Sketching for First Order Method: Efficient Algorithm for Low-Bandwidth Channel and Vulnerability

Zhao Song [1] **Yitan Wang** [2] **Zheng Yu** [3] **Lichen Zhang** [4]

Abstract

Sketching is one of the most fundamental tools in large-scale machine learning. It enables runtime and memory saving via randomly compressing the original large problem into lower dimensions. In this paper, we propose a novel sketching scheme for the first order method in large-scale distributed learning setting, such that the communication costs between distributed agents are saved while the convergence of the algorithms is still guaranteed. Given gradient information in a high dimension d, the agent passes the compressed information processed by a sketching matrix $R \in \mathbb{R}^{s \times d}$ with $s \ll d$, and the receiver de-compressed via the de-sketching matrix $R^\top$ to "recover" the information in original dimension. Using such a framework, we develop algorithms for federated learning with lower communication costs. However, such random sketching does not protect the privacy of local data directly. We show that the gradient leakage problem still exists after applying the sketching technique by presenting a specific gradient attack method. As a remedy, we prove rigorously that the algorithm will be differentially private by adding additional random noises in gradient information, which results in a both communication-efficient and differentially private first order approach for federated learning tasks. Our sketching scheme can be further generalized to other learning settings and might be of independent interest itself.

1. Introduction

Federated learning enables multiple parties to collaboratively train a machine learning model without directly exchanging training data. This has become particularly important in areas of artificial intelligence where users care about data privacy, security, and access rights, including healthcare (Li et al., 2020b; 2019), internet of things (Chen et al., 2020), and fraud detection (Zheng et al., 2020).

Given the importance and popularity of federated learning, two central aspects of this subject have been particularly studied: privacy and communication cost. The fundamental purpose of federated learning is to protect the data privacy of clients by only communicating the gradient information of a user. Unfortunately, recent studies (Geiping et al., 2020; Zhu et al., 2019; Wang et al., 2019) have demonstrated that attackers can recover the input data from the communicated gradients. The reason why these attacks work is the gradients carry important information about the training data (Ateniese et al., 2015; Fredrikson et al., 2015). A very recent work (Wang et al., 2023) demonstrates that via computationally intense approach based on tensor decomposition, one can recover the training data from a single gradient and model parameters for over-parametrized networks.

Communication efficiency is also one of the core concerns. In a typical federated learning setting, the model is trained through gathering individual information from many clients who operate under a low bandwidth network. On the other hand, the size of the gradient is usually large due to the sheer parameter count of many modern machine learning models. This becomes even more problematic when conducting federated learning on mobile and edge devices, where the bandwidth of the network is further limited. Many works try to address this challenge through local optimization methods, such as local gradient descent (GD), local stochastic gradient descent (SGD) (Konečnỳ et al., 2016; McMahan et al., 2017; Stich, 2019) and using classic data structures in streaming to compress the gradient (Rothchild et al., 2020). Despite of significant efforts on improving the communication cost of federated learning framework, none of these approaches, as we will show, are private enough to *truly* guard against gradient leakage attack.

The above two concerns allude us to ask the following question:

[1] Adobe Research [2] Yale University [3] Alibaba Inc. [4] MIT. Correspondence to: Zhao Song <zsong@adobe.com>, Lichen Zhang <lichenz@mit.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Is there an FL framework that protects the local privacy and has good performance even in low-bandwidth networks?

In this paper, we achieve these goals by using tools from randomized linear algebra — the linear sketches. Sketching matrices describe a distribution of random matrices $R : \mathbb{R}^d \rightarrow \mathbb{R}^{b_{\text{sketch}}}$ where $b_{\text{sketch}} \ll d$ and for vectors $x \in \mathbb{R}^d$ one has $\|Rx\|_2 = (1 \pm \epsilon)\|x\|_2$. While these random projections effectively reduce the dimension of the gradient, we still need to "recover" them to the original dimension for training purpose. To realize this goal, we apply the de-sketch matrix, which is essentially the transpose of R as a decoder. Instead of running the gradient descent $w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot g^{(t)}$ using true gradient $g^{(t)} \in \mathbb{R}^d$, we apply sketch and de-sketch to the gradient:

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot R^\top \cdot R \cdot g^{(t)}.$$

Here $R \in \mathbb{R}^{b_{\text{sketch}} \times d}$ denotes a sketching matrix that sketches the true gradient to a lower dimension and $R^\top \in \mathbb{R}^{d \times b_{\text{sketch}}}$ denotes the de-sketching process that maps the sketched gradient back to the true gradient dimension. To ensure that the gradient descent still has good convergence behavior under the linear map $x \mapsto R^\top Rx$, we argue that it is enough for R to satisfy the coordinate-wise embedding property (Song & Yu, 2021). This property states that $R^\top Rg^{(t)}$ is an unbiased estimator of $g^{(t)}$ and has small second moment, and many of the popular sketching matrices satisfy this property. Hence, all clients will only communicate sketched gradients to the server, the server averages the sketched gradients and broadcasts them back to all clients. Finally, each client de-sketches the received gradients and performs local updates. Since the sketching dimension is always small compared to the original dimension, we save communication costs per iteration via sketching.

While the algorithm with sketch-and-de-sketch might seem simple and elegant, it is not enough to address the privacy challenge of federated learning. At the first glance, the sketching "masks" the communicated gradients, but this can actually be leveraged by a malicious attacker to develop gradient leakage attacks. Specifically, we propose a highly-efficient attack algorithm such that the attacker only needs to observe the sketched gradient being communicated, the sketching matrix being used and the model parameters. Then, the attacker can effectively *learn* the private local data by instantiating a gradient descent on data, instead of model parameters. For attacking the sketched gradients, we show that it is no harder than that without any sketching. Our approach is based on the classical sketch-and-solve (Clarkson & Woodruff, 2013) paradigm. To the best of our knowledge, this is the first theoretical analysis on effectiveness of the gradient leakage attack using simple and standard first-order methods that are widely-observed in practice (Geiping et al., 2020; Zhu et al., 2019). Moreover, compare to the tensor decomposition-based algorithm of (Wang et al., 2023), our algorithm is much more computationally efficient and extends to a variety of models beyond over-parametrized networks. On the other hand, the (Wang et al., 2023) algorithm produces stronger guarantees than ours and works for noisy gradients. Our leakage attack algorithm and analysis not only poses privacy challenges to our sketching-based framework, but many other popular approaches building upon randomized data structures (Rothchild et al., 2020).

To circumvent this issue, we inject random Gaussian noises to the gradients-to-be-communicated to ensure they are differentially private (Dwork et al., 2006a) and therefore provably robust against the gradient leakage attack.

We summarize the contributions in this work as follows:

Our contributions: We present our main technical contributions as follows:

- We introduce the sketch-and-de-sketch framework. Unlike the classical sketch-and-solve paradigm, our iterative sketch and de-sketch method can be combined with gradient-based methods and extended to broader optimization problems.
- We apply our sketch-and-de-sketch method to federated learning, obtaining an algorithm that only needs to communicate lower-dimensional vector, which is particularly useful in low-bandwidth networks.
- By adding Gaussian noise, we show that our algorithm is differentially private.
- We present a gradient leakage attack algorithm that can recover the local data from only observing the communicated sketched gradients and sketching matrices. Our analysis extends to a large family of non-linear machine learning models.

Roadmap. In section 2, we discuss related work and define common notations. In section 3, we describe the problem setting and assumptions. In section 4, we present a federated learning framework with communication efficiency by leveraging sketching techniques. In section 5, we analyze the convergence property of our proposed framework for smooth and convex objectives. In section 6, we discuss the privacy guarantee of our framework. In section 7, we discuss the feasibility of the gradient attacking when the framework shares sketched gradient information. In section 8, we conclude the contribution and limitations of this paper.

2. Related Work

Federated Learning. Federated learning (FL) is an emerging framework in distributed deep learning. FL allows multiple parties or clients collaboratively train a model without data sharing. In this learning paradigm, local clients perform most of the computation and a central sever update the model parameters through aggregation then transfers the parameters to local models (Dean et al., 2012; Shokri & Shmatikov, 2015; McMahan et al., 2017). In this way, the details of the data are not disclosed in between each party. Unlike the standard parallel setting, FL has three unique challenge (Li et al., 2020a), including communication cost, data heterogeneity and client robustness. In our work, we focus on the first two challenges. The training data are massively distributed over an incredibly large number of devices, and the connection between the central server and a device is slow. A direct consequence is the slow communication, which motivated communication-efficient FL algorithm. Federated average (FedAvg) (McMahan et al., 2017) firstly addressed the communication efficiency problem by introducing a global model to aggregate local stochastic gradient descent updates. Later, different variations and adaptations have arisen. This encompasses a myriad of possible approaches, including developing better optimization algorithms (Wang et al., 2020a), generalizing model to heterogeneous clients under special assumptions (Zhao et al., 2018; Kairouz et al., 2021; Li et al., 2021) and utilizing succinct and randomized data structures (Rothchild et al., 2020). The work of (Li et al., 2023) provides a provable guarantee federated learning algorithm for adversarial deep neural networks training.

Sketching. Sketching is a fundamental tool in many numerical linear algebra tasks, such as linear regression, low-rank approximation (Clarkson & Woodruff, 2013; Nelson & Nguyên, 2013; Meng & Mahoney, 2013; Boutsidis & Woodruff, 2014; Song et al., 2017; Andoni et al., 2018; Makarychev et al., 2020), distributed problems (Woodruff & Zhong, 2016; Boutsidis et al., 2016), reinforcement learning (Wang et al., 2020b; Shrivastava et al., 2023), tensor decomposition (Song et al., 2019), clustering (Esfandiari et al., 2021; Deng et al., 2022), convex programming (Lee et al., 2019; Jiang et al., 2021; Song & Yu, 2021; Jiang et al., 2020; Qin et al., 2023b), gradient-based algorithm (Xu et al., 2021), online optimization problems (Reddy et al., 2022a), training neural networks (Xiao et al., 2018; Brand et al., 2021; Song et al., 2021a;b; Gao et al., 2022), submodular maximization (Qin et al., 2023a), matrix sensing (Qin et al., 2023c), relational database (Qin et al., 2022a), dynamic kernel estimation (Qin et al., 2022b), and Kronecker product regression (Reddy et al., 2022b).

Gradient Leakage Attack. A number of works (Zhu et al., 2019; Yin et al., 2021; Wei et al., 2020; Rigaki & García, 2020) have pointed out that the private information of local training data can be attacked using only the exchanged gradient information. Given the gradient of the neural network model with respect to the weights for a specific data, their method starts with a random generated dummy data and label, and its corresponding dummy gradients. By minimizing the difference between the true gradient and the dummy gradients using gradient descent, they show empirically that the dummy data and label will reveal the true data completely. The follow-up work (Zhao et al., 2020) further discuss the case of classification task with cross-entropy loss, and observe that the true label can be recovered exactly. Therefore, they only need to minimize over the dummy data and have better empirical performance. Other attack methods include but not limited to membership inference and property inference attacks (Shokri et al., 2017; Melis et al., 2019), training generative adversarial network (GAN) models (Hitaj et al., 2017; Goodfellow et al., 2014) and other learning-based methods (McPherson et al., 2016; Papernot et al., 2016). Very recently, (Wang et al., 2023) uses tensor decomposition for gradient leakage attack on over-parametrized networks with provable guarantees. However, the tensor decomposition algorithm is inherently inefficient and their analysis is restricted to over-parametrized networks.

Notations. For a positive integer n, we use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. We use $\mathbb{E}[\cdot]$ to denote expectation (if it exists), and use $\Pr[\cdot]$ to denote probability. For a vector x, we use $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$ or $\|x\|$ to denote its ℓ_2 norm. We denote $1_{\{x=l\}}$ for $l \in \mathbb{R}$ to be the indicator function which equals to 1 if $x = l$ and 0 otherwise. Let $f : A \to B$ and $g : C \to A$ be two functions, we use $f \circ g$ to denote the composition of functions f and g, i.e., for any $x \in C$, $(f \circ g)(x) = f(g(x))$. We denote I_d to be the identity mapping.

3. Problem Setup

Consider a federated learning scenario with N clients and corresponding local losses $f_c : \mathbb{R}^d \to \mathbb{R}$, our goal is to find

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{N} \sum_{c=1}^{N} f_c(w) \tag{1}$$

For the sake of discussion, we will be focusing on the classical convex and smooth setting for the objective function. Our paradigm will extends to non-convex objectives and we defer details to appendix G.

Assumption 3.1. *Assume that the set of minimizers of (1) is nonempty. Each f_c is μ-strongly convex for $\mu \geq 0$ and L-smooth. That is, for all $x, y \in \mathbb{R}^d$,*

$$\begin{aligned}\frac{\mu}{2}\|y - x\|_2^2 &\leq f_c(y) - f_c(x) + \langle y - x, \nabla f_c(x) \rangle \\ &\leq \frac{L}{2}\|y - x\|_2^2.\end{aligned}$$

Note in the case $\mu = 0$, this assumption reduces back to convexity and smoothness.

In addition to the above assumption, we allow local losses to have arbitrary heterogeneity. In other words, we allow f_c's to vary between different clients.

Our results also contain an attack algorithm, which can extract useful information by only inspecting the local gradient and model parameters. We defer those discussions to section 7.

4. Our Algorithm

In this section, we propose a federated learning framework that addresses the communication efficiency issue. When the learning gradients are of high dimension, classical federated learning framework that communicates the exact gradient could incur a heavy communication cost per round. Sketching technique, which emerges as an effective way to reduce the dimension of vector while preserving significant amount of information (Sarlós, 2006; Woodruff, 2014), is highly preferred in this setting. It enables us to compress the gradient vector into a lower dimension while preserving convergence rates, and greatly saves the communication cost per round.

Algorithm 1 Iterative sketching-based federated larning Algorithm with K local steps

```
procedure IterativeSketchingFL
    Each client initializes $w^0$ with the same seed
    for $t = 1 \to T$ do
        /* Client */
        parfor $c = 1 \to N$ do
            if $t = 1$ then
                $u_c^{t,0} \leftarrow w^0$
            else
                $u_c^{t,0} \leftarrow w^{t-1} + \mathsf{desk}_t(\Delta\widetilde{w}^{t-1})$
            end if
            $w^t \leftarrow u_c^{t,0}$
            for $k = 1 \to K$ do
                $u_c^{t,k} \leftarrow u_c^{t,k-1} - \eta_{\mathrm{local}} \cdot \nabla f_c(u_c^{t,k-1})$
            end for
            $\Delta w_c(t) \leftarrow u_c^{t,K} - w^t$
            Client $c$ sends $\mathsf{sk}_t(\Delta w_c(t))$ to server
        end parfor
        /* Server */
        $\Delta\widetilde{w}^t \leftarrow \eta_{\mathrm{global}} \cdot \frac{1}{N}\sum_{c=1}^N \mathsf{sk}_t(\Delta w_c(t))$
        Server sends $\Delta\widetilde{w}^t$ to each client
    end for
end procedure
```

Motivated by above discussion, we propose the iterative sketching-based federated learning algorithm, which builds upon vanilla local gradient descent: we start with a predetermined sequence of independent sketching matrices shared across all clients. In each round, local clients accumulate and sketch its change over K local steps, then transmit the low-dimensional sketch to the server. Server then averages the sketches and transmits them back to all clients. Upon receiving, each client de-sketches to update the local model.

We highlight several distinct features of our algorithm:

- **Communication:** In each sync step, we only communicates a low-dimensional sketched gradients, indicating a smaller communication cost per round. This property is particularly valuable in a small-bandwidth setting.
- **De-sketch:** We emphasize that unlike the classical sketch-and-solve paradigm that decreases the problem dimension, our algorithm applies sketching in each round, combined with a de-sketching process which recovers back to the true gradient dimension.
- **Simpler server task:** Server only needs to do simple averaging, indicating no need of a trustworthy party as the server.
- **Decentralization:** Our algorithm can be generalized to decentralized learning settings, where local clients can only communicate with neighboring nodes. In this case, it requires $O(\mathrm{diam})$ rounds to propagate the sketched local changes, where diam is the diameter of the network graph.
- **Linearity:** Compared to the framework of (Rothchild et al., 2020), our de-sketching operator is linear, this adds flexibility to the analysis and further extensions to the framework.

4.1. $\mathsf{sk}/\mathsf{desk}$ via Coordinate-wise Embedding

In this section, we discuss the concrete realization of the $\mathsf{sk}_t/\mathsf{desk}_t$ operators in Algorithm 1 through random sketching matrices. Note we should require any processed gradient $\mathsf{desk}_t \circ \mathsf{sk}_t(g)$ to "be close" to the true gradient g to avoid breaking the convergence property of the algorithm. To achieve this, we first introduce the following property for a broad family of sketching matrices, namely the *coordinate-wise embedding* (Song & Yu, 2021), that naturally connects with $\mathsf{sk}_t/\mathsf{desk}_t$ operators.

Definition 4.1 (a-coordinate-wise embedding). *We say a randomized matrix $R \in \mathbb{R}^{b_{sketch} \times d}$ satisfying a-coordinate wise embedding if for any vector $g, h \in \mathbb{R}^d$, we have*

- $\mathbb{E}_{R\sim\Pi}[h^\top R^\top R g] = h^\top g$;
- $\mathbb{E}_{R\sim\Pi}[(h^\top R^\top R g)^2] \leq (h^\top g)^2 + \frac{a}{b_{\mathrm{sketch}}}\|h\|_2^2 \cdot \|g\|_2^2$.

In general, well-known sketching matrices have their coordinate-wise embedding parameter a being a small constant (See appendix D). Note that if we choose h to be one-hot vector e_i, then the above conditions translate to

$$\mathop{\mathbb{E}}_{R\sim\Pi}[R^\top R g] = g$$

and

$$\mathop{\mathbb{E}}_{R\sim\Pi}[\|R^\top Rg\|_2^2] \leq (1 + a \cdot \frac{d}{b_{\text{sketch}}}) \cdot \|g\|_2^2.$$

This implies that by choosing

$$\begin{aligned}\mathsf{sk}_t &= R_t \in \mathbb{R}^{b_{\text{sketch}}\times d} \text{ (sketching)},\\ \mathsf{desk}_t &= R_t^\top \in \mathbb{R}^{d\times b_{\text{sketch}}} \text{ (de-sketching)}\end{aligned} \tag{2}$$

for any iteration $t \geq 1$, where R_t's are independent random matrices with sketching dimension b_{sketch}, we obtain an unbiased sketching/de-sketching scheme with bounded variance as state in the following Theorem 4.2.

Theorem 4.2. *Let sk_t and desk_t be defined by Eq. (2) using a sequence of independent sketching matrices $R_t \in \mathbb{R}^{b_{sketch}\times d}$ satisfying a-coordinate wise embedding property (Definition 4.1). Then the following properties hold:*

1. *Independence: Operators $(\mathsf{sk}_t, \mathsf{desk}_t)$'s are independent over different each iterations.*
2. *Linearity: Both sk_t and desk_t are linear operators.*
3. *Unbiased estimator: For any fixed vector $h \in \mathbb{R}^d$, it holds $\mathbb{E}[\mathsf{desk}_t(\mathsf{sk}_t(h))] = h$.*
4. *Bounded second moment: For any fixed vector $h \in \mathbb{R}^d$, it holds $\mathbb{E}[\|\mathsf{desk}_t(\mathsf{sk}_t(h))\|_2^2] \leq (1+\alpha)\cdot\|h\|_2^2$, where $\alpha = a \cdot d/b_{sketch}$. The value of $\alpha > 0$ is given in Table 1 for common sketching matrices.*

Sketching matrix	Definition	Param α
Random Gaussian	Def. D.2	$3d/b_{\text{sketch}}$
SRHT	Def. D.3	$2d/b_{\text{sketch}}$
AMS sketch	Def. D.4	$2d/b_{\text{sketch}}$
Count-sketch	Def. D.5	$3d/b_{\text{sketch}}$
Sparse embedding	Def. D.6,D.7	$2d/b_{\text{sketch}}$

Table 1: Sketching matrices and their coordinate-wise embedding parameter α.

Proof. Fix a vector $g \in \mathbb{R}^d$, note that condition 1 of Definition 4.1 implies that

$$\mathop{\mathbb{E}}_{R\sim\Pi}[(R^\top Rg)_j] = \mathop{\mathbb{E}}_{R\sim\Pi}[e_j^\top R^\top Rg] = g_j$$

This means that in expectation, each coordinate of $R^\top Rg$ is equal to corresponding coordinate of g, therefore, we have

$$\mathop{\mathbb{E}}_{R\sim\Pi}[R^\top Rg] = g$$

This proves the unbiased property of Theorem 4.2. For the variance bound, note that using the second condition of coordinate-wise embedding, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{R\sim\Pi}\Big[\sum_{j=1}^d (e_j^\top R^\top Rg)^2\Big] &= \mathop{\mathbb{E}}_{R\sim\Pi}\Big[\sum_{j=1}^d (R^\top Rg)_j^2\Big]\\ &= \mathop{\mathbb{E}}_{R\sim\Pi}[\|R^\top Rg\|_2^2]\\ &\leq \sum_{j=1}^d ((e_j^\top g)^2 + \frac{a}{k}\cdot\|g\|_2^2)\\ &= (1 + a\cdot\frac{d}{b_{\text{sketch}}})\cdot\|g\|_2^2\end{aligned}$$

Thus, we have proven that using $R^\top R$ as $\mathsf{desk} \circ \mathsf{sk}$, the variance parameter α is $a \cdot \frac{d}{b_{\text{sketch}}}$. By Table 1, a is a small constant (2 or 3). Hence, we conclude that $\alpha = O(\frac{d}{b_{\text{sketch}}})$.

Note that the independence property can be satisfied via choosing independent sketching matrix R at each iteration t, and linearity property is straightforward since R is a linear transform. □

We will use the above property to instantiate the convergent proof and communication complexity in section 5.

5. Convergence Analysis and Communication Complexity

In this section, we analyze the convergence property of our proposed framework for smooth and convex objectives. Our analysis builds upon showing that the extra randomness introduced by sketching and de-sketching does not affect the convergence rate much.

We first present our convergence result for strongly-convex objective.

Theorem 5.1 (Informal version of Theorem F.9)**.** *If Assumption 3.1 holds with $\mu > 0$. If $\eta_{\text{local}} \leq \frac{1}{8(1+\alpha)LK}$,*

$$\begin{aligned}&\mathbb{E}[f(w^{T+1}) - f(w^*)]\\ \leq\ &\frac{L}{2}\,\mathbb{E}[\|w^0 - w^*\|_2^2]e^{-\mu\eta_{\text{local}}T} + 4\eta_{\text{local}}^2 L^2K^3\sigma^2/\mu.\end{aligned}$$

where w^* is a minimizer of problem (1).

