Reducing Communication in Federated Learning with a Novel Single-Loop Variance Reduction Method

author names withheld

Under Review for OPT 2022

Abstract

In Federated Learning (FL), inter-client heterogeneity causes two types of errors: (i) *client drift error* which is induced by multiple local updates, (ii) *client sampling error* due to partial participation of clients at each communication. While several solutions have been offered to the former one, there is still much room of improvement on the latter one.

We provide a fundamental solution to this client sampling error. The key is a novel single-loop variance reduction algorithm, SLEDGE (Single-Loop mEthoD for Gradient Estimator), which does not require periodic computation of full gradient but achieves optimal gradient complexity in the nonconvex finite-sum setting. While sampling a small number of clients at each communication round, the proposed FL algorithm, FLEDGE, requires provably fewer or at least equivalent communication rounds compared to any existing method, for finding first and even second-order stationary points in the general nonconvex setting, and under the PL condition. Moreover, under less Hessian-heterogeneity between clients, the required number of communication rounds approaches to $\tilde{\Theta}(1)$.

1. Introduction

Federated learning (FL) is a paradigm of distributed learning, where each local client has access to local dataset to train a local model, and periodically local model parameters are exchanged to update a global model in the server [20, 29, 41]. Avoiding share of local data itself, FL provides privacy protection and encourages the usage of distributed Big Data [16, 22].

Since clients such as smartphones and organizations are physically separated, the main bottlenecks in FL are synchronization and communication between clients and the server.

To reduce the number of communication rounds, local update has been adopted [23, 27, 29]. Local update means that, between communication rounds, the parameters of a local model are updated several times inside each client using only its local data, and then aggregated in the server at the next communication round. Local update causes client drift error, but if that is carefully corrected, local update can provably reduce communication rounds, especially when clients have less heterogeneity [19, 30], which had experimentally been observed [18, 29, 47].

On the other hand, client sampling, meaning that not all but only a part of clients are sampled to participate in each communication, is also widely used to reduce the communication complexity (the total number of parameters communicated). Due to huge parameter size, a large number of clients, low bandwidth communication, and increase of communication cost for secure computation, which are characteristics of FL, client sampling has become indispensable in designing FL algorithms [16].

However, while a mass of literature has established treatments for client drift error [18, 19, 30], client sampling error has not effectively been controlled until recently. In fact, the convergence rate

© .

of FedAvg [29], one of the most famous FL algorithms, can be dominated by client sampling error rather than client drift error [13], and we can see that this is also true for other FL algorithms. Especially, in order to reduce communication rounds by taking advantage of less client heterogeneity with local update, client sampling is not allowed [30] or increasingly larger sampling size is required as heterogeneity gets smaller [19], due to the client sampling error.

Recently, FedVarp [13] tackled client sampling error applying a variance reduced method of SAGA [8, 38] to FedAvg [29]. Variance reduction is a technique to construct a gradient estimator with a smaller variance than vanilla SGD by recursively utilizing minibatch gradients at previously obtained anchor points [8, 15, 39]. Although we agree that applying variance reduction should be a right direction, a large part of the problem still remains. In fact, their algorithm requires relatively large client sample size of $O(P^{\frac{2}{3}})$ to the total number of clients P, and cannot take advantage of heterogeneity, resulting in sub-optimal communication rounds and complexity compared to the state-of-the-art FL methods [19, 30].

We considered that this is because existing variance reduction methods are not suitable for applying FL methods. In general, variance reduction methods are measured only in terms of gradient complexity. However, when applying them to FL methods, we additionally want single-loop structure (, that is, not to require periodic full or large minibatch gradient since this leads to impractical full client participation in FL), and fewer gradient complexity under less heterogeneity. From these perspectives, existing methods are not satisfactory. Indeed, SAGA satisfies single-loop structure, but not the others. We consider the limitations of the aforementioned work of Jhunjhunwala et al. [13] came from this point. As for other variance reduction methods, SARAH [33], SPIDER [9], and NestedSVRG [48] requires periodic full gradient. STORM [7] is sub-optimal in the nonconvex finite-sum setting, and its application to FL, MimeMVR [19], requires larger sampling size as heterogeneity gets smaller as mentioned; ZeroSARAH [26] cannot benefit from less heterogeneity, and its distributed version does not considers local update. See also Appendix A.1. Thus, we must design a novel variance reduction algorithm for FL.

1.1. Contributions

We consider FL problems with finite clients, and so variance reduction in the finite-sum setting.

First, we developed a novel single-loop variance reduction method called SLEDGE (Single-Loop mEthoD for Gradient Estimator) for the nonconvex finite-sum problems. SLEDGE does not require periodic computation of full gradients, and satisfies the followings: (i) nearly optimal gradient complexity of $\tilde{O}(\frac{\sqrt{n}}{\varepsilon^2})$ for finding ε-first-order stationary points with data size n, (ii) second-order optimality as the first such single-loop algorithm, (iii) exponential convergence under the Polyak-Łojasiewicz (PL) condition, and (iv) fewer complexity under the less heterogeneity assumption.

Next, we combined SLEDGE with local updates into an efficient federated algorithm, FLEDGE. FLEDGE appropriately controls client sampling error, and achieves the followings, with the inter-client heterogeneity ζ and the number of local updates K: (i) For first-order stationary points, the number of required communication rounds is $\tilde{O}(\frac{1}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2})$, when the number of sampled clients at each step p is larger than $\sqrt{P}$. (ii) Adding small perturbation, FLEDGE also can find SOSPs. (iii) Under the μ-PL condition, FLEDGE exhibits exponential convergence, and the number of communication rounds depends on μ only through $\frac{L}{K\mu} + \frac{\zeta}{\mu}$. For finding first and second order stationary points, our rates are smaller than or equivalent to those of all existing FL methods across all range of the inter-client heterogeneity ζ. As $\zeta \to 0$, these rates approach to $\tilde{O}(1)$ under taking $p \simeq \sqrt{P}$,

while previous methods require full client participation [30, 31] or increasingly larger client sample size [19]. For PL (and even strongly convex) functions, FLEDGE breaks the dependency on the condition number when $\zeta \ll L$ and $K \gg 1$, for the first time. For detailed comparison of existing FL methods and discussion on required local budget, see Appendix A.2.

2. SLEDGE: Single-Loop Method for Gradient Estimator

We formalize the finite-sum problems as follows:

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

Our goal is to find a solution x that is an ε-first-order stationary point (i.e., $\|\nabla f(x)\| \le \varepsilon$), and an (ε, δ)-second-order stationary point (SOSP; i.e., $\|\nabla f(x)\| \le \varepsilon$ and $\lambda_{\min}(\nabla^2 f(x)) \ge -\delta$).

Throughout the section, we assume that f_i is L-smooth, and f is bounded: $f(x^0) - f^* =: \Delta < \infty$ with $f^* = \inf_x f(x)$. Moreover, we assume Hessian-heterogeneity: $\|\nabla^2 f_i(x) - \nabla^2 f_j(x)\| \le \zeta$ (note: $\zeta \le 2L$ holds). For Option I, we suppose $\|\nabla f(x^0) - \nabla f(x^0)\| \le \sigma_c$. For SOSPs, let $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \le \rho \|x - y\|$ hold. The μ-PL condition means $2\mu(f(x) - f^*) \le \|\nabla f(x)\|^2$.

2.1. Algorithm description

We introduce our proposed method SLEDGE for the problem (1). Note that $v^t, v_i^t,$ and $\tilde{y}_i^t$ in parentheses are auxiliary variables with which $\frac{1}{n}\sum_{i=1}^n y_i^t$ is updated in $O(b)$ time, utilizing the following equation: $\frac{1}{n}\sum_{i=1}^n y_i^t = \frac{1}{n}\sum_{i=1}^n y_i^{t-1} + \frac{1}{n}\sum_{i \in I^t}(\frac{n}{b}\nabla f_i(x^t) - \frac{n-b}{b}\nabla f_i(x^{t-1}) - \tilde{y}_i^{t-1} - v^{t-1} + v_i^{t-1})$.

Algorithm 1 SLEDGE(x^0, η, b, T, r)

Option I: Randomly sample b data I^0 and $y_i^0 \leftarrow \frac{1}{b}\sum_{j \in I^0} \nabla f_j(x^0)$ for $i \in I$
Option II: $y_i^0 \leftarrow \nabla f_i(x^0)$ for $i \in I_0 = I$
($v^0 \leftarrow 0$ and $\tilde{y}_i^0 \leftarrow y_i^0$, $v_i^0 \leftarrow 0$ for $i \in I$)
For $t = 1$ to T **do**
  $x^t \leftarrow x^{t-1} - \frac{\eta}{n}\sum_{i=1}^n y_i^{t-1} + \xi^t$ (ξ^t follows the uniform distribution on the Euclidean ball in $\mathbb{R}^d$ with radius r)
  Randomly sample b data I^t
  $y_i^t \leftarrow \begin{cases} \nabla f_i(x^t) & \text{for } i \in I^t \\ \frac{1}{b}\sum_{j \in I^t}(\nabla f_j(x^t) - \nabla f_j(x^{t-1})) + y_i^{t-1} & \text{for } i \notin I^t \end{cases}$
  ($v^t \leftarrow \frac{1}{b}\sum_{j \in I^t}(\nabla f_j(x^t) - \nabla f_j(x^{t-1})) + v^{t-1}$, $\tilde{y}_i^t \leftarrow y_i^t$, $v_i^t \leftarrow v^t$ for $i \in I$ and $\tilde{y}_i^t \leftarrow y_i^{t-1}$, $v_i^t \leftarrow v_i^{t-1}$ for $i \notin I$)

2.2. Convergence analysis

SLEDGE is designed so that it inherits the best points of SAGA [8, 38] and SARAH [32, 33]. It removes periodic full gradient using stored past gradients as SAGA. Since SAGA's gradient complexity is suboptimal, we also import SARAH's recursive update to achieve the optimal rate. The noise ξ^t is to guarantee the second-order optimality.

The key observation in the analysis is that, for Option II, the discrepancy between the estimator and the true gradient is decomposed as

$$\frac{1}{n}\sum_{i=1}^n y_i^t - \nabla f(x^t) = \frac{1}{n}\sum_{s=1}^t \left(\frac{|\tilde{I}_s^t|}{b} \sum_{i \in I^s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) - \sum_{i \in \tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) \right).$$

Here, $\tilde{I}_s^t = [n] \setminus \bigcup_{\tau=s}^t I^t$ is the set of indexes not sampled between s and t. Conditioning on $|\tilde{I}_s^t|$, the inside of the large parentheses for each s is roughly $\zeta\sqrt{\frac{|\tilde{I}_s^t|}{b}}\|x^s - x^{s-1}\|$. The main technical difficulty is that the inside of the large parentheses is not independent each other since $|\tilde{I}_s^t|$ depends on all $I^s, \dots, I^t$. However, the correlation can be shown to be sufficiently weak.

Now we state the theoretical guarantee for SLEDGE. For the formal version, see Appendix D.

Theorem 1 *We take a step size η and a scale of noise r appropriately, and let $\mu \in (0,1)$. Then, with probability $1-\nu$, SLEDGE finds ε-first-order stationary points using*

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

stochastic gradients. For finding SOSPs with probability $1-\nu$, SLEDGE requires $b \gtrsim \sqrt{n} + \frac{\zeta^2}{\delta^2}$ and

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

stochastic gradients. Under the μ-PL condition, SLEDGE uses

$$\tilde{O}\left(\left(\frac{Lb}{\mu} \vee \frac{\zeta\sqrt{n}}{\mu} \vee n\right)\log\frac{\Delta + \sigma_c}{\varepsilon}\right) \text{(Option I)}, \quad \tilde{O}\left(\left(\frac{Lb}{\mu} \vee \frac{\zeta\sqrt{n}}{\mu} \vee n\right)\log\frac{\Delta}{\varepsilon}\right) \text{(Option II)}$$

stochastic gradients for finding ε-solutions with $f(x^t) - f^* \le \varepsilon$, with probability $1-\nu$.

Without periodic computation of full gradient, SLEDGE achieves nearly optimal gradient complexity for first-order stationary points, and can take advantage of less heterogeneity ζ. We show that the rate of $\tilde{O}(n + \frac{\Delta(\zeta\sqrt{n} \wedge Lb)}{\varepsilon^2})$ (option II) is near-optimal, see Appendix F. Moreover, SLEDGE can finds SOSPs as the first single-loop algorithm, and exponentially converge under the PL condition.

3. FLEDGE: Federated Learning Method with Gradient Estimator

Indexing (finite) clients and data by i and j, we consider the following FL problem:

$$\min_{x \in \mathbb{R}^d}\left\{f(x) := \frac{1}{P}\sum_{i=1}^P f_i(x)\right\}, \quad f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{i,j}(x). \tag{2}$$

In addition to the previous assumptions on f and f_i, we additionally assume that each $f_{i,j}$ is L-smooth. We bound the intra-client variance as $\|\nabla f_{i,j}(x) - \nabla f_i(x)\| \le \sigma$. For finding SOSPs, we suppose that $\|\nabla^2 f_{i,j}(x) - \nabla^2 f_{i,j}(y)\| \le \rho\|x-y\|$ for all i, j and x, y.

We combine SLEDGE with local update into a novel FL method, FLEDGE. Here, we only introduce FLEDGE with Option II and for the case of $\zeta > \delta$. For others, see Appendix E.

Theorem 2 *We take an appropriate step size η and a scale of noise r and let $\mu \in (0,1)$. Let $b \gtrsim \frac{\sigma^2}{PK\varepsilon^2}$. Then, with probability $1-\nu$, FLEDGE finds ε-first-order stationary points using*

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

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

for $i \in I^0 = I$ in parallel **do**
  Randomly select minibatch J_i^0 with size Kb and let $y_i^0 \leftarrow \frac{1}{bK}\sum_{j\in J_i^0}\nabla f_{i,j}(x^0)$
for $t = 1$ to T **do**
  Randomly sample one client i_t and Send $\frac{1}{P}\sum_{i=1}^P y_i^{t-1}$ 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**
    $x^{t,k} \leftarrow x^{t,k-1} - \eta(\frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k-1}) + \xi^{t,k}$
    (ξ^t follows the uniform distribution on the Euclidean ball in $\mathbb{R}^d$ with radius r)
    Randomly select minibatch $J_{i_t}^{t,k}$ with size b
    $z^{t,k} \leftarrow z^{t,k-1} + \frac{1}{b}\sum_{j\in J_{i_t}^{t,k}}(\nabla f_{i_t,j}(x^{t,k}) - \nabla f_{i_t,j}(x^{t,k-1}))$
  Randomly select p clients I^t, send $x^{t,K}$ from i_t to I^t and let $x^t \leftarrow x^{t,K}$
  for $i \in I^t$ in parallel **do**
    Randomly select minibatch J_i^t with size Kb
    $y_i^t \leftarrow \frac{1}{bK}\sum_{j\in J_i^t}\nabla f_{i,j}(x^t)$ and $\Delta y_i^t \leftarrow \frac{1}{bK}\sum_{j\in J_i^t}(\nabla f_{i,j}(x^t) - \nabla f_{i,j}(x^{t-1}))$
  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 only update $\frac{1}{P}\sum_{i=1}^P y_i^t$ in $O(p)$ time as in SLEDGE)

communication rounds. For finding SOSPs, FLEDGE requires $p \gtrsim \sqrt{P} + \frac{\zeta^2}{\delta^2}, b \geq K$, *and*

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

communication rounds, with probability $1-\nu$. *Under the* μ*-PL condition, FLEDGE uses*

$$\tilde{O}\left(1 + \left(\frac{\zeta}{\mu} \vee \frac{L}{\mu K} \vee \frac{L}{\mu\sqrt{Kb}} \vee \frac{\zeta\sqrt{P}}{\mu p} \vee \frac{P}{p}\right)\log\frac{\Delta}{\varepsilon}\right)$$

communication rounds for finding ε*-first-order stationary points, with probability* $1-\nu$.

Setting client sample size $p \geq \sqrt{P}$ and local minibatch size $b \geq K$, FLEDGE requires only $\tilde{O}(\frac{\zeta}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2})$ communication rounds for first-order stationary points. Compared to FedVarp [13], ours require smaller client sample size and can take advantage of less heterogeneity. What is more, as far as assumptions on the local minibatch size are satisfied, this rate is better than or equivalent to *any* existing federated learning algorithm and approaches $\tilde{\Theta}(1)$ when $\zeta \to 0$ and $K \to \infty$. In terms of communication complexity, $\tilde{O}(P + \frac{\zeta\sqrt{P}}{\varepsilon^2})$ is optimal in a sense that it almost matches the lower bound of gradient complexity of the finite-sum case. We will show this in Appendix F.

Moreover, FLEDGE guarantees second-order optimality. To the best of our knowledge, the only such method is BVR-P-LSGD [31], but that synchronizes all clients at every round.

Furthermore, the algorithm requires $\tilde{O}(\frac{\zeta}{\mu}\log\frac{\Delta}{\varepsilon})$ communication rounds under μ-PL condition. Thus, even if the condition number $\frac{L}{\mu}$ is bad, the required communication rounds goes to $\tilde{O}(\frac{P}{p})$ when $\zeta \to 0$. On the other hand, even in the strongly-convex case and without client sampling, all existing algorithms require $\Omega(\frac{L}{\mu}\log\frac{1}{\varepsilon})$ rounds regardless of the Hessian heterogeneity ζ.

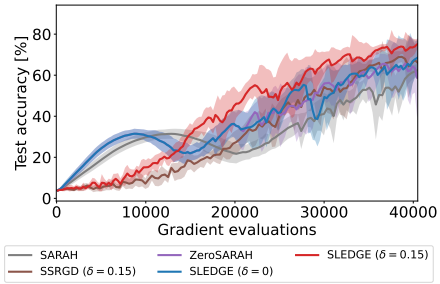


Figure 1: Comparison of the test accuracy (SLEDGE, finite-sum)

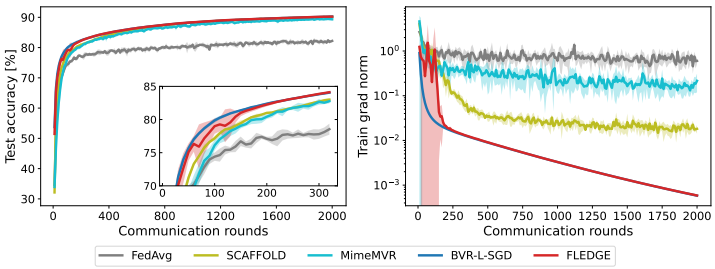


Figure 2: Comparison of the test accuracy (left) and the gradient norm (right) (FLEDGE, federated learning)

4. Experiments

Finally, we validate our theories on both SLEDGE and FLEDGE by numerical experiments. We considered classification of the capital letters using EMNIST dataset [6] for both experiments.

Escaping saddle points with SLEDGE For the finite-sum problem (1), we prepared each f_i by sampling 100 data from one class, employing a four-layer neural network as the training model, and then defining the average of the cross-entropy loss over the data as f_i. We repeated this five times for each class, and thus $n = 130$. We compared SLEDGE and its perturbed version with SARAH [32, 35], SSRGD [24], and ZeroSARAH [26], in terms of the test accuracy. We set $b = 12$, the inner-loop length of SARAH and SSRGD to 10, and $\lambda = \frac{b}{n} \fallingdotseq 0.092$ for ZeroSARAH. For SSRGD and SLEDGE, we added perturbation of $\delta = 0.15$. We tuned the learning rate for each algorithm individually. The experiment was repeated with ten different random seeds for each method.

Figure 1 shows the result. We can observe that (i) SLEDGE and ZeroSARAH require fewer gradient evaluations than SARAH to achieve the same test accuracy, owing to avoidance of periodic full gradient. Similarly, SLEDGE with small noise is faster than SSRGD. (ii) Adding small noise helps stable convergence; Although SLEDGE with $\delta = 0$ does not necessarily yield a monotonic increase in the accuracy (see around 10000-15000 gradient evaluations), SLEDGE with small noise perturbation makes the accuracy increase almost monotonically.

Faster Convergence with FLEDGE For the federated learning problem (2), we again consider the classification of the capital letters, where each f_i consists of 90% data from one class and 10% data from the other classes. This makes each f_i a little less heterogeneous. We used two-leyer neural networks. We compared FLEDGE with FedAvg [29], SCAFFOLD [18], MimeMVR [19], and BVR-L-SGD [30]. For each algorithm, we employed $P = 104$ as the total number of clients and $p = 10$ as the number of the clients used at each communication (except for BVR-L-SGD, which requires $P = p = 104$). Then, we set $b = 16$ and $K = 10$. We tuned the learning rate for each algorithm individually. The experiment was repeated with five different random seeds for each method.

Figure 2 (left) shows that FLEDGE achieves the high test accuracy with fewer communication, compared to FedAvg, SCAFFOLD, and MimeMVR. In Figure 2 (right), FLEDGE achieves the small gradient norm $\|\nabla f(x^t)\|$ and the linear convergence at the neighborhood of solutions. Moreover, we observe that FLEDGE performs similarly to BVR-L-SGD, which almost can be seen as a special case of FLEDGE with $p = P$. This means that FLEDGE 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.

References

[1] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Aleksandr Beznosikov and Martin Takáč. Random-reshuffled sarah does not need a full gradient computations. *arXiv preprint arXiv:2111.13322*, 2021.

[4] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

[5] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127, 2006.

[6] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[7] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

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

[9] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

[10] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[11] Rong Ge, Zhize Li, Weiyao Wang, and Xiang Wang. Stabilized svrg: Simple variance reduction for nonconvex optimization. In *Conference on learning theory*, pages 1394–1448. PMLR, 2019.

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

[13] Divyansh Jhunjhunwala, PRANAY SHARMA, Aushim Nagarkatti, and Gauri Joshi. Fedvarp: Tackling the variance due to partial client participation in federated learning. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[14] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[15] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

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

[17] Hamed Karimi, Julie Nutini, and Mark 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.

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

[19] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021.

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

[21] Dmitry Kovalev, Samuel Horváth, and Peter 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.

[22] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

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

[24] Zhize Li. Ssrgd: Simple stochastic recursive gradient descent for escaping saddle points. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

[26] Zhize Li, Slavomír Hanzely, and Peter Richtárik. Zerosarah: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021.

[27] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

[28] Deyi Liu, Lam M Nguyen, and Quoc Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

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

[30] Tomoya Murata and Taiji Suzuki. Bias-variance reduced local sgd for less heterogeneous federated learning. In *International Conference on Machine Learning*, pages 7872–7881. PMLR, 2021.

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

[32] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

[33] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017.

[34] Lam M Nguyen, Katya Scheinberg, and Martin Takáč. Inexact sarah algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258, 2021.

[35] Lam M Nguyen, Marten van Dijk, Dzung T Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R Kalagnanam. Finite-sum smooth optimization with sarah. *Computational Optimization and Applications*, pages 1–33, 2022.

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

[37] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

[38] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Fast incremental method for smooth nonconvex optimization. In *2016 IEEE 55th conference on decision and control (CDC)*, pages 1971–1977. IEEE, 2016.

[39] Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

[40] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, page 567, 2013.

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

[42] Terence Tao and Van Vu. Random matrices: universality of local spectral statistics of non-hermitian matrices. *The Annals of Probability*, 43(2):782–874, 2015.

[43] Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071, 2022.

[44] Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

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

[46] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[48] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *The Journal of Machine Learning Research*, 21(1):4130–4192, 2020.

Appendix

Table of Contents

1 Introduction **1**
1.1 Contributions . . . 2

2 SLEDGE: Single-Loop Method for Gradient Estimator **3**
2.1 Algorithm description . . . 3
2.2 Convergence analysis . . . 3

3 FLEDGE: Federated Learning Method with Gradient Estimator **4**

4 Experiments **6**

A Further Literature review **12**
A.1 Variance reduction . . . 12
A.2 Federated learining . . . 13

B Additional experiments **14**
B.1 Details of the experiment for Figure 1 . . . 14
B.2 Details of the experiment for Figure 2 . . . 14
B.3 Additional experiments for SLEDGE . . . 15
B.4 Additional experiments for FLEDGE . . . 17
B.5 Computing infrastructures . . . 18

C Assumptions and Tools **19**
C.1 Assumptions . . . 19
C.2 Concentration inequalities . . . 20
C.3 Proof of Proposition 6 . . . 21
C.4 Linear Algebraic Tool . . . 23

D Missing Proofs for SLEDGE **23**
D.1 Intuitition behind SLEDGE . . . 24
D.2 Finding First-order Stationary Points (Proof of Theorem 10) . . . 25
D.3 Finding Second-order Stationary Points (Proof of Theorem 11) . . . 33
D.4 Convergence under PL condition (proof of Theorem 12) . . . 43

E Missing Statements and Proofs for FLEDGE **45**
E.1 Finding First-order Stationary Points (Proof of Theorem 22) . . . 45
E.2 Finding Second-order Stationary Points (Proof of Theorem 28) . . . 54
E.3 Finding Second-Order Stationary Points When Clients are Homogeneous ($\zeta \ll \frac{1}{\delta}$) . . . 65
E.4 Convergence under PL condition . . . 66

F Lower bound **70**

The appendix consists of the following sections. In Appendix A, we provide further literature review and detailed comparison with existing algorithms. Appendix B describes details of the experiments and additional experiment results. Appendix C formalizes the assumptions and prepares mathematical tools for later use. Appendix D gives formal statements for theoretical guarantees on SLEDGE and the complete proofs. We also explain intuition behind SLEDGE estimators. Appendix E applies SLEDGE into its federated learning extension. Finally, Appendix F shows gradient complexity lower bound under the Hessian heterogeneity assumption, which proves SLEDGE's gradient complexity under the Hessian heterogeneity of ζ is optimal up to polylogarithmic factors.

For those who do not have enough time to read all the contents, we recommend to look over Appendix D.1. We expect this part gives a flavor of the core concepts of the SLEDGE estimator to achieve optimal complexity and take advantage of less Hessian heterogeneity without periodic full gradient computation.

