\sqrt{\frac{\mu\varepsilon}{8}}\wedge\sqrt{\frac{\eta\varepsilon p}{64PK}}$ *and* $PKb\ge\frac{96\sigma^2}{\mu\varepsilon}\vee\frac{2304\sigma^4}{225L^2\varepsilon^2}$ *for finding* ε*-solutions in expectation and that* $r\le\frac{\eta\varepsilon}{4}$, $PKb\ge\frac{192\sigma^2}{\varepsilon^2}$ *for finding* ε*-first-order stationary points in expectation. Then, Algorithm 2 requires*

$$T=\Omega\left(\left[\frac{L}{\mu K}\vee\frac{P}{p}\vee\frac{\zeta(\frac{\sqrt{P}}{p}\vee 1)}{\mu}\vee\frac{L(\frac{\sqrt{P}}{p}\vee 1)}{\mu\sqrt{bK}}\right]\log\frac{\Delta+\mathbb{1}[\text{Option II}]PK^2\left[\frac{\sigma_c^2}{p^2}+\frac{\sigma^2}{p^2Kb}\right]}{\varepsilon}\right).$$

communication rounds and

$$\mathbb{1}[\text{Option I}]P+Tp=\Omega\left(\left[P\vee\frac{Lp}{\mu K}\vee\frac{\zeta(\sqrt{P}\vee p)}{\mu}\vee\frac{L(\sqrt{P}\vee p)}{\mu\sqrt{bK}}\right]\log\frac{\Delta+\mathbb{1}[\text{Option II}]PK^2\left[\frac{\sigma_c^2}{p^2}+\frac{\sigma^2}{p^2Kb}\right]}{\varepsilon}\right)$$

communication complexity to find ε*-solutions in expectation. Moreover, Algorithm 2 requires*

$$T=\Omega\left(\left[\frac{L}{\mu K}\vee\frac{P}{p}\vee\frac{\zeta(\frac{\sqrt{P}}{p}\vee 1)}{\mu}\vee\frac{L(\frac{\sqrt{P}}{p}\vee 1)}{\mu\sqrt{bK}}\right]\log\frac{\Delta+\mathbb{1}[\text{Option II}]PK^2\left[\frac{\sigma_c^2}{p^2}+\frac{\sigma^2}{p^2Kb}\right]}{\varepsilon}\right).$$

communication rounds and

$$\mathbb{1}[\text{Option I}]P+Tp=\Omega\left(\left[P\vee\frac{Lp}{\mu K}\vee\frac{\zeta(\sqrt{P}\vee p)}{\mu}\vee\frac{L(\sqrt{P}\vee p)}{\mu\sqrt{bK}}\right]\log\frac{\Delta+\mathbb{1}[\text{Option II}]PK^2\left[\frac{\sigma_c^2}{p^2}+\frac{\sigma^2}{p^2Kb}\right]}{\varepsilon}\right)$$

communication complexity to find ε*-first-order stationary points in expectation.*

Proof. Similarly to the proof of Theorem 3 (31), we have

$$
\begin{aligned}
&f(x^{t,k}) - f^* + \frac{\eta}{2}\|\nabla f(x^{t,k-1})\|^2 \\
&\leq (1-\frac{\eta\mu}{2})(f(x^{t,k-1}) - f^*) + \eta\|\nabla f(x^{t,k-1}) - (\frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^{t,k} - x^{t,k-1}\|^2 + \frac{r^2}{\eta}.
\end{aligned}
$$

By multiplying this by $(1-\alpha)^{TK-(t-1)K-k}$ for some $0 < \alpha \leq \frac{\eta\mu}{2}$ and summing up this over $(t,k) = (1,1)$ to (T,K), we get

$$
\begin{aligned}
&f(x^{T,K}) - f^* + \frac{\eta}{2}\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}\|\nabla f(x^{t,k-1})\|^2 \\
&\leq (1-\frac{\eta\mu}{2})(1-\alpha)^{TK-1}(f(x^0) - f^*) - \sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}\left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\left[\|x^{t,k} - x^{t,k-1}\|^2\right] \\
&+ \eta\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}\mathbb{E}\left[\|\nabla f(x^{t,k-1}) - (\frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k-1})\|^2\right] + \frac{r^2}{\eta}\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}.
\end{aligned}
$$