We note that while standard analysis for strongly-convex and smooth objective will exhibit a linear convergence rate for gradient descent, our result is more align with that of stochastic gradient descent. In fact, our iterative sketching method can be viewed as generating a stochastic gradient that has certain low-dimensional structure. Using the property of structured random matrices, our algorithm gives a better convergence rate than standard stochastic gradient descent. This is because in the standard stochastic gradient descent analysis, one only has an absolute upper bound on the second moment:

$$\mathbb{E}_{\widetilde{g}}[\|\widetilde{g}\|_2^2] \leq C^2$$

for some parameter C, where $\widetilde{g}$ is the stochastic gradient with $\mathbb{E}_{\widetilde{g}}[\widetilde{g}] = g$. In contrast, coordinate-wise embedding guarantees that the second moment of our estimate is upper bounded *multiplicatively in terms of* $\|g\|_2^2$:

$$\mathbb{E}_R[\|R^\top Rg\|_2^2] \leq \big(1 + O(\frac{d}{b_{\text{sketch}}})\big) \cdot \|g\|_2^2,$$

this nice property enables us to obtain a more refined analysis on the convergence.

We obtain the communication cost as a direct corollary:

Corollary 5.2 (Informal version of Theorem F.10)**.** *If Assumption 3.1 holds with $\mu > 0$. Then within Algorithm 1 outputs an ϵ-optimal solution $w^T \in \mathbb{R}^d$ satisfying $\mathbb{E}[f(w^T) - f(w^*)] \leq \epsilon$ by using*

$$O((LN/\mu)\max\{d, \sqrt{\sigma^2/(\mu\epsilon)}\}\log(L\,\mathbb{E}[\|w^0 - w^*\|_2^2]/\epsilon))$$

bits of communication.

We observe that compared to vanilla approaches, our method requires a step size shrinkage by a factor of $O(\alpha)$, thus enlarge the number of rounds approximately by a factor of $O(\alpha)$. Since the iterative sketching algorithm only communicates $O(b_{\text{sketch}}/d)$ as many bits per round due to sketching, the total communication cost does not increase at all for commonly used sketching matrices, according to Theorem 4.2.

We also point out that when $\epsilon \geq \sigma^2/(\mu d^2)$, our analysis implies a linear convergence rate of local GD under only strongly-convex and smooth assumptions, which is new as far as we concern.

We also have a similar observation for convergence in the convex losses case, as well as communication cost.

Theorem 5.3 (Informal version of Theorem F.7)**.** *If Assumption 3.1 holds with $\mu = 0$. If $\eta_{\text{local}} \leq \frac{1}{8(1+\alpha)LK}$,*

$$\begin{aligned}&\mathbb{E}[f(\overline{w}^T) - f(w^*)]\\ \leq\ &\frac{4\,\mathbb{E}[\|w^0 - w^*\|_2^2]}{\eta_{\text{local}}KT} + 32\eta_{\text{local}}^2 LK^2\sigma^2,\end{aligned}$$

where

$$\overline{w}^T = \frac{1}{KT}(\sum_{t=1}^{T}\sum_{k=0}^{K-1}\overline{u}^{t,k})$$

is the average over parameters throughout the execution of Algorithm 1.

Corollary 5.4 (Informal version of Theorem F.8)**.** *If Assumption 3.1 holds with $\mu = 0$. Then Algorithm 1 outputs an ϵ-optimal solution $\overline{w}^T \in \mathbb{R}^d$ satisfying*

$$\mathbb{E}[f(\overline{w}^T) - f(w^*)] \leq \epsilon$$

by using

$$O(\mathbb{E}[\|w^0 - w^*\|_2^2]N\max\{Ld/\epsilon, \sigma\sqrt{L}/\epsilon^{3/2}\})$$

bits of communication.

We compare our communication cost with the work of (Khaled et al., 2019), which analyzes the local gradient descent using the same assumption and framework. The result of (Khaled et al., 2019) shows a communication cost of

$$O\left(\mathbb{E}[\|w^0 - w^*\|_2^2]Nd \cdot \max\{\frac{L}{\epsilon}, \frac{\sigma\sqrt{L}}{\epsilon^{3/2}}\}\right),$$

which is strictly not better than our results. This shows again our approach does not introduce extra overall communication cost.

6. Differential Privacy

In this section, we show that if each client adds a Gaussian noise corresponding to its local loss function, then the iterative sketching scheme is differentially private.

To discuss the privacy guarantee of our proposed approach, we consider that each client c trying to learn upon its local dataset $\mathcal{D}_c$ with corresponding local loss

$$f_c(x) = \frac{1}{|\mathcal{D}_c|}\sum_{z_i \in \mathcal{D}_c} f_c(x, z_i),$$

where we overload the notation f_c to denote the local loss for notation simplicity. We assume f_c is ℓ_c-Lipschitz for agent $c = 1, 2, \cdots, N$. We also assume that the dataset for each client c is disjoint.

To prove the final privacy guarantee of Algorithm 2, we employ a localized analysis by first analyzing the privacy guarantee obtained for a single step performed by a single client. We then combine different clients over all iterations via well-known composition tools: we first use Parallel Composition to compose different clients, then use Advanced Sequential Composition to compose over all iterations. We also amplify privacy via sub-sampling. We defer all proofs to appendix H.4.

Lemma 6.1 (Informal version of Lemma H.9)**.** *Let $\widehat{\epsilon}, \widehat{\delta} \in [0, 1)$, $\widehat{\epsilon} < \frac{1}{\sqrt{K}}$ and $c \in [N]$. For client c, the local-K-step stochastic gradient as in Algorithm 1 is*

$$(\sqrt{K}\cdot\widehat{\epsilon}, K\cdot\widehat{\delta})\text{–DP}.$$

Algorithm 2 Private Iterative Sketching-based Federated Learning Algorithm with K local steps

```
procedure PrivateIterativeSketchingFL
    Each client initializes $w^0$ with the same seed
    for $t = 1 \to T$ do
        /* Client */
        parfor $c = 1 \to N$ do
            if $t = 1$ then
                $u_c^{t,0} \leftarrow w^0$
            else
                $u_c^{t,0} \leftarrow w^{t-1} + \mathsf{desk}_t(\Delta\widetilde{w}^{t-1})$
            end if
            $w^t \leftarrow u_c^{t,0}$
            $\sigma^2 \leftarrow O(\log(1/\widehat{\delta})\ell_c^2/\widehat{\epsilon}^2)$
            for $k = 1 \to K$ do
                $\xi_c^{t,k} \sim \mathcal{N}(0, \sigma^2 \cdot I_d)$
                $\mathcal{D}_c^{t,k} \leftarrow$ Random batch of local data
                $u_c^{t,k} \leftarrow u_c^{t,k-1} - \eta_{\mathrm{local}} \cdot (\frac{1}{|\mathcal{D}_c^{t,k}|} \cdot \sum_{z_i \in \mathcal{D}_c^{t,k}} \nabla f_c(u_c^{t,k-1}, z_i) + \xi_c^{t,k})$
            end for
            $\Delta w_c(t) \leftarrow u_c^{t,K} - w^t$
            Client $c$ sends $\mathsf{sk}_t(\Delta w_c(t))$ to server
        end parfor
        /* Server */
        $\Delta\widetilde{w}^t \leftarrow \eta_{\mathrm{global}} \cdot \frac{1}{N} \sum_{c=1}^{N} \mathsf{sk}_t(\Delta w_c(t))$
        Server sends $\Delta\widetilde{w}^t$ to each client
    end for
end procedure
```

Theorem 6.2 (Informal version of Theorem H.11). *Let $\widehat{\epsilon}, \widehat{\delta}$ be as in Lemma 6.1. Then, Algorithm 2 is $(\epsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-DP, with*

$$\epsilon_{\mathrm{DP}} = \sqrt{TK} \cdot \widehat{\epsilon}, \quad \delta_{\mathrm{DP}} = TK \cdot \widehat{\delta}.$$

Proof Sketch. Notice that each agent c works on individual subsets of the data, therefore we can make use of Lemma H.2 to conclude that over all N agents, the process is $(\sqrt{K} \cdot \widehat{\epsilon}, K \cdot \widehat{\delta})$-DP. Finally, apply Lemma H.3 over all T iterations, we conclude that Algorithm 2 is $(\epsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-DP, with

$$\epsilon_{\mathrm{DP}} = \sqrt{TK} \cdot \widehat{\epsilon}, \quad \delta_{\mathrm{DP}} = TK \cdot \widehat{\delta}.$$

□

Compared to an iterative sketching framework we described without Gaussian noises, Algorithm 2 injects extra noises at each local step for each local client. It also performs sub-sampling. We note that the sub-sampling is essentially a form of SGD, hence, it does not affect the convergence too much. For the additive Gaussian noise, note that its parameter only mildly depends on the local Lipschitz constant ℓ_c, therefore it is unbiased and has small variance. Coupled with the convergence analysis in section 5, we obtain an algorithm that only communicates low-dimensional information, has differential privacy guarantee and provides good convergence rate.

We would also like to point out via more advanced techniques in differential privacy such as moment account or gradient clipping, the privacy-utility trade-off of Algorithm 2 can be improved. We do not aim to optimize over these perspectives in this paper since our purpose is to show the *necessity* of adapting differential privacy techniques. As we will show in section 7, without additional privacy introduced by the Gaussian noise, there exists a simple, iterative algorithm to recover the training data from communicated gradient and local parameter for a variety of loss functions.

7. Attack Sketched Gradients

To complement our algorithmic contribution, we show that under certain conditions on the loss functions f_c's for $c \in [N]$ and the local step $K = 1$, Algorithm 1 *without the additive Gaussian noise can leak information about the local data*. To achieve this goal, we present an attacking algorithm that effectively *learns* the local data through gradient descent.

7.1. Warm-up: Attacking Algorithm without Sketching

To start off, we describe an attacking algorithm without sketching being applied. We denote the loss function of the model by $F(w, x)$, where $x \in \mathbb{R}^m$ is the input and $w \in \mathbb{R}^d$ is the model parameter. We do not constrain $F(w, x)$ to be the loss of any specific model or task. $F(w, x)$ can be an ℓ_2 loss of linear regression model, a cross-entropy loss of a neural network, or any function that the clients in the training system want to minimize. In our federated learning scenario, we have $F(w, x) = \frac{1}{N} \sum_{c=1}^{N} f_c(w)$. Note that one can view the local loss function f_c being associated with the local dataset that can only be accessed by client c. During training, client c will send the gradient computed with the local training data $\nabla_w F(w, \widetilde{x}^{(c)})$ to the server where $\widetilde{x}^{(c)}$ denotes the local data.

The attacker can can observe the gradient information shared in the algorithm. For client c, the attacker could observe $g = \nabla_w F(w, \widetilde{x}^{(c)})$ and w. Intuitively, one can view attacker has hijacked one of the client and hence gaining access to the model parameter. Local data $\widetilde{x}^{(c)}$ will not be revealed to the attacker.

The attacker also has access to a gradient oracle, meaning that it can generate arbitrary data $x \in \mathbb{R}^m$ and feed into the oracle, and the oracle will return the gradient with respect to x and parameter w. The attacker will then try to find x

that minimizes

$$L(x) = \|\nabla_w F(w, x) - g\|^2$$

by running gradient descent. The attacker will start with random initialization x_0, and iterates as

$$x_{t+1} = x_t - \eta \cdot \nabla L(x_t)$$

where $\eta > 0$ is the step size chosen by the attacker.

To formalize the analysis, we introduce some key definitions. Given a function $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$, a data point $x \in \mathbb{R}^m$, a fixed (gradient) vector $g \in \mathbb{R}^d$ and a fixed (weight) vector $w \in \mathbb{R}^d$, we define the function L as follows:

$$L(x) := \|\nabla_w F(w, x) - g\|^2.$$

We consider the regime where $d \leq m$, i.e., the model is *under-parametrized*. The over-parametrized setting is studied in a recent work (Wang et al., 2023) that uses tensor decomposition to recover the data from gradients. In contrast, our approach simply applies gradient descent, therefore it can easily get stuck in a local minima, which is often the case in over-parametrized setting. However, our algorithm is notably simpler and computationally efficient.

To better illustrate properties we want on L, we define the matrix K as follows:

Definition 7.1. *Let $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$, suppose F is differentiable on both x and w, then we define pseudo-Hessian mapping $\Phi : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^{m \times d}$ as follows*

$$\Phi(x, w) = \nabla_x \nabla_w F(x, w).$$

Correspondingly, we define a pseudo-kernel $K : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ with respect to $\nabla_x F(w, x)$ as:

$$K(x, w) = \Phi(x, w)^\top \Phi(x, w).$$

Note the weight vector w is fixed in our setting, we write $K(x) = K(x, w)$ for simplicity.

For a regular Hessian matrix, one considers taking second derivative with respect to a single variable. Here, our input is $\nabla_w F(w, x)$ and we need to take gradient of the input with respect to x, hence, it is instructive to study the structure of $\nabla_x \nabla_w F(w, x)$.

We additionally introduce several key definitions that can be implied through some basic assumptions we will make. The first is a generalization of smoothness to the notion of *semi-smoothness*.

Definition 7.2 (Semi-smoothness)**.** *For any $p \in [0, 1]$, we say $L : \mathbb{R}^m \to \mathbb{R}$ is (a, b, p)-semi-smoothness if for any $x, y \in \mathbb{R}^m$, we have*

$$\begin{aligned} L(y) \leq\ & L(x) + \langle \nabla L(x), y - x \rangle \\ & + b\|y - x\|^2 + a\|x - y\|^{2-2p} L(x)^p. \end{aligned}$$

For examples, $L(x) = \|x\|^2$, $L(x) = \ln(1 + \exp(w^\top x))$, $L(x) = \tanh(w^\top x + b)$, $L(x) = \sqrt{w^\top x + b}$, $L(x) = \mathrm{sigmoid}(w^\top x + b)$, and $L(x) = \log(w^\top x)$ are semi-smooth.

Definition 7.3 (Non-critical point)**.** *We say $L : \mathbb{R}^m \to \mathbb{R}$ is (θ_1, θ_2)-non-critical point if*

$$\theta_1^2 \cdot L(x) \leq \|\nabla L(x)\|^2 \leq \theta_2^2 \cdot L(x).$$

The intuition for non-critical point property is that if $L(x)$ is large enough, then gradient descent can still make progress because $\|\nabla L(x)\|$ is lower bounded by $\theta_1^2 \cdot L(x)$. Suppose $F(w, x)$ has Lipschitz gradient and non-degenerate pseudo-kernel, then the corresponding L is semi-smooth and non-critical point:

Theorem 7.4. *If F satisfies the following properties: $\forall x \in \mathbb{R}^m$, $\nabla_w F(w, x)$ is β-Lipschitz w.r.t. x, and K's eigenvalues can be bounded by*

$$0 < \theta_1^2 \leq \lambda_1^2(x) \leq \cdots \leq \lambda_{\min(d,m)}^2(x) \leq \theta_2^2.$$

Then we have L is $(2(\beta + \theta_2), \beta, 1/2)$-semi-smooth (Def. 7.2), and L satisfies (θ_1, θ_2)-non-critical point (Def. 7.3).

We state Theorem 7.5 here and the proof is provided in appendix L.1.

Theorem 7.5. *Let*

- $\theta_1^2 > a \cdot \theta_2^{2-2p}$,
- $\eta \leq (\theta_1^2 - a \cdot \theta_2^{2-2p})/(2b \cdot \theta_2^2)$,
- $\gamma = \eta(\theta_1^2 - a \cdot \theta_2^{2-2p})/2$.

Suppose we run gradient descent algorithm to update x_{t+1} in each iteration as

$$x_{t+1} = x_t - \eta \cdot \nabla L(x)|_{x = x_t}.$$

If we assume L is (a, b, p)-semi-smooth (Def. 7.2) and (θ_1, θ_2)-non-critical point (Def. 7.3), then we have

$$L(x_{t+1}) - L(x^*) \leq (1 - \gamma)(L(x_t) - L(x^*)).$$

Theorem 7.5 states that a gradient descent that starts with a dummy data point x_0 can *converge* in the sense that it generates a point x_T whose gradient is close to the gradient of x^* we want to learn. As a direct consequence, if F has the property that similar gradients imply similar data points, then the attack algorithm truly recovers the data point it wants to learn. Such phenomenon has been widely observed in practice (Zhu et al., 2019; Yin et al., 2021; Zhao et al., 2020).

7.2. Attacking Gradients under Sketching

Now we consider the setting where *sketched* gradients are shared instead of the true gradient. Let $R : \mathbb{R}^d \to \mathbb{R}^{b_{\mathrm{sketch}}}$ be a sketching operator, then the gradient we observe becomes $R(\nabla_w F(w, x))$. Additionally, we can also observe the sketching matrix R and model parameter w. In this setting, the objective function we consider becomes

$$L_R(x) := \|R(\nabla_w F(w, x)) - R(g)\|^2.$$

It is reasonable to assume the attacker has access to R, since frameworks that make use of sketching (Rothchild et al., 2020) do so by sharing the sketching matrix across all nodes. Lemma 7.6 and Lemma 7.7 shows that with reasonable assumptions about R, which are typical properties of every popular sketch matrix, L still satisfies semi-smooth and non-critical-point condition. We defer all the proofs to appendix M.

Lemma 7.6. *If the sketching operator R satisfies $\|R(u) - R(v)\| \leq \tau\|u - v\|$ and $\|S\| \leq \gamma_2$, and F satisfies the conditions as in Theorem 7.4, then $L_R(x)$ is $(A, B, 1/2)$-semi-smooth where $A = 2\tau\beta + 2\theta_2\gamma_2$, $B = \tau^2\beta$.*

Lemma 7.7. *If the sketching operator R satisfies that the smallest singular value of $R^\top$ is at least $\gamma_1 > 0$ and F satisfies conditions as in Theorem 7.4, then $L_R(x)$ is $(2\theta_1\gamma_1, 2\theta_2\gamma_2)$-non-critical-point.*

While R itself is a short and fat matrix and is impossible to have nonzero smallest singular value, our singular value assumption is imposed on $R^\top \in \mathbb{R}^{d\times b_{\mathrm{sketch}}}$, hence reasonable. Moreover, for many sketching matrices R (such as each entry being i.i.d. Gaussian random variables), the matrix $R^\top$ is full rank almost surely. Combining Lemma 7.6 and Lemma 7.7, Theorem 7.8 shows that the system is still vulnerable to the gradient attack even for sketched gradients.

Theorem 7.8. *If the sketching operator R satisfies*

- $\|R(u) - R(v)\| \leq \tau\|u - v\|$,
- $0 < \gamma_1 \leq \sigma_1(R^\top) \leq \ldots \leq \sigma_s(R^\top) \leq \gamma_2$,

F satisfies the conditions in Theorem 7.4, then $L_R(x)$ is

- $(2\tau\beta + 2\theta_2\gamma_R, \tau^2\beta, 1/2)$*-semi-smooth,*
- $(2\theta_1\gamma_1, 2\theta_2\gamma_2)$*-non-critical-point.*

As popularized in (Rothchild et al., 2020), in federated learning, sketching can be applied to gradient vectors efficiently while squashing down the dimension of vectors being communicated. However, as indicated by our result, as soon as the attacker has access to the sketching operator, solving the sketched gradient attack problem reduces to the classical sketch-and-solve paradigm (Clarkson & Woodruff, 2013). This negative result highlights the necessity of using more complicated mechanisms to "encode" the gradients for privacy. One can adapt a cryptography-based algorithms at the expense of higher computation cost (Bonawitz et al., 2017), or alternatively, as we have shown in this paper, using differential privacy. We "mask" the gradient via Gaussian noises, so that even the attack algorithm can recover a point x that has similar gradient to the noisy gradient, it is still offset by the noise. Instead of injecting noises directly onto the gradient, one can also add noises *after* applying the sketching (Kenthapadi et al., 2013; Nikolov, 2023). We believe this approach will also lead to interesting privacy guarantees.

8. Conclusion

In this work, we propose the iterative sketch-based federated learning framework, which only communicates the sketched gradients with noises. Such a framework enjoys the benefits of both better privacy and lower communication cost per round. We also rigorously prove that the randomness from sketching will not introduce extra overall communication cost. Our approach and results can be extended to other gradient-based optimization algorithms and analysis, including but not limited to gradient descent with momentum and local stochastic gradient descent. This is because the sketched and de-sketched gradient $R^\top Rg$ is an unbiased estimator of the true gradient g with second moments being a multiplier of $\|g\|_2^2$.

By a simple modification to our algorithm with additive Gaussian noises on the gradients, we can also prove the differential privacy of our learning system by "hiding" the most important component in the system for guarding safety and privacy. This additive noise also does not affect the convergence behavior of our algorithm too much, since it does not make the estimator biased, and the additive variance can be factored into our original analysis.

To complement our algorithmic result, we also present a gradient leakage attack algorithm that can effectively learn the private data a federated learning framework wants to hide. Our gradient leakage attack algorithm is essentially that of gradient descent, but instead of optimizing over the model parameters, we try to optimize over the data points that a malicious attacker wants to learn. Even though the FL algorithm tries to "hide" information via random projections or data structures, we show that as long as the attacker has access to the sketching operator, it can still learn from gradient. Our attack algorithm is also computationally efficient.

Acknowledgement

The authors would like to thank Lianke Qin for many helpful discussions. Yitan Wang gratefully acknowledges support from ONR Award N00014-20-1-2335. Lichen Zhang is supported by NSF grant No. CCF-1955217 and NSF grant No. CCF-2022448.

References

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

Andoni, A., Lin, C., Sheng, Y., Zhong, P., and Zhong, R. Subspace embedding and linear regression with orlicz norm. In *International Conference on Machine Learning (ICML)*, pp. 224–233. PMLR, 2018.

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., and Felici, G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

Bernstein, S. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. *Practical Secure Aggregation for Privacy-Preserving Machine Learning*, pp. 1175–1191. Association for Computing Machinery, New York, NY, USA, 2017. ISBN 9781450349468.

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

Boutsidis, C. and Woodruff, D. P. Optimal cur matrix decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 353–362. ACM, 2014.

Boutsidis, C., Woodruff, D. P., and Zhong, P. Optimal principal component analysis in distributed and streaming models. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing (STOC)*, pp. 236–249, 2016.

Brand, J. v. d., Peng, B., Song, Z., and Weinstein, O. Training (over-parametrized) neural networks in near-linear time. In *ITCS*, 2021.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 634–649, Los Alamitos, CA, USA, oct 2015. IEEE Computer Society.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pp. 693–703. Springer, 2002.

Chen, M., Yang, Z., Saad, W., Yin, C., Poor, H. V., and Cui, S. A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications*, 2020.

Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pp. 81–90, 2013.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pp. 1223–1231, 2012.

Deng, Y., Song, Z., Wang, Y., and Yang, Y. A nearly optimal size coreset algorithm with nearly linear time. *arXiv preprint arXiv:2210.08361*, 2022.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–487, 2013.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60, 2010. doi: 10.1109/FOCS.2010.12.

Esfandiari, H., Mirrokni, V., and Zhong, P. Almost linear time density level set estimation via dbscan. In *AAAI*, 2021.