Appendix A. Further Literature review

A.1. Variance reduction

As explained in the main part, variance reduction is a technique in minibatch sampling to construct a gradient estimator with a smaller variance than vanilla SGD by utilizing gradients at previously obtained anchor points [8, 15, 39]. It is originally developed for (strongly)-convex optimization [8, 15, 39, 40], and thereafter extended to nonconvex settings [1, 37, 38].

One of the difficulties in obtaining an appropriate gradient estimator, especially in nonconvex settings, is that recursive update of a gradient estimator with minibatch gradients easily accumulates the error and eventually buries the correct descent directions. To address this issue, there have been two major approaches. The first approach is to explicitly store previously calculated gradients as in SAGA [8]. However, SAGA requires $O(\frac{1}{\varepsilon^2})$ times of updates with minibatch sample size with $O(n^{\frac{2}{3}})$ [38] for solving (1), which is still sub-optimal from the lower bound of [9, 25]. The second one is to use double-loop algorithms that periodically compute the full gradient or a gradient with a large minibatch to refresh a gradient estimator. These algorithms include SARAH, SPIDER, and NestedSVRG [9, 33, 48], which have the optimal rate, i.e., $O(\frac{1}{\varepsilon^2})$ times of updates with minibatch sample size with $O(\sqrt{n})$. On the other hand, this approach has an issue that the step of gradient-refreshing slows down practical computational speed and becomes a bottleneck in the application to federated learning since this leads to periodic synchronization and communication between the whole client.

Recent studies have attempted to develop methods that solve this trade-off, or namely that do not require periodic computation of gradients with a large minibatch size to achieve near-optimal rates [3, 7, 21, 26, 28, 34, 43]. Among them, Li et al. [26] introduced ZeroSARAH as a single-loop algorithm with optimal gradient complexity for nonconvex optimization. Here, we say an algorithm is single-loop when it does not require periodic full or large minibatch gradients. However, ZeroSARAH's estimator cannot take advantage of heterogeneity between f_is. (The reason will be briefly explained in Appendix D.1.) To benefit from less Hessian-heterogeneity, we need a path-integrated type estimator like Fang et al. [9], Nguyen et al. [32], Zhou et al. [48].

Moreover, these recent single-loop methods have an issue in their versatility. First, while it is usual to extend an optimization algorithm to ensure second-order optimality [10, 14, 46], and variance reduction methods also have been applied to this [2, 9, 24], no single-loop algorithm cannot

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

Algorithms	Stochastic gradient complexity			Periodic full gradient
	Nonconvex	SOSP	PL condition	
(Noisy) SGD [10, 12, 17]	$\frac{\Delta\sigma_c^2}{\varepsilon^4}$	$\mathrm{poly}(\varepsilon^{-1}, \delta^{-1}, d, \sigma, \Delta)$	$\frac{\sigma^2}{\mu^2\varepsilon}\log\varepsilon^{-1}$	Every iteration
SPIDER-SFO$^+$ [9]	$n + \frac{\Delta\sqrt{n}}{\varepsilon^2}$	$n + \Delta(\frac{\sqrt{n}}{\varepsilon^2} + \frac{1}{\varepsilon\delta^3} + \frac{1}{\delta^5})$	None	Required
SARAH [33] and its variants [24, 34]	$n + \frac{\Delta\sqrt{n}}{\varepsilon^2}$	$n + \Delta(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\delta^4} + \frac{n}{\delta^3})$	$n + \frac{\sigma^2}{\varepsilon^2}\log\varepsilon^{-1}$	Required
ZeroSARAH[26]	$\frac{(\Delta+\sigma_c^2)\sqrt{n}}{\varepsilon^2}$	None	None	Never required
	$n + \frac{\Delta\sqrt{n}}{\varepsilon^2}$	None	None	Only at x^0
PAGE [25]	$n + \frac{\Delta\sqrt{n}}{\varepsilon^2}$	None	$(n + \frac{L\sqrt{n}}{\mu})\log\varepsilon^{-1}$	Required
SLEDGE (Option I) (ours)	$\frac{(\Delta+\sigma_c^2)\sqrt{n}}{\varepsilon^2}$	$(\Delta+\sigma_c^2)(\sqrt{n}+\frac{\zeta^2}{\delta^2})(\frac{1}{\varepsilon^2}+\frac{1}{\delta^2})$	$(n + \frac{\zeta\sqrt{n}}{\mu})\log\varepsilon^{-1}$	Never required
SLEDGE (Option II) (ours)	$n + \frac{\zeta\Delta\sqrt{n}}{\varepsilon^2}$	$n+\Delta(\frac{\sqrt{n}}{\varepsilon^2}+\frac{\sqrt{n}}{\delta^4}+\frac{\zeta^2}{\varepsilon^2\delta^2}+\frac{\zeta^2}{\delta^6})$	$(n + \frac{\zeta\sqrt{n}}{\mu})\log\varepsilon^{-1}$	Only at x^0

Note: 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.
In nonconvex and SOSP problems, polylogarithmic terms are omitted. Since $\zeta \leq 2L$, SLEDGE with Option I has at most the same complexity to ZeroSARAH, and SLEDGE with Option II does to SPIDER, SARAH, PAGE, and the lower bound, up to log factors.
In PL, polylogarithmic dependency on other than ε^{-1} and doubly-logarithmic terms are omitted, thus SLEDGE has exponential convergence.

find SOSPs. 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. In addition, most of the existing single-loop methods have focused on removing full gradient computation in some specific setting. Thus none of them achieve both optimal complexity in nonconvex settings and exponential convergence in strongly-convex settings.

Note that he Polyak-Łojasiewicz (PL) condition, which we use instead of strong-convexity, is a generalization of strong convexity to nonconvex settings [36]. One of the recent lines of research is to loosen the conventional assumption of strong convexity and to show exponential convergence under the PL condition [17, 25]. For example, PAGE [25] achieves exponential convergence under the PL condition and the optimal rate in general nonconvex settings, but it should compute the full gradient at a certain probability.

We summarize the convergence rates of existing algorithms and ours in Table 1. Our gradient estimator, SLEDGE, satisfies nearly optimal complexity for general nonconvex settings, speedup under less Hessian heterogeneity, second-order optimality, and exponential convergence under the PL condition, without requiring periodic full gradient computation. For more details on SLEDGE, see Appendix D. Also, Appendix F shows that SLEDGE with Option II achieves nearly optimal gradient complexity under less Hessian heterogeneity of ζ.

A.2. Federated learining

In Table 2, we compare the required numbers of communication rounds of existing algorithms and ours.

Since inner-loop complexity is different for each algorithm but it is most usual to compare algorithms in terms of communication rounds, we listed the complexity by ignoring dependency on the intra-client variance σ. To do so, ours requires $b \gtrsim \frac{\sigma^2}{PK\varepsilon^2}$ and $b \geq K$. Note that this is quite moderate, since FedAvg [29] requires $b \gtrsim \frac{\sigma^2}{pK\varepsilon^2}$ [18] (, which is larger than $\frac{\sigma^2}{PK\varepsilon^2}$), MimeMVR [19] needs at least one local full gradient between communication rounds, and BVR-L-SGD requires $b \geq K$ and $b \geq (\frac{\sqrt{Pm}}{K} \wedge \frac{\sigma}{K\varepsilon})/\zeta$ [30].

Table 2: Comparison of communication rounds and complexity for a non-convex FL (2).

Algorithms	Communication rounds	Client sampling (other then x^0)
FedAvg (nonconvex) [18]	$\frac{\sigma_c^2}{p\varepsilon^4} + \frac{\sigma_c}{\varepsilon^3} + \frac{1}{\varepsilon^2}$	✓
SCAFFOLD (nonconvex) [18]	$\frac{1}{\varepsilon^2}(\frac{P}{p})^{\frac{2}{3}}$	✓
MimeMVR (nonconvex) [19]	$\frac{\zeta'\sigma_c}{\sqrt{p}\varepsilon^3} + \frac{\sigma_c^2}{p\varepsilon^2} + \frac{1}{K\varepsilon^2} + \frac{\zeta'}{\varepsilon^2}$	✓
BVR-L-SGD (nonconvex) [30]	$\frac{1}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2}$	×
FLEDGE (nonconvex) (ours)	$\frac{1}{K\varepsilon^2} + \frac{\zeta\sqrt{P}}{p\varepsilon^2} + \frac{\zeta}{\varepsilon^2}$	✓
BVR-L-PSGD (SOSP) [31]	$(\frac{1}{K} + \zeta)(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4})$	×
FLEDGE (SOSP) (ours)	$(\frac{1}{K} + \zeta)(\frac{1}{\varepsilon^2} + \frac{1}{\delta^4})$	✓ (requiring $p \gtrsim \sqrt{P} + \frac{\zeta^2}{\delta^2}$)
MimeSGD (PL) [19]	$\frac{\sigma_c^2}{\mu p\varepsilon^2} + \frac{L}{\mu}$	✓
FLEDGE (PL) (ours)	$\frac{L}{\mu K} + \frac{\zeta\sqrt{P}}{\mu p} + \frac{\zeta}{\mu} + \frac{P}{p}$	✓

Note: P is the number of clients, μ is the parameter for PL condition, σ_c is the variance between clients, which can be as large as $O(P)$. ζ is the Hessian-heterogeneity between clients and ζ' in MimeMVR is the Hessian-heterogeneity *between all data* (i.e., $\|\nabla^2 f_{i,j}(x) - \nabla^2 f(x)\| \leq \zeta'$). ζ' contains not only the inter-client Hessian-heterogeneity but also the intra-client Hessian-heterogeneity. Thus, $\zeta \leq \zeta'$ always holds and moreover it is possible that $\zeta \ll \zeta'$.
We here choose Option II for FLEDGE, where full participation of clients is conducted only once at x^0. Option I allows client sampling even at x^0, at the cost of additional terms, as we detail in Appendix E.

Appendix B. Additional experiments

B.1. Details of the experiment for Figure 1

We consider a classification of the capital letters using EMNIST By_Class dataset [6]. The original dataset consists of $814,255$ images of handwritten uppercase and lowercase letters and numbers 0-9. Note that the number of data points in each class is not balanced. Since 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 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 four-layer fully-connected neural network, following Murata and Suzuki [31]. We added L_2-regularizer with a regularization parameter of $\lambda = 0.01$ to the empirical risk.

As competitors, we implemented SARAH [32, 33], SSRGD [24], and ZeroSARAH [26]. We set the minibatch size to $b = 12 \fallingdotseq \sqrt{n} = \sqrt{130}$ 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 [26] 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 $\{1.0, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001\}$, so that the test accuracy after 2000 iterations is the highest. For SSRGD and noisy SLEDGE, 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.2. Details of the experiment for Figure 2

We consider a classification of the capital letters using EMNIST By_Class dataset [6] as well. However, here we divided the images in such a way that f_i is a little less heterogeneous, but still more heterogeneous than i.i.d. sampling, as follows. First, we prepared the same number of data points

for each class, and divided them into each client i by the following procedure, setting $q = 0.9$; Then, for each class, we distributed $q \times 100\%$ of the images into four clients, and the rest into the remaining 100 clients. This yields that we have $4 \times 26 = 104$ clients, each of which contains $q \times 100\%$ of the data from one class, and $(1-q) \times 100\%$ of the data from the other classes. We call this grouping as a dataset with the heterogeneity parameter of q. Then, we constructed $f_{i,j}$ with the cross-entropy loss and a two-layer fully-connedted neural network, following Murata and Suzuki [30]. L_2-regularizer with a scale of $\lambda = 0.01$ is added to the empirical risk.

We compared FLEDGE with FedAvg [29], SCAFFOLD [18], MimeMVR [19], and BVR-L-SGD [30]. For each algorithm, we set $p = 10 \fallingdotseq \sqrt{P} = \sqrt{104}$ as the number of the clients used at each communication (except for BVR-L-SGD, which requires $p = P = 104$). Then, we set the local minibatch size as $b = 16$ and the number of local update to $K = 10$. 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. Here we set the global learning rate of SCAFFOLD to $\eta = 1$, as is done in the original paper [18]. 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.

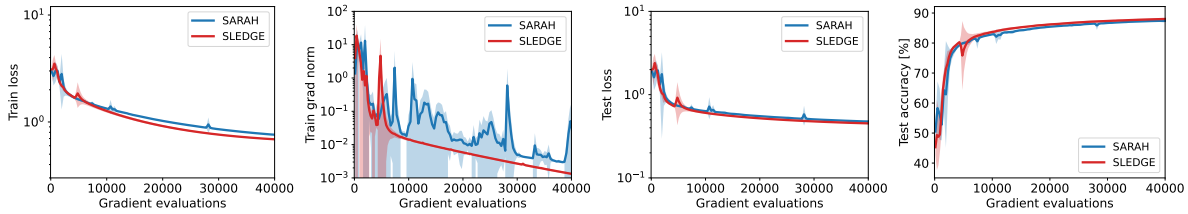
B.3. Additional experiments for SLEDGE

Comparison with SARAH by changing the learning rate Here we provide comparison of SLEDGE with SARAH [32, 33], which is one of the most prevailing variance reduction algorithm with theoretical optimal complexity of $O(\frac{\sqrt{n}}{\varepsilon^2})$.

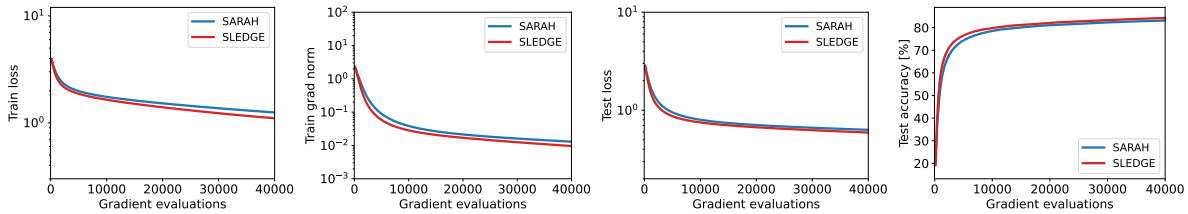
As is done in the experiment for Figure 1,we prepared f_i in the following way. We repeatedly sampled 100 data points five times per each uppercase letter, which yields $26 \times 5 = 130$ groups of sampled data. For each group i, we define f_i as the average of the cross-entropy loss between the output 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, following Murata and Suzuki [30]. We set the minibatch size to $b = 12 \fallingdotseq \sqrt{n} = \sqrt{130}$ 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 SLEDGE 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 3 shows the result. We clearly observe that the proposed algorithm SLEDGE slightly faster than SARAH in all range of learning rate η. The trajectories of SLEDGE 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 SLEDGE.

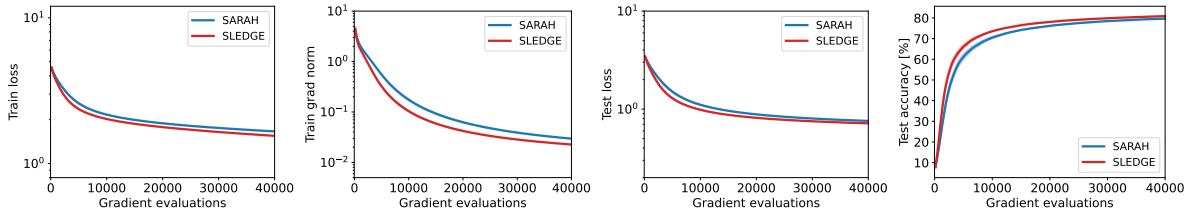
Discrepancy between the gradient estimators and the true gradient Here we compare the norm between the gradient estimators and the true gradient because this is the most essential measure that quantify the quality of the gradient estimator. The setting is completely the same as the previous experiment for Figure 3, thus $n = 130$. We compared SLEDGE estimator with SARAH and SAGA [8, 38], taking the minibatch size as $b = 12$ and the inner-loop length of SARAH to



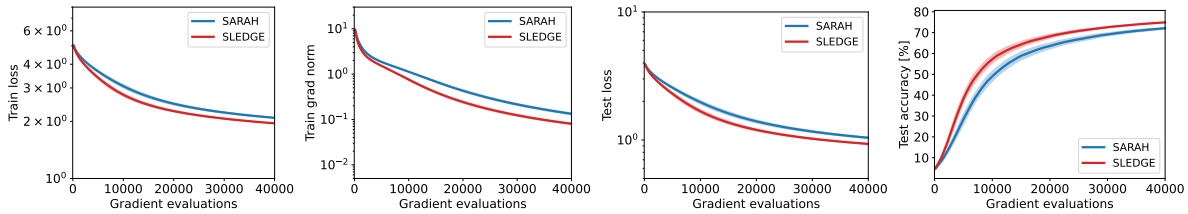
(a) Train loss ($\eta = 0.1$) (b) Gradient norm ($\eta = 0.1$) (c) Test loss ($\eta = 0.1$) (d) Test accuracy ($\eta = 0.1$)



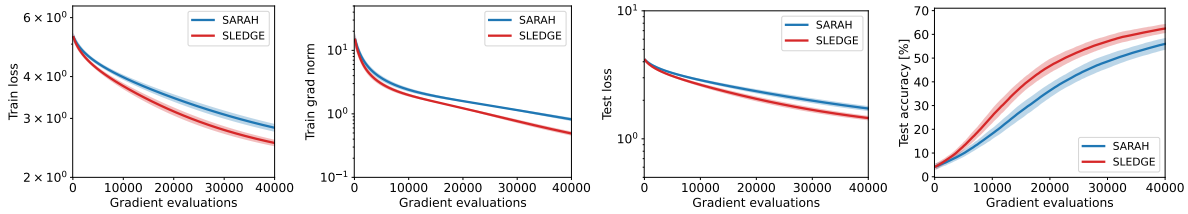
(e) Train loss ($\eta = 0.03$) (f) Gradient norm ($\eta = 0.03$) (g) Test loss ($\eta = 0.03$) (h) Test accuracy ($\eta = 0.03$)



(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 3: Comparison with SARAH by changing the learning rate

$m = \lfloor \frac{n}{b} \rfloor = 10$. 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.

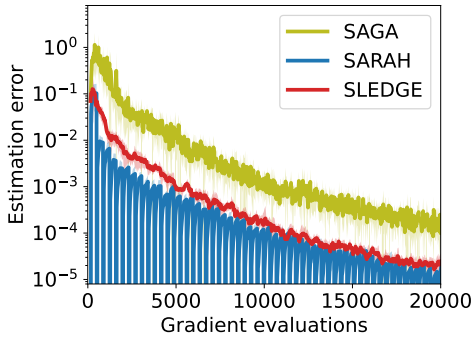


Figure 4: Accuracy of the gradient estimators

Figure 4 shows the squared norm $\|v^t - \nabla f(x^t)\|^2$ between the gradient estimator v^t of each algorithm and the true gradient $\nabla f(x^t)$ at each step t. The discrepancy of the SLEDGE estimator is clearly smaller than that of SAGA, and close to that of SARAH. Note that SARAH estimator is refreshed at every $m = 10$ steps. Remind the discussion in Subsection 3.1. The SLEDGE estimator is designed to have as small variance as that of SARAH, while removing the need of periodic full gradient computation. Therefore, this result validates that our strategy actually works well.

B.4. Additional experiments for FLEDGE

Escaping saddle points with FLEDGE Theorem 28 guarantees second-order optimality of FLEDGE. To validate this theoretical result, we considered the following experiment. We first prepared a dataset with the heterogeneity parameter of $q = 0.7$ (see Appendix B.2 for details). Then, we constructed $f_{i,j}$ with the cross-entropy loss and a three-layer fully-connedted neural network, following Murata and Suzuki [31]. L_2-regularizer with a scale of $\lambda = 0.01$ is added to the empirical risk. We compared FLEDGE with FedAvg [29], SCAFFOLD [18], MimeMVR [19], BVR-L-SGD [30], and BVR-L-PSGD [31]. Here, we set $P = 104$, $p = 10$, $b = 16$, and $K = 10$. Note that, according to Theorem 28, setting $p \simeq \sqrt{P}$ theoretically guarantees that the convergence rate of FLEDGE is not affected by the client sampling and achieves the same number of communication complexity as that of BVR-L-PSGD to find SOSPs. For FLEDGE and BVR-L-PSGD, we added small noise of $r = 0.015$. We plotted the mean of the five trials with different random seeds. We omitted the sample variance for clearer presentation.

The result is shown in Figure 5. We can clearly observe that FLEDGE with small noise and BVR-L-PSGD achieve the highest test accuracy. Note that BVR-L-PSGD is almost the same as FLEDGE with no client sampling ($p = P$). Thus, this shows that FLEDGE is not affected by the client sampling with $p \fallingdotseq \sqrt{P}$, which is consistent with the theory. Ours is as ten times efficient as BVR-L-PSGD, in terms of communication complexity (the number of gradients communicated between the clients).

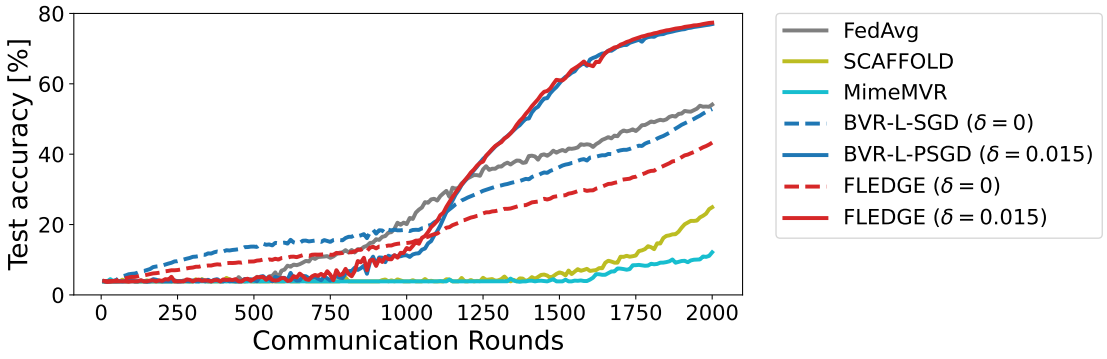


Figure 5: Small perturbation helps faster convergence

Performance under changing heterogeneity To exhibit how correctly FLEDGE can control the variance between clients, we measured the performance of FLEDGE 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 FLEDGE with FedAvg, in terms of both train and test accuracy. All other settings are the same as that of the experiment for Figure 2. Note that we chose $p = 10 \fallingdotseq \sqrt{P} = \sqrt{104}$, where the theory says that the convergence is never affected by sampling of clients. Figure 6 shows the average of five trials with different random seeds.

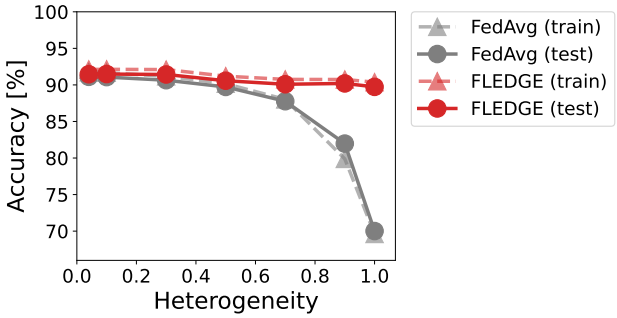


Figure 6: Performance under changing heterogeneity

According to Figure 6, while FedAvg decreases the train and test accuracy as the heterogeneity increases, the performance of FLEDGE with even $q = 1.0$ is only slightly worse than that with $q = 0.04$. The fact that FLEDGE is little affected by the strong heterogeneity in this experiment supports our theoretical guarantee (Theorems 22 and 34) on the effect from sampling of clients. That is, setting $p \geq \sqrt{P}$, our algorithm does not affected by sampling of clients and finds ε-first-order stationary points within $\tilde{O}(\frac{1}{K\varepsilon^2} + \frac{\zeta}{\varepsilon^2})$ communication rounds.

B.5. 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

Appendix C. Assumptions and Tools

In this section, we formally restate the assumptions and introduce mathematical tools we utilize in the missing proofs.

C.1. Assumptions

First, gradient Lipschitzness and boundedness are assumed as usual.

Assumption 1 (Gradient Lipschitzness) *For all $i \in [n]$, f_i is L-gradient Lipschitz, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x-y\|, \forall x, y \in \mathbb{R}^d$. For $f_{i,j}$, we also assume the same.*

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

Below, (i) inter-client gradient boundedness is assumed for SLEDGE with Option I to remove full gradient even at x^0, as in ZeroSARAH [26]. (ii) Intra-client gradient boundedness is assumed for FLEDGE. In the main text, we bounded inter-client and intra-client variances uniformly as $\|\nabla f_i(x^0) - \nabla f(x^0)\| \leq \sigma_c$ and $\|\nabla f_{i,j}(x) - \nabla f_i(x)\| \leq \sigma$. However, if the variances are small in expectation, we can loosen the uniform boundedness assumptions into the followings. From now, we assume Assumption 3 instead of $\|\nabla f_i(x^0) - \nabla f(x^0)\| \leq \sigma_c$ and $\|\nabla f_{i,j}(x) - \nabla f_i(x)\| \leq \sigma$.

Assumption 3 (Boundedness of Gradient) (i) *It holds that $\mathbb{E}_i[\|\nabla f_i(x^0) - \nabla f(x^0)\|^2] \leq \sigma_c^2$, where the expectation $\mathbb{E}_i$ is taken over the choice of i. Moreover, $\|\nabla f_i(x^0) - \nabla f(x^0)\| \leq G_c$ holds for all i.* (ii) *For all i and x, assume $\mathbb{E}_j[\|\nabla f_{i,j}(x) - \nabla f_i(x)\|^2] \leq \sigma^2$, here the expectation $\mathbb{E}_j$ is taken about the choice of j. For all i, j and x, $\|\nabla f_{i,j}(x) - \nabla f_i(x)\|^2 \leq G^2$.*

In order to give second-order optimality, Hessian Lipschiteness is usually assumed [11, 24].