Note that $\frac{r^2}{\eta}\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)} \leq \frac{r^2}{\eta\alpha}$. Applying Lemma 10 yields

$$
\begin{aligned}
&\mathbb{E}\left[f(x^{t,k}) - f^* + \frac{\eta}{2}\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}\|\nabla f(x^{t,k-1})\|^2\right] \\
&\leq (1-\frac{\eta\mu}{2})(1-\alpha)^{TK-1}(f(x^0) - f^*) \\
&\quad - \sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}\left(\frac{1}{2\eta} - \frac{L}{2} - \eta\left[\left(\sum_{s=t+1}^T (1-\frac{p}{4P})^{s-t-1}(1-\alpha)^{-(s-t)K}K\right)\left(\frac{180\zeta^2 K}{p} + \frac{12L^2}{pb}\right)\right.\right. \\
&\quad \left.\left. + (1-\alpha)^{-K}K\left(\frac{6L^2}{b} + 6\zeta^2 K\right)\right]\right)\mathbb{E}\left[\|x^{t,k} - x^{t,k-1}\|^2\right] + \eta\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}\frac{12\sigma^2}{PKb} \\
&\quad + \eta\mathbb{1}[\text{Option II}]\sum_{t=1}^T\sum_{k=1}^K (1-\alpha)^{TK-((t-1)K+k)}K(1-\frac{p}{P})^{t-1}\left[\frac{54\sigma_c^2}{p} + \frac{6\sigma^2}{pKb}\right] + \frac{r^2}{\eta\alpha}.
\end{aligned} \tag{54}
$$

We let

$$
\eta \leq \frac{1}{5L} \wedge \frac{p}{\zeta K\sqrt{14400P}} \wedge \frac{p\sqrt{b}}{L\sqrt{960KP}} \wedge \frac{\sqrt{b}}{L\sqrt{60K}} \wedge \frac{1}{\sqrt{60}\zeta K} \tag{55}
$$

and

$$
\alpha = \frac{\mu\eta}{2} \wedge \frac{p}{16PK}.
$$

Then, we have $(1-\frac{p}{4P})^{s-t-1}(1-\alpha)^{-(s-t)K}K \leq 2(1-\frac{p}{8P})^{s-t-1}K$ and $(1-\alpha)^{-K} \leq 2$. Therefore, the coefficient of $\mathbb{E}[\|x^{t,k} - x^{t,k-1}\|^2]$ in (54) is bounded by

$$
\begin{aligned}
&-\frac{1}{2\eta} + \frac{L}{2} + \eta\left[\left(\sum_{s=t+1}^T (1-\frac{p}{4P})^{s-t-1}(1-\alpha)^{-(s-t)K}K\right)\left(\frac{180\zeta^2 K}{p} + \frac{12L^2}{pb}\right) + (1-\alpha)^{-K}K\left(\frac{6L^2}{b} + 6\zeta^2 K\right)\right] \\
&\leq -\frac{1}{2\eta} + \frac{L}{2} + \frac{16\eta PK}{p}\left(\frac{180\zeta^2 K}{p} + \frac{12L^2}{pb}\right) + 2\eta K\left(\frac{6L^2}{b} + 6\zeta^2 K\right) \\
&\leq 0.
\end{aligned}
$$

Now (54) is written as

$$\begin{aligned}
&\mathbb{E}\left[f(x^{t,k})-f^*+\frac{\eta}{2}\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\alpha)^{TK-((t-1)K+k)}\|\nabla f(x^{t,k-1})\|^2\right]\\
&\le(1-\frac{\eta\mu}{2})(1-\alpha)^{TK-1}(f(x^0)-f^*)+\eta\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\alpha)^{TK-((t-1)K+k)}\frac{12\sigma^2}{PKb}\\
&\quad+\eta\mathbb{1}[\text{Option II}]\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\alpha)^{TK-((t-1)K+k)}K(1-\frac{p}{P})^{t-1}\left[\frac{54\sigma_c^2}{p}+\frac{6\sigma^2}{pKb}\right]+\frac{r^2}{\eta\alpha}\\
&\le(1-\alpha)^{TK}\Delta+\frac{\eta}{\alpha}\frac{12\sigma^2}{PKb}+\eta\mathbb{1}[\text{Option II}](1-\alpha)^{TK}\frac{2PK^2}{p}\left[\frac{54\sigma_c^2}{p}+\frac{6\sigma^2}{pKb}\right]+\frac{r^2}{\eta\alpha},
\end{aligned}\tag{56}$$

For the final inequality, we used $(1-\alpha)^{TK-((t-1)K+k)}K(1-\frac{p}{P})^{t-1}\le 2K(1-\alpha)^{TK}(1-\frac{p}{2P})^{t-1}$ when $\alpha\le\frac{p}{16P}$.