Foss, S., Korshunov, D., and Zachary, S. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

Gao, Y., Qin, L., Song, Z., and Wang, Y. A sublinear adversarial training algorithm. *arXiv preprint arXiv:2208.05395*, 2022.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting gradients–how easy is it to break privacy in federated learning? *Advances in neural information processing systems (NeurIPS)*, 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, pp. 2672–2680, 2014.

Haagerup, U. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

Hanson, D. L. and Wright, F. T. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Hitaj, B., Ateniese, G., and Perez-Cruz, F. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 603–618, 2017.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jiang, H., Lee, Y. T., Song, Z., and Wong, S. C.-w. An improved cutting plane method for convex optimization, convex-concave games and its applications. In *STOC*, 2020.

Jiang, S., Song, Z., Weinstein, O., and Zhang, H. Faster dynamic matrix inverse for faster lps. In *STOC*. arXiv preprint arXiv:2004.07470, 2021.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawit, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Theertha Suresh, A., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. *Advances and Open Problems in Federated Learning*. Now Foundations and Trends, 2021.

Kenthapadi, K., Korolova, A., Mironov, I., and Mishra, N. Privacy via the johnson-lindenstrauss transform. *Journal of Privacy and Confidentiality*, 2013.

Khaled, A., Mishchenko, K., and Richtárik, P. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

Khintchine, A. Über dyadische brüche. *Mathematische Zeitschrift*, 18(1):109–116, 1923.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *Advances in neural information processing systems (NeurIPS)*, 2016.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Lee, Y. T., Song, Z., and Zhang, Q. Solving empirical risk minimization in the current matrix multiplication time. In *COLT*, 2019.

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

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 133–141. Springer, 2019.

Li, X., Gu, Y., Dvornek, N., Staib, L., Ventola, P., and Duncan, J. S. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis*, 2020b.

Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Li, X., Song, Z., and Yang, J. Federated adversarial learning: A framework with convergence analysis. In *ICML*, 2023.

Lu, Y., Dhillon, P., Foster, D. P., and Ungar, L. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pp. 369–377, 2013.

Makarychev, K., Reddy, A., and Shan, L. Improved guarantees for k-means++ and k-means++ parallel. *Advances in Neural Information Processing Systems*, 33, 2020.

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

McPherson, R., Shokri, R., and Shmatikov, V. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706. IEEE, 2019.

Meng, X. and Mahoney, M. W. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing (STOC)*, pp. 91–100, 2013.

Nelson, J. and Nguyên, H. L. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 117–126. IEEE, 2013.

Nikolov, A. Private query release via the johnson-lindenstrauss transform. In *SODA*, 2023.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

Qin, L., Jayaram, R., Shi, E., Song, Z., Zhuo, D., and Chu, S. Adore: Differentially oblivious relational database operators. In *VLDB*, 2022a.

Qin, L., Reddy, A., Song, Z., Xu, Z., and Zhuo, D. Adaptive and dynamic multi-resolution hashing for pairwise summations. In *BigData*, 2022b.

Qin, L., Song, Z., and Wang, Y. Fast submodular function maximization. *arXiv preprint arXiv:2305.08367*, 2023a.

Qin, L., Song, Z., Zhang, L., and Zhuo, D. An online and unified algorithm for projection matrix vector multiplication with application to empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 101–156. PMLR, 2023b.

Qin, L., Song, Z., and Zhang, R. A general algorithm for solving rank-one matrix sensing. *arXiv preprint arXiv:2303.12298*, 2023c.

Reddy, A., Rossi, R. A., Song, Z., Rao, A., Mai, T., Lipka, N., Wu, G., Koh, E., and Ahmed, N. Online map inference and learning for nonsymmetric determinantal point processes. In *International Conference on Machine Learning (ICML)*, 2022a.

Reddy, A., Song, Z., and Zhang, L. Dynamic tensor product regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Rigaki, M. and García, S. A survey of privacy attacks in machine learning. *ArXiv*, abs/2007.07646, 2020.

Rothchild, D., Panda, A., Ullah, E., Ivkin, N., Stoica, I., Braverman, V., Gonzalez, J., and Arora, R. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pp. 8253–8265. PMLR, 2020.

Rudelson, M. and Vershynin, R. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Sarlós, T. Improved approximation algorithms for large matrices via random projections. In *Proceedings of 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

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

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Shrivastava, A., Song, Z., and Xu, Z. A tale of two efficient value iteration algorithms for solving linear mdps with large action space. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Song, Z. and Yu, Z. Oblivious sketching-based central path method for solving linear programming problems. In *38th International Conference on Machine Learning (ICML)*, 2021.

Song, Z., Woodruff, D. P., and Zhong, P. Low rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the 49th Annual Symposium on the Theory of Computing (STOC)*, 2017.

Song, Z., Woodruff, D. P., and Zhong, P. Relative error tensor low rank approximation. In *SODA*, 2019.

Song, Z., Yang, S., and Zhang, R. Does preprocessing help training over-parameterized neural networks? *Advances in Neural Information Processing Systems*, 34, 2021a.

Song, Z., Zhang, L., and Zhang, R. Training multi-layer over-parametrized neural network in subquadratic time. *arXiv preprint arXiv:2112.07628*, 2021b.

Stich, S. U. Local sgd converges fast and communicates little. In *ICLR*, 2019.

Tropp, J. A. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Federated learning with matched averaging. In *ICLR*, 2020a.

Wang, R., Zhong, P., Du, S. S., Salakhutdinov, R. R., and Yang, L. F. Planning with general objective functions: Going beyond total rewards. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., and Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2512–2520. IEEE, 2019.

Wang, Z., Lee, J., and Lei, Q. Reconstructing training data from model gradient, provably. In *AISTATS*, 2023.

Wei, W., Liu, L., Loper, M., Chow, K.-H., Gursoy, M., Truex, S., and Wu, Y. A framework for evaluating gradient leakage attacks in federated learning. In *ESORICS*, 2020.

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Woodruff, D. P. and Zhong, P. Distributed low rank approximation of implicit functions of a matrix. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 847–858. IEEE, 2016.

Xiao, C., Zhong, P., and Zheng, C. Bourgan: generative networks with metric embeddings. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2275–2286, 2018.

Xu, Z., Song, Z., and Shrivastava, A. Breaking the linear iteration cost barrier for some well-known conditional gradient methods using maxip data-structures. *Advances in Neural Information Processing Systems*, 34, 2021.

Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., and Molchanov, P. See through gradients: Image batch recovery via gradinversion. In *CVPR*, 2021.

Zhao, B., Mopuri, K. R., and Bilen, H. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zheng, W., Yan, L., Gou, C., and Wang, F.-Y. Federated meta-learning for fraudulent credit card detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In *NeurIPS*, pp. 14774–14784, 2019.

Appendix

Roadmap. We organize the appendix as follows. In section A, we introduce some notations and definitions that will be used across the appendix. In section B, we study several probability tools we will be using in the proof of cretain properties of various sketching matrices. In section C, we lay out some key assumptions on local objective function f_c and global objective function f, in order to proceed our discussion of convergence theory. In section D, we discuss the (α, β, δ)-coordinate wise embedding property we proposed in this work through several commonly used sketching matrices. In section E, we give complete proofs for single-step scheme. We dedicate sections F and G to illustrate formal analysis of the convergence results of Algorithm 1 under k local steps, given different assumptions of objective function f. In section H, we introduce additive noise to make our gradients differentially private, and conclude that an SGD version of our algorithm is indeed differentially private. In section I, we provide some preliminary definitions on gradient attack and elementary lemmas. In section J, we show what conditions of F would imply semi-smoothness and non-critical point of L. In section K, we prove with proper assumptions, x_t converges to the unique optimal solution x^*. In section L, we prove $L(x_t)$ converges to $L(x^*)$ under proper conditions. In section M, we extend the discussion by considering sketching and show what conditions of sketching would imply proper conditions of L.

A. Preliminary

For a positive integer n, we use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. We use $\mathbb{E}[\cdot]$ to denote expectation (if it exists), and use $\Pr[\cdot]$ to denote probability. For a function f, we use $\widetilde{O}(f)$ to denote $O(f \operatorname{poly} \log f)$. For a vector x, For a vector x, we use $\|x\|_1 := \sum_i |x_i|$ to denote its ℓ_1 norm, we use $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$ to denote its ℓ_2 norm, we use $\|x\|_\infty := \max_{i \in [n]} |x_i|$ to denote its ℓ_∞ norm. For a matrix A and a vector x, we define $\|x\|_A := \sqrt{x^\top A x}$. For a full rank square matrix A, we use A^{-1} to denote its true inverse. For a matrix A, we use $A^\dagger$ to denote its pseudo-inverse. For a matrix A, we use $\|A\|$ to denote its spectral norm. We use $\|A\|_F := (\sum_{i,j} A_{i,j}^2)^{1/2}$ to denote its Frobenius norm. We use $A^\top$ to denote the transpose of A. We denote $1_{\{x=l\}}$ for $l \in \mathbb{R}$ to be the indicator function which equals to 1 if $x = l$ and 0 otherwise. Let $f : A \to B$ and $g : C \to A$ be two functions, we use $f \circ g$ to denote the composition of functions f and g, i.e., for any $x \in C$, $(f \circ g)(x) = f(g(x))$. Given a real symmetric matrix $A \in \mathbb{R}^{d \times d}$, we use $\lambda_1(A), \ldots, \lambda_d(A)$ denote its smallest to largest eigenvalues. Given a real matrix A, we use $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ to denote its smallest and largest singular values.

B. Probability

Lemma B.1 (Chernoff bound (Chernoff, 1952)). *Let $Y = \sum_{i=1}^n Y_i$, where $Y_i = 1$ with probability p_i and $Y_i = 0$ with probability $1 - p_i$, and all Y_i are independent. Let $\mu = \mathbb{E}[Y] = \sum_{i=1}^n p_i$. Then*
1. $\Pr[Y \geq (1+\delta)\mu] \leq \exp(-\delta^2\mu/3)$, for all $\delta > 0$;
2. $\Pr[Y \leq (1-\delta)\mu] \leq \exp(-\delta^2\mu/2)$, for all $0 < \delta < 1$.

Lemma B.2 (Hoeffding bound (Hoeffding, 1963)). *Let $Z_1, \cdots, Z_n$ denote n independent bounded variables in $[a_i, b_i]$. Let $Z = \sum_{i=1}^n Z_i$, then we have*

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma B.3 (Bernstein inequality (Bernstein, 1924)). *Let $W_1, \cdots, W_n$ be independent zero-mean random variables. Suppose that $|W_i| \leq M$ almost surely, for all i. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^n W_i > t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[W_j^2] + Mt/3}\right).$$

Lemma B.4 (Khintchine's inequality, (Khintchine, 1923; Haagerup, 1981)). *Let $\sigma_1, \cdots, \sigma_n$ be i.i.d. sign random variables, and let $z_1, \cdots, z_n$ be real numbers. Then there are constants $C > 0$ so that for all $t > 0$*

$$\Pr\left[\Big|\sum_{i=1}^n z_i\sigma_i\Big| \geq t\|z\|_2\right] \leq \exp(-Ct^2).$$

Lemma B.5 (Hason-wright inequality (Hanson & Wright, 1971; Rudelson & Vershynin, 2013)). *Let $z \in \mathbb{R}^n$ denote a random vector with independent entries z_i with $\mathbb{E}[z_i] = 0$ and $|z_i| \leq K$. Let B be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\Pr[|z^\top Bz - \mathbb{E}[z^\top Bz]| > t] \leq 2 \cdot \exp(-c \min\{t^2/(K^4\|B\|_F^2), t/(K^2\|B\|)\}).$$

We state a well-know Lemma (see Lemma 1 on page 1325 in (Laurent & Massart, 2000)).

Lemma B.6 (Laurent and Massart (Laurent & Massart, 2000))**.** *Let $Z \sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero mean and σ^2 variance. Then*

$$\Pr[Z - k\sigma^2 \geq (2\sqrt{kt} + 2t)\sigma^2] \leq \exp(-t),$$
$$\Pr[k\sigma^2 - Z \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t).$$

Lemma B.7 (Tail bound for sub-exponential distribution (Foss et al., 2011))**.** *We say $X \in \mathrm{SE}(\sigma^2, \alpha)$ with parameters $\sigma > 0, \alpha > 0$ if:*

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\lambda^2\sigma^2/2), \quad \forall |\lambda| < 1/\alpha.$$

Let $X \in \mathrm{SE}(\sigma^2, \alpha)$ and $\mathbb{E}[X] = \mu$, then:

$$\Pr[|X - \mu| \geq t] \leq \exp(-0.5 \min\{t^2/\sigma^2, t/\alpha\}).$$

Lemma B.8 (Matrix Chernoff bound (Tropp, 2011; Lu et al., 2013))**.** *Let $\mathcal{X}$ be a finite set of positive-semidefinite matrices with dimension $d \times d$, and suppose that*

$$\max_{X \in \mathcal{X}} \lambda_{\max}(X) \leq B.$$

Sample $\{X_1, \cdots, X_n\}$ uniformly at random from $\mathcal{X}$ without replacement. We define $\mu_{\min}$ and $\mu_{\max}$ as follows:

$$\mu_{\min} := n \cdot \lambda_{\min}(\mathop{\mathbb{E}}_{X \sim \mathcal{X}}[X]) \text{ and } \mu_{\max} := n \cdot \lambda_{\max}(\mathop{\mathbb{E}}_{X \sim \mathcal{X}}[X]).$$

Then

$$\Pr\left[\lambda_{\min}(\sum_{i=1}^n X_i) \leq (1-\delta)\mu_{\min}\right] \leq d \cdot \exp(-\delta^2\mu_{\min}/B) \text{ for } \delta \in [0, 1),$$
$$\Pr\left[\lambda_{\max}(\sum_{i=1}^n X_i) \geq (1+\delta)\mu_{\max}\right] \leq d \cdot \exp\left(-\delta^2\mu_{\max}/(4B)\right) \text{ for } \delta \geq 0.$$

C. Optimization Backgrounds

Definition C.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function, we say f is L-smooth if for any $x, y \in \mathbb{R}^d$, we have*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

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

$$f(y) \leq f(x) + \langle y - x, \nabla f(x)\rangle + \frac{L}{2}\|y - x\|_2^2$$

Definition C.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function, we say f is convex if for any $x, y \in \mathbb{R}^d$, we have*

$$f(x) \geq f(y) + \langle x - y, \nabla f(y)\rangle$$

Definition C.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function, we say f is μ-strongly-convex if for any $x, y \in \mathbb{R}^d$, we have*

$$\|\nabla f(x) - \nabla f(y)\|_2 \geq \mu\|x - y\|_2$$

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

$$f(y) \geq f(x) + \langle y - x, \nabla f(x)\rangle + \frac{\mu}{2}\|y - x\|_2^2$$

Fact C.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth and convex function, then for any $x, y \in \mathbb{R}^d$, we have*

$$f(y) - f(x) \geq \langle y - x, \nabla f(x)\rangle + \frac{1}{2L} \cdot \|\nabla f(y) - \nabla f(x)\|_2^2$$

Fact C.5 (Inequality 4.12 in (Bottou et al., 2018))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex function. Let x^* be the minimizer of f. Then for any $x \in \mathbb{R}^d$, we have*

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

D. Sketching Matrices as Coordinate-wise Embedding

In this section, we discuss the (α, β, δ)-coordinate wise embedding property we proposed in this work through several commonly used sketching matrices.

We consider several standard sketching matrices:

1. Random Gaussian matrices.
2. Subsampled randomized Hadamard/Fourier transform matrices (Lu et al., 2013).
3. AMS sketch matrices (Alon et al., 1999), random $\{-1, +1\}$ per entry.
4. Count-Sketch matrices (Charikar et al., 2002), each column only has one non-zero entry, and is $-1, +1$ half probability each.
5. Sparse embedding matrices (Nelson & Nguyên, 2013), each column only has s non-zero entries, and each entry is $-\frac{1}{\sqrt{s}}, +\frac{1}{\sqrt{s}}$ half probability each.
6. Uniform sampling matrices.

D.1. Definition

Definition D.1 (k-wise independence)**.** *$\mathcal{H} = \{h : [m] \to [l]\}$ is a k-wise independent hash family if $\forall i_1 \neq i_2 \neq \cdots \neq i_k \in [n]$ and $\forall j_1, \cdots, j_k \in [l]$,*

$$\Pr_{h \in \mathcal{H}}[h(i_1) = j_1 \wedge \cdots \wedge h(i_k) = j_k] = \frac{1}{l^k}.$$

Definition D.2 (Random Gaussian matrix)**.** *We say $R \in \mathbb{R}^{b \times n}$ is a random Gaussian matrix if all entries are sampled from $\mathcal{N}(0, 1/b)$ independently.*

Definition D.3 (Subsampled randomized Hadamard/Fourier transform matrix (Lu et al., 2013))**.** *We say $R \in \mathbb{R}^{b \times n}$ is a subsampled randomized Hadamard transform matrix*[i] *if it is of the form $R = \sqrt{n/b}SHD$, where $S \in \mathbb{R}^{b \times n}$ is a random matrix whose rows are b uniform samples (without replacement) from the standard basis of $\mathbb{R}^n$, $H \in \mathbb{R}^{n \times n}$ is a normalized Walsh-Hadamard matrix, and $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables.*

Definition D.4 (AMS sketch matrix (Alon et al., 1999))**.** *Let $h_1, h_2, \cdots, h_b$ be b random hash functions picking from a 4-wise independent hash family $\mathcal{H} = \{h : [n] \to \{-\frac{1}{\sqrt{b}}, +\frac{1}{\sqrt{b}}\}\}$. Then $R \in \mathbb{R}^{b \times n}$ is a AMS sketch matrix if we set $R_{i,j} = h_i(j)$.*

Definition D.5 (Count-sketch matrix (Charikar et al., 2002))**.** *Let $h : [n] \to [b]$ be a random 2-wise independent hash function and $\sigma : [n] \to \{-1, +1\}$ be a random 4-wise independent hash function. Then $R \in \mathbb{R}^{b \times n}$ is a count-sketch matrix if we set $R_{h(i),i} = \sigma(i)$ for all $i \in [n]$ and other entries to zero.*

Definition D.6 (Sparse embedding matrix I (Nelson & Nguyên, 2013))**.** *We say $R \in \mathbb{R}^{b \times n}$ is a sparse embedding matrix with parameter s if each column has exactly s non-zero elements being $\pm 1/\sqrt{s}$ uniformly at random, whose locations are picked uniformly at random without replacement (and independent across columns)* [ii].

Definition D.7 (Sparse embedding matrix II (Nelson & Nguyên, 2013))**.** *Let $h : [n] \times [s] \to [b/s]$ be a a ramdom 2-wise independent hash function and $\sigma : [n] \times [s] \to \{-1, 1\}$ be a 4-wise independent. Then $R \in \mathbb{R}^{b \times n}$ is a sparse embedding matrix II with parameter s if we set $R_{(j-1)b/s + h(i,j), i} = \sigma(i,j)/\sqrt{s}$ for all $(i, j) \in [n] \times [s]$ and all other entries to zero.*[iii]

Definition D.8 (Uniform sampling matrix)**.** *We say $R \in \mathbb{R}^{b \times n}$ is a uniform sampling matrix if it is of the form $R = \sqrt{n/b}SD$, where $S \in \mathbb{R}^{b \times n}$ is a random matrix whose rows are b uniform samples (without replacement) from the standard basis of $\mathbb{R}^n$, and $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements are i.i.d. Rademacher random variables.*

[i] In this case, we require $\log n$ to be an integer.

[ii] For our purposes the signs need only be $O(\log d)$-wise independent, and each column can be specified by a $O(\log d)$-wise independent permutation, and the seeds specifying the permutations in different columns need only be $O(\log d)$-wise independent.

[iii] This definition has the same behavior as sparse embedding matrix I for our purpose.

D.2. Coordinate-wise Embedding

We define coordinate-wise embedding as follows

Definition D.9 ((α, β, δ)-coordinate-wise embedding)**.** *We say a randomized matrix $R \in \mathbb{R}^{b \times n}$ satisfying (α, β, δ)-coordinate wise embedding if*

1. $\mathbb{E}_{R \sim \Pi}[g^\top R^\top R h] = g^\top h,$
2. $\mathbb{E}_{R \sim \Pi}[(g^\top R^\top R h)^2] \leq (g^\top h)^2 + \frac{\alpha}{b} \|g\|_2^2 \|h\|_2^2,$
3. $\Pr_{R \sim \Pi}\left[|g^\top R^\top R h - g^\top h| \geq \frac{\beta}{\sqrt{b}} \|g\|_2 \|h\|_2 \right] \leq \delta.$

Remark D.10. *Given a randomized matrix $R \in \mathbb{R}^{b \times n}$ satisfying (α, β, δ)-coordinate wise embedding and any orthogonal projection $P \in \mathbb{R}^{n \times n}$, above definition implies*

1. $\mathbb{E}_{R \sim \Pi}[P R^\top R h] = P h,$
2. $\mathbb{E}_{R \sim \Pi}[(P R^\top R h)_i^2] \leq (P h)_i^2 + \frac{\alpha}{b} \|h\|_2^2,$
3. $\Pr_{R \sim \Pi}\left[|(P R^\top R h)_i - (P h)_i| \geq \frac{\beta}{\sqrt{b}} \|h\|_2 \right] \leq \delta.$

since $\|P\|_2 \leq 1$ implies $\|P_{i,:}\|_2 \leq 1$ for all $i \in [n]$.