Assumption 4 (Hessian Lipschitzness) *$\{f_i\}_{i=1}^n$ is ρ-Hessian Lipschitz, i.e., $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq \rho\|x-y\|$, $\forall i \in [n]$ and $x, y \in \mathbb{R}^d$.*

For federated learning, we solely assume inter-client Hessian-heterogeneity to show the efficiency of the proposed method in a less heterogeneous setting. It has previously appeared in Mime [19] (but intra-client Hessian-heterogeneity was assumed at the same time) and BVR-L-SGD [30].

Assumption 5 (Hessian-heterogeneity) *$\{f_i\}_{i=1}^n$ is Hessian-heterogeneous with ζ, i.e., for any $i, j \in [n]$ and $x \in \mathbb{R}^d$, $\|\nabla^2 f_i(x) - \nabla^2 f_j(x)\| \leq \zeta$.*

Finally, we explain the PL condition [36]. It is easy to see a μ-strongly convex function satisfies this with μ.

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

C.2. Concentration inequalities

Here, we prepare concentration inequalities for later use. We first present Bernstein-type bounds.

Proposition 3 (Matrix Bernstein inequality [45]) *Let $X_1, \cdots, X_k$ be a finite sequence of independent random matrices with dimension $d_1 \times d_2$. Assume each random matrix satisfies*

$$\mathbb{E}[X_i] = 0 \quad \textit{and} \quad \|X_i\| \leq R \quad \textit{almost surely.}$$

Define

$$\sigma^2 = \max\left\{\left\|\sum_i \mathbb{E}[X_i X_i^\top]\right\|, \left\|\sum_i \mathbb{E}[X_i^\top X_i]\right\|\right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left[\left\|\sum_i X_i\right\| \geq t\right] \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

In this paper, we deal only with the vector case. In that case, the inequality is rewritten for bounds with high probability, as follows.

Proposition 4 (Vector Bernstein inequality) *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 probability at least $1 - \nu/\mathrm{poly}(n, P, T, K)$,

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

where $C_1 = O\left(\log(\nu^{-1} + n + P + d + T + K)\right) = \tilde{O}(1)$.

Remark 5 *Here we do not specify $\mathrm{poly}(n, P, T, K)$ to apply different polynomials to later. Whenever we use this inequality with different $\mathrm{poly}(n, m, T)$, we will reuse C_1 for the notational simplicity. We also use this constant C_1 in the following parts to denote constants as large as $O(\log(\nu^{-1} + n + P + d + T + K))$, with a slight abuse of notations.*

Moreover, a similar inequality holds when we consider sampling without replacement. To our knowledge, Bernstein inequality without replacement for vectors has not been rigorously proven and we attach its complete proof at the next subsection.

Proposition 6 (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*

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

Define

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

Then, for each $t \geq 0$ and $l < k$,

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

Moreover, for each $l < k$, with probability at least $1 - \nu/\mathrm{poly}(n, P, T, K)$,

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

where $C_1 = O\left(\log(n + P + d + T + K)\right) = \tilde{O}(1)$.

Finally, we need a high-probability version of Azuma-Hoeffding inequality.

Proposition 7 (Azuma-Hoeffding inequality with high probability [5, 42]) *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 probability at least $1 - \nu/\mathrm{poly}(n, P, T, K) - \sum_{i=1}^k \nu_i$,

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

where $C_1 = O\left(\log(\nu^{-1} + n + P + d + T + K)\right) = \tilde{O}(1)$.

C.3. Proof of Proposition 6

In order to show Proposition 6, we use the Martingale counterpart of Bernstein's Inequality for random matrix. The following is a slightly weaker version of Tropp [44].

Proposition 8 (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

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

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

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

Proof of Proposition 6 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

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

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 8 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 8 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 $t = O\left((l\sigma^2 + R)\log(\nu^{-1} + n + m + d + T)\right) = C_1 \cdot (l\sigma^2 + R)$.

■

C.4. Linear Algebraic Tool

The following lemma is due to Murata and Suzuki [31]. We provide its proof here.

Lemma 9 (Murata and Suzuki [31]) *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. ■

Appendix D. Missing Proofs for SLEDGE

This section provides the missing proofs in Section 2 about the convergence property of SLEDGE. First, we divide Theorem 1 into the following three formal theorems.

Theorem 10 *Under Assumptions 1, 2 and 5, and 3-(i) for Option I, if we choose $\eta = \tilde{\Theta}(\frac{1}{L} \wedge \frac{b}{\zeta\sqrt{n}})$ and $r \leq \frac{\eta\varepsilon}{2}$, Algorithm 1 finds ε-first-order stationary points using*

$$\tilde{O}\left(\frac{\Delta\left(\zeta\sqrt{n} \vee Lb\right) + \frac{n}{b}\sigma_c^2}{\varepsilon^2}\right) \quad \text{(Option I)},$$

$$\tilde{O}\left(n + \frac{\Delta\left(\zeta\sqrt{n} \vee Lb\right)}{\varepsilon^2}\right) \quad \text{(Option II)}$$

stochastic gradients with probability at least $1-\nu$.

Theorem 11 *Assume Assumptions 1, 2, 4 and 5, 3-(i) for Option I. Let $b \gtrsim \sqrt{n} + \frac{\zeta^2}{\delta^2}$, $\eta = \tilde{\Theta}(\frac{1}{L})$, $r \lesssim \tilde{O}\left(\frac{\varepsilon}{L}\right)$, and $\nu \in (0,1)$. Then, Algorithm 1 finds (ε, δ)-SOSPs using*

$$\tilde{O}\left(\left(L\Delta + \sigma_c^2\right)\left(\frac{1}{\varepsilon^2} + \frac{\rho^2}{\delta^4}\right) b\right) \quad \text{(Option I)},$$

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

stochastic gradients, with probability at least $1-\nu$.

Theorem 12 *Assume Assumptions 1, 2, 3-(i), 5, and 6. If* $\eta = \tilde{\Theta}(\frac{1}{L} \wedge \frac{b}{\zeta\sqrt{n}} \wedge \frac{b}{\mu n})$, *and* $r \leq \eta\sqrt{\frac{\varepsilon\mu}{3}}$, *Algorithm 1 finds an* ε*-solution with* $f(x^t) - f^* \leq \varepsilon$ *using*

$$\tilde{O}\left(\left(\frac{Lb}{\mu} \vee \frac{\zeta\sqrt{n}}{\mu} \vee n\right)\log\frac{\Delta + \sigma_c}{\varepsilon}\right) \quad \text{(Option I)},$$
$$\tilde{O}\left(\left(\frac{Lb}{\mu} \vee \frac{\zeta\sqrt{n}}{\mu} \vee n\right)\log\frac{\Delta}{\varepsilon}\right) \quad \text{(Option II)}$$

stochastic gradients with probability at least $1 - \nu$. $\tilde{O}$ *hides at most* $\log^{5.5}(n + \mu^{-1} + \nu^{-1})$ *and* polyloglog *factors.*

Below, we first describe basic idea for designing SLEDGE. Then,appendix D.2 presents the first-order optimality (Theorem 10). Then we prove that SLEDGE can escape saddle points and find SOSPs (Theorem 11) in appendix D.3. Finally, exponential convergence under the PL condition (Theorem 12) is shown in appendix D.4.

D.1. Intuitition behind SLEDGE

SLEDGE is designed so that it inherits the best points of SAGA [8, 38] and SARAH [32, 33]. We first compare gradient estimators of SAGA and SARAH to explain why SAGA is suboptimal. Then, we also note SARAH's estimator gets more accurate under less heterogeneity.

According to SAGA's update rule, the discrepancy of the gradient estimator from the true gradient at a step t can be decomposed as

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

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

On the other hand, the difference between SARAH's gradient estimator, which computes the full gradient periodically, 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. We can interpret this scheme as it decomposes $\nabla f(x^t) - \nabla f(x^{T(t)})$ into the sum of $\nabla f(x^s) - \nabla f(x^{s-1})$, and each term is approximated by an independent minibatch with size b. Then, the variance is bounded by $\frac{L}{b}\sum_{s=T(t)+1}^{t}\|x^s - x^{s-1}\|^2$, meaning that SARAH's estimator is better than that of SAGA by the $t - T(t)$ factor, which can be as large as $O(\frac{n}{b})$.

Moreover, in the perspective of application to FL, whether an algorithm can take advantage of less heterogeneity becomes important. We consider the difference of the probabilistic term fo

In SAGA's update, the variance cannot be bounded by a factor of ζ, since $x^{T(t,i)}$ is different for different i. On the other hand, we can see that, for SARAH's path-integrated type estimator, the variance can be bounded by $\frac{\zeta}{b}\sum_{s=T(t)+1}^{t}\|x^s - x^{s-1}\|^2$ under the Hessian heterogeneity of ζ. Thus, the path-integrated estimator of SARAH (, which also appears in SPIDER [9] or NestedSVRG [48]), is crucial in taking advantage of less heterogeneity. This is also the reason for why ZeroSARAH [26] fails to utilize less Hessian heterogeneity, since at step t they estimate $\nabla f(x^{t-1})$ based on naive application of SAGA.

Based on the above discussion, we first decompose SAGA's approximation target $\frac{1}{n}\sum_{i=1}^{n}\nabla(f_i(x^t) - \nabla f_i(x^{T(t,i)}))$ into the sum of $\nabla f_i(x^s) - \nabla f_i(x^{s-1})$, each of which is approximated in SARAH's manner. Namely, the decomposed form is written as follows:

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

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. Then, we approximate $\sum_{i\in\tilde{I}_s^t}^{n}(f_i(x^s) - f_i(x^{s-1}))$ with $\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}^{n}(f_i(x^s) - f_i(x^{s-1}))$. This procedure yields SLEDGE, and the following error bound on the SLEDGE estimator.

Lemma 13 (Informal) *Let $\nu \in (0,1)$, $T_1 = \tilde{O}(\frac{n}{b})$, and $C = \tilde{O}(1)$. We have that, ignoring the initialization error, with probability $1-\nu$ for all $t = 1, \cdots, T$,*

$$\left\|\nabla f(x^t) - \frac{1}{n}\sum_{i=1}^{n} y_i^t\right\|^2 \leq \frac{C\zeta^2}{b}\sum_{s=1\vee(t-T_1+1)}^{t}\|x^s - x^{s-1}\|^2.$$

Here, T_1 is defined so that $T_1 \geq t - T(t)$ holds with high probability. This lemma tells us that our gradient estimator has comparable quality to SARAH without computing full gradient. Moreover, this lemma explicitly states that the variance of SLEDGE estimator is quadratically bounded with ζ, meaning that we require fewer gradients when $\zeta \ll L$, which is later exploited for federated learning application.

While the development is intuitively straightforward, we have the technical difficulty to evaluate the error, that $|\tilde{I}_s^t|$ depends not only on I^s but also on $I^{s+1}, \ldots, I^t$. In other words, unlike SARAH, the discrepancy cannot be decomposed into completely independent terms about I^s, which prevents us from using a usual expectation bound. To address this, we prepared vector Bernstein inequality without replacement (Proposition 6) to give a high probability bound on the discrepancy.

D.2. Finding First-order Stationary Points (Proof of Theorem 10)

In this subsection, we show that SLEDGE finds first-order stationary points with high probability (Theorem 10). For the proof of Theorem 10, we use the following classical argument (e.g. Ge et al. [11], Li [24], Li et al. [25]), which ensures decrease of the function values.

Lemma 14 *Let f be an L-gradient Lipschitz function and $x^t := x^{t-1} - \eta v^{t-1} + \xi^{t-1}$ with $\|\xi^{t-1}\| \leq r$. Then,*

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

holds.

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

$$
\begin{aligned}
&f(x^t) \\
&\leq f(x^{t-1}) + \langle \nabla f(x^{t-1}), x^t - x^{t-1} \rangle + \frac{L}{2}\|x^t - x^{t-1}\|^2 \\
&= f(x^{t-1}) + \left\langle \nabla f(x^{t-1}) - v^{t-1} + \frac{\xi^{t-1}}{\eta}, x^t - x^{t-1} \right\rangle + \left\langle v^{t-1} - \frac{\xi^{t-1}}{\eta}, x^t - x^{t-1} \right\rangle \\
&\quad + \frac{L}{2}\|x^t - x^{t-1}\|^2 \\
&= f(x^{t-1}) + \left\langle \nabla f(x^{t-1}) - v^{t-1} + \frac{\xi^{t-1}}{\eta}, x^t - x^{t-1} \right\rangle - \left(\frac{1}{\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 \\
&= f(x^{t-1}) + \frac{\eta}{2}\left\|\nabla f(x^{t-1}) - v^{t-1} + \frac{\xi^{t-1}}{\eta}\right\|^2 - \frac{\eta}{2}\|\nabla f(x^{t-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 \quad (3) \\
&\leq f(x^{t-1}) + \eta\|\nabla f(x^{t-1}) - v^{t-1}\|^2 - \frac{\eta}{2}\|\nabla f(x^{t-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 \quad (4) \\
&\quad + \frac{\left\|\xi^{t-1}\right\|^2}{\eta} \\
&\leq f(x^{t-1}) + \eta\|\nabla f(x^{t-1}) - v^{t-1}\|^2 - \frac{\eta}{2}\|\nabla f(x^{t-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 + \frac{r^2}{\eta}, \quad (5)
\end{aligned}
$$

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 (3), $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ for (4), and $\|\xi^{t-1}\| \leq r$ for (5). ■

Our algorithm uses $v^t = \frac{1}{n}\sum_{i=1}^n y_i^t$ as an estimator of $\nabla f(x^t)$. To apply Lemma 14 for our algorithm, we need to evaluate the term $\|v^t - \nabla f(x^t)\|^2$, the variance of the gradient estimator. The next lemma provides its upper bound that holds with high probability.

Lemma 13 *Let $v^t = \frac{1}{n}\sum_{i=1}^n y_i^t$ and all the other variables be as stated in Algorithm 1. Then, by taking $T_1 = \frac{n}{b}C_1$,*

$$
\begin{aligned}
&\|v^t - \nabla f(x^t)\|^2 \\
&\quad \leq \begin{cases} \dfrac{15C_1^8\zeta^2}{b} \displaystyle\sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2 + \dfrac{12C_1^2\mathbb{1}[t < T_1]}{b} \cdot \left(\sigma_c^2 + \dfrac{G_c^2}{b}\right) & (\textit{Option I}) \\ \dfrac{15C_1^8\zeta^2}{b} \displaystyle\sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2 & (\textit{Option II}) \end{cases}
\end{aligned}
$$

holds for all $t = 1, \cdots, T$ with probability at least $1 - 3\nu$.

We decompose $\|v^t - \nabla f(x^t)\|$ into three parts to each of which one of the following lemmas is applied. Below, for each $1 \leq s \leq t$, we let $\tilde{I}_s^t = [n] \setminus \bigcup_{\tau=s}^t I^t$, which is a set of indexes that are not selected between $s + 1$ and t.

Lemma 15 *The following holds uniformly for all $1 \leq t \leq T$ with probability at least $1-\nu$:*

$$\left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \leq \frac{C_1^3 \zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2.$$

Lemma 16 *The following inequality holds uniformly for all $1 \leq t \leq T$ with probability at least $1-\nu$:*

$$\left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \frac{|\tilde{I}_s^t|}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \leq \frac{4C_1^8 \zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2. \tag{6}$$

Lemma 17 *The following inequality holds with probability at least $1-\nu$ uniformly over $1 \leq t \leq T$:*

$$\left\| \frac{1}{n} \sum_{i \in \tilde{I}_1^t} (y_i^0 - \nabla f_i(x^0)) \right\|^2 \leq \begin{cases} \frac{4C_1^2}{b}\left(\sigma_c^2 + \frac{G_c^2}{b}\right) & (\textit{Option I}) \\ 0 & (\textit{Option II}) \end{cases}.$$

Proof of Lemma 15 First, we have that

$$\begin{aligned} (\mathrm{a}) &:= \left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\ &\leq \frac{t - \max\{1, t-T_1+1\} + 1}{n^2} \times \\ &\qquad \sum_{s=\max\{1,t-T_1+1\}}^{t} \left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\ &\leq \frac{T_1}{n^2} \sum_{s=\max\{1,t-T_1+1\}}^{t} \left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2, \end{aligned} \tag{7}$$

where we use $\| \sum_{i=1}^m a_i \|^2 \leq m \sum_{i=1}^m \|a_i\|^2$ for the first inequality. For each s, from Assumption 5,

$$\|\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))\| \leq \zeta \|x^s - x^{s-1}\|$$

holds for all $i \in [n]$. By vector Bernstein inequality without replacement (Proposition 6), for each $t \geq 1$ and s satisfying $\max\{1, t-T_1+1\} \leq s \leq t$, we have that

$$\left\| \sum_{i \in \tilde{I}_s^t} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \leq C_1^2 |\tilde{I}_s^t| \zeta^2 \|x^s - x^{s-1}\|^2 \tag{8}$$

holds with probability at least $1-\frac{\nu}{T^2}$. Thus, in (7), (8) holds uniformly for all t and s with probability at least $1-\nu$. Applying this bound to (7) yields

$$\text{(a)} \leq \frac{T_1}{n^2} \sum_{s=\max\{1,t-T_1+1\}}^{t} C_1^2 |\tilde{I}_s^t| \zeta^2 \|x^s - x^{s-1}\|^2 \leq \frac{C_1^3 \zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2,$$

where the second inequality follows from $|\tilde{I}_s^t| \leq n$ and $T_1 = \frac{n}{b} C_1$. ■

Proof of Lemma 16 Since $|\tilde{I}_s^t|$ depends not only on I^s but also on $I^s, I^{s+1}, \cdots, I^t$, the left-hand side of (6) is not a sum of martingale variables with respect to the filtration $\{\sigma(I^1, \cdots, I^s)\}_{s=1}^t$. Thus, we consider $\mathbb{E}[|\tilde{I}_s^t|]$ instead of $|\tilde{I}_s^t|$ and validate the difference between them later. We decompose (6) as

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \frac{|\tilde{I}_s^t|}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \\
&\leq 2 \left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2 \\
&\quad + 2 \left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \left(|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|] \right)^2 \frac{1}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2 \\
&\leq 2 \left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2 \qquad (9) \\
&\quad + \frac{2T_1}{n^2} \sum_{s=\max\{1,t-T_1+1\}}^{t} \left(|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|] \right)^2 \left\| \frac{1}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1})) \right\|^2. \qquad (10)
\end{aligned}$$

First, we bound the term (9). We can see that $\frac{\mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1})))$ is a martingale difference sequence. Moreover, by the vector Bernstein inequality without replacement (Proposition 6) and Assumption 5, we have

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

with probability at least $1 - \frac{\nu}{5T^2}$ for each t and s in (9). This allows us to use the Azuma-Hoeffding inequality with high probability (Proposition 7). Consequently, with probability at least $1 - \frac{\nu}{5T} -$

$T \cdot \frac{\nu}{5T^2} = 1 - \frac{2\nu}{5T}$, it holds that

$$\left\| \frac{1}{n} \sum_{s=\max\{1,t-T_1+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{b} \sum_{i \in I^s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1}))) \right\|^2 \leq \frac{C_1^2 \zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2 \tag{11}$$

for each $t \in [T]$. Therefore, (11) holds for all t with probability $1 - \frac{2\nu}{5}$.

As for the term (10), by the Bernstein inequality without replacement (Proposition 6) and Assumption 5, we have

$$\left\| \frac{1}{b} \sum_{i \in I_s} (\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) + \nabla f(x^{s-1}))) \right\|^2 \leq \frac{C_1^2 \zeta^2}{b} \|x^s - x^{s-1}\|^2, \tag{12}$$

for all s with probability at least $1 - \frac{\nu}{5}$. We move to bound the difference $\left(|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|]\right)^2$. For this purpose, we regard this as a function of (at most) T_1 variables $I^t, \cdots, I^s$ and prepare a "reverse" filtration $\tilde{\mathcal{F}} = \{\tilde{\mathcal{F}}_s^t\}_{s=t}^{\max\{1,t-T_1+1\}}$ with $\tilde{\mathcal{F}}_s^t = \sigma(I_t, I_{t-1}, \cdots, I_s)$. Then, the sequence $\{|\tilde{I}_s^t|\}_{s=\max\{1,t-T_1+1\}}^{t}$ is a measurable process with respect to $\tilde{\mathcal{F}}$. We consider the conditional expectation of $|\tilde{I}_s^t| - |\tilde{I}_{s+1}^t|$ with respect to $\tilde{\mathcal{F}}$. When samples in $\tilde{I}_{s+1}^t$ are not chosen between t to $s+1$, each of them is chosen with probability $\frac{b}{n}$ for the first time at step s. Thus, we have

$$\mathbb{E}_s \left[|\tilde{I}_{s+1}^t| - |\tilde{I}_s^t| \mid \tilde{\mathcal{F}}_{s+1}^t \right] = \frac{b}{n} |\tilde{I}_{s+1}^t|,$$

which leads to $\mathbb{E}_s \left[|\tilde{I}_s^t| \middle| \tilde{\mathcal{F}}_{s+1}^t \right] = \left(1 - \frac{b}{n}\right) |\tilde{I}_{s+1}^t|$. Hence, the process $\{u_s^t := |\tilde{I}_s^t| - \left(1 - \frac{b}{n}\right) |\tilde{I}_{s+1}^t| \mid t > s \geq t - T_1 + 1\}$ 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 6, we have $\|u_s^t\| \leq C_1 \sqrt{b}$ with probability at least $1 - \frac{1}{5T^2}$. Finally, we apply Proposition 7 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

$$\left| |\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|] \right|^2 \leq C_1^2 \sum_{\tau=t}^{s} C_1^2 \left(1 - \frac{b}{n}\right)^{(\tau - s)} b \leq C_1^4 b T_1 = C_1^5 n \tag{13}$$

with probability at least $1 - \frac{\nu}{5T} - T \cdot \frac{\nu}{5T^2} = \frac{2\nu}{5T}$.

Combining (12) and (13), with probability $1-\frac{\nu}{5}$, we get

$$\frac{T_1}{n^2}\sum_{s=\max\{1,t-T_1+1\}}^{t}\left(|\tilde{I}_s^t|-\mathbb{E}[|\tilde{I}_s^t|]\right)^2\left\|\frac{1}{b}\sum_{i\in I^s}(\nabla f_i(x^s)-\nabla f_i(x^{s-1})-(\nabla f(x^s)+\nabla f(x^{s-1})))\right\|^2 \leq \frac{C_1^8\zeta^2}{b^2}\sum_{s=\max\{1,t-T_1+1\}}^{t}\|x^s-x^{s-1}\|^2, \tag{14}$$

by letting $T_1=\frac{n}{b}C_1$. Finally, we get the assertion by combining (11) and (14), and applying $\frac{2C_1^2}{b}+\frac{2C_1^8}{b^2}\leq\frac{4C_1^8}{b}$.

■

Proof of Lemma 17 As for the Option II, the assertion directly follows from the definition $y_i^0=\nabla f_i(x^0)$ $(i=1,\cdots,n)$. Henceforth, we prove the bound

$$\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0-\nabla f_i(x^0))\right\|^2\leq\frac{4C_1^2}{b}\left(\sigma_c^2+\frac{G_c^2}{b}\right)$$

when we use Option I. To this end, we decompose $\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0-\nabla f_i(x_0))\right\|^2$ as

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0-\nabla f_i(x^0))\right\|^2&=\left\|\frac{1}{n}\left(\frac{|\tilde{I}_1^t|}{b}\sum_{i\in I^1}(\nabla f_i(x^0)-\nabla f(x^0))+\sum_{i\in\tilde{I}_1^t}(\nabla f(x^0)-\nabla f_i(x^0))\right)\right\|^2\\&\leq 2\left\|\frac{1}{n}\frac{|\tilde{I}_1^t|}{b}\sum_{i\in I^1}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2+2\left\|\frac{1}{n}\sum_{i\in\tilde{I}_1^t}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2,\end{aligned} \tag{15}$$

where we use the inequality $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$. For the first term in (15), Proposition 6 and Assumption 3 imply that

$$\left\|\frac{1}{n}\frac{|\tilde{I}_1^t|}{b}\sum_{i\in I^1}(\nabla f_i(x^0)-\nabla f(x^0))\right\|^2\leq\frac{|\tilde{I}_1^t|^2}{n^2b^2}C_1^2b\left(\sigma_c^2+\frac{G_c^2}{b}\right)\leq\frac{C_1^2}{b}\left(\sigma_c^2+\frac{G_c^2}{b}\right), \tag{16}$$

holds with probability at least $1-\frac{\nu}{2}$ for all t. For the second term in (15), we have

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

from Proposition 6 and Assumption 3, with probability at least $1-\frac{\nu}{2T}$ for each t and at least $1-\frac{\nu}{2}$ uniformly over all t.

Substituting (15) and (16) to (14), we obtain the desired bound.

■

Proof of Lemma 13 We first observe that $v^t - \nabla f(x^t)$ is written as

$$v^t - \nabla f(x^t) = \frac{1}{n}\sum_{s=1}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) - \sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}))\right) + \frac{1}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0 - \nabla f_i(x^0)).$$

We can ensure that if $t-s$ is sufficiently large, every f_i is sampled at least once between $s+1$ and t with high probability. Indeed, for each f_i and $\nu > 0$, the probability that f_i is not sampled between $s+1$ and t is bounded as

$$\left(1-\frac{1}{n}\right)^{\sum_{s=\max\{1,t-s+1\}}^{t} b} \le \left(1-\frac{1}{n}\right)^{\sum_{s=\max\{1,t-T_1+1\}}^{t} b} \le \left(1-\frac{1}{n}\right)^{nC_1} \le \exp\left(-C_1\right), \quad (17)$$

where we use $\sum_{s=\max\{1,t-T_1+1\}}^{t} b \le T_1 b = nC_1$ in the second inequality. By taking $C_1 = \Omega\left(\log\frac{nT}{\nu}\right)$, the right-hand side of (17) is bounded by $\frac{\nu}{nT}$. In other words, $\tilde{I}_s^t = \emptyset$ with probability at least $1-\nu$ for every t and $s \le t - T_1$. Henceforth, we assume $\tilde{I}_s^t = \emptyset$ and focus on the errors between $\max\{1, t-T_1+1\} \le s \le t$.