In order to find an ε-solution (i.e., $f(x^{t,k})-f^*\le\varepsilon$) in expectation, we take $r\le\sqrt{\frac{\alpha\eta\varepsilon}{4}}=\eta\sqrt{\frac{\mu\varepsilon}{8}}\wedge\sqrt{\frac{\eta\varepsilon p}{64PK}}$, $PKb\ge\frac{48\eta\sigma^2}{\alpha\varepsilon}$ (, which holds when $PKb\ge\frac{96\sigma^2}{\mu\varepsilon}\vee\frac{2304\sigma^4}{225L^2\varepsilon^2}$ with $\eta\le\frac{p\sqrt{b}}{L\sqrt{960KP}}$), and

$$T\ge\frac{1}{K\alpha}\log\frac{4\Delta}{\varepsilon}+\frac{\mathbb{1}[\text{Option II}]}{K\alpha}\log\frac{8PK^2\left[\frac{54\sigma_c^2}{p^2}+\frac{6\sigma^2}{p^2Kb}\right]}{\varepsilon}.$$

Especially, when the equality in (55) hold, i.e.,

$$\eta=\Theta\left(\frac{1}{L}\wedge\frac{(\frac{p}{\sqrt{P}}\wedge 1)}{\zeta K}\wedge\frac{(\frac{p}{\sqrt{P}}\wedge 1)\sqrt{b}}{L\sqrt{K}}\right),\tag{57}$$

T should be taken as

$$T=\Omega\left(\left[\frac{L}{\mu K}\vee\frac{P}{p}\vee\frac{\zeta(\frac{\sqrt{P}}{p}\vee 1)}{\mu}\vee\frac{L(\frac{\sqrt{P}}{p}\vee 1)}{\mu\sqrt{bK}}\right]\log\frac{\Delta+\mathbb{1}[\text{Option II}]PK^2\left[\frac{\sigma_c^2}{p^2}+\frac{\sigma^2}{p^2Kb}\right]}{\varepsilon}\right).$$

Moreover, to find ε-first-order stationary points, we transform (56) as

$$\begin{aligned}
&\mathbb{E}\left[\min_{\substack{1\le t\le T\\1\le k\le K}}\|\nabla f(x^{t,k-1})\|^2\right]\\
&\le 4\eta^{-1}\alpha(1-\alpha)^{TK}\Delta+\frac{48\sigma^2}{PKb}+4\alpha\mathbb{1}[\text{Option II}](1-\alpha)^{TK}\frac{2PK^2}{p}\left[\frac{54\sigma_c^2}{p}+\frac{6\sigma^2}{pKb}\right]+\frac{4r^2}{\eta^2}.
\end{aligned}$$

Note that we used $\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\alpha)^{TK-((t-1)K+k)}\|\nabla f(x^{t,k-1})\|^2\ge\frac{1}{2\alpha}\min_{1\le t\le T,1\le k\le K}\|\nabla f(x^{t,k-1})\|^2$ when $TK\ge\alpha$ and $\alpha\le\frac{\mu}{10L}\le\frac{1}{10}$ (we have already let η satisfy $\eta\le\frac{1}{5L}$ and will later let T to satisfy this below). We let $r\le\frac{\eta\varepsilon}{4}$, $PKb\ge\frac{192\sigma^2}{\varepsilon^2}$, and

$$T\ge\frac{1}{\alpha K}\log\frac{32\alpha\Delta}{\varepsilon\eta}+\frac{\mathbb{1}[\text{Option II}]}{\alpha K}\log\frac{32\alpha PK^2\left[\frac{54\sigma_c^2}{p^2}+\frac{6\sigma^2}{p^2Kb}\right]}{\varepsilon}.$$

(Note that this implies $TK\ge\alpha$.) When we take η as (57), T becomes

$$T=\Omega\left(\left[\frac{L}{\mu K}\vee\frac{P}{p}\vee\frac{\zeta(\frac{\sqrt{P}}{p}\vee 1)}{\mu}\vee\frac{L(\frac{\sqrt{P}}{p}\vee 1)}{\mu\sqrt{bK}}\right]\log\frac{\Delta+\mathbb{1}[\text{Option II}]PK^2\left[\frac{\sigma_c^2}{p^2}+\frac{\sigma^2}{p^2Kb}\right]}{\varepsilon}\right).$$

□

F. Lower bound

Proposition 1 can be derived by using the bounds of Zhou and Gu (2019); Fang et al. (2018); Li et al. (2021a). First, we give a definition of a linear-span first-order algorithm.

Definition 1 (Linear-span first-order algorithm)**.** *Fix some x^0. Let $\mathcal{A}$ be a (randomized) algorithm with the initial point x^0, and x^t be the point at the t-th iteration. We assume $\mathcal{A}$ select one individual function i_t at each iteration t and computes $\nabla f_{i_t}(x^t)$. Then $\mathcal{A}$ is called a linear-span first-order algorithm if*

$$x^t \in \mathrm{span}\{x^0, x^1, \cdots, x^{t-1}, \nabla f_{i_0}(x^0), \nabla f_{i_1}(x^1), \cdots, \nabla f_{i_{t-1}}(x^t)\}$$

holds for all t with probability one.