D.3. Expectation and Variance

Lemma D.11. *Let $R \in \mathbb{R}^{b \times n}$ denote any of the random matrix in Definition D.2, D.3, D.4, D.6, D.7, D.8. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\mathbb{E}_{R \sim \Pi}[g^\top R^\top R h] = g^\top h$$

Proof.

$$\mathbb{E}_{R \sim \Pi}[g^\top R^\top R h] = g^\top \mathbb{E}_{R \sim \Pi}[R^\top R] h = g^\top I h = g^\top h.$$

□

Lemma D.12. *Let $R \in \mathbb{R}^{b \times n}$ denote a subsampled randomized Hadamard transform or AMS sketch matrix as in Definition D.3, D.4. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\mathbb{E}_{R \sim \Pi}[(g^\top R^\top R h)^2] \leq (g^\top h)^2 + \frac{2}{b} \|g\|_2^2 \cdot \|h\|_2^2.$$

Proof. If $\mathbb{E}_a[a] = b$, it is easy to see that

$$\mathbb{E}_a[(a-b)^2] = \mathbb{E}_a[a^2 - 2ab + b^2] = \mathbb{E}_a[a^2 - b^2]$$

We can rewrite it as follows:

$$\mathbb{E}_{R \sim \Pi}[(g^\top R^\top R h)^2 - (g^\top h)^2] = \mathbb{E}_{R \sim \Pi}[(g^\top (R^\top R - I) h)^2],$$

It can be bounded as follows:

$$\mathbb{E}_{R \sim \Pi}[(g^\top (R^\top R - I) h)^2]$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}(Rg)_k(Rh)_k - g^\top h\right)^2\right] \\
&= \mathop{\mathbb{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \cdot \sum_{j\in[n]\backslash\{i\}} R_{k,j}h_j\right)^2\right] \\
&= \mathop{\mathbb{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \cdot \sum_{j\in[n]\backslash\{i\}} R_{k,j}h_j\right)\cdot\left(\sum_{k'=1}^{b}\sum_{i'=1}^{n} R_{k',i'}g_{i'} \cdot \sum_{j'\in[n]\backslash\{i'\}} R_{k',j'}h_{j'}\right)\right] \\
&= \mathop{\mathbb{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^2g_i^2 \cdot \sum_{j\in[n]\backslash\{i\}} R_{k,j}^2h_j^2\right)+\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^2g_ih_i \cdot \sum_{j\in[n]\backslash\{i\}} R_{k,j}^2g_jh_j\right)\right] \\
&= \frac{1}{b}\left(\sum_{i=1}^{n} g_i^2 \sum_{j\in[n]\backslash\{i\}} h_j^2\right)+\frac{1}{b}\left(\sum_{i=1}^{n} g_ih_i \sum_{j\in[n]\backslash\{i\}} g_jh_j\right) \\
&\leq \frac{2}{b}\|g\|_2^2\|h\|_2^2,
\end{aligned}
$$

where the second step follows from $R_{k,i}^2 = 1/b$, $\forall k, i \in [b]\times[n]$, the forth step follows from $\mathbb{E}[R_{k,i}R_{k,j}R_{k',i'}R_{k',j'}] \neq 0$ only if $i = i'$, $j = j'$, $k = k'$ or $i = j'$, $j = i'$, $k = k'$, the fifth step follows from $R_{k,i}$ and $R_{k,j}$ are independent if $i \neq j$ and $R_{k,i}^2 = R_{k,j}^2 = 1/b$, and the last step follows from Cauchy-Schwartz inequality.

Therefore,

$$
\mathop{\mathbb{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2 - (g^\top h)^2] = \mathop{\mathbb{E}}_{R\sim\Pi}[(g^\top(R^\top R - I)h)^2] \leq \frac{2}{b}\|g\|_2^2\|h\|_2^2.
$$

□

Lemma D.13. *Let $R \in \mathbb{R}^{b\times n}$ denote a random Gaussian matrix as in Definition D.2. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$
\mathop{\mathbb{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \leq (g^\top h)^2 + \frac{3}{b}\|g\|_2^2\cdot\|h\|_2^2.
$$

Proof. Note

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{R\sim\Pi}[(g^\top R^\top Rh)^2] \\
&= \mathop{\mathbb{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \cdot \sum_{j=1}^{n} R_{k,j}h_j\right)^2\right] \\
&= \mathop{\mathbb{E}}_{R\sim\Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}g_i \cdot \sum_{j=1}^{n} R_{k,j}h_j\right)\cdot\left(\sum_{k'=1}^{b}\sum_{i'=1}^{n} R_{k',i'}g_{i'} \cdot \sum_{j'=1}^{n} R_{k',j'}h_{j'}\right)\right] \\
&= \mathop{\mathbb{E}}_{R\sim\Pi}\Big[\left(\sum_{k=1}^{b}\sum_{k'\in[b]\backslash\{k\}}\sum_{i=1}^{n}\sum_{i'=1}^{n} R_{k,i}^2R_{k',i'}^2g_ih_ig_{i'}h_{i'}\right)+\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^4g_i^2h_i^2\right) \\
&\quad+\left(\sum_{k=1}^{b}\sum_{i=1}^{n}\sum_{j\in[n]\backslash\{i\}} R_{k,i}^2R_{k,j}^2g_i^2h_j^2\right)+\left(\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{i'\in[n]\backslash\{i\}} R_{k,i}^2R_{k,i'}^2g_ih_ig_{i'}h_{i'}\right) \\
&\quad+\left(\sum_{k=1}^{b}\sum_{i=1}^{n}\sum_{j\in[n]\backslash\{i\}} R_{k,i}^2R_{k,j}^2g_ih_jg_jh_i\right)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{b-1}{b} \sum_{i=1}^{n} \sum_{i'=1}^{n} g_i h_i g_{i'} h_{i'} + \frac{3}{b} \sum_{i=1}^{n} g_i^2 h_i^2 \\
&\quad + \frac{1}{b} \sum_{i=1}^{n} \sum_{j \in [n] \setminus [i]} g_i^2 h_j^2 + \frac{1}{b} \sum_{i=1}^{n} \sum_{i' \in [n] \setminus [i]} g_i h_i g_{i'} h_{i'} + \frac{1}{b} \sum_{i=1}^{n} \sum_{j \in [n] \setminus [i]} g_i h_j g_j h_i \\
&\leq (g^\top h)^2 + \frac{3}{b} \|g\|_2^2 \|h\|_2^2,
\end{aligned}
$$

where the third step follows from that for independent entries of a random Gaussian matrix, $\mathbb{E}[R_{k,i} R_{k,j} R_{k',i'} R_{k',j'}] \neq 0$ only if 1. $k \neq k'$, $i = j$, $i' = j'$, or 2. $k = k'$, $i = i' = j = j'$, or 3. $k = k'$, $i = i' \neq j = j'$, or 4. $k = k'$, $i = j \neq i' = j'$, or 5. $k = k'$, $i = j' \neq i' = j$, the fourth step follows from $\mathbb{E}[R_{k,i}^2] = 1/b$ and $\mathbb{E}[R_{k,i}^4] = 3/b^2$, and the last step follows from Cauchy-Schwartz inequality. □

Lemma D.14. *Let $R \in \mathbb{R}^{b \times n}$ denote a count-sketch matrix as in Definition D.5. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$
\mathop{\mathbb{E}}_{R \sim \Pi}[(g^\top R^\top R h)^2] \leq (g^\top h)^2 + \frac{3}{b} \|g\|_2^2 \|h\|_2^2.
$$

Proof. Note

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{R \sim \Pi}[(g^\top R^\top R h)^2] \\
={}& \mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k=1}^{b} \sum_{i=1}^{n} R_{k,i} g_i \sum_{j=1}^{n} R_{k,j} h_j \right)^2 \right] \\
={}& \mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k=1}^{b} \sum_{i=1}^{n} R_{k,i} g_i \sum_{j=1}^{n} R_{k,j} h_j \right) \cdot \left(\sum_{k'=1}^{b} \sum_{i'=1}^{n} R_{k',i'} g_{i'} \sum_{j'=1}^{n} R_{k',j'} h_{j'} \right) \right] \\
={}& \mathop{\mathbb{E}}_{R \sim \Pi}\Big[\left(\sum_{k=1}^{b} \sum_{k' \in [b] \setminus \{k\}} \sum_{i=1}^{n} \sum_{i' \in [n] \setminus \{i\}}^{n} R_{k,i}^2 R_{k',i'}^2 g_i h_i g_{i'} h_{i'} \right) + \left(\sum_{k=1}^{b} \sum_{i=1}^{n} R_{k,i}^4 g_i^2 h_i^2 \right) \\
&+ \left(\sum_{k=1}^{b} \sum_{i=1}^{n} \sum_{j \in [n] \setminus \{i\}} R_{k,i}^2 R_{k,j}^2 g_i^2 h_j^2 \right) + \left(\sum_{k=1}^{n} \sum_{i=1}^{n} \sum_{i' \in [n] \setminus \{i\}}^{n} R_{k,i}^2 R_{k,i'}^2 g_i h_i g_{i'} h_{i'} \right) \\
&+ \left(\sum_{k=1}^{b} \sum_{i=1}^{n} \sum_{j \in [n] \setminus \{i\}} R_{k,i}^2 R_{k,j}^2 g_i h_j g_j h_i \right) \Big] \\
={}& \frac{b-1}{b} \sum_{i=1}^{n} \sum_{i' \in [n] \setminus i} g_i h_i g_{i'} h_{i'} + \sum_{i=1}^{n} g_i^2 h_i^2 \\
&+ \frac{1}{b} \sum_{i=1}^{n} \sum_{j \in [n] \setminus \{i\}} g_i^2 h_j^2 + \frac{1}{b} \sum_{i=1}^{n} \sum_{i' \in [n] \setminus \{i\}} g_i h_i g_{i'} h_{i'} + \frac{1}{b} \sum_{i=1}^{n} \sum_{j \in [n] \setminus \{i\}} g_i h_j g_j h_i \\
\leq{}& (g^\top h)^2 + \frac{3}{b} \|g\|_2^2 \|h\|_2^2,
\end{aligned}
$$

where in the third step we are again considering what values of k, k', i, i', j, j' that makes
$\mathbb{E}[R_{k,i} R_{k,j} R_{k',i'} R_{k',j'}] \neq 0$. Since the hash function $\sigma(\cdot)$ of the count-sketch matrix is 4-wise independent, $\forall k, k'$, when $i \neq i' \neq j \neq j'$, or $i = i' = j \neq j'$ (and the other 3 symmetric cases), we have that $\mathbb{E}[R_{k,i} R_{k,j} R_{k',i'} R_{k',j'}] = 0$. Since the count-sketch matrix has only one non-zero entry in every column, when $k \neq k'$, if $i = i'$ or $i = j'$ or $j = i'$ or $j = j'$, we also have $\mathbb{E}[R_{k,i} R_{k,j} R_{k',i'} R_{k',j'}] = 0$. Thus we only need to consider the cases: 1. $k \neq k'$, $i = j \neq i' = j'$, or 2. $k = k'$, $i = i' = j = j'$, or 3. $k = k'$, $i = i' \neq j = j'$, or 4. $k = k'$, $i = j \neq i' = j'$, or 5. $k = k'$, $i = j' \neq i' = j$. And the fourth step follows from $\mathbb{E}[R_{k,i}^2] = 1/b$ and $\mathbb{E}[R_{k,i}^4] = 1/b$, and the last step follows from Cauchy-Schwartz inequality. □

Lemma D.15. *Let $R \in \mathbb{R}^{b \times n}$ denote a sparse embedding matrix as in Definition D.6, D.7. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$2. \mathop{\mathbb{E}}_{R \sim \Pi}[(g^\top R^\top R h)^2] \le (g^\top h)^2 + \frac{2}{b}\|g\|_2^2 \cdot \|h\|_2^2.$$

Proof. Note

$$\begin{aligned}
&\mathop{\mathbb{E}}_{R \sim \Pi}[(g^\top R^\top R h)^2] \\
= &\mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i} g_i \sum_{j=1}^{n} R_{k,j} h_j\right)^2\right] \\
= &\mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i} g_i \sum_{j=1}^{n} R_{k,j} h_j\right) \cdot \left(\sum_{k'=1}^{b}\sum_{i'=1}^{n} R_{k',i'} g_{i'} \sum_{j'=1}^{n} R_{k',j'} h_{j'}\right)\right] \\
= &\mathop{\mathbb{E}}_{R \sim \Pi}\Big[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^2 g_i^2 \sum_{j \in [n] \setminus \{i\}} R_{k,j}^2 h_j^2\right) + \left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i}^2 g_i h_i \sum_{j \in [n] \setminus \{i\}} R_{k,j}^2 g_j h_j\right) \\
&+ \left(\sum_{k}\sum_{i \ne i'} R_{k,i}^2 R_{k,i'}^2 g_i h_i g_{i'} h_{i'}\right) + \left(\sum_{k}\sum_{i} R_{k,i}^4 g_i^2 h_i^2\right) + \left(\sum_{k \ne k'}\sum_{i \ne i'} R_{k,i}^2 R_{k',i'}^2 g_i h_i g_{i'} h_{i'}\right) \\
&+ \left(\sum_{k \ne k'}\sum_{i} R_{k,i}^2 R_{k',i}^2 g_i^2 h_i^2\right)\Big] \\
= &\frac{1}{b}\sum_{i \ne j} g_i^2 h_j^2 + \frac{1}{b}\sum_{i \ne j} g_i h_i g_j h_j + \frac{1}{b}\sum_{i \ne i'} g_i h_i g_{i'} h_{i'} + \frac{1}{s}\sum_{i} g_i^2 h_i^2 + \frac{b-1}{b}\sum_{i \ne i'} g_i h_i g_{i'} h_{i'} + \frac{s-1}{s}\sum_{i} g_i^2 h_i^2 \\
\le &(g^\top h)^2 + \frac{2}{b}\|g\|_2^2 \|h\|_2^2,
\end{aligned}$$

where the third step follows from the fact that the sparse embedding matrix has independent columns and s non-zero entry in every column, the fourth step follows from $\mathbb{E}[R_{k,i}^2] = 1/b$, $\mathbb{E}[R_{k,i}^4] = 1/(bs)$, and $\mathbb{E}[R_{k,i}^2 R_{k',i}^2] = \frac{s(s-1)}{b(b-1)} \cdot \frac{1}{s^2}, \forall k \ne k'$ and the last step follows from Cauchy-Schwartz inequality. □

Lemma D.16. *Let $R \in \mathbb{R}^{b \times n}$ denote a uniform sampling matrix as in Definition D.8. Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$2. \mathop{\mathbb{E}}_{R \sim \Pi}[(g^\top R^\top R h)^2] \le (g^\top h)^2 + \frac{n}{b}\|g\|_2^2 \|h\|_2^2.$$

Proof. Note

$$\begin{aligned}
&\mathop{\mathbb{E}}_{R \sim \Pi}[(g^\top R^\top R h)^2] \\
= &\mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i} g_i \sum_{j=1}^{n} R_{k,j} h_j\right)^2\right] \\
= &\mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k=1}^{b}\sum_{i=1}^{n} R_{k,i} g_i \sum_{j=1}^{n} R_{k,j} h_j\right) \cdot \left(\sum_{k'=1}^{b}\sum_{i'=1}^{n} R_{k',i'} g_{i'} \sum_{j'=1}^{n} R_{k',j'} h_{j'}\right)\right] \\
= &\mathop{\mathbb{E}}_{R \sim \Pi}\left[\left(\sum_{k}\sum_{i} R_{k,i}^4 g_i^2 h_i^2\right) + \left(\sum_{k \ne k'}\sum_{i \ne i'} R_{k,i}^2 R_{k',i'}^2 g_i h_i g_{i'} h_{i'}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n}{b} \sum_i g_i^2 h_i^2 + \frac{(b-1)n}{(n-1)b} \sum_{i \neq i'} g_i h_i g_{i'} h_{i'} \\
&\leq (g^\top h)^2 + \frac{n}{b} \|g\|_2^2 \|h\|_2^2,
\end{aligned}
$$

where the third step follows from the fact that the random sampling matrix has one non-zero entry in every row, the fourth step follows from $\mathbb{E}[R_{k,i}^2 R_{k',i'}^2] = n/((n-1)b^2)$ for $k \neq k',\ i \neq i'$ and $\mathbb{E}[R_{k,i}^4] = n/b^2$. □

Remark D.17. *Lemma D.16 indicates that uniform sampling fails in bounding variance in some sense, since the upper bound give here involves n.*

D.4. Bounding Inner Product

Lemma D.18 (Gaussian)**.** *Let $R \in \mathbb{R}^{b \times n}$ be a random Gaussian matrix (Definition D.2). Then we have:*

$$
\Pr\Big[\max_{i \neq j} |\langle R_{*,i}, R_{*,j} \rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \Theta(\delta).
$$

Proof. Note for $i \neq j$, $R_{*,i}, R_{*,j} \sim \mathcal{N}(0, \frac{1}{b} I_b)$ are two independent Gaussian vectors. Let $z_k = R_{k,i} R_{k,j}$ and $z = \langle R_{*,i}, R_{*,j} \rangle$. Then we have for any $|\lambda| \leq b/2$,

$$
\mathbb{E}[e^{\lambda z_k}] = \frac{1}{\sqrt{1 - \lambda^2/b^2}} \leq \exp(\lambda^2/b^2),
$$

where the first step follows from $z_k = \frac{1}{4}(R_{k,i} + R_{k,j})^2 + \frac{1}{4}(R_{k,i} - R_{k,j})^2 = \frac{b}{2}(Q_1 - Q_2)$ where $Q_1, Q_2 \sim \chi_1^2$, and $\mathbb{E}[e^{\lambda Q}] = \frac{1}{\sqrt{1-2\lambda}}$ for any $Q \sim \chi_1^2$.

This implies $z_k \in \mathrm{SE}(2/b^2, 2/b)$ is a sub-exponential random variable. Thus, we have $z = \sum_{k=1}^b z_k \in \mathrm{SE}(2/b, 2/b)$, by sub-exponential concentration Lemma B.7 we have

$$
\Pr[|z| \geq t] \leq 2\exp(-bt^2/4)
$$

for $0 < t < 1$. Picking $t = \sqrt{\log(n^2/\delta)/b}$, we have

$$
\Pr\Big[|\langle R_{*,i}, R_{*,j} \rangle| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \delta/n^2.
$$

Taking the union bound over all $(i,j) \in [n] \times [n]$ and $i \neq j$, we complete the proof. □

Lemma D.19 (SRHT)**.** *Let $R \in \mathbb{R}^{b \times n}$ be a subsample randomized Hadamard transform (Definition D.3). Then we have:*

$$
\Pr\Big[\max_{i \neq j} |\langle R_{*,i}, R_{*,j} \rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \Theta(\delta).
$$

Proof. For fixed $i \neq j$, let $X = [R_{*,i}, R_{*,j}] \in \mathbb{R}^{b \times 2}$. Then $X^\top X = \sum_{k=1}^b G_k$, where

$$
G_k = [R_{k,i}, R_{k,j}]^\top [R_{k,i}, R_{k,j}] = \begin{bmatrix} \frac{1}{b} & R_{k,i} R_{k,j} \\ R_{k,i} R_{k,j} & \frac{1}{b} \end{bmatrix}.
$$

Note the eigenvalues of G_k are 0 and $\frac{2}{b}$ and $\mathbb{E}[X^\top X] = b \cdot \mathbb{E}[G_k] = I_2$ for all $k \in [b]$. Thus, applying matrix Chernoff bound B.8 to $X^\top X$ we have

$$
\begin{aligned}
&\Pr\Big[\lambda_{\max}(X^\top X) \leq 1 - t\Big] \leq 2\exp\left(-t^2 b/2\right) \text{ for } t \in [0,1), \text{ and} \\
&\Pr\Big[\lambda_{\max}(X^\top X) \geq 1 + t\Big] \leq 2\exp\left(-t^2 b/8\right) \text{ for } t \geq 0.
\end{aligned}
$$

which implies the eigenvalues of $X^\top X$ are between $[1-t, 1+t]$ with probability $1 - 4\exp(-\frac{t^2 b}{8})$. So the eigenvalues of $X^\top X - I_2$ are between $[-t, t]$ with probability $1 - 4\exp(-\frac{t^2 b}{8})$. Picking $t = \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}$, we have

$$\Pr\Big[\|X^\top X - I_2\| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \frac{\delta}{n^2}.$$

Note

$$X^\top X - I_2 = \begin{bmatrix} 0 & \langle R_{*,i}, R_{*,j}\rangle \\ \langle R_{*,i}, R_{*,j}\rangle & 0 \end{bmatrix},$$

whose spectral norm is $|\langle R_{*,i}, R_{*,j}\rangle|$. Thus, we have

$$\Pr\Big[|\langle R_{*,i}, R_{*,j}\rangle| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \delta/n^2.$$

Taking a union bound over all pairs $(i,j) \in [n] \times [n]$ and $i \neq j$, we complete the proof. □

Lemma D.20 (AMS)**.** *Let $R \in \mathbb{R}^{b\times n}$ be a random AMS matrix (Definition D.4). Let $\{\sigma_i,\ i \in [n]\}$ be independent Rademacher random variables and $\overline{R} \in \mathbb{R}^{b\times n}$ with $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$. Then we have:*

$$\Pr\Big[\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}\Big] \leq \Theta(\delta).$$

Proof. Note for any fixed $i \neq j$, $\overline{R}_{*,i}$ and $\overline{R}_{*,j}$ are independent. By Hoeffding inequality (Lemma B.2), we have

$$\Pr\Big[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq t\Big] \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^b (\frac{1}{b} - (-\frac{1}{b}))^2}\Big) \leq 2e^{-t^2 b/2}$$

Choosing $t = \sqrt{2\log(2n^2/\delta)}/\sqrt{b}$, we have

$$\Pr\Big[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \sqrt{2\log(2n^2/\delta)}/\sqrt{b}\Big] \leq \frac{\delta}{n^2}.$$

Taking a union bound over all pairs $(i,j) \in [n] \times [n]$ and $i \neq j$, we complete the proof. □

Lemma D.21 (Count-Sketch)**.** *Let $R \in \mathbb{R}^{b\times n}$ be a count-sketch matrix (Definition D.5). Let $\{\sigma_i,\ i \in [n]\}$ be independent Rademacher random variables and $\overline{R} \in \mathbb{R}^{b\times n}$ with $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$. Then we have:*

$$\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \leq 1.$$

Proof. Directly follow the definition of count-sketch matrices. □

Lemma D.22 (Sparse embedding)**.** *Let $R \in \mathbb{R}^{b\times n}$ be a sparse embedding matrix with parameter s (Definition D.6 and D.7). Let $\{\sigma_i,\ i \in [n]\}$ be independent Rademacher random variables and $\overline{R} \in \mathbb{R}^{b\times n}$ with $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$. Then we have:*

$$\Pr\Big[\max_{i\neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{s}}\Big] \leq \Theta(\delta).$$

Proof. Note for fixed $i \neq j$, $\overline{R}_{*,i}$ and $\overline{R}_{*,j}$ are independent. Assume $R_{*,i}$ and $R_{*,j}$ has u non-zero elements at the same positions, where $0 \leq u \leq s$, then by Hoeffding inequality (Lemma B.2), we have

$$\Pr[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq t] \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^u (\frac{1}{s} - (-\frac{1}{s}))^2}\Big) \leq 2\exp(-t^2 s^2/(2u)) \tag{3}$$

Let $t = \sqrt{(2u/s^2)\log(2n^2/\delta)}$, we have

$$\begin{aligned}\Pr\Big[|\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \geq \sqrt{2s^{-1}\log(2n^2/\delta)}\Big] &\leq \Pr\Big[|\langle R_{*,i}, R_{*,j}\rangle| \geq \sqrt{2us^{-2}\log(2n^2/\delta)}\Big] \\ &\leq \delta/n^2 \end{aligned} \tag{4}$$

since $u \leq s$. By taking a union bound over all $(i,j) \in [n] \times [n]$ and $i \neq j$, we complete the proof. □

D.5. Infinite Norm Bound

Lemma D.23 (SRHT and AMS)**.** *Let $R \in \mathbb{R}^{b \times n}$ denote a subsample randomized Hadamard transform (Definition D.3) or AMS sketching matrix (Definition D.4). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top Rh) - (g^\top h)| > \frac{\log^{1.5}(n/\delta)}{\sqrt{b}}\|g\|_2\|h\|_2\right] \leq \Theta(\delta).$$

Proof. We can rewrite $(g^\top R^\top Rh) - (g^\top h)$ as follows:,

$$\begin{aligned}(g^\top R^\top Rh) - (g^\top h) &= \sum_{i=1}^n \sum_{j \in [n] \backslash i}^n g_i h_j \langle R_{*,i}, R_{*,j}\rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1) \\ &= \sum_{i=1}^n \sum_{j \in [n] \backslash i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j}\rangle.\end{aligned}$$

where σ_i's are independent Rademacher random variables and $\overline{R}_{*,i} = \sigma_i R_{*,i},\ \forall i \in [n]$, and the second step follows from $\|R_{*,i}\|_2^2 = 1, \forall i \in [n]$.