When $\tilde{I}_s^t = \emptyset$ holds for $s \le t - T_1$, the variance term $\|v^t - \nabla f(x^t)\|^2$ is decomposed as

$$\begin{aligned}
&\|v^t - \nabla f(x^t)\|^2 \\
&= \left\|\frac{1}{n}\sum_{s=\max\{1,t-T_1+1\}}^{t}\left(\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I_s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) - \sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}))\right)\right. \\
&\qquad \left. + \frac{\mathbb{1}[t\le T_1]}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0 - \nabla f_i(x^0))\right\|^2 \\
&\le 3\underbrace{\left\|\frac{1}{n}\sum_{s=\max\{1,t-T_1+1\}}^{t}\sum_{i\in\tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))\right\|^2}_{\text{(a)}} \\
&\quad + 3\underbrace{\left\|\frac{1}{n}\sum_{s=\max\{1,t-T_1+1\}}^{t}\frac{|\tilde{I}_s^t|}{b}\sum_{i\in I^s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1}) - (\nabla f(x^s) - \nabla f(x^{s-1})))\right\|^2}_{\text{(b)}} \\
&\quad + 3\underbrace{\left\|\frac{\mathbb{1}[t\le T_1]}{n}\sum_{i\in\tilde{I}_1^t}(y_i^0 - \nabla f_i(x^0))\right\|^2}_{\text{(c)}}, \qquad (18)
\end{aligned}$$

by the inequality $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$. Then, we give the bound of (18) for Option I and Option II, respectively.

Option I According to Lemmas 15 to 17, we have

$$\begin{aligned}
&\|v^t - \nabla f(x^t)\|^2 \\
&\leq \underbrace{\frac{3C_1^3\zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2}_{\text{(a)}} + \underbrace{\frac{12C_1^8\zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2}_{\text{(b)}} \\
&\quad + \underbrace{\frac{12C_1^2\left(\sigma_c^2 + \frac{G_c^2}{b}\right)\mathbb{1}[t \leq T_1]}{b}}_{\text{(c)}} \\
&\leq \left(3C_1^3 + 12C_1^8\right)\frac{\zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2 + \frac{12C_1^2\left(\sigma_c^2 + \frac{G_c^2}{b}\right)\mathbb{1}[t \leq T_1]}{b}
\end{aligned}$$

with probability at least $1 - 3\nu$ uniformly over all t.

Option II Almost as well as the previous case, we have

$$\begin{aligned}
&\|v_t - \nabla f(x^t)\|^2 \\
&\leq \underbrace{\frac{3C_1^3\zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2}_{\text{(a)}} + \underbrace{\frac{12C_1^8\zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2}_{\text{(b)}} + \underbrace{0}_{\text{(c)}} \\
&\leq \left(3C_1^3 + 12C_1^8\right)\frac{\zeta^2}{b} \sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2
\end{aligned}$$

with probability at least $1 - 2\nu$ uniformly over all t.

By replacing C_1 with $C_1 \vee 1$ and applying $3C_1^3 + 12C_1^8 \leq 15C_1^8$, we obtain the desired bound for both cases. ■

Now, we are ready to prove the first-order convergence of SLEDGE.

Proof of Theorem 10

Summing up (5) over all $t = 1, 2, \cdots, T$ and arranging the terms, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 \\
&\leq \frac{2}{\eta}\left[\left(f(x^0) - f(x^t)\right) - \sum_{t=1}^{T}\left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 + \eta\sum_{t=1}^{T}\|\nabla f(x^{t-1}) - v^{t-1}\|^2\right] + \frac{2Tr^2}{\eta^2}.
\end{aligned}$$

Applying Lemma 13 to this, we obtain that

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 \leq \frac{2}{\eta}\left[\left(f(x^0) - f(x^t)\right) - \left(\frac{1}{2\eta} - \frac{L}{2} - \frac{15C_1^9\eta\zeta^2 n}{b^2}\right)\sum_{t=1}^{T}\|x^t - x^{t-1}\|^2\right] + \frac{2Tr^2}{\eta^2} + \begin{cases} \frac{2}{\eta}\frac{12\eta C_1^3 n}{b^2}\left(\sigma_c^2 + \frac{G_c^2}{b}\right) & \text{(Option I)} \\ 0 & \text{(Option II)} \end{cases}$$

with probability at least $1 - 3\nu$.

Option I We take $\eta = \min\left\{\frac{1}{2L}, \frac{b}{C_1^4\zeta\sqrt{60n}}\right\} = \tilde{\Theta}\left(\frac{b}{Lb\vee\zeta\sqrt{n}}\right)$, which implies $\frac{1}{2\eta} - \frac{L}{2} - \frac{15C_1^9\eta\zeta^2 n}{b^2} \geq 0$. We have $\frac{2Tr^2}{\eta^2} \leq \frac{T\varepsilon^2}{2}$ by letting $r \leq \frac{\eta\varepsilon}{2}$. Also, we have $f(x_0) - f(x^t) \leq \Delta$ by Assumption 2. Summarizing these, we get

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 \leq \frac{2\Delta}{\eta} + \frac{12C_1^2 n}{b^2}\left(\sigma_c^2 + \frac{G_c}{b}\right) + \frac{T\varepsilon^2}{2}.$$

Thus, by setting $T \geq \left(\frac{2\Delta}{\eta} + \frac{12C_1^3 n}{b^2}\left(\sigma_c^2 + \frac{G_c^2}{b}\right)\right)\frac{2}{\varepsilon^2} = \tilde{O}\left(\frac{\left(L + \frac{\zeta\sqrt{n}}{b}\right)\Delta + \frac{n}{b^2}\left(\sigma_c^2 + \frac{G_c^2}{b}\right)}{\varepsilon^2}\right)$, we obtain $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x^{t-1})\|^2 \leq \varepsilon^2$ with probability at least $1 - 3\nu$, which implies that there exists some t such that $0 \leq t \leq T-1$ and $\|\nabla f(x^t)\|^2 \leq \varepsilon^2$. From this, the desired conclusion is obtained.

Option II As well as Option I, we take $\eta = \min\left\{\frac{1}{2L}, \frac{b}{C_1^4\zeta\sqrt{60n}}\right\} = \tilde{\Theta}\left(\frac{b}{Lb\vee\zeta\sqrt{n}}\right)$, which implies $\frac{1}{2\eta} - \frac{L}{2} - \frac{15C_1^9\eta\zeta^2 n}{b^2} \geq 0$. In addition, by letting $r \leq \frac{\eta\varepsilon}{2}$, we obtain $\sum_{t=1}^{T}\|\nabla f(x^{t-1})\|^2 \leq \frac{2\Delta}{\eta} + \frac{T\varepsilon^2}{2}$. Then, by taking $T \geq \frac{4\Delta}{\eta\varepsilon^2} = \tilde{O}\left(\frac{L\Delta}{\varepsilon^2}\right)$, there exists some t such that $0 \leq t \leq T-1$ and $\|\nabla f(x^t)\|^2 \leq \varepsilon^2$, with probability at least $1 - 3\nu$. Accordingly, we get the assertion for Option II.

■

D.3. Finding Second-order Stationary Points (Proof of Theorem 11)

The goal of this subsection is to show that SLEDGE is the single-loop algorithm with theoretical guarantee for finding second-order stationary points.

The argument follows that of [11, 14, 24]; 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 SLEDGE 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 proof outline has been popular, we face the difficulties arising from the single-loop structure of the algorithm. 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. This is the first difficulty, and we found that taking minibatch size as large as $b \gtrsim \sqrt{n} + \frac{\zeta^2}{\delta^2}$ is sufficient. When $\delta = O(\sqrt{\varepsilon})$, $\frac{\zeta^2}{\delta^2}$ is about $O(\sqrt{n} + \frac{1}{\varepsilon})$, which is usually assumed in existing literature [11, 24]. Secondly, our estimator is more correlated due to the $|\tilde{I}_s^t|$ term, thus requiring more delicate analysis than that for SSRGD [24], which is based on SARAH [32, 33].

We formalize the exponential separation of two sequences in the following lemma.

Lemma 18 (Small stuck region) *Let $\{x^t\}$ be a sequence generated by SLEDGE 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 SLEDGE 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 $r_e \geq \frac{r\nu}{T\sqrt{d}}$. Let $w^t = x^t - \tilde{x}^t$, $v^t = \frac{1}{n}\sum_{i=1}^n y_i^t$, $\tilde{v}^t = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i^t$, and $g^t = v^t - \nabla f(x^t) - (\tilde{v}^t - \nabla f(\tilde{x}^t))$. Here $\tilde{y}_i^t$ is the counterpart of y_i^t and corresponds to $\{\tilde{x}^t\}$.*

Then, there exists constants $C_2 = \tilde{O}(1), C_3 = O(1)$, such that if we take $b \geq \sqrt{n} + \frac{C_2^2\zeta^2}{\delta^2}$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_2 = \frac{C_3 \log \frac{\delta}{C_2 \rho r_e}}{\eta\gamma} \leq \tilde{O}\left(\frac{L}{\delta}\right)$, with probability at least $1 - \frac{\nu}{T}$ ($\nu \in (0, T)$), it holds that

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

In order to show Lemma 18, we need the following lemma, which is analogous to Lemmas 15 to 17.

Lemma 19 *Under the same assumption as that of Lemma 18, we 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 probability at least $1 - \frac{\nu}{T}$:*

$$\|g^t\| \leq \begin{cases} 0 & (t < \tau) \\ \frac{C_4 \zeta r_e}{\sqrt{b}} & (t = \tau) \\ \frac{C_4\zeta r_e}{\sqrt{b}} + \frac{C_4\zeta}{\sqrt{b}}\sqrt{\sum_{s=\max\{\tau+1, t-T_1+1\}}^t \|w^s - w^{s-1}\|^2} + \frac{C_4\delta}{C_2\sqrt{b}}\sqrt{\sum_{s=\max\{\tau, t-T_1+1\}}^t \|w^s\|^2} & (\textit{otherwise}), \end{cases}$$

where $T_1 = \frac{n}{b}C_1$. Here $C_4 = \tilde{O}(1)$ is a sufficiently large constant.

Proof of Lemma 19 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

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

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

$$\begin{aligned} \|u_i^s\| &= \|(\nabla f_i(x^s) - \nabla f_i(\tilde{x}^s)) - (\nabla f_i(x^{s-1}) - \nabla f_i(\tilde{x}^{s-1})) \\ &\quad - (\nabla f(x^s) - \nabla f(\tilde{x}^s)) + (\nabla f(x^{s-1}) - \nabla f(\tilde{x}^{s-1}))\| \\ &= \Big\| \int_0^1 \nabla^2 f_i(\tilde{x}^s - \theta(x^s - \tilde{x}^s))(x^s - \tilde{x}^s)\mathrm{d}\theta - \int_0^1 \nabla^2 f_i(\tilde{x}^{s-1} - \theta(x^{s-1} - \tilde{x}^{s-1}))(x^{s-1} - \tilde{x}^{s-1})\mathrm{d}\theta \\ &\quad - \int_0^1 \nabla^2 f(\tilde{x}^s - \theta(x^s - \tilde{x}^s))(x^s - \tilde{x}^s)\mathrm{d}\theta + \int_0^1 \nabla^2 f(\tilde{x}^{s-1} - \theta(x^{s-1} - \tilde{x}^{s-1}))(x^{s-1} - \tilde{x}^{s-1})\mathrm{d}\theta \Big\| \\ &= \|(H_i + \mathrm{d}H_i^s)w^s - (H_i + \mathrm{d}H_i^{s-1})w^{s-1} - (H + \mathrm{d}H^s)w^s + (H + \mathrm{d}H^{s-1})w^{s-1}\| \\ &\leq \|H_i - H\|\|w^s - w^{s-1}\| + (\|\mathrm{d}H_i^s\| + \|\mathrm{d}H^s\|)\|w^s\| + (\|\mathrm{d}H_i^{s-1}\| + \|\mathrm{d}H^{s-1}\|)\|w^{s-1}\| \\ &\leq \zeta\|w^s - w^{s-1}\| + 2\rho \max_{0\leq\theta\leq 1}\{\|\tilde{x}^s - \theta(x^s - \tilde{x}^s) - x^\tau\|\}\|w^s\| \\ &\quad + 2\rho \max_{0\leq\theta\leq 1}\{\|\tilde{x}^{s-1} - \theta(x^{s-1} - \tilde{x}^{s-1}) - x^\tau\|\}\|w^{s-1}\| \\ &= \zeta\|w^s - w^{s-1}\| + 2\rho\max\{\|x^s - x^\tau\|, \|\tilde{x}^s - x^\tau\|\}\|w^s\| \\ &\quad + 2\rho\max\{\|x^{s-1} - x^\tau\|, \|\tilde{x}^{s-1} - x^\tau\|\}\|w^{s-1}\| \\ &< \zeta\|w^s - w^{s-1}\| + \frac{2\delta}{C_2}\|w^s\| + \frac{2\delta}{C_2}\|w^{s-1}\|, \end{aligned} \tag{19}$$

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

Recall the discussion in Lemma 13, we have

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

As for the first term in both cases, we have

$$
\left\| \frac{1}{n} \left(\frac{|\tilde{I}_\tau^t|}{b} \sum_{i \in I^\tau} u_i^\tau - \sum_{i \in \tilde{I}_\tau^t} u_i^\tau \right) \right\| \leq \left\| \frac{1}{n} \frac{|\tilde{I}_\tau^t|}{b} \sum_{i \in I^\tau} u_i^\tau \right\| + \left\| \frac{1}{n} \sum_{i \in \tilde{I}_\tau^t} u_i^\tau \right\| \leq \frac{|\tilde{I}_\tau^t|}{n} \frac{2C_1 \zeta r_e}{\sqrt{b}} \leq \frac{2C_1 \zeta r_e}{\sqrt{b}}, \tag{20}
$$

by using Proposition 6 and $\|u_i^\tau\| \leq 2\zeta r_e$, with probability at least $1 - \frac{\nu}{4T}$ for all t.

For the second term in the case $t \geq \tau + 1$, we follow the same line as the proof of Lemma 16. 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 (19) to obtain that

$$
\begin{aligned}
&\left\| \frac{1}{n} \sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} \frac{|\tilde{I}_s^t|}{b} \sum_{i \in I^s} u_i^s \right\| \\
&\leq \frac{2C_1^4}{\sqrt{b}} \sqrt{\sum_{s=\max\{1, t-T_1+1\}}^{t} \left(\zeta \|w^s - w^{s-1}\| + \frac{2\delta}{C_2} \|w^s\| + \frac{2\delta}{C_2} \|w^{s-1}\| \right)^2} \\
&\leq \frac{2C_1^4}{\sqrt{b}} \sqrt{\sum_{s=\max\{1, t-T_1+1\}}^{t} \left(\zeta \|w^s - w^{s-1}\| + \frac{2\delta}{C_2} \|w^s\| + \frac{2\delta}{C_2} \|w^{s-1}\| \right)^2}.
\end{aligned} \tag{21}
$$

with probability at least $1 - \frac{\nu}{4T}$ for all t.

Finally, we bound the last term in the case $t \geq \tau + 1$. By using Proposition 6, we obtain

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} \sum_{i \in \tilde{I}_s^t} u_i^s \right\| \\
&\leq \frac{\sqrt{T_1}}{n} \sqrt{\sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} \left\| \sum_{i \in \tilde{I}_s^t} u_i^s \right\|^2} \\
&\leq \frac{C_1^{\frac{1}{2}}}{\sqrt{nb}} \sqrt{\sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} C_1^2 b \left(\zeta \|w^s - w^{s-1}\| + \frac{2\delta}{C_2} \|w^s\| + \frac{2\delta}{C_2} \|w^{s-1}\| \right)^2} \\
&\leq \frac{C_1^{\frac{3}{2}}}{\sqrt{b}} \sqrt{\sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} \left(\zeta \|w^s - w^{s-1}\| + \frac{2\delta}{C_2} \|w^s\| + \frac{2\delta}{C_2} \|w^{s-1}\| \right)^2} \qquad (22)
\end{aligned}$$

with probability at least $1 - \frac{\nu}{4T}$ for all t.

Combining (20), (21), and (22), we have

$$\begin{aligned}
\|g^t\| &\leq \frac{2C_1 \zeta r_e}{\sqrt{b}} + \frac{2C_1^4 + C_1^{\frac{3}{2}}}{\sqrt{b}} \sqrt{\sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} \left(\zeta \|w^s - w^{s-1}\| + \frac{2\delta}{C_2} \|w^s\| + \frac{2\delta}{C_2} \|w^{s-1}\| \right)^2} \\
&\leq \frac{C_4 \zeta r_e}{\sqrt{b}} + \frac{C_4 \zeta}{\sqrt{b}} \sqrt{\sum_{s=\max\{\tau+1, t-T_1+1\}}^{t} \|w^s - w^{s-1}\|^2} + \frac{C_4 \delta}{C_2 \sqrt{b}} \sqrt{\sum_{s=\max\{\tau, t-T_1+1\}}^{t} \|w^s\|^2}
\end{aligned}$$

with probability at least $1 - \frac{\nu}{T}$ for all $t > \tau$. Here we take $C_4 = \tilde{O}(1)$. For $t = \tau$, (20) directly implies the desired bound. ■

Now, we move to prove Lemma 18.

Proof of Lemma 18 We assume the contrary, i.e., $\max_{\tau_0 \leq t \leq \tau + T_2} \{\|x^t - x^\tau\|, \|\tilde{x}^t - x^\tau\|\} < \frac{\delta}{C_2 \rho}$, and show the following by induction: for $\tau \leq t \leq \tau + T_2$,

$$\begin{aligned}
&\text{(a)} \quad \frac{1}{2}(1+\eta\gamma)^{t-\tau} r_e \leq \|w^t\| \leq 2(1+\eta\gamma)^{t-\tau} r_e \\
&\text{(b)} \quad \|w^t - w^{t-1}\| \leq \begin{cases} r_e & (\text{for } t = \tau) \\ 3\eta\gamma(1+\eta\gamma)^{t-\tau} r_e & (\text{for } t \geq \tau + 1) \end{cases} \\
&\text{(c)} \quad \|g^t\| \leq \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2} (1+\eta\gamma)^{t-\tau} r_e.
\end{aligned}$$

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

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

It is easy to check (a) and (b) for $t = \tau$. As for (c), by taking $b \geq \frac{\zeta^2}{\delta^2}$, $\|g^t\| \leq C_4 \delta r_e \leq C_4 \gamma (1 + \eta\gamma)^{t-\tau} r_e$ holds with probability at least $1 - \frac{\nu}{T}$ by Lemma 19.

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 $t \geq \tau + 1$, we can decompose w^t as

$$
\begin{aligned}
& w^t \\
&= w^{t-1} - \eta \left(v^{t-1} - \tilde{v}^{t-1} \right) \\
&= w^{t-1} - \eta \left(\nabla f(x^{t-1}) - \nabla f(\tilde{x}^{t-1}) + v^{t-1} - \nabla f(x^{t-1}) - \tilde{v}^{t-1} + \nabla f(\tilde{x}^{t-1}) \right) \\
&= w^{t-1} - \\
&\quad \eta \left(\int_0^1 \nabla^2 f(\tilde{x}^{t-1} + \theta(x^{t-1} - \tilde{x}^{t-1}))(x^{t-1} - \tilde{x}^{t-1}) \mathrm{d}\theta + v^{t-1} - \nabla f(x^{t-1}) - \tilde{v}^{t-1} + \nabla f(\tilde{x}^{t-1}) \right) \\
&= w^{t-1} - \eta \left((\mathrm{d}H^{t-1} + H) w^{t-1} + v^{t-1} - \nabla f(x^{t-1}) - \tilde{v}^{t-1} + \nabla f(\tilde{x}^{t-1}) \right) \\
&= (I - \eta H) w^{t-1} - \eta (\mathrm{d}H^{t-1} w^{t-1} + g^{t-1}) \\
&= (I - \eta H)^{t-\tau} w^{\tau} - \eta \sum_{s=\tau}^{t-1} (I - \eta H)^{t-1-s} (\mathrm{d}H^s w^s + g^s) \\
&= (1 + \eta\gamma)^{t-\tau} r_e \mathsf{e} - \eta \sum_{s=\tau}^{t-1} (I - \eta H)^{t-1-s} (\mathrm{d}H^s w^s + g^s),
\end{aligned} \tag{23}
$$

where we use the same notation as the proof of Lemma 19. According to this decomposition, we verify (a), (b), and (c).

Verifying (a) The first term of (23) 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 $\eta \sum_{s=\tau}^{t-1} (I - \eta H)^{t-1-s} (\mathrm{d}H_s w_s + y_s)$ by $\frac{1}{2}(1 + \eta\gamma)^{t-\tau} r_e$. We have

$$
\begin{aligned}
\left\| \eta \sum_{s=\tau}^{t} (I - \eta H)^{t-s} \mathrm{d}H_s w_s \right\| &\leq \eta \sum_{s=\tau}^{t} \|I - \eta H\|^{t-s} \|\mathrm{d}H_s\| \|w_s\| \\
&\leq \eta (1 + \eta\gamma)^{t-\tau} r_e \sum_{s=\tau}^{t} \|\mathrm{d}H_s\| && (24) \\
&\leq \eta (1 + \eta\gamma)^{t-\tau} r_e T_2 \frac{\delta}{C_2} && (25) \\
&\leq \frac{\delta \eta T_2}{C_2} (1 + \eta\gamma)^{t-\tau} r_e && (26) \\
&\leq \frac{1}{4} (1 + \eta\gamma)^{t-\tau} r_e. && (27)
\end{aligned}
$$

For (24), we used the facts that the maximum eigenvalue of ηH is at most $\eta L \leq 1$ when $\eta \leq \frac{1}{L}$ and that the minimum eigenvalue is $-\eta\gamma$, which imply $\|I - \eta H\| \leq 1 + \eta\gamma$. (25) follows from the

assumptions on $\|w_s\|$. For (26), we use $t \leq \tau + T_2$ and

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

where the first inequality follows from the hessian Lipschitzness (Assumption 4). The final inequality (27) holds when we take C_2 as $C_2 \leq 4C_3 \log \frac{\delta}{C_2\rho r_e}$ (this is satisfied by taking $C_2 = \tilde{O}(C_3) = \tilde{O}(1)$).

In addition, we have

$$\begin{aligned}\left\| \eta \sum_{s=\tau}^{t} (I - \eta H)^{t-s} g^s \right\| &\leq \eta \sum_{s=\tau}^{t} \|I - \eta H\|^{t-s} \|g^s\| \\ &\leq \eta \sum_{s=\tau}^{t-1} (1+\eta\gamma)^{t-s} \frac{3C_4\gamma}{C_2} (1+\eta\gamma)^{s-\tau} r_e \qquad (28) \\ &= \eta T_2 \frac{3C_4\gamma}{C_2} (1+\eta\gamma)^{t-\tau} r_e \\ &\leq \frac{3C_4 C_3 \log \frac{\delta}{C_2\rho r_e}}{C_2} (1+\eta\gamma)^{t-\tau} (C_4\delta + \gamma) r_e \\ &\leq \frac{1}{4}(1+\eta\gamma)^{t-\tau} r_e. \qquad (29)\end{aligned}$$

Note that (28) can be checked by the same argument as (24) and the inductive hypothesis. (29) holds when we take C_2 sufficiently large such that $\frac{3C_4C_3 \log \frac{\delta}{C_2\rho r_e}}{C_2} \leq \frac{1}{4}$ holds.

Combining (27) and (29), we can bound the second term of (23) as desired, which concludes (a) holds for $t \geq \tau + 1$.

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

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

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

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

The norm of the second term can be bounded by using (a) and (b) for $\tau+1, \cdots, t-1$ and Lemma 9 as follows:

$$\begin{aligned}
&\left\| \eta \sum_{s=\tau}^{t-1} \eta H (I-\eta H)^{t-1-s} (\mathrm{d}H_s w_s + y_s) \right\| \\
&\leq \sum_{s=\tau}^{t-1} \eta \left\| \eta H (I-\eta H)^{t-1-s} \right\| \left(\|\mathrm{d}H_s\| \|w_s\| + \|y_s\| \right) \\
&\leq \sum_{s=\tau}^{t-1} \eta \left\| \eta H (I-\eta H)^{t-1-s} \right\| \left(\frac{\delta}{C_2} (1+\eta\gamma)^{s-\tau} r_e + \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2} (1+\eta\gamma)^{s-\tau} r_e \right) \\
&\leq \sum_{s=\tau}^{t-1} \eta \left\| \eta H (I-\eta H)^{t-1-s} \right\| \left(\frac{\delta}{C_2} + \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2} \right) (1+\eta\gamma)^{s-\tau} r_e \\
&\leq \sum_{s=\tau}^{t-1} \eta \left(\eta\gamma (1+\eta\gamma)^{t-1-s} + \frac{1}{t-s} \right) \left(\frac{\delta}{C_2} + \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2} \right) (1+\eta\gamma)^{s-\tau} r_e \\
&\leq \eta \left(\eta\gamma T_2 + \log T_2 \right) \left(\frac{\delta}{C_2} + \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2} \right) (1+\eta\gamma)^{t-\tau} r_e.
\end{aligned}$$

Since $T_2 = \tilde{O}\left(\frac{1}{\eta\delta}\right)$ and $\gamma \geq \delta$, setting $C_2 = \tilde{O}(1)$ and $\eta = \tilde{O}\left(\frac{1}{L}\right)$ with sufficiently large hidden constants yields $(\eta\gamma T_2 + \log T_2)\left(\frac{\delta}{C_2} + \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2}\right) \leq \gamma$. Thus, the second term is bounded by $\eta\gamma(1+\eta\gamma)^{t-\tau} r_e$.