Note that this definition includes minibatch updade, by letting $x^{sb} = x^{sb+1} =, \cdots, = x^{(s+1)b-1}$ with the minibatch size b.

We also define problem classes $\mathcal{F}^L_{n,\Delta}$ and $\mathcal{F}^{L,\zeta}_{n,\Delta}$ for (1), as follows.

Definition 2 (A class of finite-sum optimization problems)**.** *Fix some x^0. For an integer n, $L > 0$, we define a problem class $\mathcal{F}^L_n$ as*

$$\mathcal{F}^L_{n,\Delta} = \left\{ f = \frac{1}{n}\sum_{i=1} f_i \colon \mathbb{R}^d \to \mathbb{R} \;\middle|\; \begin{array}{l} d \in \mathbb{N}.\ \text{Each } f_i \colon \mathbb{R}^d \to \mathbb{R} \text{ is } L\text{-gradient Lipschitz (Assumption 1-(b)),} \\ \text{and } f(x^0) - \inf_x f(x) = \Delta. \end{array} \right\}$$

Moreover, for an integer n, $L > 0$, and $\zeta > 0$, a problem class $\mathcal{F}^{L,\zeta}_n$ is defined as

$$\mathcal{F}^{L,\zeta}_n = \left\{ f = \frac{1}{n}\sum_{i=1}^n f_i \colon \mathbb{R}^d \to \mathbb{R} \;\middle|\; \begin{array}{l} d \in \mathbb{N}.\ \text{Each } f_i \colon \mathbb{R}^d \to \mathbb{R} \text{ is } L\text{-gradient Lipschitz (Assumption 1-(b))} \\ \text{and Hessian-heterogeneous with } \zeta \text{ (Assumption 4-(b)),} \\ \text{and } f(x^0) - \inf_x f(x) = \Delta. \end{array} \right\}.$$

Note that function-wise gradient Lipschitzness (Assumption 1-(b)) and Hessian Heterogeneity (Assumption 4-(b)) are stronger than averaged gradient Lipschitzness (Assumption 1-(a)) and Hessian Heterogeneity (Assumption 4-(a)).

Carmon et al. (2020) proved the following lower bound.

Proposition 6 (Carmon et al. (2020))**.** *Fix x^0. For any $L > 0$, $\Delta > 0$, and $\varepsilon > 0$, there exists a function $f \in \mathcal{F}^L_{1,\Delta}$ such that any linear-span first-order algorithm requires $\Omega\left(\frac{\Delta L}{\varepsilon^2}\right)$ stochastic gradient accesses in order to find ε-first-order stationary points.*

Zhou and Gu (2019); Fang et al. (2018); Li et al. (2021a) extended this to the lower bound on the finite-sum optimization problem.

Proposition 7 (Zhou and Gu (2019); Fang et al. (2018); Li et al. (2021a))**.** *Fix x^0. For $n > 0$, $L > 0$, $\Delta > 0$, and $\varepsilon > 0$, there exists a function $f \in \mathcal{F}^L_{n,\Delta}$ such that any linear-span first-order algorithm requires $\Omega\left(n + \frac{\Delta L\sqrt{n}}{\varepsilon^2}\right)$ stochastic gradient accesses in order to find ε-first-order stationary points.*

Based on these, we give the lower bound under the additional assumption of ζ-Hessian-heterogeneity.

Proposition 1. *Suppose that 1-(b), 2, 5-(b) hold. For any $L > 0$, $\Delta > 0$, and $\varepsilon > 0$, there exists a function $f \in \mathcal{F}^{L,\zeta}_{n,\Delta}$ such that any linear-span first-order algorithm requires*

$$\Omega\left(n + \frac{\Delta(\zeta\sqrt{n} + L)}{\varepsilon^2}\right)$$

stochastic gradient accesses in order to find ε-first-order stationary points.

Proof. It is easy to see that the lower bound of Proposition 6 also applies to $\mathcal{F}^L_{n,\Delta}$, by letting $f_1 = f_2 = \cdots = f_n = f^*$ where f^* is the function that gives the bound of Proposition 6. On the other hand, we have $\mathcal{F}^{\frac{\zeta}{2}}_{n,\Delta} \subseteq \mathcal{F}^{L,\zeta}_{n,\Delta}$. Thus, Proposition 7 yields that there exists a function $f \in \mathcal{F}^{\frac{\zeta}{2}}_{n,\Delta} \subseteq \mathcal{F}^{L,\zeta}_{n,\Delta}$ that requires $\Omega\left(n + \frac{\Delta\zeta\sqrt{n}}{\varepsilon^2}\right)$ stochastic gradients to find ε-first-order stationary points. Therefore, by combining these two bounds, we have the desired lower bound. □

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]