We define matrix $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ as follows:

$$\begin{aligned}A_{i,j} &= g_i h_j \cdot \langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle, && \forall i \in [n], j \in [n] \\ B_{i,j} &= g_i h_j \cdot \max_{i' \neq j'} |\langle \overline{R}_{*,i'}, \overline{R}_{*,j'}\rangle| && \forall i \in [n], j \in [n]\end{aligned}$$

We define $A^\circ \in \mathbb{R}^{n \times n}$ to be the matrix $A \in \mathbb{R}^{n \times n}$ with removing diagonal entries, applying Hason-wright inequality (Lemma B.5), we have

$$\Pr_\sigma[|\sigma^\top A^\circ \sigma| \geq \tau] \leq 2 \cdot \exp(-c \min\{\tau^2/\|A^\circ\|_F^2, \tau/\|A^\circ\|\})$$

We can upper bound $\|A^\circ\|$ and $\|A^\circ\|_F$.

$$\begin{aligned}\|A^\circ\| &\leq \|A^\circ\|_F \\ &\leq \|A\|_F \\ &\leq \|B\|_F \\ &= \|g\|_2 \cdot \|h\|_2 \cdot \max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \\ &\leq \|g\|_2 \cdot \|h\|_2 \cdot \max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle|.\end{aligned}$$

where the forth step follows from B is rank-1.

For SRHT, note $\overline{R}$ has the same distribution as R. By Lemma D.19 (for AMS, we use Lemma D.20) with probability at least $1 - \Theta(\delta)$, we have :

$$\max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j}\rangle| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}.$$

Conditioning on the above event holds.

Choosing $\tau = \|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(n/\delta)/\sqrt{b}$, we can show that

$$\Pr\left[\left|(g^\top R^\top Rh) - (g^\top h)\right| \geq \|g\|_2 \cdot \|h\|_2 \frac{\log^{1.5}(n/\delta)}{\sqrt{b}}\right] \leq \Theta(\delta).$$

Thus, we complete the proof. □

Lemma D.24 (Random Gaussian)**.** *Let $R \in \mathbb{R}^{b \times n}$ denote a random Gaussian matrix (Definition D.2). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\Pr_{R \sim \Pi}\Big[|(g^\top R^\top R h) - (g^\top h)| > \frac{\log^{1.5}(n/\delta)}{\sqrt{b}} \|g\|_2 \|h\|_2 \Big] \leq \Theta(\delta).$$

Proof. We follow the same procedure as proving Lemma D.23.

We can rewrite $(g^\top R^\top R h) - (g^\top h)$ as follows:,

$$\begin{aligned}(g^\top R^\top R h) - (g^\top h) &= \sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle R_{*,i}, R_{*,j} \rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1) \\ &= \sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j} \rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1). \end{aligned} \tag{5}$$

where σ_i's are independent Rademacher random variables and $\overline{R}$ has the same distribution as R.

To bound the first term $\sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j} \rangle$, we define matrix $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ as follows:

$$\begin{aligned} A_{i,j} &= g_i h_j \cdot \langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle, & \forall i \in [n], j \in [n] \\ B_{i,j} &= g_i h_j \cdot \max_{i' \neq j'} |\langle \overline{R}_{*,i'}, \overline{R}_{*,j'} \rangle| & \forall i \in [n], j \in [n] \end{aligned}$$

We define $A^\circ \in \mathbb{R}^{n \times n}$ to be the matrix $A \in \mathbb{R}^{n \times n}$ with removing diagonal entries, applying Hason-wright inequality (Lemma B.5), we have

$$\Pr_\sigma[|\sigma^\top A^\circ \sigma| \geq \tau] \leq 2 \cdot \exp(-c \min\{\tau^2 / \|A^\circ\|_F^2, \tau / \|A^\circ\|\})$$

We can upper bound $\|A^\circ\|$ and $\|A^\circ\|_F$.

$$\begin{aligned} \|A^\circ\| &\leq \|A^\circ\|_F \\ &\leq \|A\|_F \\ &\leq \|B\|_F \\ &= \|g\|_2 \cdot \|h\|_2 \cdot \max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle| \\ &\leq \|g\|_2 \cdot \|h\|_2 \cdot \max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle|. \end{aligned}$$

where the forth step follows from B is rank-1.

Using Lemma D.18 with probability at least $1 - \Theta(\delta)$, we have :

$$\max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle| \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b}}.$$

Conditioning on the above event holds.

Choosing $\tau = \|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(n/\delta)/\sqrt{b}$, we can show that

$$\Pr\Big[\Big| \sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle \sigma_i \overline{R}_{*,i}, \sigma_j \overline{R}_{*,j} \rangle \Big| \geq \|g\|_2 \cdot \|h\|_2 \frac{\log^{1.5}(n/\delta)}{\sqrt{b}} \Big] \leq \Theta(\delta). \tag{6}$$

To bound the second term $\sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1)$, note that $b\|R_{*,i}\|_2^2 \sim \chi_b^2$ for every $i \in [n]$. Applying Lemma B.6, we have

$$\Pr\Big[\Big| \|R_{*,i}\|_2^2 - 1 \Big| \geq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}} \Big] \leq \delta / n.$$

which implies

$$\Pr\left[\sum_{i=1}^{n} g_i h_i \Big| \|R_{*,i}\|_2^2 - 1 \Big| \geq \|g\|_2 \|h\|_2 \frac{c\sqrt{\log(n/\delta)}}{\sqrt{b}}\right] \leq \Theta(\delta). \tag{7}$$

Plugging the bounds Eq. (6) and (7) back to Eq. (5), we complete the proof. □

Lemma D.25 (Count-sketch)**.** *Let $R \in \mathbb{R}^{b \times n}$ denote a count-sketch matrix (Definition D.5). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| \geq \log(1/\delta) \|g\|_2 \|h\|_2\right] \leq \Theta(\delta).$$

Proof. We follow the identical procedure as proving Lemma D.23 to apply Hason-wright inequality (Lemma B.5).

Then note Lemma D.21 shows

$$\max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle| \leq 1$$

Thus, choosing $\tau = c\|g\|_2 \cdot \|h\|_2 \cdot \log(1/\delta)$, we can show that

$$\Pr\left[|(g^\top R^\top R h) - (g^\top h)| \geq c\|g\|_2 \cdot \|h\|_2 \log(1/\delta)\right] \leq \delta.$$

which completes the proof. □

Lemma D.26 (Count-sketch 2)**.** *Let $R \in \mathbb{R}^{b \times n}$ denote a count-sketch matrix (Definition D.5). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| \geq \frac{1}{\sqrt{b\delta}} \|g\|_2 \|h\|_2\right] \leq \Theta(\delta).$$

Proof. It is known that a count-sketch matrix with $b = \epsilon^{-2}\delta^{-1}$ rows satisfies the $(\epsilon, \delta, 2)$-JL moment property (see e.g. Theorem 14 of (Woodruff, 2014)). Using Markov's inequality, $(\epsilon, \delta, 2)$-JL moment property implies

$$\Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| \geq \epsilon \|g\|_2 \|h\|_2\right] \leq \Theta(\delta),$$

where $\epsilon = \frac{1}{\sqrt{b\delta}}$. □

Lemma D.27 (Sparse embedding)**.** *Let $R \in \mathbb{R}^{b \times n}$ denote a sparse-embedding matrix (Definition D.6 and D.7). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$3. \Pr_{R \sim \Pi}\left[|(g^\top R^\top R h) - (g^\top h)| > \frac{\log^{1.5}(n/\delta)}{\sqrt{s}} \|g\|_2 \|h\|_2\right] \leq \Theta(\delta).$$

Proof. We follow the identical procedure as proving Lemma D.23 to apply Hason-wright inequality (Lemma B.5).

Then note Lemma D.22 shows with probability at least $1 - \delta$ we have

$$\max_{i \neq j} |\langle \overline{R}_{*,i}, \overline{R}_{*,j} \rangle| \leq \frac{c\sqrt{\log(n/\delta)}}{\sqrt{s}}.$$

Conditioning on the above event holds, choosing $\tau = c'\|g\|_2 \cdot \|h\|_2 \cdot \log^{1.5}(1/\delta)$, we can show that

$$\Pr\left[|(g^\top R^\top R h) - (g^\top h)| \geq \frac{c' \log^{1.5}(n/\delta)}{\sqrt{s}} \|g\|_2 \cdot \|h\|_2\right] \leq \Theta(\delta).$$

Thus, we complete the proof. □

Lemma D.28 (Uniform sampling)**.** *Let $R \in \mathbb{R}^{b \times n}$ denote a uniform sampling matrix (Definition D.8). Then for any fixed vector $h \in \mathbb{R}^n$ and any fixed vector $g \in \mathbb{R}^n$, the following properties hold:*

$$3. |(g^\top R^\top R h) - (g^\top h)| \leq (1 + \frac{n}{b}) \|g\|_2 \|h\|_2$$

where $I \subset [n]$ be the subset of indexes chosen by the uniform sampling matrix.

Proof. We can rewrite $(g^\top R^\top R h) - (g^\top h)$ as follows:,

$$\begin{aligned}(g^\top R^\top R h) - (g^\top h) &= \sum_{i=1}^n \sum_{j \in [n] \setminus i}^n g_i h_j \langle R_{*,i}, R_{*,j} \rangle + \sum_{i=1}^n g_i h_i (\|R_{*,i}\|_2^2 - 1) \\ &= \frac{n}{b} \sum_{i \in I} g_i h_i - \sum_{i=1}^n g_i h_i.\end{aligned}$$

where the second step follows from the uniform sampling matrix has only one nonzero entry in each row.

Let $I \subset [n]$ be the subset chosen by the uniform sampling matrix, then $\|R_{*,i}\|_2^2 = n/b$ for $i \in I$ and $\|R_{*,i}\|_2^2 = 0$ for $i \in [n] \setminus I$. So we have

$$\begin{aligned}|(g^\top R^\top R h) - (g^\top h)| &= \Big| \sum_{i \in I} g_i h_i (\frac{n}{b} - 1) - \sum_{i \in [n] \setminus I} g_i h_i \Big| \\ &\leq (1 + \frac{n}{b}) \|g\|_2 \|h\|_2.\end{aligned}$$

□

E. Analysis of Convergence: Single-step Scheme

E.1. Preliminary

Throughout the proof of convergence, we will use $\mathcal{F}_t$ to denote the sequence $w_{t-1}, w_{t-2}, \ldots, w_0$. Also, we use η as a shorthand for $\eta_{\text{global}} \cdot \eta_{\text{local}}$.

E.2. Strongly-convex f Convergence Analysis

Theorem E.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumption 3.1 with $\mu > 0$. Let $w^* \in \mathbb{R}^d$ be the optimal solution to f and assume* $\mathsf{sk}/\mathsf{desk}$ *functions satisfying Theorem 4.2. Suppose $\eta := \eta_{\text{global}} \cdot \eta_{\text{local}}$ has the property that $\eta \leq \frac{1}{(1+\alpha)L}$, then*

$$\mathbb{E}[f(w^{t+1})] - f(w^*) \leq (1 - \mu\eta)^t \cdot (f(w^0) - f(w^*))$$

Proof. We shall first bound $f(w^{t+1}) - f(w^t)$:

$$\begin{aligned}f(w^{t+1}) - f(w^t) \leq & \langle w^{t+1} - w^t, \nabla f(w^t) \rangle + \frac{L}{2} \|w^{t+1} - w^t\|_2^2 \\ = & \langle \mathsf{desk}_t(\Delta \widetilde{w}^t), \nabla f(w^t) \rangle + \frac{L}{2} \|\mathsf{desk}_t(\Delta \widetilde{w}^t)\|_2^2 \\ = & - \langle \eta_{\text{global}} \cdot \mathsf{desk}_t(\frac{1}{N} \sum_{c=1}^N \mathsf{sk}_t(\eta_{\text{local}} \cdot \nabla f_c(w^t))), \nabla f(w^t) \rangle \\ & + \frac{L}{2} \|\eta_{\text{global}} \cdot \mathsf{desk}_t(\frac{1}{N} \sum_{c=1}^N \mathsf{sk}_t(\eta_{\text{local}} \cdot \nabla f_c(w^t)))\|_2^2 \\ = & - \eta_{\text{global}} \cdot \eta_{\text{local}} \cdot \langle \mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t))), \nabla f(w^t) \rangle \\ & + (\eta_{\text{global}} \cdot \eta_{\text{local}})^2 \cdot \|\mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t)))\|_2^2\end{aligned}$$

where the first step uses the L-smoothness condition of f, and the last step uses the linearity property of $\mathsf{sk}/\mathsf{desk}$ functions.

Taking expectation over iteration t conditioning on $\mathcal{F}_t$ and note that only w^{t+1} depends on randomness at t, we get

$$\begin{aligned}
& \mathbb{E}[f(w^{t+1}) - f(w^t) \mid \mathcal{F}_t] \\
\leq & -\eta \cdot \langle \mathbb{E}[\mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t))) \mid \mathcal{F}_t], \nabla f(w^t) \rangle + \frac{L\eta^2}{2} \mathbb{E}[\|\mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t)))\|_2^2 \mid \mathcal{F}_t] \\
\leq & -\eta \cdot \langle \nabla f(w^t), \nabla f(w^t) \rangle + \frac{L\eta^2}{2}(1+\alpha) \cdot \|\nabla f(w^t)\|_2^2 \\
\leq & -\frac{\eta}{2} \cdot \|\nabla f(w^t)\|_2^2 \\
\leq & -\mu\eta \cdot (f(w^t) - f(w^*))
\end{aligned} \tag{8}$$

where the second step comes from the fact that $\mathsf{desk}_t(\mathsf{sk}_t(h))$ is an unbiased estimator for any fixed $h \in \mathbb{R}^d$ and the bound on its variance, the third step comes from $\eta \leq \frac{1}{(1+\alpha)L}$, and the last step comes from Fact C.5.

Upon rearranging and subtracting both sides by $f(w^*)$, we get

$$\mathbb{E}[f(w^{t+1})] - f(w^*) \mid \mathcal{F}_t] \leq (1 - \mu\eta) \cdot (f(w^t) - f(w^*)) \tag{9}$$

Note that if we apply expectation over $\mathcal{F}_t$ on both sides of Eq. (9) we can get

$$\mathbb{E}[f(w^{t+1})] - f(w^*) \leq (1 - \mu\eta) \cdot (\mathbb{E}[f(w^t)] - f(w^*)) \tag{10}$$

Notice since $1 - \mu\eta \leq 1$, this is a contraction map, if we iterate this recurrence relation, we will finally get

$$\mathbb{E}[f(w^{t+1}) - f(w^*)] \leq (1 - \mu\eta)^t \cdot (f(w^0) - f(w^*)). \tag{11}$$

□

E.3. Convex f Convergence Analysis

Assume f is a convex function, we obtain a convergence bound in terms of the average of all parameters.

Theorem E.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumption 3.1 with $\mu = 0$. Suppose $\mathsf{sk}/\mathsf{desk}$ functions satisfying Theorem 4.2. If $\eta := \eta_{\mathrm{global}} \cdot \eta_{\mathrm{local}} \leq \frac{1}{2(1+\alpha)L}$, then*

$$\mathbb{E}[f(\overline{w}^T) - f(w^*)] \leq \frac{\mathbb{E}[\|w^0 - w^*\|_2^2]}{\eta \cdot (T+1)}$$

where $\overline{w}^T := \frac{1}{T+1}\sum_{t=0}^T w^t$ and $w^* \in \mathbb{R}^d$ is the optimal solution.

Proof. We shall first compute the gap between w^{t+1} and w^*:

$$\begin{aligned}
& \|w^{t+1} - w^*\|_2^2 \\
= & \|w^t - \mathsf{desk}_t(\Delta\widetilde{w}^t) - w^*\|_2^2 \\
= & \|w^t - \eta \cdot \mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t))) - w^*\|_2^2 \\
= & \|w^t - w^*\|_2^2 + \eta^2 \cdot \|\mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t)))\|_2^2 - 2\eta \cdot \langle w^t - w^*, \mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t))) \rangle
\end{aligned} \tag{12}$$

By unbiasedness of $\mathsf{desk}_t \circ \mathsf{sk}_t$, we have

$$\mathbb{E}[\langle w^t - w^*, \mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t))) \rangle \mid \mathcal{F}_t] = \mathbb{E}[\langle w^t - w^*, \nabla f(w^t) \rangle \mid \mathcal{F}_t] \tag{13}$$

Taking total expectation of Eq. (12) and plug in Eq. (13), we get

$$\begin{aligned}
& \mathbb{E}[\|w^{t+1} - w^*\|_2^2 \mid \mathcal{F}_t] \\
= & \mathbb{E}[\|w^t - w^*\|_2^2 \mid \mathcal{F}_t] + \eta^2 \cdot \mathbb{E}[\|\mathsf{desk}_t(\mathsf{sk}_t(\nabla f(w^t)))\|_2^2 \mid \mathcal{F}_t] - 2\eta \cdot \mathbb{E}[\langle w^t - w^*, \nabla f(w^t) \rangle \mid \mathcal{F}_t]
\end{aligned}$$

$$\begin{aligned}&\leq \mathbb{E}[\|w^t - w^*\|_2^2 \mid \mathcal{F}_t] + \eta^2 \cdot (1+\alpha) \cdot \mathbb{E}[\|\nabla f(w^t)\|_2^2 \mid \mathcal{F}_t] + 2\eta \cdot \mathbb{E}[\langle w^* - w^t, \nabla f(w^t)\rangle \mid \mathcal{F}_t] \\ &\leq \mathbb{E}[\|w^t - w^*\|_2^2 \mid \mathcal{F}_t] + \eta^2 \cdot (1+\alpha) \cdot \mathbb{E}[\|\nabla f(w^t)\|_2^2 \mid \mathcal{F}_t] + 2\eta \cdot \mathbb{E}[f(w^*) - f(w^t) \mid \mathcal{F}_t] \qquad (14)\end{aligned}$$

where the second step follows from the variance of $\mathsf{desk}_t \circ \mathsf{sk}_t$, and the last step follows from the convexity of f. Taking the expectation over $\mathcal{F}_t$ and re-organizing the above equation, we can get

$$\begin{aligned}2\eta \cdot \mathbb{E}[f(w^t) - f(w^*)] &\leq \mathbb{E}[\|w^t - w^*\|_2^2] - \mathbb{E}[\|w^{t+1} - w^*\|_2^2] + \eta^2 \cdot (1+\alpha) \cdot \mathbb{E}[\|\nabla f(w^t)\|_2^2] \\ &\leq \mathbb{E}[\|w^t - w^*\|_2^2] - \mathbb{E}[\|w^{t+1} - w^*\|_2^2] + \eta^2 \cdot (1+\alpha) \cdot 2L \cdot \mathbb{E}[f(w^t) - f(w^*)]\end{aligned}$$

where the second step follows from the convexity and L-smoothness of f. Rearrange the above inequality, we have

$$(2\eta - \eta^2 \cdot (1+\alpha) \cdot 2L) \cdot \mathbb{E}[f(w^t) - f(w^*)] \leq \mathbb{E}[\|w^t - w^*\|_2^2] - \mathbb{E}[\|w^{t+1} - w^*\|_2^2]$$

Note $\eta \leq \frac{1}{2(1+\alpha)L}$, we have

$$\eta \cdot \mathbb{E}[f(w^t) - f(w^*)] \leq \mathbb{E}[\|w^t - w^*\|_2^2] - \mathbb{E}[\|w^{t+1} - w^*\|_2^2]$$

Sum over all T iterations, we arrive at

$$\eta \cdot \sum_{t=0}^{T} \mathbb{E}[f(w^t) - f(w^*)] \leq \mathbb{E}[\|w^0 - w^*\|_2^2] - \mathbb{E}[\|w^{T+1} - w^*\|_2^2] \leq \mathbb{E}[\|w^0 - w^*\|_2^2] \qquad (15)$$

Let $\overline{w}^T = \frac{1}{T+1}\sum_{t=0}^{T} w^t$ denote the average of parameters across iterations, then by convexity of f, we conclude:

$$\mathbb{E}[f(\overline{w}^T) - f(w^*)] \leq \frac{\mathbb{E}[\|w^0 - w^*\|_2^2]}{\eta \cdot (T+1)}$$

□

E.4. Non-convex f Convergence Analysis

Next, we prove a version when f is not even a convex function, due to loss of convexity, we can no longer bound the gap between $\mathbb{E}[f(w^t)]$ and $f(w^*)$, but we can instead bound the minimum (or average) expected gradient.

Theorem E.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function (Def. C.1) and* $\mathsf{sk}/\mathsf{desk}$ *functions satisfying Theorem 4.2, let $w^* \in \mathbb{R}^d$ be the optimal solution to f. Suppose $\eta := \eta_{\mathrm{local}} \cdot \eta_{\mathrm{global}} \leq \frac{1}{(1+\alpha)L}$, then*

$$\min_{t \in [T]} \mathbb{E}[\|\nabla f(w^t)\|_2^2] \leq \frac{2}{\eta(T+1)}(\mathbb{E}[f(w^0)] - f(w^*))$$

Proof. Note that the only place we used strongly-convex assumption in the proof of Theorem E.1 is Eq. (8), so by the same analysis, we can get

$$\mathbb{E}[f(w^{t+1}) - f(w^t) \mid \mathcal{F}_t] \leq -\frac{\eta}{2} \cdot \|\nabla f(w^t)\|_2^2$$

Rearranging and taking total expectation over $\mathcal{F}_t$, we get

$$\mathbb{E}[\|\nabla f(w^t)\|_2^2] \leq \frac{2}{\eta}(\mathbb{E}[f(w^t)] - \mathbb{E}[f(w^{t+1})])$$

Averaging over all T iterations, we get

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}[\|\nabla f(w^t)\|_2^2] &\leq \frac{2}{\eta(T+1)}\sum_{t=0}^{T}(\mathbb{E}[f(w^t)] - \mathbb{E}[f(w^{t+1})]) \\ &= \frac{2}{\eta(T+1)}(\mathbb{E}[f(w^0)] - \mathbb{E}[f(w^T)])\end{aligned}$$

$$\leq \frac{2}{\eta(T+1)}(\mathbb{E}[f(w^0)] - f(w^*))$$

This implies our final result:

$$\min_{t \in [T]} \mathbb{E}[\|\nabla f(w^t)\|_2^2] \leq \frac{2}{\eta(T+1)}(\mathbb{E}[f(w^0)] - f(w^*))$$

□

Remark E.4. *Notice due to the structure of* $\mathsf{sk}/\mathsf{desk}$ *functions, i.e., their variance is bounded in terms of true gradient, the convergence rate depends completely on the term* $\frac{1}{(1+\alpha)L}$*. If it's a constant, then we essentially recover a convergence rate of gradient descent. On the other hand, if* $\frac{1}{(1+\alpha)L} \leq \frac{1}{\sqrt{T}}$*, then we get a similar convergence rate as SGD. One clear advantage of our* $\mathsf{sk}/\mathsf{desk}$ *functions is they don't introduce extra noise term as in SGD, since we can choose appropriate step size to absorb the variance term.*

F. k-step Convex & Strongly-convex f_c Analysis

F.1. Preliminary

In this section, we assume each f_c satisfies Assumption 3.1 and $\eta_{\mathrm{global}} = 1$. For notation simplicity, we also denote $u_c^{t,-1} = u_c^{t-1,K-1}$ for $t \geq 2$.