Finally, we consider the third term. We have $\|\mathrm{d}H^t w^t\| \leq \frac{\delta}{C_2} r_e (1+\eta\gamma)^{t-\tau} r_e$ and $\|g^t\| \leq \frac{3C_1^{\frac{1}{2}} C_4 \gamma}{C_2} (1+\eta\gamma)^{t-\tau} r_e$ 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 19 and the inductive hypothesis, we have

$$\begin{aligned}
\|g^{t+1}\| &\leq \frac{C_4 \zeta r_e}{\sqrt{b}} + \frac{C_4 \zeta}{\sqrt{b}} \sqrt{\sum_{s=\max\{\tau, t-T_1+1\}}^{t} \|w^s - w^{s-1}\|^2} + \frac{C_4 \delta}{C_2 \sqrt{b}} \sqrt{\sum_{s=\max\{\tau, t-T_1+1\}}^{t} \|w^s\|^2} \\
&\leq \frac{C_4 \zeta}{\sqrt{b}} r_e + \frac{3C_1 C_4 \zeta \sqrt{n} \eta\gamma}{b} (1+\eta\gamma)^{t-\tau} r_e + \frac{C_1^{\frac{1}{2}} C_4 \sqrt{n} \delta}{C_2 b} (1+\eta\gamma)^{t-\tau} r_e \\
&\leq \left(\frac{C_4 \zeta}{\sqrt{b}} + 3C_1 C_4 \zeta \eta\gamma + \frac{C_1^{\frac{1}{2}} C_4 \delta}{C_2} \right) (1+\eta\gamma)^{t-\tau} r_e
\end{aligned}$$

with probability at least $1 - \frac{\nu}{T}$ for all t. Taking $b \geq \frac{C_2^2 \zeta^2}{\delta^2}$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $C_2 = O(C_1 C_4) = \tilde{O}(1)$ gives $\frac{C_4 \zeta}{\sqrt{b}} + C_1 C_4 \zeta \eta\gamma + \frac{C_1^{\frac{1}{2}} C_4 \delta}{C_2} \leq \frac{3C_1^{\frac{1}{2}} C_4}{C_2} \gamma$. Thus, we obtain that (c) holds for $t+1$.

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

From Lemma 18, we can ensure that SLEDGE escapes saddle points with high probability.

Lemma 20 *Let $\{x^t\}$ be a sequence generated by SLEDGE and $\tau_0(\geq 0)$ be a step where $-\gamma := \lambda_{\min}(\nabla^2 f(x^{\tau_0})) \leq -\delta$ holds. We denote the eigenvector with the eigenvalue $\lambda_{\min}(\nabla^2 f(x^{\tau_0}))$ by* e. *We take $b \geq \sqrt{n} + \frac{\zeta^2}{\delta^2}$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_2 = \frac{C_3 \log \frac{\delta}{C_2 \rho r_e}}{\eta\gamma} \lesssim \tilde{O}\left(\frac{L}{\delta}\right)$ with a constant $C_3 = O(1)$. Then, for arbitrary $\tau > \tau_0$, it holds that*

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

Proof

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

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

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 $r_e \geq \frac{r\nu}{T\sqrt{d}}$, can be elements of A at the same time since by Lemma 18, it holds that

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

with probability at least $1 - \frac{\nu}{T}$. Let $V_d(r)$ be the volume of Euclidean ball with radius r in $\mathbb{R}^d$. Then, we have

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

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

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

which gives the conclusion. ■

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

Proof of Theorem 11 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 20 holds.

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

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

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

According to Lemma 20, with probability $1-2\nu$ over all τ, it holds that if $a^\tau = 2$ then that phase successes escaping saddle points; i.e., there exists $2\tau T_2 \le t < (2\tau+1)T_2$ such that

$$\max_{t\le s\le t+T_2} \|x^s - x^t\| > \frac{\delta}{C_2\rho} \tag{30}$$

holds. (30) further leads to

$$T_2 \sum_{t=2\tau T_2}^{2(\tau+1)T_2-1} \|x^{t+1} - x^t\|^2 > \left(\frac{\delta}{C_2\rho}\right)^2 \left(\Longleftrightarrow \sum_{t=2\tau T_2}^{2(\tau+1)T_2-1} \|x^{t+1} - x^t\|^2 > \frac{\delta^2}{T_2 C_2^2 \rho^2} \right). \tag{31}$$

On the other hand, in Theorem 10, we derived that

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 \le \frac{2}{\eta}\left[\left(f(x^0) - f(x^t)\right) - \left(\frac{1}{2\eta} - \frac{L}{2} - \frac{15C_1^9\eta\zeta^2 n}{b^2}\right) \sum_{t=1}^{T} \|x^t - x^{t-1}\|^2 \right] + \frac{2Tr^2}{\eta^2}$$
$$+ \begin{cases} \dfrac{2C_1^3 T_1}{b}\left(\sigma_c^2 + \dfrac{G_c^2}{b}\right) & \text{(Option I)} \\ 0 & \text{(Option II)} \end{cases}$$

with probability $1-3\nu$. By taking $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, applying $b \ge \sqrt{n}$ and $f(x^0) - f(x^t) \le \Delta$, and rearranging terms, we obtain

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T} \|x^t - x^{t-1}\|^2$$
$$\le \begin{cases} \dfrac{2}{\eta}\left[\Delta + \dfrac{12\eta C_1^3 n}{b^2}\left(\sigma_c^2 + \dfrac{G_c^2}{b}\right)\right] + \dfrac{2Tr^2}{\eta^2} & \text{(Option I)}, \\ \dfrac{2\Delta}{\eta} + \dfrac{2Tr^2}{\eta^2} & \text{(Option II)}. \end{cases}$$

From the definition of $a^\tau = 1$ and (31), that the left-hand side is bounded as

$$\sum_{t=1}^{T} \|\nabla f(x^{t-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T} \|x^t - x^{t-1}\|^2 \ge N_1 T_2 \varepsilon^2 + \frac{\delta^2 N_2}{2\eta^2 T_2 C_2^2 \rho^2}.$$

Thus, it holds that

$$\max\left\{N_1 T_2 \varepsilon^2, N_2 T_2 \cdot \frac{\delta^2}{2\eta^2 T_2^2 C_2^2 \rho^2}\right\} \le \begin{cases} \dfrac{2}{\eta}\left[\Delta + \dfrac{12\eta C_1^3 n}{b^2}\left(\sigma_c^2 + \dfrac{G_c^2}{b}\right)\right] + \dfrac{2Tr^2}{\eta^2} & \text{(Option I)}, \\ \dfrac{2\Delta}{\eta} + \dfrac{2Tr^2}{\eta^2} & \text{(Option II)}. \end{cases} \tag{32}$$

By the parameter settings, we have $\frac{2\eta^2 T_2^2 C_2^2 \rho^2}{\delta^2} = \tilde{O}\left(\frac{\rho^2}{\delta^4}\right)$. 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 (32)). Taking $T \geq 2(N_1+N_2+1)T_2$, there exists τ such that $a^\tau = 3$, which concludes the proof. ■

Remark 21 *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 18, if we carefully check the proof, we can see that the condition $b \gtrsim \sqrt{n} + \frac{\zeta^2}{\delta^2}$ is needed only for the step τ. On the other hand, for all $\tau_0 \leq t \leq \tau + T_2$ except for $t = \tau$, $b = \sqrt{n}$ is sufficient.*

If we take $b = \sqrt{n} + \frac{\zeta^2}{\delta^2}$ only at $t = (2\tau+1)T_2$ $\left(\tau = 0, \cdots, \frac{T}{2T_2} - 1\right)$ and $b = \sqrt{n}$ at the other steps, the above argument still holds with a slight modification. Then, the gradient complexity is reduced to

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

In the classical setting $\delta = O(\sqrt{\rho\varepsilon})$, this bound is no worse than SPIDER-SFO$^+$(+Neon2) [2, 9], no matter what n and δ are.

Finally, we note that if δ is too small, $\frac{L^2}{\delta^2}$ can be as large as n. In such case, it is more efficient to replace sampling such number of samples is replaced by full gradient computation. Then, the complexity gets

$$\tilde{O}\left(\left(L\Delta + \sigma_c^2 + \frac{G}{b}\right)\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\rho^2\sqrt{n}}{\delta^4} + \frac{n\delta}{L\varepsilon^2} + \frac{n\rho^2}{L\delta^3}\right)\right) \quad \text{(Option I)},$$
$$\tilde{O}\left(n + L\Delta\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\rho^2\sqrt{n}}{\delta^4} + \frac{n\delta}{L\varepsilon^2} + \frac{n\rho^2}{L\delta^3}\right)\right) \quad \text{(Option II)}.$$

When $\delta = O(\sqrt{\rho\varepsilon})$, this bound is no worse than NestedSVRG+Neon2 [2, 48]. However, it is unusual to assume $\frac{L^2}{\delta^2} = n$ in the first place. In fact, carefully looking the proof of SSRGD [24], we find that they implicitly limits their analysis to the case of $\frac{L^2}{\delta^2} \lesssim n$.

D.4. Convergence under PL condition (proof of Theorem 12)

In this subsection, we provide the proof of the convergence under Assumption 6, i.e., PL condition holds for the objective function.

Proof According to the descent lemma (Lemma 14) and PL condition (Assumption 6), we have that

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

Rearranging the terms yields

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

By applying Lemma 13 to this, we obtain that with probability at least $1 - 3\nu$,

$$\begin{aligned}&f(x^t) - f^* \\ &\leq (1-\eta\mu)(f(x^{t-1}) - f^*) - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^t - x^{t-1}\|^2 + \frac{r^2}{\eta} \\ &\quad + \begin{cases} \dfrac{15C_1^8\eta\zeta^2}{b} \displaystyle\sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2 + \dfrac{12C_1^2\eta\mathbb{1}[t<T_1]}{b}\cdot\left(\sigma_c^2 + \dfrac{G_c^2}{b}\right) & \text{(Option I)} \\ \dfrac{15C_1^8\eta\zeta^2}{b} \displaystyle\sum_{s=\max\{1,t-T_1+1\}}^{t} \|x^s - x^{s-1}\|^2 & \text{(Option II)} \end{cases}\end{aligned}$$

holds for all $t = 1, \cdots, T$. Multiplying both sides by $(1-\eta\mu)^{T-t}$ and summing up over all $t = 1, 2, \cdots, T$ and arranging the terms, we get

$$\begin{aligned}&f(x^T) - f^* \\ &\leq (1-\eta\mu)^T(f(x^0) - f^*) + \sum_{t=1}^{T}(1-\eta\mu)^{T-t}\frac{r^2}{\eta} \\ &\quad - \sum_{t=1}^{T}(1-\eta\mu)^{T-t}\left(\frac{1}{2\eta} - \frac{L}{2} - \frac{15C_1^9\eta\zeta^2 n(1-\eta\mu)^{-T_1}}{b^2}\right)\|x^t - x^{t-1}\|^2 \\ &\quad + \begin{cases} (1-\eta\mu)^{T-T_1}\dfrac{12C_1^3\eta n}{b^2}\left(\sigma_c^2\dfrac{G_c^2}{b}\right) & \text{(Option I)} \\ 0 & \text{(Option II)}. \end{cases}\end{aligned}$$

Note that $T_1 = \frac{n}{b}C_1$. According to this, we take η as

$$\eta = \Theta\left(\frac{1}{L} \wedge \frac{b}{C_1^{4.5}\zeta\sqrt{n}} \wedge \frac{b}{\mu C_1 n}\right)$$

so that $\frac{1}{2\eta} - \frac{L}{2} - \frac{15C_1^9\eta\zeta^2 n(1-\eta\mu)^{-T_1}}{b^2} \geq 0$ holds. Then, we have that

$$\begin{aligned}f(x^t) - f^* &\leq (1-\eta\mu)^T(f(x^0) - f^*) + \sum_{t=1}^{T}(1-\eta\mu)^{T-t}\frac{r^2}{\eta} \\ &\quad + \begin{cases} (1-\eta\mu)^{T-T_1}\dfrac{12C_1^3\eta n}{b^2}\left(\sigma_c^2 + \dfrac{G_c^2}{b}\right) & \text{(Option I)} \\ 0 & \text{(Option II)} \end{cases}\end{aligned}$$

The first term $(1-\eta\mu)^T(f(x^0)-f^*)$ is smaller than $\frac{\varepsilon}{3}$ if we take $T = O\left(\frac{1}{\eta\mu}\log\frac{\Delta}{\varepsilon}\right)$. The second term is bounded by $\frac{r^2}{\eta^2\mu}$, which is smaller than $\frac{\varepsilon}{3}$ if we take $r \le \eta\sqrt{\frac{\varepsilon\mu}{3}}$. The third term for Option I, $(1-\eta\mu)^{T-T_1}\frac{12\eta C_1^3 n}{b^2}\left(\sigma_c^2+\frac{G}{b}\right)$, is also bounded by $\frac{\varepsilon}{3}$, if we take $T = T_1 + O\left(\frac{1}{\eta\mu}\log\frac{\frac{C_1^3\eta n}{b^2}\left(\sigma_c^2+\frac{G_c^2}{b}\right)}{\varepsilon}\right) = O\left(\frac{n}{b}C_1 + \frac{C_1}{\eta\mu}\log\frac{(\sigma_c+G_c)}{\varepsilon}\right)$.

Thus, for Option I, taking

$$T = O^*\left(\frac{n}{b}C_1 + C_1\left(\frac{L}{\mu}\vee\frac{C_1^{4.5}\zeta\sqrt{n}}{\mu b}\vee\frac{C_1 n}{b}\right)\log\frac{\Delta+\sigma_c+G_c}{\varepsilon}\right),$$

yields the desired bound with probability at least $1-3\nu$.

And for Option II, taking

$$T = O\left(\left(\frac{L}{\mu}\vee\frac{C_1^{4.5}\zeta\sqrt{n}}{\mu b}\vee\frac{C_1 n}{b}\right)\log\frac{\Delta}{\varepsilon}\right)$$

yields the desired bound.

Note that T depends on ε^{-1} only logarithmically, which means that C_1 depends on ε^{-1} in only $\log\log$ order and $C_1 = O^*(\log(n+\mu^{-1}+\nu^{-1}))$, where O^* suppresses $\log\log$ factors. ■

Appendix E. Missing Statements and Proofs for FLEDGE

This section provides the missing information of FLEDGE that we abbreviate in section 3 and gives the proofs of the theorems on FLEDGE about the convergence property of FLEDGE. First, we provide the full version of FLEDGE in the following, including Option I. Note that $B(0,r)$ is the uniform distribution on the Euclidean ball in $\mathbb{R}^d$ with radius r. As the case of SLEDGE, we divide Theorem 2 into Theorem 22 (first-order optimality), Theorem 28 (second-order optimality), Theorem 34 (exponential convergence under the PL condition), each of which will be presented in one of the following subsections in order.

E.1. Finding First-order Stationary Points (Proof of Theorem 22)

In this subsection, we show that Algorithm 2 finds first-order stationary points with high probability. First, we describe the formal statement of Theorem 22.

Theorem 22 *Let $r \le \frac{\eta\varepsilon}{2\sqrt{2}}$ and $PKb \ge \tilde{\Omega}\left(\frac{\sigma^2}{\varepsilon^2}+\frac{G}{\varepsilon}\right)$. Under Assumptions 1 to 3 and 5, if we choose*

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

Algorithm 3 FLEDGE$(x^0, \eta, p, b, T, K, r)$ (formal)

Option I:
  Randomly select p agents I^0
  for $i \in I^0$ in parallel **do**
    Randomly select bK samples J_i^0
    $y_i^0 \leftarrow \frac{1}{bK}\sum_{j\in J_i^0} \nabla f_{i,j}(x^0)$
  Communicate $\{y_i^0\}_{i\in I^0}$ between I^0
  $y_i^0 \leftarrow \frac{1}{p}\sum_{i\in I^0} y_i^0 \ (i = 1, \cdots, n)$ // we do not need to explicitly communicate this between all the clients
Option II:
  for $i \in I^0 = I$ in parallel **do**
    Randomly select bK samples J_i^0
    $y_i^0 \leftarrow \frac{1}{bK}\sum_{j\in J_i^0} \nabla f_{i,j}(x^0)$
for $t = 1$ to T **do**
  Randomly sample one agent i_t
  Communicate $\{\frac{1}{P}\sum_{i=1}^P y_i^{t-1}\}$, and x^{t-1} between $I^{t-1} \cup \{i^t\}$ and the server
  $x^{t,0} \leftarrow x^{t-1},\ z^{t,0} \leftarrow 0$
  for $k = 1$ to K **do**
    $$x^{t,k} \leftarrow x^{t,k-1} - \left(\frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k-1}\right) + \xi^{t,k} \quad (\xi^{t,k} \sim B(0, r))$$
    randomly select b samples $J_{i_t}^{t,k}$
    $z^{t,k} \leftarrow z^{t,k-1} + \frac{1}{b}\sum_{j\in J_{i_t}^{t,k}}(\nabla f_{i_t,j}(x^{t,k}) - \nabla f_{i_t,j}(x^{t,k-1}))$
  $x^t \leftarrow x^{t,K}$
  Randomly select p agents I^t
  Communicate x^t between $I \cup \{i_t\}$
  for $i \in I^t$ in parallel **do**
    Randomly select b samples J_i^t
    $y_i^t \leftarrow \frac{1}{bK}\sum_{j\in J_i^t} \nabla f_{i,j}(x^t)$
    $\Delta y_i^t \leftarrow \frac{1}{bK}\sum_{j\in J_i^t}(\nabla f_{i,j}(x^t) - f_{i,j}(x^{t-1}))$
  Communicate $\{\Delta y_i^t\}_{i\in I^t}$ between I^t and 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 only $\frac{1}{P}\sum_{i=1}^P y_i^t$ in the server in $O(p)$ time.

Algorithm 2 with Option I finds an ε-first-order stationary point for problem (2) by using

$$\tilde{O}\left(\left(L \vee \frac{\zeta\sqrt{P}K}{p} \vee \frac{\zeta\sqrt{PK}}{p\sqrt{b}} \vee \zeta K \vee \frac{L\sqrt{K}}{\sqrt{b}}\right)\frac{\Delta pb}{\varepsilon^2} \wedge \left(\frac{\sigma^2 P}{p^2 b} \vee \frac{PG^2}{p^3 K b^2} \vee \frac{PK\sigma_c^2}{p^2} \vee \frac{PKG_c^2}{p^3}\right)\frac{pb}{\varepsilon^2}\right)$$

stochastic gradients and

$$\tilde{O}\left(\left(\frac{L}{K} \vee \frac{\zeta\sqrt{P}}{p} \vee \frac{\zeta\sqrt{P}}{p\sqrt{Kb}} \vee \zeta \vee \frac{L}{\sqrt{Kb}}\right)\frac{\Delta}{\varepsilon^2} \wedge \left(\frac{\sigma^2 P}{p^2 K b} \vee \frac{PG^2}{p^3 K^2 b^2} \vee \frac{P\sigma_c^2}{p^2} \vee \frac{PG_c^2}{p^3}\right)\frac{1}{\varepsilon^2}\right)$$

communication rounds

with probability at least $1 - 8\nu$.

Moreover, under the same assumptions, Algorithm 2 with Option II finds an ε-first-order stationary point for problem (2) by using

$$\tilde{O}\left(PKb + \left(L \vee \frac{\zeta\sqrt{P}K}{p} \vee \frac{\zeta\sqrt{PK}}{p\sqrt{b}} \vee \zeta K \vee \frac{L\sqrt{K}}{\sqrt{b}}\right)\frac{\Delta pb}{\varepsilon^2}\right) \text{ stochastic gradients and}$$

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

with probability at least $1 - 8\nu$.

Let $v^{t,k-1} = \frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} + z^{t,k-1}$ and $K(t)$ be the last inner loop step in the t-th outer loop as stated in the algorithm. The descent lemma (Lemma 14) also works here: as was discussed for Algorithm 1, for each t and k $(1 \le t \le T, 1 \le k \le K(t))$, it holds that

$$\begin{aligned} f(x^{t,k}) \le f(x^{t,k-1}) &+ \eta\|\nabla f(x^{t,k-1}) - v^{t,k-1}\|^2 \\ &- \frac{\eta}{2}\|\nabla f(x^{t,k-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x^{t,k} - x^{t,k-1}\|^2 + \frac{r^2}{\eta}. \end{aligned} \tag{33}$$

Our strategy is to bound the variance term $\|v^{t,k} - \nabla f(x^{t,k})\|^2$ with high probability, as summarized in the following lemma.

Lemma 23 *Let $v^{t,k} = \frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} + z^{t,k}$ and all the other variables be as stated in algorithm 3. Then, with taking $T_3 = \frac{P}{p}C_1$, we have*

$$\begin{aligned} &\left\|v^{t,k} - \nabla f(x^{t,k})\right\|^2 \\ &\le \left(\frac{120C_1^8\zeta^2K}{p} + \frac{32C_1^{10}\zeta^2}{pb}\right)\sum_{s=\max\{1,t-T_3\}}^{t-1}\sum_{l=1}^{K}\|x^{s,l} - x^{s,l-1}\|^2 \\ &\quad + \left(4\zeta^2K + \frac{4C_1^2L^2}{b}\right)\sum_{l=1}^{k}\|x^{t,l} - x^{t,l-1}\|^2 \\ &\quad + \frac{8C_1^2}{PKb}\left(\sigma^2 + \frac{G^2}{PKb}\right) + \begin{cases} 96C_1^2\mathbb{1}[t \le T_3]\left(\frac{\sigma^2}{pKb} + \frac{G^2}{p^2K^2b^2} + \frac{\sigma_c^2}{p} + \frac{G_c^2}{p^2}\right) & \text{(Option I)} \\ 0 & \text{(Option II)} \end{cases} \end{aligned}$$

for all t, k $(1 \le t \le T, 0 \le k \le K-1)$, with probability at least $1 - 8\nu$.

For the proof of Lemma 23, we utilize the four following auxiliary lemmas. Below, we define $\tilde{y}_i^0$ by

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

As well as the previous section, we define $\tilde{I}_s^t := [n] \setminus \bigcup_{\tau=s}^{t} I^t$ for $1 \le s \le t$. In addition, for each s, t $(s \le t)$ and $i \in [P]$, we let $T_4(t,i)$ as $T_4(t,i) := \max\{s \mid s = 0 \text{ or } 1 \le s \le t \text{ with } s \in I^s\}$, i.e., the last step when y_i^s is updated before t. We remark that the setting $T_3 = \frac{P}{p}C_1$ gives $\tilde{I}_s^t = \emptyset$ with probability at least $1 - \nu$ for all t and $s \le t - T_3$.

Lemma 24 *With probability at least $1-\nu$, the following holds for all $t = 1, \cdots, T$:*

$$\left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{|\tilde{I}_s^t|}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \leq \frac{4C_1^{10}\zeta^2}{pKb} \sum_{s=\max\{1,t-T_3+1\}}^{t} \|x^s - x^{s-1}\|^2.$$

Proof First, we decompose the left hand side as

$$\begin{aligned}
&\left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{|\tilde{I}_s^t|}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&\leq 2 \left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&\quad + 2 \left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|]}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&\leq 2 \left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&+ 2 \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{T_3 \left(|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|]\right)^2}{P^2p^2K^2b^2} \left\| \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&= 2 \left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&+ 2 \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{C_1 \left(|\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|]\right)^2}{Pp^3K^2b^2} \left\| \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2.
\end{aligned} \tag{34}$$

To bound the first term, by applying Proposition 4 to the choice of I^s and J_i^s, we have

$$\left\| \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \leq C_1^4\zeta^2 pKb\|x^s - x^{s-1}\|^2. \tag{35}$$

with probability at least $1 - \frac{\nu}{4T^2}$. Then, we use Proposition 7 to obtain

$$\begin{aligned}
&\left\| \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{\mathbb{E}[|\tilde{I}_s^t|]}{PpKb} \sum_{i\in I^s} \sum_{j\in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2 \\
&\leq \frac{C_1^6\zeta^2}{pKb}\|x^s - x^{s-1}\|^2
\end{aligned} \tag{36}$$

with probability at least $1 - \frac{\nu}{4T} - T \cdot \frac{\nu}{4T^2} = 1 - \frac{\nu}{2T}$.

For the second term, following the same argument in Lemma 16, we can show that $|\tilde{I}_s^t - \mathbb{E}[\tilde{I}_s^t]|^2 \leq C_1^5 P$ with probability at least $1 - \frac{\nu}{4T^2}$ for each s, t. Combining this with (35), we have

$$\sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{||\tilde{I}_s^t| - \mathbb{E}[|\tilde{I}_s^t|]|^2 C_1}{Pp^3K^2b^2} \left\| \sum_{i \in I^s} \sum_{j \in J_i^s} (\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1}))) \right\|^2$$
$$\leq \sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{C_1^{10}\zeta^2 P}{p^2 Kb} \|x^s - x^{s-1}\|^2 \tag{37}$$

with probability at least $1 - T \cdot \frac{\nu}{4T^2} - T \cdot \frac{\nu}{4T^2} = 1 - \frac{\nu}{2T}$.