Definition F.1. *Let* $(t,k) \in \{1, \cdots, T+1\} \times \{-1, 0, 1, \cdots, K-1\}$*, we define the following terms for iteration* (t,k)*:*

$$\overline{u}^{t,k} := \frac{1}{N}\sum_{c=1}^{N} u_c^{t,k}, \quad r^{t,k} := \overline{u}^{t,k} - w^*$$

to be the average of local parameters and its distance to the optimal solution,

$$g_c^{t,k} := \nabla f_c(u_c^{t,k}), \quad \overline{g}^{t,k} := \frac{1}{N}\sum_{c=1}^{N} \nabla f_c(u_c^{t,k})$$

to be the local gradient and its average,

$$V^{t,k} := \frac{1}{N}\sum_{c=1}^{N} \|u_c^{t,k} - \overline{u}^{t,k}\|_2^2$$

to be the variances of local updates,

$$\sigma^2 = \frac{1}{N}\sum_{c=1}^{N} \|\nabla f_c(w^*)\|^2$$

to be a finite constant that characterize the heterogeneity of local objectives.

We also define the following indicator function: let $l \in \mathbb{R}$*, then we define* $1_{\{x=l\}}$ *to be*

$$1_{\{x=l\}} = \begin{cases} 1 & \textit{if } x = l, \\ 0 & \textit{otherwise}. \end{cases}$$

F.2. Unifying the Update Rule of Algorithm 1

Lemma F.2. *We have the following facts for* $u_c^{t,k}$ *and* $\widetilde{u}^{t,k}$*:*

$$u_c^{t,0} = \overline{u}^{t,0}$$
$$u_c^{t,k} = u_c^{t,k-1} - \eta_{\mathrm{local}} \cdot g_c^{t,k-1}, \ \forall k \geq 1$$
$$\overline{u}^{t,k} = \overline{u}^{t,k-1} - \eta_{\mathrm{local}} \cdot \overline{g}^{t,k-1} + 1_{\{k=0\}} \cdot \eta_{\mathrm{local}} \cdot (I_d - \mathsf{desk}_t \circ \mathsf{sk}_t)(\sum_{i=0}^{K-1} \overline{g}^{t-1,i}), \ \forall (t,k) \neq (1,0)$$

where $I_d : \mathbb{R}^d \to \mathbb{R}^d$ *is the identity function.*

Proof. First two equation directly follows from the update rule of Algorithm 1.
For $k = 1, 2, \cdots, K-1$, taking the average of the second equation we obtain:

$$\overline{u}^{t,k} = \overline{u}^{t,k-1} - \eta_{\mathrm{local}} \cdot \overline{g}^{t,k-1}$$

For $k = 0$ and $t \geq 2$, we have

$$\begin{aligned}
\overline{u}^{t,0} &= \overline{u}^{t-1,0} - \eta_{\mathrm{local}} \cdot \mathsf{desk}_t(\mathsf{sk}_t(\sum_{i=0}^{K-1} \overline{g}^{t-1,i})) \\
&= \overline{u}^{t-1,0} - \eta_{\mathrm{local}} \sum_{i=0}^{K-1} \overline{g}^{t-1,i} + \eta_{\mathrm{local}} \sum_{i=0}^{K-1} \overline{g}^{t-1,i} - \eta_{\mathrm{local}} \cdot \mathsf{desk}_t(\mathsf{sk}_t(\sum_{i=0}^{K-1} \overline{g}^{t-1,i})) \\
&= \overline{u}^{t-1,K-1} - \eta_{\mathrm{local}} \cdot \overline{g}^{t-1,K-1} + \eta_{\mathrm{local}} \cdot (I_d - \mathsf{desk}_t \circ \mathsf{sk}_t)(\sum_{i=0}^{K-1} \overline{g}^{t-1,i})
\end{aligned}$$

Combining above results together, we prove the third equation. □

F.3. Upper Bounding $\|\overline{g}^{t,k}\|_2^2$

Lemma F.3. *Suppose for any $c \in [N]$, $f_c : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth. Then*

$$\|\overline{g}^{t,k}\|_2^2 \leq \ 2L^2 V^{t,k} + 4L(f(\overline{u}^{t,k}) - f(w^*))$$

Proof. By triangle inequality and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\|\overline{g}^{t,k}\|_2^2 &= \|\overline{g}^{t,k} - \nabla f(\overline{u}^{t,k}) + \nabla f(\overline{u}^{t,k})\|_2^2 \\
&\leq 2\|g^{t,k} - \nabla f(\overline{u}^{t,k})\|_2^2 + 2\|\nabla f(\overline{u}^{t,k})\|_2^2
\end{aligned}$$

where the first term can be bounded as

$$\begin{aligned}
\|\overline{g}^{t,k} - \nabla f(\overline{u}^{t,k})\|_2^2 &= \|\frac{1}{N}\sum_{c=1}^{N} \nabla f_c(u_c^{t,k}) - \frac{1}{N}\sum_{c=1}^{N} \nabla f_c(\overline{u}^{t,k})\|_2^2 \\
&\leq \frac{1}{N}\sum_{c=1}^{N} \|\nabla f_c(u_c^{t,k}) - f_c(\overline{u}^{t,k})\|_2^2 \\
&\leq \frac{L^2}{N}\sum_{c=1}^{N} \|u_c^{t,k} - \overline{u}^{t,k}\|_2^2
\end{aligned}$$

and the second term can be bounded as follows:

$$\begin{aligned}
\|\nabla f(\overline{u}^{t,k})\|_2^2 &= \|\nabla f(\overline{u}^{t,k}) - \nabla f(w^*)\|_2^2 \\
&\leq 2L(f(\overline{u}^{t,k}) - f(w^*))
\end{aligned}$$

where the last step follows from that f is L-smooth and Fact C.4.

Combining bounds on these two terms, we get

$$\begin{aligned}
\|\overline{g}^{t,k}\|_2^2 &\leq \frac{2L^2}{N}\sum_{c=1}^{N} \|u_c^{t,k} - \overline{u}^{t,k}\|_2^2 + 2L^2\|\overline{u}^{t,k} - w^*\|_2^2 \\
&\leq 2L^2 V^{t,k} + 4L(f(\overline{u}^{t,k}) - f(w^*))
\end{aligned}$$

□

F.4. Lower Bounding $\langle \overline{u}^{t,k} - w^*, \overline{g}^{t,k} \rangle$

Lemma F.4. *Suppose each f_c satisfies Assumption 3.1 with $\mu \geq 0$, then*

$$\langle \overline{u}^{t,k} - w^*, \overline{g}^{t,k} \rangle \geq f(\overline{u}^{t,k}) - f(w^*) - \frac{L}{2} V^{t,k} + \frac{\mu}{2} \| \overline{u}^{t,k} - w^* \|_2^2$$

Proof. We will provide a lower bound on this inner product:

$$\langle \overline{u}^{t,k} - w^*, \overline{g}^{t,k} \rangle = \frac{1}{N} \sum_{c=1}^{N} \langle \overline{u}^{t,k} - w^*, \nabla f_c(u_c^{t,k}) \rangle$$

It suffices to consider each term separately:

$$\begin{aligned} \langle \overline{u}^{t,k} - w^*, \nabla f_c(u_c^{t,k}) \rangle &= \langle \overline{u}^{t,k} - u_c^{t,k} + u_c^{t,k} - w^*, \nabla f_c(u_c^{t,k}) \rangle \\ &= \langle \overline{u}^{t,k} - u_c^{t,k}, \nabla f_c(u_c^{t,k}) \rangle + \langle u_c^{t,k} - w^*, \nabla f_c(u_c^{t,k}) \rangle \end{aligned}$$

The first term can be lower bounded via L-smoothness:

$$\langle \overline{u}^{t,k} - u_c^{t,k}, \nabla f_c(u_c^{t,k}) \rangle \geq f_c(\overline{u}^{t,k}) - f_c(u_c^{t,k}) - \frac{L}{2} \| \overline{u}^{t,k} - u_c^{t,k} \|_2^2$$

The second term can be lower bounded via convexity:

$$\langle u_c^{t,k} - w^*, \nabla f_c(u_c^{t,k}) \rangle \geq f_c(u_c^{t,k}) - f_c(w^*) + \frac{\mu}{2} \| u_c^{t,k} - w^* \|_2^2$$

Combining these two bounds and average them, we get a lower bound:

$$\begin{aligned} \langle \overline{u}^{t,k} - w^*, g^{t,k} \rangle &\geq \frac{1}{N} \sum_{c=1}^{N} (f_c(\overline{u}^{t,k}) - f_c(w^*) - \frac{L}{2} \| \overline{u}^{t,k} - u_c^{t,k} \|_2^2 + \frac{\mu}{2} \| u_c^{t,k} - w^* \|_2^2) \\ &\geq \frac{1}{N} \sum_{c=1}^{N} (f_c(\overline{u}^{t,k}) - f_c(w^*)) - \frac{L}{2} V^{t,k} + \frac{\mu}{2} \| \overline{u}^{t,k} - w^* \|_2^2 \\ &= f(\overline{u}^{t,k}) - f(w^*) - \frac{L}{2} V^{t,k} + \frac{\mu}{2} \| \overline{u}^{t,k} - w^* \|_2^2 \end{aligned}$$

□

F.5. Upper Bounding Variance within K Local Steps

Lemma F.5. *Suppose each f_c is convex and L-smooth. Assume $\eta_{\mathrm{local}} \leq \frac{1}{8LK}$. Then for any $t \geq 0$,*

$$\sum_{k=0}^{K-1} V^{t,k} \leq 8 \eta_{\mathrm{local}}^2 L K^2 \sum_{k=0}^{K-1} (f(\overline{u}^{t,k}) - f(w^*)) + 4 \eta_{\mathrm{local}}^2 K^3 \sigma^2$$

Proof. By Lemma F.2, we know $V^{t,0} = 0$ for any $t \geq 0$. Consider $k \in \{1, 2, \cdots, K-1\}$, we have

$$\begin{aligned} V^{t,k} &= \frac{1}{N} \sum_{c=1}^{N} \| u_c^{t,k} - \overline{u}^{t,k} \|_2^2 \\ &= \frac{1}{N} \sum_{c=1}^{N} \| u_c^{t,0} - \sum_{i=0}^{k-1} \eta_{\mathrm{local}} \cdot g_c^{t,i} - \overline{u}^{t,0} + \sum_{i=0}^{k-1} \eta_{\mathrm{local}} \cdot \overline{g}^{t,i} \|_2^2 \\ &= \frac{\eta_{\mathrm{local}}^2}{N} \sum_{c=1}^{N} \| \sum_{i=0}^{k-1} (\overline{g}^{t,i} - g_c^{t,i}) \|_2^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\eta_{\text{local}}^2 k}{N} \sum_{c=1}^{N} \sum_{i=0}^{k-1} \|\overline{g}^{t,i} - g_c^{t,i}\|_2^2 \\
&\leq \frac{\eta_{\text{local}}^2 K}{N} \sum_{c=1}^{N} \sum_{i=0}^{k-1} \|g_c^{t,i}\|_2^2 \qquad (16)
\end{aligned}
$$

where the second step follows from Lemma F.2, the last step follows from $\overline{g}^{t,i}$ being the average of $g_c^{t,i}$. By Cauchy-Schwartz inequality, we further have:

$$
\begin{aligned}
\|g_c^{t,i}\|_2^2 &\leq 3\|g_c^{t,i} - \nabla f_c(\overline{u}^{t,i})\|_2^2 + 3\|\nabla f_c(\overline{u}^{t,i}) - \nabla f_c(w^*)\|_2^2 + 3\|\nabla f_c(w^*)\|_2^2 \\
&\leq 3L^2\|u_c^{t,i} - \overline{u}^{t,i}\|_2^2 + 6L(f_c(\overline{u}^{t,i}) - f_c(w^*) + \langle w^* - \overline{u}^{t,0}, \nabla f_c(w^*)\rangle) + 3\|\nabla f_c(w^*)\|_2^2.
\end{aligned}
$$

where the last step follows from applying L-smoothness to the first and second term.

Averaging with respect to c,

$$
\frac{1}{N}\sum_{c=1}^{N} \|g_c^{t,i}\|_2^2 \leq 3L^2 V^{t,i} + 6L(f(\overline{u}^{t,i}) - f(w^*)) + 3\sigma^2.
$$

Note that the inner product term vanishes since $\frac{1}{N}\sum_{c=1}^{N} \nabla f_c(w^*) = \nabla f(w^*) = 0$.

Plugging back to Eq. (16), we obtain

$$
\begin{aligned}
V^{t,k} &\leq \frac{\eta_{\text{local}}^2 K}{N} \sum_{c=1}^{N} \sum_{i=0}^{k-1} \|g_c^{t,i}\|_2^2 \\
&\leq \eta_{\text{local}}^2 K \sum_{i=0}^{k-1} (3L^2 V^{t,i} + 6L(f(\overline{u}^{t,i}) - f(w^*)) + 3\sigma^2).
\end{aligned}
$$

Summing up above inequality as k varies from 0 to $K-1$,

$$
\begin{aligned}
\sum_{k=0}^{K-1} V^{t,k} &\leq \eta_{\text{local}}^2 K \sum_{k=0}^{K-1} \sum_{i=0}^{k-1} (3L^2 V^{t,i} + 6L(f(\overline{u}^{t,i}) - f(w^*)) + 3\sigma^2) \\
&\leq \eta_{\text{local}}^2 K \sum_{k=0}^{K-1} \sum_{i=0}^{K-1} (3L^2 V^{t,i} + 6L(f(\overline{u}^{t,i}) - f(w^*)) + 3\sigma^2) \\
&= 3\eta_{\text{local}}^2 L^2 K^2 \sum_{i=0}^{K-1} V^{t,i} + 6\eta_{\text{local}}^2 L K^2 \sum_{i=0}^{K-1} (f(\overline{u}^{t,i}) - f(w^*)) + 3\eta_{\text{local}}^2 K^3 \sigma^2
\end{aligned}
$$

Rearranging terms we obtain:

$$
(1 - 3\eta_{\text{local}}^2 L^2 K^2) \sum_{k=0}^{K-1} V^{t,k} \leq 6\eta_{\text{local}}^2 L K^2 \sum_{i=0}^{K-1} (f(\overline{u}^{t,i}) - f(w^*)) + 3\eta_{\text{local}}^2 K^3 \sigma^2
$$

Since $\eta_{\text{local}} \leq \frac{1}{8LK}$, we have $1 - 3\eta_{\text{local}}^2 L^2 K^2 \geq \frac{3}{4}$, implying

$$
\sum_{k=0}^{K-1} V^{t,k} \leq 8\eta_{\text{local}}^2 L K^2 \sum_{i=0}^{K-1} (f(\overline{u}^{t,i}) - f(w^*)) + 4\eta_{\text{local}}^2 K^3 \sigma^2
$$

□

F.6. Bounding the Expected Gap Between $\overline{u}^{t,k}$ and w^*

Lemma F.6. *Suppose each f_c satisfies Assumption 3.1 with $\mu \geq 0$. If* $\mathsf{sk}/\mathsf{desk}$ *satisfying Theorem 4.2 and $\eta_{\mathrm{local}} \leq \frac{1}{4L}$, then for any $(t,k) \neq (1,0)$, we have*

$$\begin{aligned}\mathbb{E}[\|\overline{u}^{t,k} - w^*\|_2^2] \leq (1-\mu\eta_{\mathrm{local}})\,\mathbb{E}[\|\overline{u}^{t,k-1} - w^*\|_2^2] + \frac{3}{2}\eta_{\mathrm{local}} L\,\mathbb{E}[V^{t,k-1}] - \eta_{\mathrm{local}}\,\mathbb{E}[f(\overline{u}^{t,k-1}) - f(w^*)] \\ + 1_{\{k=0\}}\eta_{\mathrm{local}}^2 \alpha K\Big(2L^2\sum_{i=0}^{K-1}\mathbb{E}[V^{t-1,i}] + 4L\sum_{i=0}^{K-1}\mathbb{E}[f(\overline{u}^{t-1,i}) - f(w^*)]\Big)\end{aligned}$$

Proof. By Lemma F.2, we have for any $(t,k) \neq (1,0)$,

$$\overline{u}^{t,k} = \overline{u}^{t,k-1} - \eta_{\mathrm{local}} \cdot \overline{g}^{t,k-1} + 1_{\{k=0\}} \cdot \eta_{\mathrm{local}} \cdot (I_d - \mathsf{desk}_t \circ \mathsf{sk}_t)(\sum_{i=0}^{K-1} \overline{g}^{t-1,i})$$

Therefore, denoting $h^t := (I_d - \mathsf{desk}_t \circ \mathsf{sk}_t)(\sum_{i=0}^{K-1} \overline{g}^{t-1,i})$, we have

$$\begin{aligned}\|\overline{u}^{t,k} - w^*\|_2^2 &= \|\overline{u}^{t,k-1} - w^* - \eta_{\mathrm{local}} \cdot \overline{g}^{t,k-1} + 1_{\{k=0\}}\eta_{\mathrm{local}} \cdot h^t\|_2^2 \\ &= \|\overline{u}^{t,k-1} - w^*\|_2^2 + \eta_{\mathrm{local}}^2 \cdot \|\overline{g}^{t,k-1}\|_2^2 - 2\eta_{\mathrm{local}}\langle \overline{u}^{t,k-1} - w^*, \overline{g}^{t,k-1}\rangle \\ &\quad + 2\eta_{\mathrm{local}} 1_{\{k=0\}}\langle \overline{u}^{t,k-1} - w^*, h^t\rangle - 2\eta_{\mathrm{local}}^2 1_{\{k=0\}}\langle \overline{g}^{t,k-1}, h^t\rangle \\ &\quad + \eta_{\mathrm{local}}^2 1_{\{k=0\}} \cdot \|h^t\|_2^2 \end{aligned} \tag{17}$$

Note by Theorem 4.2, we have:

$$\mathbb{E}[\mathsf{desk}_t(\mathsf{sk}_t(h))] = h, \qquad \mathbb{E}[\|\mathsf{desk}_t(\mathsf{sk}_t(h))\|_2^2] \leq (1+\alpha) \cdot \|h\|_2^2$$

hold for any vector h. Hence, by taking expectation over Eq. (17),

$$\begin{aligned}\mathbb{E}[\|\overline{u}^{t,k} - w^*\|_2^2 | \mathcal{F}_t] =\ & \mathbb{E}[\|\overline{u}^{t,k-1} - w^*\|_2^2 | \mathcal{F}_t] + \eta_{\mathrm{local}}^2 \cdot \mathbb{E}[\|\overline{g}^{t,k-1}\|_2^2 | \mathcal{F}_t] \\ & - 2\eta_{\mathrm{local}}\,\mathbb{E}[\langle \overline{u}^{t,k-1} - w^*, \overline{g}^{t,k-1}\rangle | \mathcal{F}_t] + 1_{\{k=0\}} \cdot \eta_{\mathrm{local}}^2 \cdot \mathbb{E}[\|h^t\|_2^2 | \mathcal{F}_t]\end{aligned}$$

Note that since $\mathbb{E}[h^t \mid \mathcal{F}_t] = 0$, so the two inner products involving h^t vanishes.

Since

$$\begin{aligned}\mathbb{E}[\|h^t\|_2^2 | \mathcal{F}_t] =\ & \mathbb{E}[\|(I_d - \mathsf{desk}_t \circ \mathsf{sk}_t)(\sum_{i=0}^{K-1} \overline{g}^{t-1,i})\|_2^2 | \mathcal{F}_t] \\ \leq\ & \alpha\,\mathbb{E}[\|\sum_{i=0}^{K-1} \overline{g}^{t-1,i}\|_2^2 | \mathcal{F}_t] \\ \leq\ & \alpha K \sum_{i=0}^{K-1} \mathbb{E}[\|\overline{g}^{t-1,i}\|_2^2 | \mathcal{F}_t]\end{aligned}$$

Taking total expectation, we have

$$\begin{aligned}&\mathbb{E}[\|\overline{u}^{t,k} - w^*\|_2^2] \\ \leq\ & \mathbb{E}[\|\overline{u}^{t,k-1} - w^*\|_2^2] + \eta_{\mathrm{local}}^2 \cdot \mathbb{E}[\|\overline{g}^{t,k-1}\|_2^2] - 2\eta_{\mathrm{local}}\,\mathbb{E}[\langle \overline{u}^{t,k-1} - w^*, \overline{g}^{t,k-1}\rangle] \\ & + 1_{\{k=0\}} \cdot \eta_{\mathrm{local}}^2 \cdot \alpha K \sum_{i=0}^{K-1} \mathbb{E}[\|\overline{g}^{t-1,i}\|_2^2] \\ \leq\ & \mathbb{E}[\|\overline{u}^{t,k-1} - w^*\|_2^2] + \eta_{\mathrm{local}}^2 \cdot \mathbb{E}[2L^2 V^{t,k-1} + 4L(f(\overline{u}^{t,k-1}) - f(w^*))] \\ & - 2\eta_{\mathrm{local}}\,\mathbb{E}[f(\overline{u}^{t,k-1}) - f(w^*) - \frac{L}{2}V^{t,k-1} + \frac{\mu}{2}\|\overline{u}^{t,k-1} - w^*\|_2^2]\end{aligned}$$

$$
\begin{aligned}
&+ 1_{\{k=0\}} \cdot \eta_{\mathrm{local}}^2 \cdot \alpha K \sum_{i=0}^{K-1} \mathbb{E}[2L^2 V^{t-1,i} + 4L(f(\overline{u}^{t-1,i}) - f(w^*))] \\
\le\ & (1-\mu\eta_{\mathrm{local}})\,\mathbb{E}[\|\overline{u}^{t,k-1} - w^*\|_2^2] + \eta_{\mathrm{local}} \cdot L \cdot (1+2\eta_{\mathrm{local}} L) \cdot \mathbb{E}[V^{t,k-1}] \\
&- 2\eta_{\mathrm{local}} \cdot (1-2\eta_{\mathrm{local}} L) \cdot \mathbb{E}[f(\overline{u}^{t,k-1}) - f(w^*)] \\
&+ 1_{\{k=0\}} \cdot \eta_{\mathrm{local}}^2 \cdot \alpha K \cdot \Big(2L^2 \sum_{i=0}^{K-1} \mathbb{E}[V^{t-1,i}] + 4L \sum_{i=0}^{K-1} \mathbb{E}[f(\overline{u}^{t-1,i}) - f(w^*)]\Big)
\end{aligned}
$$

where the second step follows from Lemma F.3 and Lemma F.4. Since $\eta_{\mathrm{local}} \le \frac{1}{4L}$, we have

$$
\begin{aligned}
\mathbb{E}[\|\overline{u}^{t,k} - w^*\|_2^2] \le\ & (1-\mu\eta_{\mathrm{local}})\,\mathbb{E}[\|\overline{u}^{t,k-1} - w^*\|_2^2] + \frac{3}{2}\eta_{\mathrm{local}} L\,\mathbb{E}[V^{t,k-1}] - \eta_{\mathrm{local}}\,\mathbb{E}[f(\overline{u}^{t,k-1}) - f(w^*)] \\
&+ 1_{\{k=0\}}\eta_{\mathrm{local}}^2 \alpha K\Big(2L^2 \sum_{i=0}^{K-1} \mathbb{E}[V^{t-1,i}] + 4L \sum_{i=0}^{K-1} \mathbb{E}[f(\overline{u}^{t-1,i}) - f(w^*)]\Big)
\end{aligned}
$$

□

F.7. Main Result: Convex Case

Theorem F.7 (Formal version of Theorem 5.3)**.** *Assume each f_c is convex and L-smooth. If Theorem 4.2 holds and $\eta_{\mathrm{local}} \le \frac{1}{8(1+\alpha)LK}$,*

$$
\mathbb{E}[f(\overline{w}^T) - f(w^*)] \le \frac{4\,\mathbb{E}[\|w^0 - w^*\|_2^2]}{\eta_{\mathrm{local}} KT} + 32\eta_{\mathrm{local}}^2 L K^2 \sigma^2,
$$

where $\overline{w}^T = \frac{1}{KT}(\sum_{t=1}^T \sum_{k=0}^{K-1} \overline{u}^{t,k})$ is the average over parameters throughout the execution of Algorithm 1.

Proof. Summing up Lemma F.6 as t varies from 1 to T and k varies from 0 to $K-1$,

$$
\begin{aligned}
&\mathbb{E}[\|\overline{u}^{T+1,0} - w^*\|_2^2] - \mathbb{E}[\|w^0 - w^*\|_2^2] \\
\le\ & \frac{3}{2}\eta_{\mathrm{local}} L \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[V^{t,k}] - \eta_{\mathrm{local}} \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)] \\
&+ \sum_{t=1}^T \sum_{k=0}^{K-1} 1_{\{k=0\}} \eta_{\mathrm{local}}^2 \alpha K\Big(2L^2 \sum_{i=0}^{K-1} \mathbb{E}[V^{t,i}] + 4L \sum_{i=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,i}) - f(w^*)]\Big) \\
=\ & \frac{3}{2}\eta_{\mathrm{local}} L \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[V^{t,k}] - \eta_{\mathrm{local}} \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)] \\
&+ \eta_{\mathrm{local}}^2 \alpha K\Big(2L^2 \sum_{t=1}^T \sum_{i=0}^{K-1} \mathbb{E}[V^{t,i}] + 4L \sum_{t=1}^T \sum_{i=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,i}) - f(w^*)]\Big) \\
=\ & \eta_{\mathrm{local}} L(\frac{3}{2} + 2\eta_{\mathrm{local}} \alpha L K) \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[V^{t,k}] \\
&- \eta_{\mathrm{local}}(1 - 4\eta_{\mathrm{local}} \alpha L K) \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)] \\
\le\ & 2\eta_{\mathrm{local}} L \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[V^{t,k}] - \frac{1}{2}\eta_{\mathrm{local}} \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)] \\
\le\ & 2\eta_{\mathrm{local}} L \sum_{t=1}^T (8\eta_{\mathrm{local}}^2 L K^2 \sum_{i=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,i}) - f(w^*)] + 4\eta_{\mathrm{local}}^2 K^3 \sigma^2) - \frac{1}{2}\eta_{\mathrm{local}} \sum_{t=1}^T \sum_{k=0}^{K-1} \mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)]
\end{aligned}
$$

$$\leq -\frac{1}{4}\eta_{\rm local}\sum_{t=1}^{T}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k})-f(w^*)]+8\eta_{\rm local}^3 LK^3T\sigma^2$$

where the fourth step follows from $\eta_{\rm local}\leq\frac{1}{8\alpha LK}$, the last step follows from $\eta_{\rm local}\leq\frac{1}{8LK}$. Rearranging the terms, we obtain

$$\frac{1}{KT}\sum_{t=1}^{T}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k})-f(w^*)]\leq\frac{4\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\eta_{\rm local}KT}+32\eta_{\rm local}^2 LK^2\sigma^2$$

Finally, by the convexity of f we complete the proof. □

Now we are ready to answer the question: how much communication cost is sufficient to guarantee $\mathbb{E}[f(\overline{w}^T)-f(w^*)]\leq\epsilon$? we have the following communication cost result:

Theorem F.8 (Formal version of Theorem 5.4)**.** *Assume each f_c is convex and L-smooth. If Theorem 4.2 holds. With $O\left(\mathbb{E}[\|w^0-w^*\|_2^2]N\max\{\frac{Ld}{\epsilon},\frac{\sigma\sqrt{L}}{\epsilon^{3/2}}\}\right)$ bits of communication cost, Algorithm 1 outputs an ϵ-optimal solution $\overline{w}^T$ satisfying:*

$$\mathbb{E}[f(\overline{w}^T)-f(w^*)]\leq\epsilon,$$

where $\overline{w}^T=\frac{1}{KT}(\sum_{t=1}^{T}\sum_{k=0}^{K-1}\overline{u}^{t,k})$.

Proof. To calculate the communication complexity, we first note communication only happens in sync steps. Specifically, in each sync step, the algorithm requires $O(Nb_{\rm sketch})$ bits of communication cost, where $b_{\rm sketch}$ denotes the sketching dimension. Therefore, the total cost of communication is given by $O(Nb_{\rm sketch}T)$. To obtain the optimal communication cost for ϵ-optimal solution, we choose $T,K,\eta_{\rm local}$ and $b_{\rm sketch}$ by solving the following optimization problem:

$$\begin{aligned}\min_{T,K,\eta_{\rm local},b_{\rm sketch},\alpha}\quad & Nb_{\rm sketch}T\\ \text{s.t.}\quad & 0<\eta_{\rm local}\leq\frac{1}{8(1+\alpha)LK}\\ & \frac{4\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\eta_{\rm local}KT}\leq\frac{\epsilon}{2}\\ & 32\eta_{\rm local}^2LK^2\sigma^2\leq\frac{\epsilon}{2}\\ & d\geq b_{\rm sketch}=O(\frac{d}{\alpha})\geq 1\end{aligned}$$

where d is the parameter dimension and the last constraint is due to Theorem 4.2. Above constraints imply:

$$T\geq\frac{8\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\eta_{\rm local}K\epsilon},\quad ,K\eta_{\rm local}\leq\min\{\frac{1}{8(1+\alpha)L},\frac{1}{8\sigma}\sqrt{\frac{\epsilon}{L}}\}$$

Therefore, when $\epsilon\geq\frac{\sigma^2}{(1+\alpha)^2L}$, the optimal solution is given by

$$K\eta_{\rm local}=\frac{1}{8(1+\alpha)L},\ T=\frac{64\,\mathbb{E}[\|w^0-w^*\|_2^2](1+\alpha)L}{\epsilon},\ b_{\rm sketch}=O(\frac{d}{\alpha})$$

and the corresponding optimal communication cost is $O(\frac{\mathbb{E}[\|w^0-w^*\|_2^2LNd}{\epsilon})$.

when $\epsilon<\frac{\sigma^2}{(1+\alpha)^2L}$, the optimal solution is given by

$$K\eta_{\rm local}=\frac{1}{8\sigma}\sqrt{\frac{\epsilon}{L}},\ T=\frac{64\,\mathbb{E}[\|w^0-w^*\|_2^2\sigma\sqrt{L}]}{\epsilon^{3/2}},\ b_{\rm sketch}=O(\frac{d}{\alpha})$$

and the corresponding optimal communication cost is $O(\frac{\mathbb{E}[\|w^0-w^*\|_2^2\sigma\sqrt{L}Nd}{\alpha\epsilon^{3/2}})$.

Combining above two cases, the optimal α is given by $O(d)$, and the corresponding optimal communication cost will be $O(\mathbb{E}[\|w^0-w^*\|_2^2]N\max\{\frac{Ld}{\epsilon},\frac{\sigma\sqrt{L}}{\epsilon^{3/2}}\})$. □

F.8. Main Result: Strongly-convex Case

Theorem F.9 (Formal version of Theorem 5.1)**.** *Assume each f_c is μ-strongly convex and L-smooth. If Theorem 4.2 holds and $\eta_{\mathrm{local}} \leq \frac{1}{8(1+\alpha)LK}$,*

$$\mathbb{E}[f(w^{T+1}) - f(w^*)] \leq \frac{L}{2}\,\mathbb{E}[\|w^0 - w^*\|_2^2] e^{-\mu\eta_{\mathrm{local}}T} + 4\eta_{\mathrm{local}}^2 L^2 K^3 \sigma^2/\mu.$$

Proof. Summing up Lemma F.6 as k varies from 0 to $K-1$, then we have for any $t \geq 1$,

$$\begin{aligned}
&(\mathbb{E}[\|\overline{u}^{t+1,0} - w^*\|_2^2] + \sum_{k=1}^{K-1}\mathbb{E}[\|\overline{u}^{t,k} - w^*\|_2^2]) - (1-\mu\eta_{\mathrm{local}})\sum_{k=0}^{K-1}\mathbb{E}[\|\overline{u}^{t,k} - w^*\|_2^2])\\
\leq{}& \frac{3}{2}\eta_{\mathrm{local}} L \sum_{k=0}^{K-1}\mathbb{E}[V^{t,k}] - \eta_{\mathrm{local}}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)]\\
&+ \sum_{k=0}^{K-1} 1_{\{k=0\}}\eta_{\mathrm{local}}^2 \alpha K\Big(2L^2\sum_{i=0}^{K-1}\mathbb{E}[V^{t,i}] + 4L\sum_{i=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,i}) - f(w^*)]\Big)\\
={}& \frac{3}{2}\eta_{\mathrm{local}} L \sum_{k=0}^{K-1}\mathbb{E}[V^{t,k}] - \eta_{\mathrm{local}}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)]\\
&+ \eta_{\mathrm{local}}^2 \alpha K\Big(2L^2\sum_{i=0}^{K-1}\mathbb{E}[V^{t,i}] + 4L\sum_{i=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,i}) - f(w^*)]\Big)\\
={}& \eta_{\mathrm{local}} L(\frac{3}{2} + 2\eta_{\mathrm{local}}\alpha LK)\sum_{k=0}^{K-1}\mathbb{E}[V^{t,k}] - \eta_{\mathrm{local}}(1 - 4\eta_{\mathrm{local}}\alpha LK)\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)]\\
\leq{}& 2\eta_{\mathrm{local}} L\sum_{k=0}^{K-1}\mathbb{E}[V^{t,k}] - \frac{1}{2}\eta_{\mathrm{local}}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)]\\
\leq{}& 2\eta_{\mathrm{local}} L(8\eta_{\mathrm{local}}^2 LK^2\sum_{i=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,i}) - f(w^*)] + 4\eta_{\mathrm{local}}^2 K^3\sigma^2) - \frac{1}{2}\eta_{\mathrm{local}}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)]\\
\leq{}& -\frac{1}{4}\eta_{\mathrm{local}}\sum_{k=0}^{K-1}\mathbb{E}[f(\overline{u}^{t,k}) - f(w^*)] + 8\eta_{\mathrm{local}}^3 LK^3\sigma^2
\end{aligned}$$

where the fourth step follows from $\eta_{\mathrm{local}} \leq \frac{1}{8\alpha LK}$, the last step follows from $\eta_{\mathrm{local}} \leq \frac{1}{8LK}$. Rearranging the terms, we obtain

$$\mathbb{E}[\|\overline{u}^{t+1,0} - w^*\|_2^2] \leq (1-\mu\eta_{\mathrm{local}})\,\mathbb{E}[\|\overline{u}^{t,0} - w^*\|_2^2] + 8\eta_{\mathrm{local}}^3 LK^3\sigma^2$$

implying

$$\mathbb{E}[\|\overline{u}^{t+1,0} - w^*\|_2^2] - 8\eta_{\mathrm{local}}^2 LK^3\sigma^2/\mu \leq (1-\mu\eta_{\mathrm{local}})(\mathbb{E}[\|\overline{u}^{t,0} - w^*\|_2^2] - 8\eta_{\mathrm{local}}^2 LK^3\sigma^2/\mu).$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\|w^{T+1} - w^*\|_2^2] - 8\eta_{\mathrm{local}}^2 LK^3\sigma^2/\mu &\leq (1-\mu\eta_{\mathrm{local}})^T(\mathbb{E}[\|w^0 - w^*\|_2^2] - 8\eta_{\mathrm{local}}^2 LK^3\sigma^2/\mu)\\
&\leq \mathbb{E}[\|w^0 - w^*\|_2^2] e^{-\mu\eta_{\mathrm{local}}T}
\end{aligned}$$

Finally, by L-smoothness of function f, we obtain

$$\mathbb{E}[f(w^{T+1}) - f(w^*)] \leq \frac{L}{2}\,\mathbb{E}[\|w^{T+1} - w^*\|_2^2] \leq \frac{L}{2}\,\mathbb{E}[\|w^0 - w^*\|_2^2] e^{-\mu\eta_{\mathrm{local}}T} + 4\eta_{\mathrm{local}}^2 L^2 K^3\sigma^2/\mu.$$

□

Theorem F.10 (Formal version of Corollary 5.2)**.** *Assume each f_c is μ-strongly convex and L-smooth. If Theorem 4.2 holds. With $O\left(\frac{LN}{\mu}\max\{d,\sqrt{\frac{\sigma^2}{\mu\epsilon}}\}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon})\right)$ bits of communication cost, Algorithm 1 outputs an ϵ-optimal solution w^T satisfying:*

$$\mathbb{E}[f(w^T)-f(w^*)]\le\epsilon.$$

Proof. To calculate the communication complexity, we first note communication only happens in sync steps. Specifically, in each sync step, the algorithm requires $O(Nb_{\text{sketch}})$ bits of communication cost, where b_{sketch} denotes the sketching dimension. Therefore, the total cost of communication is given by $O(Nb_{\text{sketch}}T)$. To obtain the optimal communication cost for ϵ-optimal solution, we choose $T, K, \eta_{\text{local}}$ and b_{sketch} by solving the following optimization problem:

$$\begin{aligned}\min_{T,K,\eta_{\text{local}},b_{\text{sketch}},\alpha}\quad & Nb_{\text{sketch}}T\\ \text{s.t.}\quad & 0<\eta_{\text{local}}\le\frac{1}{8(1+\alpha)LK}\\ & \frac{L}{2}\,\mathbb{E}[\|w^0-w^*\|_2^2]e^{-\mu\eta_{\text{local}}T}\le\frac{\epsilon}{2}\\ & 4\eta_{\text{local}}^2L^2K^3\sigma^2/\mu\le\frac{\epsilon}{2}\\ & d\ge b_{\text{sketch}}=O(\frac{d}{\alpha})\ge 1\end{aligned}$$

where d is the parameter dimension and the last constraint is due to Theorem 4.2. Above constraints imply:

$$T\ge\frac{1}{\mu\eta_{\text{local}}}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon}),\quad \eta_{\text{local}}\le\min\{\frac{1}{8(1+\alpha)LK},\frac{1}{2LK\sigma}\sqrt{\frac{\mu\epsilon}{2K}}\}$$

Therefore, the optimal value is given when $K=1$. When $\epsilon\ge\frac{\sigma^2}{16(1+\alpha)^2\mu}$, the optimal solution is given by

$$\eta_{\text{local}}=\frac{1}{8(1+\alpha)L},\ T=\frac{8(1+\alpha)L}{\mu}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon}),\ b_{\text{sketch}}=O(\frac{d}{\alpha})$$

and the corresponding optimal communication cost is $O(\frac{LNd}{\mu}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon}))$.

when $\epsilon<\frac{\sigma^2}{16(1+\alpha)^2\mu}$, the optimal solution is given by

$$\eta_{\text{local}}=\frac{1}{2L\sigma}\sqrt{\frac{\mu\epsilon}{2}},\ T=\frac{2L\sigma}{\mu^{3/2}}\sqrt{\frac{2}{\epsilon}}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon}),\ b_{\text{sketch}}=O(\frac{d}{\alpha})$$

and the corresponding optimal communication cost is $O(\frac{\sigma LNd}{\alpha\mu^{3/2}\sqrt{\epsilon}}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon}))$.

Combining above two cases, the optimal α is given by $O(d)$, and the corresponding optimal communication cost will be $O(\frac{LN}{\mu}\max\{d,\sqrt{\frac{\sigma^2}{\mu\epsilon}}\}\log(\frac{L\,\mathbb{E}[\|w^0-w^*\|_2^2]}{\epsilon}))$. □

G. k-step Non-convex f Convergence Analysis

In this section, we present convergence result for non-convex f case in the k-local-step regime. In order for the proof to go through, we assume that for any $c\in[N]$ and any $w\in\mathbb{R}^d$, there exists a universal constant G such that

$$\|\nabla f_c(w)\|_2\le G.$$

Throughout the proof, we will use $\mathcal{F}_t$ to denote the sequence $w_{t-1},w_{t-2},\dots,w_0$. Also, we use η as a shorthand for $\eta_{\text{global}}\cdot\eta_{\text{local}}$.

Note that in k-local-step scheme, the average of local gradients is no longer the true gradient, therefore, we can no longer bound everything using the true gradients. This means it's necessary to introduce the gradient norm upper bound assumption.

Lemma G.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumption 3.1 and $\mathsf{sk}/\mathsf{desk}$ functions satisfying Theorem 4.2. Further, assume $\eta_{\mathrm{local}} \le \frac{1}{2LK}$. Then*

$$\mathbb{E}[f(w^{t+1}) - f(w^t) \mid \mathcal{F}_t] \le \, -\eta_{\mathrm{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta \cdot L \cdot K^2 \cdot G^2 \cdot \Big(\eta_{\mathrm{local}} + \frac{\eta}{2} \cdot (1+\alpha)\Big)$$

Proof. We start by bounding $f(w^{t+1}) - f(w^t)$ without taking conditional expectation:

$$\begin{aligned} & f(w^{t+1}) - f(w^t) \\ \le\ & \langle w^{t+1} - w^t, \nabla f(w^t)\rangle + \frac{L}{2}\|w^{t+1} - w^t\|_2^2 \\ =\ & \langle \mathsf{desk}_t(\Delta\widetilde{w}^t), \nabla f(w^t)\rangle + \frac{L}{2}\|\mathsf{desk}_t(\Delta\widetilde{w}^t)\|_2^2 \\ =\ & A + \frac{L}{2}B \end{aligned}$$

where

$$\begin{aligned} A :=&\ -\langle \eta_{\mathrm{global}} \cdot \mathsf{desk}_t(\frac{1}{N}\sum_{c=1}^N \mathsf{sk}_t(\sum_{k=0}^{K-1} \eta_{\mathrm{local}} \cdot \nabla f_c(u_c^{t,k}))), \nabla f(w^t)\rangle \\ B :=&\ \|\eta_{\mathrm{global}} \cdot \mathsf{desk}_t(\frac{1}{N}\sum_{c=1}^N \mathsf{sk}_t(\sum_{k=1}^{K} \eta_{\mathrm{local}} \cdot \nabla f_c(u_c^{t,k})))\|_2^2 \end{aligned}$$

Bounding $\mathbb{E}[A \mid \mathcal{F}_t]$ Using the fact that $\mathsf{sk}_t/\mathsf{desk}_t$ are linear functions and $\mathbb{E}[\mathsf{desk}_t(\mathsf{sk}_t(h))] = h$, we get

$$\begin{aligned} \mathbb{E}[A \mid \mathcal{F}_t] =&\ -\langle \eta_{\mathrm{global}} \cdot \frac{1}{N}\sum_{c=1}^N \sum_{k=0}^{K-1} \eta_{\mathrm{local}} \cdot \nabla f_c(u_c^{t,k}), \nabla f(w^t)\rangle \\ =&\ -\eta_{\mathrm{global}} \cdot \langle \frac{1}{N}\sum_{c=1}^N \Big(\sum_{k=0}^{K-1} \eta_{\mathrm{local}} \cdot \nabla f_c(u_c^{t,k}) - \nabla f_c(w^t) + \nabla f_c(w^t)\Big), \nabla f(w^t)\rangle \\ =&\ -\eta_{\mathrm{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta_{\mathrm{global}} \cdot \eta_{\mathrm{local}} \cdot \frac{1}{N}\sum_{c=1}^N \sum_{k=0}^{K-1} \langle \nabla f_c(u_c^{t,k}) - \nabla f_c(w^t), \nabla f(w^t)\rangle \end{aligned}$$

It suffices to bound the inner product, notice for $k = 0$, the inner product is 0, so assume $k \ge 1$:

$$\begin{aligned} & \langle \nabla f_c(u_c^{t,k}) - \nabla f_c(w^t), \nabla f(w^t)\rangle \\ \le\ & \|\nabla f_c(u_c^{t,k}) - \nabla f_c(w^t)\|_2 \cdot \|\nabla f(w^t)\|_2 \\ \le\ & L \cdot \|u_c^{t,k} - w^t\|_2 \cdot \|\nabla f(w^t)\|_2 \end{aligned} \tag{18}$$

where the gap between $u_c^{t,k}$ and w^t can be further expanded:

$$\begin{aligned} \|u_c^{t,k} - w^t\|_2 =&\ \|u_c^{t,k} - u_0^{t,k}\|_2 \\ =&\ \|\eta_{\mathrm{local}} \sum_{i=0}^{k-1} \nabla f_c(u_c^{t,i})\|_2 \\ \le&\ \eta_{\mathrm{local}} \sum_{i=0}^{k-1} \|\nabla f_c(u_c^{t,i})\|_2 \\ \le&\ \eta_{\mathrm{local}} \cdot k \cdot G \end{aligned} \tag{19}$$

Plug in Eq. (19) to Eq. (18), we get

$$\langle \nabla f_c(u_c^{t,k}) - \nabla f_c(w^t), \nabla f(w^t)\rangle \le L \cdot \eta_{\mathrm{local}} \cdot k \cdot G^2$$

Recall that $\eta = \eta_{\text{global}} \cdot \eta_{\text{local}}$. Put things together, we finally obtain a bound on $\mathbb{E}[A \mid \mathcal{F}_t]$:

$$\begin{aligned}\mathbb{E}[A \mid \mathcal{F}_t] \leq & -\eta_{\text{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta \cdot \eta_{\text{local}} \cdot L \cdot (\sum_{k=0}^{K-1} k) \cdot G^2 \\ \leq & -\eta_{\text{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta \cdot \eta_{\text{local}} \cdot L \cdot K^2 \cdot G^2 \end{aligned} \tag{20}$$