Finally, substituting (36) and (37) for (34), we obtain the assertion. ■

Lemma 25 *With probability at least $1 - \nu$, the following holds for all $t = 1, \cdots, T$:*

$$\left\| \frac{1}{PbK} \sum_{i=1}^{P} \mathbb{1}[T_4(t,i) \geq t - T_3] \sum_{j \in J_i^{T_4(t,i)}} (\nabla f_{i,j}(x^{T_4(t,i)}) - \nabla f_i(x^{T_4(t,i)})) \right\|^2 \leq \frac{C_1^2}{PKb} \left(\sigma^2 + \frac{G^2}{PKb} \right).$$

Proof We condition the events on $\{T_4(i,s)\}$ and apply the Bernstein's inequality to obtain the desired bound. ■

Lemma 26 *With probability at least $1 - \nu$, the following holds for all $t = 1, \cdots, T$:*

$$\left\| \frac{\mathbb{1}[t \leq T_3]}{P} \sum_{i \in \tilde{I}_1^t} (y_i^0 - \tilde{y}_i^0) \right\|^2 \leq \begin{cases} \dfrac{C_1^2 \mathbb{1}[t \leq T_3]}{pKb} \left(\sigma^2 + \dfrac{G^2}{pKb} \right) & \text{(Option I)} \\ \dfrac{C_1^2}{PKb} \left(\sigma^2 + \dfrac{G^2}{PKb} \right) & \text{(Option II)} \end{cases}$$

Proof Recall the definition of $\tilde{y}_i^0$:

Option I By conditioning I^0, Proposition 4 yields that

$$\left\| \frac{\mathbb{1}[t \leq T_3]}{P} \sum_{i \in \tilde{I}_1^t} (y_i^0 - \tilde{y}_i^0) \right\|^2 = \left\| \frac{|\tilde{I}_1^t| \mathbb{1}[t \leq T_3]}{P} \cdot \frac{1}{pKb} \sum_{i \in I^0} \sum_{j \in J_i^0} (\nabla f_{i,j}(x^0) - \nabla f_i(x^0)) \right\|^2$$
$$\leq \frac{C_1^2 \mathbb{1}[t \leq T_3]}{pKb} \left(\sigma^2 + \frac{G^2}{pKb} \right),$$

with probability at least $1 - \frac{\nu}{T}$ for each t.

Option II In this case, Proposition 4 directly yields that

$$\left\|\frac{\mathbb{1}[t\le T_3]}{P}\sum_{i\in\tilde{I}_1^t}(y_i^0-\tilde{y}_i^0)\right\|^2=\left\|\frac{|\tilde{I}_1^t|\mathbb{1}[t\le T_3]}{P}\cdot\frac{1}{PKb}\sum_{i\in I}\sum_{j\in J_i^0}(\nabla f_{i,j}(x^0)-\nabla f_i(x^0))\right\|^2$$
$$\le\frac{C_1^2}{PKb}\left(\sigma^2+\frac{G^2}{PKb}\right),$$

with probability at least $1-\frac{\nu}{T}$ for each t.

■

Lemma 27 *With probability at least $1-\nu$, the following holds for all $t=1,\cdots,T-1$ and $k=1,\cdots,K-1$:*

$$\|z^{t,k}-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))\|^2\le\left(2\zeta^2K+\frac{2C_1^2L^2}{b}\right)\sum_{l=1}^k\|x^{t,l}-x^{t,l-1}\|^2$$

Proof We decompose the $z^{t,k}-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))$ as

$$\begin{aligned}
&\|z^{t,k}-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))\|^2\\
&\le 2\|z^{t,k}-(\nabla f_{i_t}(x^{t,k})-\nabla f_{i_t}(x^{t,0}))\|^2+2\|\nabla f_{i_t}(x^{t,k})-\nabla f_{i_t}(x^{t,0})-(\nabla f(x^{t,k})-\nabla f(x^{t,0}))\|^2\\
&\le 2\left\|\sum_{l=1}^k\frac{1}{b}\sum_{j\in J_{i_t}^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))\right\|^2+2\zeta^2\|x^{t,k}-x^{t,0}\|^2\\
&\le 2\left\|\sum_{l=1}^k\frac{1}{b}\sum_{j\in J_{i_t}^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))\right\|^2+2\zeta^2K\sum_{l=1}^k\|x^{t,l}-x^{t,l-1}\|^2,
\end{aligned}\tag{38}$$

where we use Assumption 5 for the second inequality.

We apply Proposition 4 to $\frac{1}{b}\sum_{j\in J_{i_t}^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))$ and obtain that

$$\left\|\frac{1}{b}\sum_{j\in J_{i_t}^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))\right\|^2\le\frac{C_1^2L^2}{b}\|x^{t,l}-x^{t,l-1}\|^2$$

with probability at least $1-\frac{1}{2TK^2}$ for each (t,l). Using Proposition 7, with probability $1-K\cdot\frac{1}{2TK^2}-\frac{1}{2TK}=1-\frac{1}{TK}$, we have

$$\left\|\sum_{l=1}^k\frac{1}{b}\sum_{j\in J_{i_t}^{t,l}}(\nabla f_{i_t,j}(x^{t,l})-\nabla f_{i_t,j}(x^{t,l-1}))\right\|^2\le\frac{C_1^4L^2}{b}\sum_{l=1}^k\|x^{t,l}-x^{t,l-1}\|^2\tag{39}$$

for each (t, k).

By substituting (39) to (38), we obtain the desired bound. ■

Proof of Lemma 23 First, we observe that

$$\begin{aligned}
&\left\| \frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} + z^{t,k} - \nabla f(x^{t,k}) \right\|^2 \\
&= \left\| \frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} - \nabla f(x^{t-1}) + z^{t,k} - (\nabla f(x^{t,k}) - \nabla f(x^{t,0})) \right\|^2 \\
&\leq 2\left\| \frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} - \nabla f(x^{t-1}) \right\|^2 + 2\left\| \nabla f(x^{t-1}) + z^{t,k} - (\nabla f(x^{t,k}) - \nabla f(x^{t,0})) \right\|^2
\end{aligned} \tag{40}$$

We first bound $\left\| \frac{1}{P}\sum_{i=1}^{P} y_i^{t-1} - \nabla f(x^{t-1}) \right\|^2$.

Similarly to Lemma 13, with probability at least $1-\nu$, $\tilde{I}_s^t = \emptyset$ holds for all $s \leq t - T_3$ and we can expand $\frac{1}{P}\sum_{i=1}^{P} y_i^t - \nabla f(x^t)$ as

$$\begin{aligned}
&\frac{1}{P}\sum_{i=1}^{P} y_i^t - \nabla f(x^t) \\
&= \underbrace{\frac{1}{P}\sum_{s=\max\{1,t-T_3+1\}}^{t} \left(\frac{|\tilde{I}_s^t|}{p}\sum_{i\in I^s}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) - \sum_{i\in \tilde{I}_s^t}(\nabla f_i(x^s) - \nabla f_i(x^{s-1})) \right)}_{\text{(a)}} \\
&\quad + \underbrace{\frac{\mathbb{1}[t \leq T_3]}{P}\sum_{i\in \tilde{I}_1^t}(\tilde{y}_i^0 - \nabla f_i(x^0))}_{\text{(a)}} \\
&\quad + \underbrace{\frac{1}{P}\sum_{s=\max\{1,t-T_3+1\}}^{t} \frac{|\tilde{I}_s^t|}{pKb}\sum_{i\in I^s}\sum_{j\in J_i^s}(\nabla f_{i,j}(x^s) - \nabla f_{i,j}(x^{s-1}) - (\nabla f_i(x^s) - \nabla f_i(x^{s-1})))}_{\text{(b)}} \\
&\quad + \underbrace{\frac{1}{PKb}\sum_{i=1}^{P}\mathbb{1}[T_4(t,i) \geq t - T_3]\sum_{j\in J_i^{T_4(t,i)}}(\nabla f_{i,j}(x^{T_4(t,i)}) - \nabla f_i(x^{T_4(t,i)}))}_{\text{(c)}} \\
&\quad + \underbrace{\frac{\mathbb{1}[t \leq T_3]}{P}\sum_{i\in \tilde{I}_1^t}(y_i^0 - \tilde{y}_i^0)}_{\text{(d)}}
\end{aligned}$$

for all t.

The norm of the part (a) can be bounded by using Lemma 13, just replacing n by P, i.e.,

$$\|(\mathrm{a})\|^2 \leq \frac{15C_1^8\zeta^2}{p} \sum_{s=\max\{1,t-T_3+1\}} \|x^s - x^{s-1}\|^2 + \frac{12C_1^2 \mathbb{1}[t \leq T_3]}{p}\left(\sigma_c^2 + \frac{G_c^2}{p}\right)$$

for Option I and

$$\|(\mathrm{a})\|^2 \leq \frac{15C_1^8\zeta^2}{p} \sum_{s=\max\{1,t-T_3+1\}} \|x^s - x^{s-1}\|^2$$

for Option II, with probability at least $1-3\nu$ for all t. For the bound of (b), (c) and (d), we apply Lemma 24, Lemma 25, and Lemma 26, respectively.

Then, by summarizing all these and using $\|x^s - x^{s-1}\|^2 \leq K\sum_{l=1}^K \|x^{s,l} - x^{s,l-1}\|^2$, we get

$$\begin{aligned}
&\left\|\frac{1}{P}\sum_{i=1}^P y_i^t - \nabla f(x^t)\right\|^2 \\
&\leq 4\|(\mathrm{a})\|^2 + 4\|(\mathrm{b})\|^2 + 4\|(\mathrm{c})\|^2 + 4\|(\mathrm{d})\|^2 \\
&\leq \left(\frac{60C_1^8\zeta^2}{p} + \frac{16C_1^{10}\zeta^2}{pKb}\right) \sum_{s=\max\{1,t-T_3+1\}}^{t} \|x^s - x^{s-1}\|^2 + \frac{4C_1^2}{PKb}\left(\sigma^2 + \frac{G^2}{PKb}\right) \\
&\qquad + 48C_1^2\mathbb{1}[t \leq T_3]\left(\frac{\sigma^2}{pKb} + \frac{G^2}{p^2K^2b^2} + \frac{\sigma_c^2}{p} + \frac{G_c^2}{p^2}\right) \\
&\leq \left(\frac{60C_1^8\zeta^2K}{p} + \frac{16C_1^{10}\zeta^2}{pb}\right) \sum_{s=\max\{1,t-T_3+1\}}^{t} \sum_{l=1}^{K} \|x^{s,l} - x^{s,l-1}\|^2 + \frac{4C_1^2}{PKb}\left(\sigma^2 + \frac{G^2}{PKb}\right) \\
&\qquad + 48C_1^2\mathbb{1}[t \leq T_3]\left(\frac{\sigma^2}{pKb} + \frac{G^2}{p^2K^2b^2} + \frac{\sigma_c^2}{p} + \frac{G_c^2}{p^2}\right)
\end{aligned}$$

for Option I and

$$\begin{aligned}
&\left\|\frac{1}{P}\sum_{i=1}^P y_i^t - \nabla f(x^t)\right\|^2 \\
&\leq \left(\frac{60C_1^8\zeta^2}{p} + \frac{16C_1^{10}\zeta^2}{pKb}\right) \sum_{s=\max\{1,t-T_3+1\}}^{t} \|x^s - x^{s-1}\|^2 + \frac{4C_1^2}{PKb}\left(\sigma^2 + \frac{G^2}{PKb}\right) \\
&\leq \left(\frac{60C_1^8\zeta^2K}{p} + \frac{16C_1^{10}\zeta^2}{pb}\right) \sum_{s=\max\{1,t-T_3+1\}}^{t} \sum_{l=1}^{K} \|x^{s,l} - x^{s,l-1}\|^2 + \frac{4C_1^2}{PKb}\left(\sigma^2 + \frac{G^2}{PKb}\right)
\end{aligned}$$

for Option II, with probability $1-7\nu$ for all t.

Also, we have $\|z^{t,k} - (\nabla f(x^{t,k} - \nabla f(x^{t,0}))\|^2 \leq \left(2\zeta^2K + \frac{2C_1^2L^2}{b}\right)\sum_{l=1}^k \|x^{t,l} - x^{t,l-1}\|^2$ by Lemma 27, with probability $1-\nu$ over all t, k.

By substituting these bound to (40), we obtain the desired bound.

■

Now, we are ready to prove the first-order convergence of FLEDGE.

Proof of Theorem 22 Summing up (33) over all t and k and rearranging the terms, we get

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2 \le \frac{2}{\eta}\left[\left(f(x^0)-f(x^T)\right)-\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,k}-x^{t,k-1}\|^2+\eta\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})-v_{t,k-1}\|^2+\frac{2TKr^2}{\eta^2}\right].$$

Applying Lemma 23 to this, we have that

$$\begin{aligned}\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2 &\le \frac{2}{\eta}\left(f(x^0)-f(x^T)\right)\\ &-\frac{2}{\eta}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{2\eta}-\frac{L}{2}-\eta\left(\frac{120C_1^9\zeta^2PK^2}{p^2}+\frac{128C_1^{11}\zeta^2PK}{p^2b}+4\zeta^2K^2+\frac{4C_1^2L^2K}{b}\right)\right)\|x^{t,k}-x^{t,k-1}\|^2\\ &+\frac{16C_1^2T}{Pb}\left(\sigma^2+\frac{G^2}{PKb}\right)+\frac{2TKr^2}{\eta^2}\\ &+\begin{cases}192C_1^3\left(\frac{\sigma^2P}{p^2b}+\frac{PG^2}{p^3Kb^2}+\frac{PK\sigma_c^2}{p^2}+\frac{PKG_c^2}{p^3}\right) & \text{(Option I)}\\ 0 & \text{(Option II)}\end{cases}\end{aligned}$$

with probability at least $1-8\nu$.

Option I We set η as

$$\begin{aligned}\eta &= \min\left\{\frac{1}{2L},\left(\frac{480C_1^9\zeta^2P}{p^2}+\frac{604C_1^{11}\zeta^2PK}{p^2b}+16\zeta^2K^2+\frac{16C_1^2L^2K}{b}\right)^{-\frac{1}{2}}\right\}\\ &= \tilde{\Theta}\left(\frac{1}{L}\wedge\frac{p}{\zeta\sqrt{P}K}\wedge\frac{p\sqrt{b}}{\zeta\sqrt{PK}}\wedge\frac{1}{\zeta K}\wedge\frac{\sqrt{b}}{L\sqrt{K}}\right),\end{aligned}$$

so that $\frac{1}{2\eta}-\frac{L}{2}-\eta\left(\frac{152C_1^{11}\zeta^2PK}{p^2b}+4\zeta^2K^2+\frac{4C_1^2L^2K}{b}\right)\ge 0$ holds. By taking $r\le\frac{\eta\varepsilon}{2\sqrt{2}}$ and $PKb\ge\frac{128C_1^2\sigma^2}{\varepsilon^2}+\frac{8\sqrt{2}C_1G}{\varepsilon}$, we obtain $\frac{2TKr^2}{\eta^2}\le\frac{TK\varepsilon^2}{4}$ and $\frac{16C_1^2T}{Pb}\left(\sigma^2+\frac{G^2}{PKb}\right)\le\frac{TK\varepsilon^2}{4}$, respectively. Moreover, we apply $f(x^0)-f(x^t)\le\Delta$. Summarizing these, we get

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2\le\frac{2\Delta}{\eta}+\frac{TK\varepsilon^2}{2}+192C_1^3\left(\frac{\sigma^2P}{p^2b}+\frac{PG^2}{p^3Kb^2}+\frac{PK\sigma_c^2}{p^2}+\frac{PKG_c^2}{p^3}\right).$$

Hence, taking

$$\begin{aligned}TK &\ge \frac{4\Delta}{\eta\varepsilon^2}+\frac{384C_1^3}{\varepsilon^2}\left(\frac{\sigma^2P}{p^2b}+\frac{PG^2}{p^3Kb^2}+\frac{PK\sigma_c^2}{p^2}+\frac{PKG_c^2}{p^3}\right)\\ &= \tilde{O}\left(\left(L\vee\frac{\zeta\sqrt{P}K}{p}\vee\frac{\zeta\sqrt{PK}}{p\sqrt{b}}\vee\zeta K\vee\frac{L\sqrt{K}}{\sqrt{b}}\right)\frac{\Delta}{\varepsilon^2}\wedge\left(\frac{\sigma^2P}{p^2b}\vee\frac{PG^2}{p^3Kb^2}\vee\frac{PK\sigma_c^2}{p^2}\vee\frac{PKG_c^2}{p^3}\right)\frac{1}{\varepsilon^2}\right)\end{aligned}$$

results in

$$\frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2 \leq \varepsilon^2,$$

which implies that FLEDGE can find ε-first order stationary points with probability at least $1-8\nu$. Thus, the gradient complexity and the communication complexity are bounded as stated.

Option II We set η as the same as that for Option I, so that $\frac{1}{2\eta}-\frac{L}{2}-\eta\left(\frac{152C_1^{11}\zeta^2PK}{p^2b}+4\zeta^2K^2+\frac{4C_1^2L^2K}{b}\right) \geq 0$ holds as well. We take $r \leq \frac{\eta\varepsilon}{2\sqrt{2}}$ and $PKb \geq \frac{128C_1^2\sigma^2}{\varepsilon^2}+\frac{8\sqrt{2}C_1G}{\varepsilon}$. Then, we get

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\|\nabla f(x^{t,k-1})\|^2 \leq \frac{2\Delta}{\eta}+\frac{TK\varepsilon^2}{2}.$$

Therefore, by the similar argument to Option I, taking $TK \geq \frac{2\Delta}{\eta\varepsilon^2} = \tilde{O}\left(\left(L \vee \frac{\zeta\sqrt{P}K}{p} \vee \frac{\zeta\sqrt{PK}}{p\sqrt{b}} \vee \zeta K \vee \frac{L\sqrt{K}}{\sqrt{b}}\right)\frac{\Delta}{\varepsilon^2}\right)$ ensures that FLEDGE finds ε-first order stationary points with probability at least $1-8\nu$. ■

E.2. Finding Second-order Stationary Points (Proof of Theorem 28)

Here, we show that FLEDGE can efficiently find second-order stationary points. With a slight abuse of notations, we sometimes identify (t,k) with (t',k') when $tK+k=t'K+k'$ holds. Moreover, we say $(t_1,k_1) > (t_2,k_2)$ when $t_1K+k_1 > t_2K+k_2$.

First, we state the formal theorem as follows:

Theorem 28 *We assume Assumptions 1 to 5, and $\delta < \frac{1}{\zeta}$. Let $p \geq \sqrt{P}+\frac{C_5^2\zeta^2}{\delta^2}+\frac{C_5^2L^2}{Kb\delta^2}$ with $C_5=\tilde{O}(1)$, $b \geq K$, $K=\tilde{O}\left(\frac{L}{\zeta}\right)$, $\eta=\tilde{\Theta}\left(\frac{1}{L}\right)$, $r=\tilde{O}\left(\frac{\varepsilon}{L}\right)$, $PKb \geq O\left(\frac{\sigma^2}{\varepsilon^2}+\frac{G}{\varepsilon}\right)$ and $\nu \in (0,1)$. Then, FLEDGE with Option I finds (ε,δ)-second-order stationary points using*

$$\tilde{O}\left(\left(L\Delta+\left(\frac{\sigma^2}{pb}+\frac{G^2}{pKb^2}+K\sigma_c^2+\frac{KG_c^2}{p}\right)\right)\left(\frac{1}{K\varepsilon^2}+\frac{\rho^2}{K\delta^4}\right)pKb\right) \quad \textit{stochastic gradients and}$$
$$\tilde{O}\left(\left(L\Delta+\left(\frac{\sigma^2}{pb}+\frac{G^2}{pKb^2}+K\sigma_c^2+\frac{KG_c^2}{p}\right)\right)\left(\frac{1}{K\varepsilon^2}+\frac{\rho^2}{K\delta^4}\right)\right) \quad \textit{communication rounds},$$

with probability at least $1-12\nu$.

Moreover, FLEDGE with Option II finds (ε,δ)-second-order stationary points using

$$\tilde{O}\left(PKb+L\Delta\left(\frac{1}{\varepsilon^2}+\frac{\rho^2}{\delta^4}\right)pKb\right) \quad \textit{stochastic gradients and}$$
$$\tilde{O}\left(L\Delta\left(\frac{1}{\varepsilon^2}+\frac{\rho^2}{\delta^4}\right)\right) \quad \textit{communication rounds},$$

with probability at least $1-12\nu$.

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

Lemma 29 (Small Stuck Region) *Assume $\delta < \frac{1}{\zeta}$. Let $\{x^{t,k}\}$ be a sequence generated by FLEDGE and (τ_0, κ_0) $(0 \le \kappa_0 < K)$ be a step where $-\gamma := \lambda_{\min}(\nabla^2 f(x^{\tau_0,\kappa_0})) \le -\delta$ 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 FLEDGE 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 $r_e \ge \frac{r\nu}{TK\sqrt{d}}$. Let $w^{t,k} = x^{t,k} - \tilde{x}^{t,k}$, $w^t = x^t - \tilde{x}^t$, $v^{t,k} = \frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k}$, $\tilde{v}^{t,k} = \frac{1}{P}\sum_{i=1}^P \tilde{y}_i^{t-1} + z^{t,k}$, $g^t = \frac{1}{P}\sum_{i=1}^P y_i^t - \nabla f(x^t) - \left(\frac{1}{P}\sum_{i=1}^P \tilde{y}_i^t - \nabla f(\tilde{x}^t)\right)$, 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, $v^{t,k} - \nabla f(x^{t,k}) - (\tilde{v}^{t,k} - \nabla f(\tilde{x}^{t,k})) = g^{t-1} + h^{t,k}$ holds.*

Then, there exists a sufficiently large constants $C_5 = \tilde{O}(1)$ and $C_6 = O(1)$ with which the following holds: If we take $p \ge \sqrt{P} + \frac{C_5^2\zeta^2}{\delta^2} + \frac{C_5^2 L^2}{Kb\delta^2}, b \ge K, K = O(\frac{L}{\zeta})$, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_5 = \frac{C_6 \log \frac{\delta}{C_5\rho r_e}}{\eta\gamma} \lesssim \tilde{O}\left(\frac{L}{\delta}\right)$, with probability $1 - \frac{3\nu}{TK}$ $(\nu \in (0,1))$, we have

$$\max_{(\tau_0,\kappa_0)\le(t,k)<(\tau_0+1,T_5)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} \ge \frac{\delta}{C_5\rho}.$$

In order to show Lemma 29, we prepare the two following lemmas, which bound the difference between gradient estimation errors of the two sequence.

Lemma 30 *Under the same assumption as that of Lemma 29, we assume $\max_{(\tau_0,\kappa_0)\le(t,k)<(\tau_0+1,T_5)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} < \frac{\delta}{C_5\rho}$. Then, the following holds uniformly for all $(\tau_0, \kappa_0) \le (t,k) \le (\tau_0+1, T_5)$ with probability at least $1 - \frac{\nu}{TK}$:*

$$\|g^t\| \le \begin{cases} 0 \qquad (t < \tau_0), \\ \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}}\right) C_7 r_e \qquad (t = \tau_0), \\ \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}}\right) C_7 r_e + \left(\frac{\zeta\sqrt{K}}{\sqrt{p}} + \frac{L}{\sqrt{pb}}\right) C_7 \sqrt{\sum_{s=\max\{\tau_0+1,t-T_3+1\}}^t \sum_{k=1}^K \|w^{s,k} - w^{s,k-1}\|^2} \\ \qquad + \frac{C_7\delta}{C_5\sqrt{p}} \sqrt{\sum_{s=\max\{\tau_0,t-T_3+1\}}^t \|w^s\|^2} \quad (t \ge \tau_0 + 1), \end{cases}$$

where $T_3 = \frac{P}{p} C_1$, and $C_7 = \tilde{O}(1)$ is a sufficiently large constant.

Lemma 31 *Under the same assumption as that of Lemma 29, the following holds uniformly for all $t \ge \tau_0 + 1$ and $k \ge 0$ with probability at least $1 - \frac{2\nu}{TK}$:*

$$\left\|h^{t,k}\right\| \le \zeta \sum_{l=1}^k \|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_5}\|w^{t,k}\| + \frac{2\delta}{C_5}\|w^{t,0}\| + \frac{C_1^2}{\sqrt{b}} \sqrt{\sum_{l=1}^k \left(L\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_5}\|w^{t,l}\| + \frac{2\delta}{C_5}\|w^{t,l-1}\|\right)^2}.$$

For $t < \tau_0 + 1$, we have $\left\|h^{t,k}\right\| = 0$.

Proof of Lemma 30 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

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

and

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

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 Assumptions 1, 4 and 5 and $\max_{(\tau_0,\kappa_0)\le(t,k)<(\tau_0+1,T_5)}\{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} < \frac{\delta}{C_5\rho}$, we can derive that

$$\begin{aligned}
&\|u_i^s\| \le \zeta\|w^s - w^{s-1}\| + \frac{2\delta}{C_5}\|w^s\| + \frac{2\delta}{C_2}\|w^{s-1}\| \quad \text{and}\\
&\|u_{i,j}^s\| \le L\|w^s - w^{s-1}\| + \frac{2\delta}{C_5}\|w^s\| + \frac{2\delta}{C_2}\|w^{s-1}\|
\end{aligned}$$

for $s \ge \tau_0 + 1$, by similar argument to the proof of Lemma 19. For $t = \tau_0$, we have $\|u_i^{\tau_0}\| = \|(\nabla f_i(x^{\tau_0}) - \nabla f_i(\tilde{x}^{\tau_0})) - (\nabla f(x^{\tau_0}) - \nabla f(\tilde{x}^{\tau_0}))\| \le \zeta\|x^{\tau_0} - \tilde{x}^{\tau_0}\| = \zeta r_e$ and $\|u_{i,j}^{\tau_0}\| = \|(\nabla f_{i,j}(x^{\tau_0}) - \nabla f_{i,j}(\tilde{x}^{\tau_0})) - (\nabla f(x^{\tau_0}) - \nabla f(\tilde{x}^{\tau_0}))\| \le L\|x^{\tau_0} - \tilde{x}^{\tau_0}\| = Lr_e$.

As we did in Lemma 19, for $t \geq \tau_0 + 1$, we have

$$g^t = \underbrace{\frac{1}{P}\left(\frac{|\tilde{I}^t_{\tau_0}|}{p}\sum_{i\in I^{\tau_0}} u_i^{\tau_0} - \sum_{i\in \tilde{I}^t_{\tau_0}} u_i^{\tau_0}\right)}_{(a)} + \underbrace{\frac{1}{PKb}\left(\frac{|\tilde{I}^{\tau_0}_{\tau_0}|}{p}\sum_{i\in I^{\tau_0}}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0} - \sum_{i\in \tilde{I}^{\tau_0}_{\tau_0}}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0}\right)}_{(b)}$$

$$+ \underbrace{\frac{1}{P}\sum_{s=\max\{\tau_0+1,t-T_1+1\}}^{t}\left(\frac{|\tilde{I}^t_s|}{p}\sum_{i\in I^s} u_i^s\right) - \frac{1}{P}\sum_{s=\max\{\tau_0+1,t-T_1+1\}}^{t}\sum_{i\in \tilde{I}^t_s} u_i^s}_{(c)}$$

$$+ \underbrace{\frac{1}{PKb}\sum_{s=\max\{\tau_0+1,t-T_1+1\}}^{t}\left(\frac{|\tilde{I}^t_s|}{p}\sum_{i\in I^s}\sum_{j\in J_i^s} u_{i,j}^s\right) - \frac{1}{PKb}\sum_{s=\max\{\tau_0+1,t-T_1+1\}}^{t}\sum_{i\in \tilde{I}^t_s}\sum_{j\in J_i^s} u_{i,j}^s}_{(d)}$$

with probability $1 - \frac{\nu}{8TK}$ for all t. For $t = \tau_0$, $g^{\tau_0} = (\mathrm{a}) + (\mathrm{b})$ holds.

Recall the argument in Lemma 19. We have that

$$\|(\mathrm{a})\| \leq \frac{2C_1\zeta r_e}{\sqrt{p}} \quad \text{and}$$

$$\|(\mathrm{c})\| \leq \frac{2C_1^4 + C_1^{\frac{3}{2}}}{\sqrt{p}}\sqrt{\sum_{s=\max\{\tau_0+1,t-T_3+1\}}\left(\zeta\|w^s - w^{s-1}\| + \frac{2\delta}{C_5}\|w^s\| + \frac{2\delta}{C_5}\|w^{s-1}\|\right)^2}$$

hold with probability at least $1 - \frac{1}{8TK}$ for all t.

Moreover, observe that (b) and (d) are obtained just by replacing u_i^t in (a) and (c) by $\frac{1}{Kb}\sum_{j\in J_i^t} u_{i,j}^s$. Note that $\frac{1}{Kb}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0}$ is mean-zero and its norm is bounded by $\frac{C_1 L r_e}{\sqrt{Kb}}$ for $s = \tau_0$, with probability $1 - \frac{1}{8T^2K}$. Thus, Proposition 7 yields that

$$\|(\mathrm{b})\| \leq \left\|\frac{1}{P}\frac{|\tilde{I}^t_{\tau_0}|}{p}\sum_{i\in I^{\tau_0}}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0}\right\| + \left\|\frac{1}{P}\sum_{i\in \tilde{I}^t_{\tau_0}}\sum_{j\in J_i^{\tau_0}} u_{i,j}^{\tau_0}\right\| \leq \frac{|\tilde{I}^t_{\tau_0}|}{P}\frac{C_1^2 L r_e}{\sqrt{pKb}} + \frac{\sqrt{|\tilde{I}^t_{\tau_0}|}C_1^2 L r_e}{P\sqrt{Kb}} \leq \frac{2C_1^2 L r_e}{\sqrt{pKb}},$$

with probability $1 - \frac{1}{8TK} - T\cdot\frac{1}{8T^2K} = 1 - \frac{1}{4TK}$, where we use $|\tilde{I}^t_{\tau_0}| \leq P$ and $p \leq P$ for the last inequality.

For the first term of (d), we first observe that $\frac{1}{Kb}\sum_{i\in I^s}\sum_{j\in J_i^t} u_{i,j}^s$ is mean-zero and its norm is bounded by $\frac{C_1\sqrt{p}}{\sqrt{Kb}}L\|w^s - w^{s-1}\| + \frac{2C_1\delta\sqrt{p}}{C_5\sqrt{Kb}}\|w^s\| + \frac{2C_1\delta\sqrt{p}}{C_5\sqrt{Kb}}\|w^{s-1}\|$ (for $s \geq \tau_0 + 1$) with probability at least$1 - \frac{1}{8T^2K}$. Then, we apply the same argument as Lemma 16. This yields

$$\left\|\frac{1}{PKb}\sum_{s=\max\{\tau_0+1,t-T_1+1\}}^{t}\frac{|\tilde{I}^t_s|}{p}\sum_{i\in I^s}\sum_{j\in J_i^s} u_{i,j}^s\right\|$$

$$\leq \frac{2C_1^4}{\sqrt{pKb}}\sqrt{\sum_{s=\max\{\tau_0+1,t-T_1+1\}}^{t}\left(L\|w^s - w^{s-1}\| + \frac{2\delta}{C_5}\|w^s\| + \frac{2\delta}{C_5}\|w^{s-1}\|\right)^2}.$$

with probability $1 - T \cdot \frac{\nu}{8T^2K} - \frac{\nu}{8TK} = 1 - \frac{\nu}{4TK}$ for all t. As for the second term of (d), by applying Proposition 6, we get

$$\begin{aligned}
&\left\| \frac{1}{PKb} \sum_{s=\max\{\tau_0+1, t-T_1+1\}}^{t} \sum_{i \in \tilde{I}_s^t} \sum_{j \in J_i^s} u_{i,j}^s \right\| \\
&\leq \frac{\sqrt{T_3}}{P} \sqrt{\sum_{s=\max\{\tau_0+1, t-T_1+1\}}^{t} \left\| \frac{1}{Kb} \sum_{i \in \tilde{I}_s^t} \sum_{j \in J_i^s} u_{i,j}^s \right\|^2} \\
&\leq \frac{C_1^{\frac{1}{2}}}{\sqrt{Pp}} \sqrt{\sum_{s=\max\{\tau_0+1, t-T_1+1\}}^{t} \frac{C_1^2 p}{Kb} \left(L\|w^s - w^{s-1}\| + \frac{2\delta}{C_5}\|w^s\| + \frac{2\delta}{C_5}\|w^{s-1}\| \right)^2} \\
&\leq \frac{C_1^{\frac{3}{2}}}{\sqrt{pKb}} \sqrt{\sum_{s=\max\{\tau_0+1, t-T_1+1\}}^{t} \left(L\|w^s - w^{s-1}\| + \frac{2\delta}{C_5}\|w^s\| + \frac{2\delta}{C_5}\|w^{s-1}\| \right)^2}
\end{aligned}$$

with probability $1 - T \cdot \frac{\nu}{8T^2K} - \frac{\nu}{8TK} = 1 - \frac{\nu}{4TK}$ for all t.

By combining all these, we have

$$\begin{aligned}
\|g^t\| &\leq \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}} \right) C_7 \zeta r_e + \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}} \right) C_7 \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3+1\}}^{t} \|w^s - w^{s-1}\|^2} \\
&\quad + \frac{C_7 \delta}{C_5} \sqrt{\sum_{s=\max\{\tau_0, t-T_3+1\}}^{t} \|w^s\|^2} \\
&\leq \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}} \right) C_7 \zeta r_e + \left(\frac{\zeta\sqrt{K}}{\sqrt{p}} + \frac{L}{\sqrt{pb}} \right) C_7 \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3+1\}}^{t} \sum_{k=1}^{K} \|w^{s,k} - w^{s,k-1}\|^2} \\
&\quad + \frac{C_7 \delta}{C_5 \sqrt{p}} \sqrt{\sum_{s=\max\{\tau_0, t-T_3+1\}}^{t} \|w^s\|^2}
\end{aligned}$$

with probability at least $1 - \frac{\nu}{TK}$ for all $t \geq \tau_0 + 1$. Here we take $C_7 = \tilde{O}(1)$, which is independent of C_5. 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 31 Let

$$\begin{aligned}
u_i^{t,l} := &(\nabla f_i(x^{t,l}) - \nabla f_i(\tilde{x}^{t,l})) - (\nabla f_i(x^{t,0}) - \nabla f_i(\tilde{x}^{t,0})) \\
&- (\nabla f(x^{t,l}) - \nabla f(\tilde{x}^{t,l})) + (\nabla f(x^{t,0}) - \nabla f(\tilde{x}^{t,0}))
\end{aligned}$$

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

$$\begin{aligned}\|u_i^{t,k}\| &\le \zeta\|w^{t,k} - w^{t,0}\| + \frac{2\delta}{C_5}\|w^{t,k}\| + \frac{2\delta}{C_5}\|w^{t,0}\| \\ &\le \zeta\sum_{l=1}^{k}\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_5}\|w^{t,k}\| + \frac{2\delta}{C_5}\|w^{t,0}\| \end{aligned} \tag{41}$$

and

$$\|u_{i_t}^{t,l}\| \le L\|w^{t,l} - w^{t,l}\| + \frac{2\delta}{C_5}\|w^{t,l}\| + \frac{2\delta}{C_5}\|w^{t,l-1}\|.$$

Thus, applying Proposition 4 and Proposition 7 to $\frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}} u_{i_t,j}^{t,l}$, we get

$$\left\|\frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}} u_{i_t,j}^{t,l}\right\| \le \frac{C_1^2}{\sqrt{b}}\sqrt{\sum_{l=1}^{k}\left(L\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_5}\|w^{t,l}\| + \frac{2\delta}{C_5}\|w^{t,l-1}\|\right)^2} \tag{42}$$

with probability at least $1 - \frac{1}{TK}$ for all t and K.

Substituting (41) and (42) to $h^{t,k} = u_{i_t}^{t,l} + \frac{1}{b}\sum_{l=1}^{k}\sum_{j\in J_{i_t}} u_{i_t,j}^{t,l}$, we get the desired bound. ■

Now, we are ready to prove Lemma 29.

Proof of Lemma 29 We assume the contrary and show the following by induction, for $(\tau_0 + 1, 0) \le (t,k) \le (\tau_0 + 1, T_5)$:

(a) $\frac{1}{2}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e \le \|w^{t,k}\| \le 2(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e$

(b) $\|w^{t,k} - w^{t,k-1}\| \le \begin{cases} r_e & (\text{for } (t,k) = (\tau_0+1, 0)) \\ 3\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e & (\text{for } (t,k) > (\tau_0+1,0)) \end{cases}$

(c) $\|g^{t-1} + h^{t,k}\| \le \frac{2C_8\gamma}{C_5}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e.$

Here $C_8 = \tilde{O}(1)$ is a sufficiently large constant independent of C_5. Then, (a) yields contradiction by taking $(t,k) - (\tau_0+1,0) = T_5 = O\left(1 + \frac{\log \frac{\delta}{C_2 \rho r_e}}{\eta\delta K}\right)$ 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 30, taking $p \ge \frac{\zeta^2}{\delta^2} + \frac{L^2}{\delta^2 K b}$, $\|g^{\tau_0}\| \le 2C_7\delta r_e \le 2C_7\gamma r_e$ 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:

$$\begin{aligned} & w^{t,k+1} \\ &= w^{t,k} - \eta\left(v^{t,k} - \tilde{v}^{t,k}\right) \\ &= (1+\eta\gamma)^{(t-\tau_0-1)K+k+1} r_e \mathsf{e} - \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k)} (I - \eta H)^{(t-s)K+(k-l)}(\mathrm{d}H^{s,l} w^{s,l} + g^{s-1} + h^{s,l}), \end{aligned} \tag{43}$$

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

$$\begin{aligned}
&\left\|\eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k)}(I-\eta H)^{(t-s)K+(k-l)}\mathrm{d}H^{s,l}w^{s,l}\right\| \\
&\leq \eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k)}\|I-\eta H\|^{(t-s)K+(k-l)}\left\|\mathrm{d}H^{s,l}\right\|\left\|w^{s,l}\right\| \\
&\leq 2\eta(1+\eta\gamma)^{(t-s)K+(k-l)+(s-\tau_0-1)K+l}r_e\sum_{(s,l)=(\tau_0+1,0)}^{t,k}\left\|\mathrm{d}H^{s,l}\right\| \\
&\leq 2\eta(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_eT_5K\frac{\delta}{C_5} \\
&\leq \frac{2\eta\delta T_5}{C_5}(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e \\
&\leq \frac{1}{4}(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e.
\end{aligned} \tag{44}$$

The last inequality follows from the definition of $T_5 = \frac{C_6\log\frac{\delta}{C_5\rho r_e}}{\eta\gamma}$ and sufficiently large C_5.

In addition, we have

$$\begin{aligned}
&\left\|\eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k)}(I-\eta H)^{(t-s)K+(k-l)}(g^{s-1}+h^{s,l})\right\| \\
&\leq \eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k)}\|I-\eta H\|^{(t-s)K+(k-l)}\left\|g^{s-1}+h^{s,l}\right\| \\
&\leq \eta\sum_{(s,l)=(\tau_0+1,0)}^{(t,k)}(1+\eta\gamma)^{(t-s)K+(k-l)}\frac{2C_8\gamma}{C_5}(1+\eta\gamma)^{(s-\tau_0-1)K+l} \\
&\leq \frac{2\eta\gamma T_5}{C_5}(1+\eta\gamma)^{(t-\tau_0-1)K+k} \\
&\leq \frac{1}{4}(1+\eta\gamma)^{(t-\tau_0-1)K+k}r_e.
\end{aligned} \tag{45}$$

For the final inequality, we again use $T_5 = \frac{C_6\log\frac{\delta}{C_5\rho r_e}}{\eta\gamma}$ with sufficiently large C_5.

Combining (44) and (45), we get (a) for $(t, k+1)$ as desired.

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

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

As for the first term, we can bound it as

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

Evaluating the second term requires (a) and (b) for $(\tau_0+1,0), \cdots, (t,k-1)$ and Lemma 9:

$$
\begin{aligned}
& \left\| \eta \sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)} \eta H (I-\eta H)^{(t-s)K+(k-l)} (\mathrm{d}H^{s,l} w^{s,l} + g^{s-1} + h^{s,l}) \right\| \\
&\leq \sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)} \eta \left\| \eta H (I-\eta H)^{(t-s)K+(k-l)} \right\| \left(\|\mathrm{d}H^{s,l}\| \|w^{s,l}\| + \|g^{s-1} + h^{s,l}\| \right) \\
&\leq \sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)} \eta \left\| \eta H (I-\eta H)^{(t-s)K+(k-l)} \right\| \left(\frac{\delta}{C_5} (1+\eta\gamma)^{(s-\tau_0-1)K+l} r_e + \frac{2C_8\gamma}{C_5} (1+\eta\gamma)^{(s-\tau_0-1)K+l} r_e \right) \\
&\leq \sum_{(s,l)=(\tau_0+1,0)}^{(t,k-1)} \eta \left(\eta\gamma(1+\eta\gamma)^{(t-s)K+(k-l)} + \frac{1}{(t-s)K+(k-l)} \right) \left(\frac{\delta}{C_5} + \frac{2C_8\gamma}{C_5} \right) (1+\eta\gamma)^{(s-\tau_0-1)K+l} r_e \\
&\leq \eta \left(\eta\gamma T_5 + \log T_5 \right) \left(\frac{\delta}{C_5} + \frac{2C_8\gamma}{C_5} \right) (1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e.
\end{aligned}
$$

Since $T_5 = \tilde{O}\left(\frac{1}{\eta\delta}\right)$ and $\gamma \geq \delta$, setting $C_5 = \tilde{O}(1)$ with sufficiently large C_5 yields $(\eta\gamma T_5 + \log T_5)\left(\frac{\delta}{C_5} + \frac{2C_8\gamma}{C_5}\right) \leq \gamma$. 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 $\|\mathrm{d}H^{t,k} w^{t,k}\| \leq \frac{\delta}{C_5} r_e (1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e$ and $\|g^{t-1} + h^{t,k}\| \leq \frac{2C_8\gamma}{C_5}(1+\eta\gamma)^{(t-\tau_0-1)K+k} r_e$. Thus, by taking C_5 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 30 and assumptions, we have

$$
\begin{aligned}
&\|g^{t+1}\| \\
&\leq \left(\frac{\zeta}{\sqrt{p}} + \frac{L}{\sqrt{pKb}}\right) C_7 r_e + \left(\frac{\zeta\sqrt{K}}{\sqrt{p}} + \frac{L}{\sqrt{pb}}\right) C_7 \sqrt{\sum_{s=\max\{\tau_0+1, t-T_3+1\}}^{t} \sum_{k=1}^{K} \|w^{s,k} - w^{s,k-1}\|^2} \\
&\qquad + \frac{C_7\delta}{C_5\sqrt{p}} \sqrt{\sum_{s=\max\{\tau_0, t-T_3+1\}}^{t} \|w^s\|^2} \\
&\leq \left[\frac{\zeta C_7}{\sqrt{p}} + \frac{LC_7}{\sqrt{pKb}} + \left(\frac{C_7\zeta K T_3^{\frac{1}{2}}}{\sqrt{p}} + \frac{C_7 L K^{\frac{1}{2}} T_3^{\frac{1}{2}}}{\sqrt{pb}}\right) 3\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+K}\right] r_e \\
&\quad + \frac{2C_7 T_3^{\frac{1}{2}}\delta}{C_5\sqrt{pK}}(1+\eta\gamma)^{(t-\tau_0-1)K+K} r_e \\
&= \left[\frac{\zeta C_7}{\sqrt{p}} + \frac{LC_7}{\sqrt{pKb}} + \left(\frac{C_1^{\frac{1}{2}} C_7 \zeta P^{\frac{1}{2}} K}{p} + \frac{C_1^{\frac{1}{2}} C_7 L\sqrt{PK}}{p\sqrt{b}}\right) 3\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+K}\right] r_e \\
&\quad + \frac{2C_1^{\frac{1}{2}} C_7\sqrt{P}\delta}{C_5 p}(1+\eta\gamma)^{(t-\tau_0-1)K+K} r_e
\end{aligned}
$$

with probability at least $1 - \frac{\nu}{TK}$ for all t. Taking $p \geq \sqrt{P} + \frac{C_5^2\zeta^2}{\delta^2} + \frac{C_5^2 L^2}{\delta^2 Kb}$, $\eta = \Theta(\frac{1}{L})$, $b \geq K$, $K = O\left(\frac{L}{\zeta}\right)$, and $\|g^{t+1}\| \leq \frac{C_8\gamma}{C_5}(1+\eta\gamma)^{(t-\tau_0)K}$ with sufficiently large constant C_8, that only depends on C_1, C_7, and sufficiently small $\eta = \tilde{\Theta}(\frac{1}{L})$.

Moreover, Lemma 31 states that, for $k < K$,

$$
\begin{aligned}
&\left\|h^{t,k+1}\right\| \\
&\leq \zeta \sum_{l=1}^{k+1} \|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_5}\|w^{t,k+1}\| + \frac{2\delta}{C_5}\|w^{t,0}\| \\
&\quad + \frac{C_1^2}{\sqrt{b}} \sqrt{\sum_{l=1}^{k+1} \left(L\|w^{t,l} - w^{t,l-1}\| + \frac{2\delta}{C_5}\|w^{l,k}\| + \frac{2\delta}{C_5}\|w^{t,l-1}\|\right)^2}
\end{aligned}
$$

holds with probability at least $1 - \frac{\nu}{TK}$. If (a) and (b) hold for all $(s,l) \leq (t, k+1)$, then we have

$$
\begin{aligned}
\left\|h^{t,k+1}\right\| &\leq 3\zeta K\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k+1} + \frac{8\delta}{C_5}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1} \\
&\quad + \frac{3C_1^2\sqrt{K}}{\sqrt{b}} L\eta\gamma(1+\eta\gamma)^{(t-\tau_0-1)K+k+1} + \frac{8C_1^2\sqrt{K}\delta}{\sqrt{b}}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}.
\end{aligned}
$$

Taking $b \geq K$ and $K = O\left(\frac{L}{\zeta}\right)$, with sufficiently large C_8 and sufficiently small η, we have $\|h^{t,k+1}\| \leq \frac{C_8\gamma}{C_5}(1+\eta\gamma)^{(t-\tau_0-1)K+k+1}$.

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 $T_5 = \frac{C_6 \log \frac{\delta}{C_5 \rho r_e}}{\eta\gamma}$. Taking C_6 sufficiently large, we have $\frac{1}{2}(1+\eta\gamma)^{(\tau_0+1-\tau_0-1)K+T_5} r_e \geq \frac{\delta}{C_5\rho}$. This yields contradiction against the assumption and the desired assertion follows.

■

From Lemma 29, we can show that FLEDGE escapes saddle points with high probability. We have the following lemma, and the proof is essentially the same as that of Lemma 20.

Lemma 32 *Let $\{x^{t,k}\}$ be a sequence generated by FLEDGE 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 $p \geq \sqrt{P} + \frac{C_5^2\zeta^2}{\delta^2} + \frac{C_5^2 L^2}{Kb\delta^2}, b \geq \sqrt{K}$ and, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, and $T_5 = \frac{C_6 \log \frac{\delta}{C_5\rho r_e}}{\eta\gamma} \lesssim \tilde{O}\left(\frac{L}{\delta}\right)$, with sufficiently large $C_5, C_6 = \tilde{O}(1)$. Then,*

$$\mathbb{P}\left[\max_{(\tau_0,\kappa_0)\leq(t,k)<(\tau_0+1,T_5)} \|x^{t,k} - x^{\tau_0,\kappa_0+1}\| \geq \frac{\delta}{C_5\rho} \mid I^0, \cdots, I^\tau, i_0, \cdots, i_{\tau_0}, \xi^{0,0}, \cdots, \xi^{\tau_0,\kappa_0}\right] \geq 1 - \frac{4\nu}{TK}.$$

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 28 Since $T_5 = \frac{C_6 \log \frac{\delta}{C_5\rho r_e}}{\eta\gamma}$ depends on x^{τ_0}, we take $T_5 = \frac{C_6 \log \frac{\delta}{C_5\rho r_e}}{\eta\delta}$ from now instead. This change does not affect whether Lemma 32 holds. Also, we let $T_6 = \lceil 1 + \frac{T_5}{K} \rceil$.

We divide $\{t = 0, 1, \cdots, T-1\}$ into the following $\lfloor \frac{T}{2T_6} \rfloor$ *phases*: $P^\tau = \{2\tau T_6 \leq t < 2(\tau+1)T_6\}$ $\left(\tau = 0, \cdots, \lfloor \frac{T}{2T_6} \rfloor - 1\right)$. For each phase, we define a^τ as a random variable taking values

$$a^\tau = \begin{cases} 1 & \left(\text{if } \sum_{t\in P^\tau}\sum_{k=0}^K \mathbb{1}[\|\nabla f(x^{t,k})\| > \varepsilon] > KT_6\right) \\ 2 & \left(\begin{array}{l}\text{if there exists } t \text{ such that } (2\tau T_6, 0) \leq (t,k) < ((2\tau+1)T_6, 0), \|\nabla f(x^{t,k})\| \leq \varepsilon \\ \text{and } \lambda_{\min}(\nabla^2 f(x^{t,k})) \leq -\delta\end{array}\right) \\ 3 & \left(\begin{array}{l}\text{if there exists } t \text{ such that } (2\tau T_6, 0) \leq (t,k) < ((2\tau+1)T_6, 0), \|\nabla f(x^{t,k})\| \leq \varepsilon \\ \text{and } \lambda_{\min}(\nabla^2 f(x^{t,k})) > -\delta\end{array}\right) \end{cases}.$$

Note that $\mathbb{P}[a^\tau = 1, 2, 3] = 1$ for each τ. This is because if there does not exist t between $(2\tau T_6, 0) \leq (t,k) < ((2\tau+1)T_6, 0)$ such that $\|\nabla f(x^{t,k})\| \leq \varepsilon$ (i.e., neither $a^\tau = 2$ nor 3), then we have $\sum_{t\in P^\tau}\sum_{k=0}^K \mathbb{1}[\|\nabla f(x^{t,k})\| > \varepsilon] \geq \sum_{t=2\tau T_2}^{(2\tau+1)T_6-1}\sum_{k=0}^K \mathbb{1}[\|\nabla f(x^{t,k})\| > \varepsilon] = T_6 K$, meaning $a^\tau = 1$. We denote $N_1 = \sum_{\tau=0}^{\lfloor \frac{T}{2T_6} \rfloor} \mathbb{1}[a^\tau = 1]$, $N_2 = \sum_{\tau=0}^{\lfloor \frac{T}{2T_6} \rfloor} \mathbb{1}[a^\tau = 2]$, and $N_3 = \sum_{\tau=0}^{\lfloor \frac{T}{2T_6} \rfloor} \mathbb{1}[a^\tau = 3]$.

According to Lemma 32, with probability $1 - 4\nu$ over all τ, it holds that if $a^\tau = 2$ then that phase successes escaping saddle points; i.e., there exists $(2\tau T_6, 0) \leq (t,k) < ((2\tau+1)T_6, 0)$ and

$$\max_{(t,k)\leq(s,l)<((2\tau+2)T_6,0)} \|x^{s,l} - x^{t,k}\| > \frac{\delta}{C_5\rho} \tag{46}$$

holds. Eq. (46) further leads to

$$T_6K \sum_{t=2\tau T_6}^{2(\tau+1)T_6-1} \sum_{k=1}^{K} \|x^{t,k} - x^{t,k-1}\|^2 > \left(\frac{\delta}{C_5\rho}\right)^2 \Leftrightarrow \sum_{t=2\tau T_6}^{2(\tau+1)T_6-1} \sum_{k=1}^{K} \|x^{t,k} - x^{t,k-1}\|^2 > \frac{\delta^2}{T_6KC_5^2\rho^2}. \tag{47}$$

On the other hand, in Theorem 22, we derived that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{k=1}^{K} \|\nabla f(x^{t,k-1})\|^2 \\
&\leq \frac{2}{\eta}\left(f(x^0) - f(x^T)\right) \\
&\quad - \frac{2}{\eta}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\frac{1}{2\eta} - \frac{L}{2} - \eta\left(\frac{120C_1^9\zeta^2PK^2}{p^2} + \frac{128C_1^{11}\zeta^2PK}{p^2b} + 4\zeta^2K^2 + \frac{4C_1^2L^2K}{b}\right)\right)\|x^{t,k} - x^{t,k-1}\|^2 \\
&\quad + \begin{cases} \frac{16C_1^2T}{Pb}\left(\sigma^2 + \frac{G^2}{PKb}\right) + \frac{2TKr^2}{\eta^2} + 192C_1^3\left(\frac{\sigma^2P}{p^2b} + \frac{PG^2}{p^3Kb^2} + \frac{PK\sigma_c^2}{p^2} + \frac{PKG_c^2}{p^3}\right) & \text{(Option I)} \\ \frac{16C_1^2T}{Pb}\left(\sigma^2 + \frac{G^2}{PKb}\right) + \frac{2r^2}{\eta^2} & \text{(Option II)} \end{cases}
\end{aligned}$$

with probability $1 - 8\nu$. Taking $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$ sufficiently small, applying $p \geq \sqrt{P}$, $K = O\left(\frac{L}{\zeta}\right)$, $K \leq b$ and $f(x^0) - f(x^t) \leq \Delta$, and arranging terms yields

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \|\nabla f(x^{t,k-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\|x^{t,k} - x^{t,k-1}\|^2 \tag{48}$$

$$\leq \frac{2\Delta}{\eta} + \begin{cases} \frac{16C_1^2T}{Pb}\left(\sigma^2 + \frac{G^2}{PKb}\right) + \frac{2TKr^2}{\eta^2} + 192C_1^3\left(\frac{\sigma^2}{b} + \frac{G^2}{pKb^2} + K\sigma_c^2 + \frac{KG_c^2}{p}\right) & \text{(Option I)} \\ \frac{16C_1^2T}{Pb}\left(\sigma^2 + \frac{G^2}{PKb}\right) + \frac{2TKr^2}{\eta^2} & \text{(Option II)} \end{cases} \tag{49}$$

From the definition of $a^\tau = 1$ and (47), We know that (48) is bounded as

$$\sum_{t=1}^{T}\sum_{k=1}^{K} \|\nabla f(x^{t,k-1})\|^2 + \frac{1}{2\eta^2}\sum_{t=1}^{T}\sum_{k=1}^{K}\|x^{t,k} - x^{t,k-1}\|^2 \geq N_1T_6K\varepsilon^2 + \frac{\delta^2N_2}{2\eta^2T_6KC_5^2\rho^2}.$$

Thus, $N_1T_6K \leq \frac{1}{\varepsilon^2} \times$ (the RHS of (49)) and $N_2T_6K \leq \frac{2\eta^2C_2^2\rho^2K^2T_6^2}{\delta^2} \times$ (the RHS of (32)) holds.

Here, $\frac{2\eta^2T_6^2K^2C_5^2\rho^2}{\delta^2} = \tilde{O}\left(\frac{\rho^2}{\delta^4} + \frac{\eta^2K^2}{\delta^2}\right) \lesssim \tilde{O}\left(\frac{\rho^2}{\delta^4}\right)$, when $K = O\left(\frac{L}{\zeta}\right) \leq O\left(\frac{L}{\delta}\right)$. From this, $(N_1 + N_2)T_6 \leq \tilde{O}\left(\frac{1}{K\varepsilon^2} + \frac{\rho^2}{K\delta^4}\right) \times$ (the right-hand side of (49)). Taking $T \geq 2(N_1 + N_2 + 1)T_6$, there exists τ such that $a^\tau = 3$, 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 $T_7 = \tilde{\Theta}\left(\frac{L}{\delta}\right)$ with a sufficiently large hidden constant. In line 18-19 of FLEDGE, when $k \equiv T_7$, we randomly select $\frac{C_5^2 L^2}{\delta^2} + b$ (not b) samples $J_{i_t}^{t,k}$, 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 $\tilde{O}(K/(L/\delta)) \times \tilde{O}(L^2/\delta^2) = \tilde{O}(KL/\delta) \lesssim \tilde{O}(K^2) \lesssim \tilde{O}(Kb)$. Thus, this does not affect the inner-loop complexity more than by constant factors.

Then, the following lemma holds, which stands as generalization of Lemma 29.

Lemma 33 (Small stuck region) *Let $\{x^{t,k}\}$ be a sequence generated by FLEDGE 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 FLEDGE 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 $\kappa \equiv T_7$;* We let $\tilde{\xi}^{\tau,\kappa} = \xi^{\tau,\kappa} - r_e\mathsf{e}$ with $r_e \geq \frac{r\nu}{TK\sqrt{d}}$. Let $w^{t,k} = x^{t,k} - \tilde{x}^{t,k}$, $w^t = x^t - \tilde{x}^t$, $v^{t,k} = \frac{1}{P}\sum_{i=1}^P y_i^{t-1} + z^{t,k}$, $\tilde{v}^t = \frac{1}{P}\sum_{i=1}^P \tilde{y}_i^{t-1} + z^{t,k}$, $g^t = \frac{1}{P}\sum_{i=1}^P y_i^t - \nabla f(x^t) - \left(\frac{1}{P}\sum_{i=1}^P \tilde{y}_i^t - \nabla f(\tilde{x}^t)\right)$, 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, $v^{t,k} - \nabla f(x^{t,k}) - (\tilde{v}^{t,k} - \nabla f(\tilde{x}^{t,k})) = g^{t-1} + h^{t,k}$.

There exists a sufficiently large constants $C_5 = \tilde{O}(1), C_6 = O(1)$, with which the following holds: If we take $p \geq \sqrt{P} + \frac{C_5^2\zeta^2}{\delta^2} + \frac{C_5^2 L^2}{Kb\delta^2}, b \geq \sqrt{K}$ and, $\eta = \tilde{\Theta}\left(\frac{1}{L}\right)$, with probability $1 - \frac{3\nu}{TK}$ ($\nu \in (0,1)$), we have

$$\max_{(\tau_0,\kappa_0) \leq (t,k) < (\tau_0,\kappa_0+3T_7)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} \geq \frac{\delta}{C_5\rho}.$$

Proof of Lemma 33 We assume K is at least as large as $3T_7$. When $K - 2T_7 \leq \kappa_0 < K - 1$, taking $T_7 \geq T_5$ yields the assertion, considering the two coupled sequence initialized at (κ_0, K), according to a slight modification of Lemma 29.

Otherwise, we let (τ, κ) as the first step after (τ_0, κ_0) with $\kappa \equiv T_7$. Then, it suffice to show that, with probability at least $1 - \frac{3\nu}{TK}$,

$$\max_{(\tau,\kappa) \leq (t,k) < (\tau,\kappa+T_7)} \{\|x^{\tau,k} - x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} \geq \frac{\delta}{C_5\rho}. \tag{50}$$

Since $K \geq 3T_7$ and $\kappa_0 < K - 2T_7$ imply $g^{t-1} = 0$ for all $(\tau, \kappa) \leq (t, k) < (\tau, \kappa + T_7)$, $g^{t-1} + h^{t,k} = h^{t,k}$ 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 4. Moreover, for $(\tau, k) > (\tau, \kappa)$, when we assume $\max_{(\tau,\kappa) \leq (t,k) < (\tau,\kappa+T_7)}\{\|x^{\tau,k} -$

$x^{\tau_0,\kappa_0}\|, \|\tilde{x}^{\tau,\kappa} - x^{\tau_0,\kappa_0}\|\} < \frac{\delta}{C_5\rho}$,

$$\begin{aligned}
&\|h^{\tau,k}\| \\
&= \left\| u_{i_\tau}^{\tau,k} + \sum_{l=\kappa}^{k} \frac{1}{|J_{i_t}^{\tau,l}|} \sum_{j \in J_{i_t}^{\tau,\kappa}} u_{i_t,j}^{\tau,l} \right\| \\
&\leq \zeta \sum_{l=\tau}^{k} \|w^{\tau,l} - w^{\tau,l-1}\| + \frac{2\delta}{C_5}\|w^{\tau,k}\| + \frac{2\delta}{C_5}\|w^{\tau,0}\| + \delta r_e \\
&\quad + \frac{C_1^2}{\sqrt{b}} \sqrt{\sum_{l=1}^{k} \left(L\|w^{\tau,k} - w^{\tau,k-1}\| + \frac{2\delta}{C_5}\|w^{\tau,k}\| + \frac{2\delta}{C_5}\|w^{\tau,k-1}\| \right)^2}.
\end{aligned}$$

Assuming that (a) $\frac{1}{2}(1+\eta\gamma)^{k-\kappa} r_e \leq \|w^{t,k}\| \leq 2(1+\eta\gamma)^{k-\kappa} r_e$ and (b) $\|w^{t,k} - w^{t,k-1}\| \leq 3\eta\gamma(1+\eta\gamma)^{k-\kappa} r_e$ for $(\tau,\kappa) < (t,k) < (\tau, \kappa + T_7)$, we get $\|h^{t,k}\| \leq \frac{2C_8\gamma}{C_5}(1+\eta\gamma)^{k-\kappa}$. Thus, following the discussion in Lemma 29 and taking T_7 similarly to T_5, we have (50). ■

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 $O(T_5)$ steps for the last local step. Therefore, taking $K \geq T_5$ was useless to reduce the number of communication rounds. On the other hand, based on Lemma 33, when FLEDGE comes to a saddle point, FLEDGE does not need to wait next communication, and can escape the stack region within $2T_7$ local steps, even if $T_7 \ll K$. 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 28.

E.4. Convergence under PL condition

Theorem 34 *Under Assumptions 1 to 3, 5 and 6, if we choose* $PKb \geq \Omega\left(\frac{C_1^2\sigma^2}{\varepsilon^2} + \frac{C_1 G}{\varepsilon}\right)$ *and* $r \leq \frac{\varepsilon\sqrt{\eta}}{8}$, $\eta = \tilde{\Theta}\left(\frac{1}{L} \wedge \frac{p\sqrt{b}}{\zeta\sqrt{PK}} \wedge \frac{p}{\mu PK} \wedge \frac{1}{\zeta K} \wedge \frac{\sqrt{b}}{L\sqrt{K}}\right)$, *Algorithm 3 with Option I finds an* ε*-first-order stationary points for problem (2) using*

$$\tilde{O}\left(PKb + \left(\frac{Lpb}{\mu} \wedge \frac{\zeta\sqrt{P}Kb}{\mu} \wedge PKb \wedge \frac{\zeta pKb}{\mu} \wedge \frac{Lp\sqrt{Kb}}{\mu}\right) \log \frac{\Delta + \sigma + G + \sigma_c + G_c}{\varepsilon}\right)$$

stochastic gradients and

$$\tilde{O}\left(\frac{P}{p} + \left(\frac{L}{\mu K} \wedge \frac{\zeta\sqrt{P}}{\mu p} \wedge \frac{P}{p} \wedge \frac{\zeta}{\mu} \wedge \frac{L}{\mu\sqrt{Kb}}\right) \log \frac{\Delta + \sigma + G + \sigma_c + G_c}{\varepsilon}\right)$$

communication rounds

with probability at least $1-8\nu$*. Moreover, under the same conditions, algorithm 3 with Option II finds an* ε*-first-order stationary points for problem* (2) *using*

$$\tilde{O}\left(PKb+\left(\frac{Lpb}{\mu}\wedge\frac{\zeta\sqrt{P}Kb}{\mu}\wedge PKb\wedge\frac{\zeta pKb}{\mu}\wedge\frac{Lp\sqrt{Kb}}{\mu}\right)\log\frac{\Delta}{\varepsilon}\right)\quad\text{stochastic gradients and}$$

$$\tilde{O}\left(1+\left(\frac{L}{\mu K}\wedge\frac{\zeta\sqrt{P}}{\mu p}\wedge\frac{P}{p}\wedge\frac{\zeta}{\mu}\wedge\frac{L}{\mu\sqrt{Kb}}\right)\log\frac{\Delta}{\varepsilon}\right)\quad\text{communication rounds}$$

with probability at least $1-8\nu$*. Here* $\tilde{O}$ *hides only at most* $\log^{6.5}(P+K+\mu^{-1}+\nu^{-1})$ *and* polyloglog *factors.*

Proof According to eq. (33) and PL condition,

$$\begin{aligned}
&f(x^{t,k})\\
&\le f(x^{t,k-1})+\eta\|\nabla f(x^{t,k-1})-v^{t,k-1}\|^2-\frac{\eta}{2}\|\nabla f(x^{t,k-1})\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,k}-x^{t,k-1}\|^2+\frac{r^2}{\eta}\\
&\le f(x^{t,k-1})+\eta\|\nabla f(x^{t,k-1})-v^{t,k-1}\|^2\\
&\qquad-\frac{\eta}{4}\|\nabla f(x^{t,k-1})\|^2-\frac{\eta\mu}{2}(f(x^{t,k-1})-f^*)-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,k-1}-x^{t,k-1}\|^2+\frac{r^2}{\eta}.
\end{aligned}$$

Rearranging the above yields that

$$\begin{aligned}
&f(x^{t,k})-f^*+\frac{\eta}{4}\|\nabla f(x^{t,k-1})\|^2 \qquad (51)\\
&\le\left(1-\frac{\eta\mu}{2}\right)(f(x^{t,k-1})-f^*)+\eta\|\nabla f(x^{t,k-1})-v^{t,k-1}\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,k}-x^{t,k-1}\|^2+\frac{r^2}{\eta}.
\end{aligned}$$

holds for all t,k $(1\le t\le T, 0\le k\le K-1)$ with probability at least $1-8\nu$.

Applying Lemma 23 to this, for all $t=1,\cdots,T$ with probability at least $1-8\nu$,

$$\begin{aligned}
&f(x^t)-f(x^*)+\frac{\eta}{4}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{K-k}\|\nabla f(x^{t,k-1})\|^2\\
&\le(1-\frac{\eta\mu}{2})^K(f(x^{t-1})-f(x^*))-\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{K-k}\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x^{t,l}-x^{t,l-1}\|^2\\
&\quad+\eta\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{K-k}\left(\left(\frac{120C_1^8\zeta^2K}{p}+\frac{32C_1^{10}\zeta^2}{pb}\right)\sum_{s=\max\{1,t-T_3\}}^{t-1}\sum_{l=1}^{K}\|x^{s,l}-x^{s,l-1}\|^2\right.\\
&\qquad\left.+\left(4\zeta^2K+\frac{4C_1^2L^2}{b}\right)\sum_{l=1}^{K-1}\|x^{t,l}-x^{t,l-1}\|^2\right)\\
&\quad+\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{K-k}\left(\frac{r^2}{\eta}+\frac{8C_1^2\eta}{PKb}\left(\sigma^2+\frac{G^2}{PKb}\right)\right)\\
&\quad+\begin{cases}\eta\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{K-k}96C_1^2\mathbb{1}[t\le T_3]\left(\frac{\sigma^2}{pKb}+\frac{G^2}{p^2K^2b^2}+\frac{\sigma_c^2}{p}+\frac{G_c^2}{p^2}\right) & \text{(Option I)}\\ 0 & \text{(Option II)}\end{cases}
\end{aligned}$$

By using this bound repeatedly, we get

$$
\begin{aligned}
&f(x^T) - f(x^*) + \frac{\eta}{4}\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)-k}\|\nabla f(x^{t,k-1})\|^2 \\
&\leq (1-\frac{\eta\mu}{2})^{TK}(f(x^0)-f(x^*)) \\
&\quad -\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)K-k}\left(\frac{1}{2\eta}-\frac{L}{2}-\sum_{s=1}^{T_3}(1-\frac{\eta\mu}{2})^{-(s+1)K}\left(\frac{120C_1^8\zeta^2\eta K^2}{p}+\frac{32C_1^{10}\zeta^2\eta K}{pb}\right)\right. \\
&\quad \left.-\eta\sum_{l=1}^{K}(1-\frac{\eta\mu}{2})^{l-K}\left(4\zeta^2K+\frac{4C_1^2L^2}{b}\right)\right)\|x^{t,k}-x^{t,k-1}\|^2 \\
&\quad +\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)K-k}\left(\frac{r^2}{\eta}+\frac{8C_1^2\eta}{PKb}\left(\sigma^2+\frac{G^2}{PKb}\right)\right) \\
&\quad +\begin{cases}\eta\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)K-k}96C_1^2\mathbb{1}[t\leq T_3]\left(\frac{\sigma^2}{pKb}+\frac{G^2}{p^2K^2b^2}+\frac{\sigma_c^2}{p}+\frac{G_c^2}{p^2}\right) & \text{(Option I)} \\ 0 & \text{(Option II)}.\end{cases}
\end{aligned}
$$

We take η as

$$
\eta = \Theta\left(\frac{1}{L}\wedge\frac{p}{C_1^{4.5}\zeta\sqrt{P}K}\wedge\frac{p\sqrt{b}}{C_1^{5.5}\zeta\sqrt{PK}}\wedge\frac{p}{\mu C_1PK}\wedge\frac{1}{\zeta K}\wedge\frac{\sqrt{b}}{C_1L\sqrt{K}}\right)
$$

so that $\frac{1}{2\eta}-\frac{L}{2}-\sum_{s=1}^{T_3}(1-\frac{\eta\mu}{2})^{-(s+1)K}\left(\frac{120C_1^8\zeta^2\eta K^2}{p}+\frac{32C_1^{10}\zeta^2\eta K}{pb}\right)-\eta\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{k-K}$ $\left(4\zeta^2K+\frac{4C_1^2L^2}{b}\right)\geq 0$ holds. We also take $r\leq\frac{\varepsilon\sqrt{\eta}}{8}$ and $PKb\geq\frac{512C_1^2\sigma^2}{\varepsilon^2}+\frac{64C_1G}{\varepsilon}$, then $\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)K-k}\left(\frac{r^2}{\eta}+\frac{8C_1^2\eta}{PKb}\left(\sigma^2+\frac{G^2}{PKb}\right)\right)\leq\frac{\varepsilon^2}{8\mu}$ holds.

Then, we have that

$$
\begin{aligned}
&f(x^T) - f(x^*) + \frac{\eta}{4}\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)-k}\|\nabla f(x^{t,k-1})\|^2 \\
&\leq \frac{\varepsilon^2}{8}+(1-\frac{\eta\mu}{2})^{TK}(f(x^0)-f^*) \\
&\quad +\begin{cases}(1-\frac{\eta\mu}{2})^{(T-t+1-T_3)K-k}96C_1^3\left(\frac{\sigma^2P}{p^2b}+\frac{G^2P}{p^3Kb^2}+\frac{\sigma_c^2PK}{p^2}+\frac{G_c^2PK}{p^3}\right) & \text{(Option I)} \\ 0 & \text{(Option II)}\end{cases}
\end{aligned}
$$

For Option I, the first term $(1-\frac{\eta\mu}{2})^{TK}(f(x^0)-f(x^*))$ is smaller than $\frac{\varepsilon^2}{32}$ if we take $TK = O\left(\frac{1}{\eta\mu}\log\frac{\Delta}{\varepsilon}\right)$. The third term is bounded by $\frac{\varepsilon^2}{32}$, if we take $T = T_3 + O(\frac{1}{\eta\mu K}$ $\log\frac{C_1^3(\frac{\sigma^2P}{p^2b}+\frac{G^2P}{p^3Kb^2}+\frac{\sigma_c^2PK}{p^2}+\frac{G_c^2PK}{p^3})}{\varepsilon}) = O\left(\frac{P}{p}C_1+\frac{C_1}{\eta\mu K}\log\frac{\sigma+G+\sigma_c+G_c}{\varepsilon}\right)$. Moreover, note that $f(x^T)-f(x^*)+\frac{\eta}{4}\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\frac{\eta\mu}{2})^{(T-t+1)-k}\|\nabla f(x^{t,k-1})\|^2\leq\frac{6}{32\mu}\min_{t,k}\|\nabla f(x^{t,k-1})\|^2$ holds when we take $T = O\left(\frac{1}{\eta\mu K}\right)$.

Thus, for Option I, if we take

$$T = O\left(\frac{P}{p}C_1 + C_1\left(\frac{L}{\mu K} \wedge \frac{C_1^{4.5}\zeta\sqrt{P}}{\mu p} \wedge \frac{C_1^{5.5}\zeta\sqrt{P}}{\mu p\sqrt{bK}} \wedge \frac{C_1 P}{p} \wedge \frac{\zeta}{\mu} \wedge \frac{C_1 L}{\mu\sqrt{Kb}}\right)\log\frac{\Delta+\sigma+G+\sigma_c+G_c}{\varepsilon}\right),$$

we obtain the desired bound with probability at least $1-8\nu$.

For Option II, taking

$$T = O\left(\left(\frac{L}{\mu K} \wedge \frac{C_1^{4.5}\zeta\sqrt{P}}{\mu p} \wedge \frac{C_1^{5.5}\zeta\sqrt{P}}{\mu p\sqrt{bK}} \wedge \frac{C_1 P}{p} \wedge \frac{\zeta}{\mu} \wedge \frac{C_1 L}{\mu\sqrt{Kb}}\right)\log\frac{\Delta}{\varepsilon}\right),$$

yields the desired bound.

Note that T depends on ε^{-1} only logarithmically, which means that C_1 depends on ε^{-1} in only $\log\log$ order and $C_1 = O^*(\log(P+K+\mu^{-1}+\nu^{-1}))$ (where O^* suppresses $\log\log$ factors). ■

Remark 22 *In order to find ε-solutions (i.e., $f(x^{t,k}) - f^* \le \varepsilon$), the same statement holds, except for slight change on the assumptions on PKb and r: $PKb \ge \Omega\left(\frac{C_1^2\sigma^2}{\mu\varepsilon} + \frac{C_1 G}{\varepsilon\sqrt{\varepsilon\mu}}\right)$ and $r \le \frac{\eta\sqrt{\varepsilon\mu}}{2}$.*

In fact, we can derive

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

similarly to (51)*, and using this, we have*

$$\begin{aligned}
&f(x^t) - f(x^*) \\
&\le (1-\eta\mu)^{TK}(f(x^0) - f(x^*)) \\
&\quad - \sum_{t=1}^{T}\sum_{k=1}^{K}(1-\eta\mu)^{(T-t+1)K-k}\left(\frac{1}{2\eta} - \frac{L}{2} - \sum_{s=1}^{T_3}(1-\eta\mu)^{-(s+1)K}\left(\frac{120C_1^8\zeta^2\eta K^2}{p} + \frac{32C_1^{10}\zeta^2\eta K}{pb}\right)\right. \\
&\quad \left. -\eta\sum_{l=1}^{K}(1-\eta\mu)^{l-K}\left(4\zeta^2 K + \frac{4C_1^2L^2}{b}\right)\right)\|x^{t,k} - x^{t,k-1}\|^2 \\
&\quad + \sum_{t=1}^{T}\sum_{k=1}^{K}(1-\eta\mu)^{(T-t+1)K-k}\left(\frac{r^2}{\eta} + \frac{8C_1^2\eta}{PKb}\left(\sigma^2 + \frac{G^2}{PKb}\right)\right) \\
&\quad + \begin{cases} \eta\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\eta\mu)^{(T-t+1)K-k}96C_1^2\mathbb{1}[t \le T_3]\left(\frac{\sigma^2}{pKb} + \frac{G^2}{p^2K^2b^2} + \frac{\sigma_c^2}{p} + \frac{G_c^2}{p^2}\right) & (\textit{Option I}) \\ 0 & (\textit{Option II}) \end{cases}
\end{aligned}$$

Taking η similary to the previous theorem, $r \le \frac{\eta\sqrt{\varepsilon\mu}}{2}$ and $PKb \ge \Omega\left(\frac{C_1^2\sigma^2}{\mu\varepsilon} + \frac{C_1 G}{\sqrt{\mu\varepsilon}}\right)$ yields

$\sum_{t=1}^{T}\sum_{k=1}^{K}(1-\eta\mu)^{(T-t+1)K-k}(\frac{r^2}{\eta}+\frac{8C_1^2\eta}{PKb}(\sigma^2+\frac{G^2}{PKb}))\leq\frac{\varepsilon}{2}$. *Thus, we finally have the following:*

$$
\begin{aligned}
&f(x^t)-f(x^*)\\
&\leq\frac{\varepsilon}{2}+(1-\eta\mu)^{TK}(f(x^0)-f^*)\\
&\quad+\begin{cases}(1-\eta\mu)^{(T-t+1-T_3)K-k}96C_1^3\left(\dfrac{\sigma^2P}{p^2b}+\dfrac{G^2P}{p^3Kb^2}+\dfrac{\sigma_c^2PK}{p^2}+\dfrac{G_c^2PK}{p^3}\right) & (\textit{Option I})\\ 0 & (\textit{Option II}).\end{cases}
\end{aligned}
$$

Now it is trivial to see that the desired bound holds.

Appendix F. Lower bound

In this section, we provide the gradient complexity lower bound of $O(n+\frac{\Delta(\zeta\sqrt{n}\vee L)}{\varepsilon^2})$ under Hessian heterogeneity, which recovers the usual lower bound for L-smooth functions by setting $\zeta=2L$. Note that the gradient complexity, or the total number of gradient communicated, of FLEDGE is $\tilde{O}(P+\frac{\zeta\sqrt{P}}{\varepsilon^2})$. Thus, this almost matches the lower bound of gradient complexity of the finite-sum case if we identify P with n.

First, we give a definition of the linear-span first-order algorithms.

Definition 36 (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 $\mathcal{F}_{n,\Delta}^{L}$ and $\mathcal{F}_{n,\Delta}^{L,\zeta}$ for (1), as follows.

Definition 37 (A class of finite-sum optimization problems) *Fix some x^0. For an integer n, $L>0$, we define a problem class $\mathcal{F}_n^L$ as*

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

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

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

We now state our lower bound theorem as follows.

Proposition 35 *Assume Assumptions 1, 2 and 5. 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 to find ε-first-order stationary points of the problem (1).

Proposition 35 can be derived by using the bounds of Carmon et al. [4], Fang et al. [9], Li et al. [25].

Carmon et al. [4] proved the following lower bound.

Proposition 38 (Carmon et al. [4]) *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.*

Fang et al. [9], Li et al. [25] extended this to the lower bound on the finite-sum optimization problem.

Proposition 39 (Fang et al. [9], Li et al. [25]) *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.

Proof It is easy to see that the lower bound of Proposition 38 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 38. On the other hand, we have $\mathcal{F}^{\frac{\zeta}{2}}_{n,\Delta} \subseteq \mathcal{F}^{L,\zeta}_{n,\Delta}$. Thus, Proposition 39 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. ■