Bounding $\mathbb{E}[B \mid \mathcal{F}_t]$ Using the fact that $\mathsf{sk}_t / \mathsf{desk}_t$ are linear functions, we get

$$\begin{aligned} B &= \eta_{\text{global}}^2 \cdot \eta_{\text{local}}^2 \cdot \frac{1}{N^2} \cdot \| \sum_{c=1}^{N} \sum_{k=0}^{K-1} \mathsf{desk}_t(\mathsf{sk}_t(\nabla f_c(u_c^{t,k}))) \|_2^2 \\ &\leq \eta_{\text{global}}^2 \cdot \eta_{\text{local}}^2 \cdot \frac{1}{N^2} \cdot N \cdot K \sum_{c=1}^{N} \sum_{k=0}^{K-1} \cdot \| \mathsf{desk}_t(\mathsf{sk}_t(\nabla f_c(u_c^{t,k}))) \|_2^2 \\ &= \eta^2 \cdot \frac{K}{N} \cdot \sum_{c=1}^{N} \sum_{k=0}^{K-1} \| \mathsf{desk}_t(\mathsf{sk}_t(\nabla f_c(u_c^{t,k}))) \|_2^2 \end{aligned}$$

Using variance bound of $\mathsf{desk}_t(\mathsf{sk}_t(h))$, we get

$$\begin{aligned} \mathbb{E}[B \mid \mathcal{F}_t] &\leq \eta^2 \cdot \frac{K}{N} \cdot (1+\alpha) \cdot \sum_{c=1}^{N} \sum_{k=0}^{K-1} \|\nabla f_c(u_c^{t,k})\|_2^2 \\ &\leq \eta^2 \cdot \frac{K}{N} \cdot (1+\alpha) \cdot \sum_{c=1}^{N} \sum_{k=0}^{K-1} G^2 \\ &= \eta^2 \cdot K^2 \cdot (1+\alpha) \cdot G^2 \end{aligned} \tag{21}$$

Put things together Put the bound on $\mathbb{E}[A \mid \mathcal{F}_t]$ and the bound on $\mathbb{E}[B \mid \mathcal{F}_t]$, we get

$$\begin{aligned} & \mathbb{E}[f(w^{t+1}) - f(w^t) \mid \mathcal{F}_t] \\ \leq & -\eta_{\text{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta \cdot \eta_{\text{local}} \cdot L \cdot K^2 \cdot G^2 + \frac{L}{2} \cdot \eta^2 \cdot K^2 \cdot (1+\alpha) \cdot G^2 \\ = & -\eta_{\text{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta \cdot L \cdot K^2 \cdot G^2 \cdot \Big(\eta_{\text{local}} + \frac{\eta}{2} \cdot (1+\alpha)\Big) \end{aligned}$$

□

Theorem G.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth. Let $w^* \in \mathbb{R}^d$ be the optimal solution to f and assume $\mathsf{sk}/\mathsf{desk}$ functions satisfying Theorem 4.2. Then*

$$\min_{t \in [T]} \mathbb{E}[\|\nabla f(w^t)\|_2^2] \leq \frac{1}{(T+1)\eta_{\text{global}}} \cdot (\mathbb{E}[f(w^0)] - f(w^*)) + \eta_{\text{local}} \cdot LK^2G^2 \cdot \Big(\eta_{\text{local}} + \frac{\eta}{2} \cdot (1+\alpha)\Big)$$

Proof. By Lemma G.1, we know that

$$\mathbb{E}[f(w^{t+1}) \mid \mathcal{F}_t] - f(w^t) \leq -\eta_{\text{global}} \cdot \|\nabla f(w^t)\|_2^2 + \eta \cdot L \cdot K^2 \cdot G^2 \cdot \Big(\eta_{\text{local}} + \frac{\eta}{2} \cdot (1+\alpha)\Big)$$

Rearranging the inequality and taking total expectation, we get

$$\mathbb{E}[\|\nabla f(w^t)\|_2^2] \leq \frac{1}{\eta_{\text{global}}} \cdot (\mathbb{E}[f(w^t)] - \mathbb{E}[f(w^{t+1})]) + \eta_{\text{local}} \cdot LK^2G^2 \cdot \Big(\eta_{\text{local}} + \frac{\eta}{2} \cdot (1+\alpha)\Big)$$

Sum over all T iterations and averaging, we arrive at

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla f(w^t)\|_2^2]$$

$$\le \frac{1}{(T+1)\eta_{\rm global}} \cdot (\mathbb{E}[f(w^0)] - \mathbb{E}[f(w^T)]) + \eta_{\rm local} \cdot LK^2G^2 \cdot \Big(\eta_{\rm local} + \frac{\eta}{2} \cdot (1+\alpha)\Big)$$
$$\le \frac{1}{(T+1)\eta_{\rm global}} \cdot (\mathbb{E}[f(w^0)] - f(w^*)) + \eta_{\rm local} \cdot LK^2G^2 \cdot \Big(\eta_{\rm local} + \frac{\eta}{2} \cdot (1+\alpha)\Big)$$

□

H. Differential Privacy

In this section, we consider a special case where each agent c trying to learn upon its local dataset $\mathcal{D}_c$ with corresponding local loss $f_c(x) = \frac{1}{|\mathcal{D}_c|}\sum_{z_i \in \mathcal{D}_c} f_c(x, z_i)$, where we overload the notation f_c to denote the local loss for notation simplicity. We assume f_c is ℓ_c-Lipschitz for agent $c = 1, 2, \cdots, N$. We also assume that the dataset for each agent c is disjoint.

H.1. Differentially Private Algorithm

Algorithm 3 Private Iterative Sketching-based Federated Learning Algorithm with K local steps

```
procedure PRIVATEITERATIVESKETCHINGFL
    Each client initializes $w^0$ using the same set of random seed
    for $t = 1 \to T$ do                                        ▷ $T$ denotes the total number of global steps
        /* Client */
        parfor $c = 1 \to N$ do                                 ▷ $N$ denotes the total number of clients
            if $t = 1$ then
                $u_c^{t,0} \leftarrow w^0$                                      ▷ $u_c^{t,0} \in \mathbb{R}^d$
            else
                $u_c^{t,0} \leftarrow w^{t-1} + \mathsf{desk}_t(\Delta\widetilde{w}^{t-1})$      ▷ $\mathsf{desk}_t : \mathbb{R}^{b_{\rm sketch}} \to \mathbb{R}^d$ de-sketch the change
            end if
            $w^t \leftarrow u_c^{t,0}$
            $\sigma^2 \leftarrow O(\log(1/\widehat{\delta})\ell_c^2/\widehat{\epsilon}^2)$
            for $k = 1 \to K$ do
                $\xi_c^{t,k} \sim \mathcal{N}(0, \sigma^2 \cdot I_{d\times d}) \leftarrow$ Independent Gaussian noise
                $\mathcal{D}_c^{t,k} \leftarrow$ Sample random batch of local data points
                $u_c^{t,k} \leftarrow u_c^{t,k-1} - \eta_{\rm local} \cdot (\frac{1}{|\mathcal{D}_c^{t,k}|} \cdot \sum_{z_i \in \mathcal{D}_c^{t,k}} \nabla f_c(u_c^{t,k-1}, z_i) + \xi_c^{t,k})$
            end for
            $\Delta w_c(t) \leftarrow u_c^{t,K} - w^t$
            Client $c$ sends $\mathsf{sk}_t(\Delta w_c(t))$ to server          ▷ $\mathsf{sk}_t : \mathbb{R}^d \to \mathbb{R}^{b_{\rm sketch}}$ sketch the change
        end parfor
        /* Server */
        $\Delta\widetilde{w}^t \leftarrow \eta_{\rm global} \cdot \frac{1}{N}\sum_{c=1}^N \mathsf{sk}_t(\Delta w_c(t))$                ▷ $\Delta\widetilde{w}^t \in \mathbb{R}^d$
        Server sends $\Delta\widetilde{w}^t$ to each client
    end for
end procedure
```

H.2. Preliminary

We define (ϵ, δ)-differential privacy (Dwork et al., 2006b;a) as

Definition H.1. *Let ϵ, δ be positive real number and $\mathcal{M}$ be a randomized mechanism that takes a dataset as input (representing the actions of the trusted party holding the data). Let im$(\mathcal{M})$ denote the image of $\mathcal{M}$. The algorithm $\mathcal{M}$ is said to provide ϵ, δ-differential privacy if, for all datasets D_1 and D_2 that differ on a single element (i.e., the data of one person), and all subsets S of im$(\mathcal{M})$:*

$$\Pr[\mathcal{M}(D_1) \in S] \le \exp(\epsilon) \cdot \Pr[\mathcal{M}(D_2) \in S] + \delta$$

where the probability is taken over the randomness used by the algorithm.

Lemma H.2 (Parallel Composition)**.** *Let $\mathcal{M}_i$ be an (ϵ_i, δ_i)-DP mechanism and each $\mathcal{M}_i$ operates on disjoint subsets of the private database, then $\mathcal{M}_1 \circ \ldots \circ \mathcal{M}_k$ is $(\max_{i\in[k]} \epsilon_i, \max_{i\in[k]} \delta_i)$-DP.*

Lemma H.3 (Advanced Composition (Dwork et al., 2010))**.** *Let $\epsilon, \delta' \in (0, 1]$ and $\delta \in [0, 1]$. If $\mathcal{M}_1, \ldots, \mathcal{M}_k$ are each (ϵ, δ)-DP mechanisms, then $\mathcal{M}_1 \circ \ldots \circ \mathcal{M}_k$ is $(\epsilon', \delta' + k\delta)$-DP where*

$$\epsilon' = \sqrt{2k \log(1/\delta')} \cdot \epsilon + 2k\epsilon^2.$$

Lemma H.4 (Amplification via Sampling (Lemma 4.12 of (Bun et al., 2015)))**.** *Let $\mathcal{M}$ be an (ϵ, δ)-DP mechanism where $\epsilon \leq 1$. Let $\mathcal{M}'$ be the mechanism that, given a database S of size n, first constructs a database $T \subset S$ by sub-sampling with repetition $k \leq n/2$ rows from S then return $\mathcal{M}(T)$. Then, $\mathcal{M}'$ is $(\frac{6\epsilon k}{n}, \exp(\frac{6\epsilon k}{n})\frac{4k}{n} \cdot \delta)$-DP.*

Lemma H.5 (Post-processing (Proposition 2.1 in (Dwork & Roth, 2013)))**.** *Let $\mathcal{M}$ be an (ϵ, δ)-DP mechanism whose image is R. Let $f : R \to \widetilde{R}$ be an arbitrary randomized mapping. Then $f \circ \mathcal{M}$ is (ϵ, δ)-DP.*

As the noise we consider follows from a Gaussian distribution, it is necessary to include notions related to the Gaussian mechanism.

Definition H.6 (ℓ_2 Sensitivity)**.** *Let $f : \mathcal{X} \to \mathbb{R}^d$, the ℓ_2 sensitivity of f is*

$$\Delta_2^{(f)} = \max_{S,S'} \|f(S) - f(S')\|_2,$$

where S, S' are neighboring databases.

It is folklore that adding Gaussian noise with appropriate σ^2 will provide DP guarantee we needed.

Lemma H.7 (Gaussian Mechanism)**.** *Let $f : \mathcal{X} \to \mathbb{R}^d$ and Δ_2 denote its ℓ_2 sensitivity. Suppose we define $\mathcal{M}(Y) = f(Y) + z$, where $z \sim \mathcal{N}(0, 2\log(1.25/\delta)\Delta_2^2/\epsilon^2 \cdot I)$. Then $\mathcal{M}$ is (ϵ, δ)-DP.*

H.3. ℓ_2 Sensitivity of the Stochastic Gradient

In this section, we bound the ℓ_2 sensitivity of the stochastic gradient, the proof relies on the assumption that each f_c is ℓ_c-Lipschitz.

Lemma H.8. *Consider the stochastic gradient $\frac{|\mathcal{D}_c|}{|\mathcal{D}_c^{t,k}|} \cdot \sum_{z_i \in \mathcal{D}_c^{t,k}} \nabla f_c(u_c^{t,k-1}, z_i)$ as in Algorithm 1. Assume that f_c is ℓ_c-Lipschitz. Then we have*

$$\left\| \frac{1}{|\mathcal{D}_c^{t,k}|} \cdot \sum_{z_i \in \mathcal{D}_c^{t,k}} \nabla f_c(u_c^{t,k-1}, z_i) \right\|_2 \leq \ell_c.$$

Proof. We first note that, since f_c is ℓ_c-Lipschitz, we automatically have that

$$\|\nabla f_c(u_c^{t,k-1}, z_i)\|_2 \leq \ell_c.$$

Hence, we can bound the target quantity via triangle inequality:

$$\begin{aligned}\left\| \frac{1}{|\mathcal{D}_c^{t,k}|} \cdot \sum_{z_i \in \mathcal{D}_c^{t,k}} \nabla f_c(u_c^{t,k-1}, z_i) \right\|_2 &\leq \frac{1}{|\mathcal{D}_c^{t,k}|} \cdot \sum_{z_i \in \mathcal{D}_c^{t,k}} \|\nabla f_c(u_c^{t,k-1}, z_i)\|_2 \\ &\leq \ell_c,\end{aligned}$$

as desired. □

H.4. Privacy guarantee of our algorithm

In this section, we provide a formal analysis on the privacy guarantee of Algorithm 3. We will first analyze the privacy property for a single agent, then combine them via composition lemma.

Lemma H.9 (Formal version of Lemma 6.1)**.** *Let $\widehat{\epsilon}, \widehat{\delta} \in [0,1)$, $\widehat{\epsilon} < \frac{1}{\sqrt{K}}$ and $c \in [N]$. For agent c, the local-K-step stochastic gradient as in Algorithm 1 is $(\sqrt{K} \cdot \widehat{\epsilon}, K \cdot \widehat{\delta})$-DP.*

Proof. First, we note that σ^2 is chosen as $O(\log(1/\widehat{\delta})\ell_c^2/\widehat{\epsilon}^2)$, hence, by Lemma H.7, we know that one step of stochastic gradient is $(\widehat{\epsilon}, \widehat{\delta})$-DP. Since we run the local SGD for K steps, by Lemma H.3, we have the process is

$$(O(\sqrt{K} \cdot \widehat{\epsilon} + K\widehat{\epsilon}^2), O(K\widehat{\delta}))$$

DP. Finally, since $\widehat{\epsilon} \leq \frac{1}{\sqrt{K}}$, we conclude that the local-K-step for agent c is $(O(\sqrt{K} \cdot \widehat{\epsilon}), O(K \cdot \widehat{\delta}))$-DP. □

Remark H.10. *We want to point out that although we perform sketching on the sum of the local gradients, by Lemma H.5, this does not change the privacy guarantee at all.*

Theorem H.11. *Let $\widehat{\epsilon}, \widehat{\delta}$ be as in Lemma 6.1. Then, Algorithm 1 is $(\epsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-DP, with*

$$\epsilon_{\mathrm{DP}} = \sqrt{TK} \cdot \widehat{\epsilon}, \, \delta_{\mathrm{DP}} = TK \cdot \widehat{\delta}.$$

Proof. Notice that each agent c works on individual subsets of the data, therefore we can make use of Lemma H.2 to conclude that over all N agents, the process is $(\sqrt{K} \cdot \widehat{\epsilon}, K \cdot \widehat{\delta})$-DP. Finally, apply Lemma H.3 over all T iterations, we conclude that Algorithm 3 is $(\epsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-DP, while

$$\epsilon_{\mathrm{DP}} = \sqrt{TK} \cdot \widehat{\epsilon}, \, \delta_{\mathrm{DP}} = TK \cdot \widehat{\delta}.$$

□

I. Preliminary on Gradient Attack

Throughout this section to the remainder of the paper, we use $F(x; w)$ to denote the loss function of the model, where $x \in \mathbb{R}^m$ is the data point and $w \in \mathbb{R}^d$ is the model parameter.

I.1. Definitions

We start with defining some conditions we will later study:

Definition I.1. *Let $F : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$, we define the loss function L to be*

$$L(x) := \|\nabla_w F(x; w) - g\|^2.$$

Definition I.2 (Smoothness)**.** *We say $L : \mathbb{R}^m \to \mathbb{R}$ is b-smooth if for any $x, y \in \mathbb{R}^m$, we have*

$$L(y) \leq L(x) + \langle \nabla L(x), y - x \rangle + b\|y - x\|^2.$$

Definition I.3 (Lipschitz)**.** *We say $L : \mathbb{R}^m \to \mathbb{R}$ is β-Lipschitz if for any $x, y \in \mathbb{R}^m$, we have*

$$\|L(x) - L(y)\|^2 \leq \beta^2 \|x - y\|^2.$$

Definition I.4 (Semi-smoothness)**.** *For any $p \in [0,1]$, we say L is (a, b, p)-semi-smoothness if*

$$\begin{aligned} L(y) \leq L(x) + \langle \nabla L(x), y - x \rangle + b\|y - x\|^2 \\ + a\|x - y\|^{2-2p} L(x)^p \end{aligned}$$

Definition I.5 (Semi-Lipschitz)**.** *For any $p \in [0,1]$, we say function L is (α, β, p)-semi-Lipschitz if*

$$(L(x) - L(y))^2 \leq \beta^2 \|x - y\|^2 + \alpha^2 \|x - y\|^{2-2p} \cdot L(x)^p.$$

Specifically, we say function L has (α, β, p)-semi-Lipschitz gradient, or L satisfies (α, β, p)-semi-Lipschitz gradient condition, if

$$\begin{aligned} \|\nabla L(x) - \nabla L(y)\|^2 \leq \beta^2 \|x - y\|^2 \\ + \alpha^2 \|x - y\|^{2-2p} \cdot L(x)^p. \end{aligned}$$

Definition I.6 (Non-critical point)**.** *We say L is (θ_1, θ_2)-non-critical point if*

$$\theta_1^2 \cdot L(x) \leq \|\nabla L(x)\|^2 \leq \theta_2^2 \cdot L(x).$$

Definition I.7 (Pseudo-Hessian)**.** *Let $F : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$, suppose F is differentiable on both x and w, then we define pseudo-Hessian mapping $\Phi : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^{d \times m}$ as follows*

$$\Phi(x, w) = \nabla_x \nabla_w F(x; w).$$

Correspondingly, we define a pseudo-kernel $K : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ with respect to $\nabla_x F(x; w)$ as:

$$K(x, w) = \Phi(x, w)^\top \Phi(x, w).$$

Note the weight vector w is fixed in our setting, we write $K(x) = K(x, w)$ for simplicity.

I.2. Useful Lemmas

We prove two useful lemmas regarding Lipschitz gradient and smoothness, and extend this result to semi-Lipschitz gradient and semi-smoothness.

Lemma I.8 (folklore)**.** *Suppose $L : \mathbb{R}^m \to \mathbb{R}$ has β-Lipschitz gradient, then L is b-smooth, where $b = \beta/2$.*

Proof. Suppose $L(x)$ has β-Lipschitz gradient. This means that for any $x, y \in \mathbb{R}^m$, we have $\|\nabla L(x) - \nabla L(y)\| \leq \beta \|x - y\|$.

By Cauchy-Schwartz,

$$\langle \nabla L(x) - \nabla L(y), x - y \rangle \leq \beta \|x - y\|^2.$$

Hence function $G(x) = \frac{\beta}{2}\|x\|^2 - L(x)$ is convex. So

$$G(y) \geq G(x) + \langle \nabla G(x), y - x \rangle,$$

which implies

$$L(y) \leq L(x) + \frac{\beta}{2} \langle \nabla L(x), y - x \rangle.$$

Thus $L(x)$ is also b-smooth where $b = \frac{\beta}{2}$. □

Lemma I.9. *Suppose L satisfies (α, β, p)-semi-Lipschitz gradient (Def. I.5), then L is also $(\alpha, \frac{\beta}{2}, p/2)$-semi-smooth (Def. I.4).*

Proof. First we can bound the inner product term

$$\begin{aligned}
&\langle \nabla L(x) - \nabla L(y), x - y \rangle \\
\leq\ & \|\nabla L(x) - \nabla L(y)\| \cdot \|x - y\| \\
\leq\ & \sqrt{\beta^2 \|x - y\|^2 + \alpha^2 \|x - y\|^{2-2p} \cdot L(x)^p} \cdot \|x - y\| \\
\leq\ & \big(\beta \|x - y\| + \alpha \|x - y\|^{(1-p)} L(x)^{p/2}\big) \cdot \|x - y\| \\
=\ & \beta \|x - y\|^2 + \alpha \|x - y\|^{2-p} L(x)^{p/2}.
\end{aligned} \tag{22}$$

The first step is Cauchy-Schwartz, the second step is the definition of (α, β, p)-semi-Lipschitz, and the third step is the fact $\sqrt{a^2 + b^2} \leq a + b$ for non-negative a and b.

Let $G(x) = \frac{\beta}{2}\|x\|^2 - L(x)$. We could verify that

$$\begin{aligned}
&\langle \nabla G(y) - \nabla G(x), y - x \rangle \\
=\ & \langle \beta y - \nabla L(y) - \beta x + \nabla L(x), y - x \rangle
\end{aligned}$$

$$\begin{aligned} &= \beta\|y-x\|^2 - \langle \nabla L(y) - \nabla L(x), y-x\rangle \\ &\geq -\alpha\|x-y\|^{2-p} L(x)^{p/2}. \end{aligned} \tag{23}$$

The first step is derived from the definition of gradient and the third step is by plugging in Eq. (22).

Let $\phi(t) = G(x + t(y-x))$. Notice that $G(y) - G(x) = \phi(1) - \phi(0) = \int_0^1 \frac{\mathrm{d}\phi}{\mathrm{d}t}\mathrm{d}t$, hence we have

$$\begin{aligned} & G(y) - G(x) \\ = & \int_0^1 \langle \nabla G(x + t(y-x)), y-x\rangle \mathrm{d}t \\ = & \int_0^1 \langle \nabla G(x + t(y-x)) - \nabla G(x), y-x\rangle \mathrm{d}t \\ & + \int_0^1 \langle \nabla G(x), y-x\rangle \mathrm{d}t \\ \geq & \int_0^1 \big(\langle \nabla G(x), y-x\rangle - t^{1-p}\cdot\alpha\|x-y\|^{2-p}L(x)^{p/2}\big)\mathrm{d}t \\ \geq & \int_0^1 \big(\langle \nabla G(x), y-x\rangle - \alpha\|x-y\|^{2-p}L(x)^{p/2}\big)\mathrm{d}t \\ = & \langle \nabla G(x), y-x\rangle - \alpha\|x-y\|^{2-p}L(x)^{p/2}. \end{aligned}$$

The third step follows from Eq. (23) and the fourth step follows from $p \in (0,1)$.

Hence,

$$\begin{aligned} G(y) \geq & G(x) + \langle \nabla G(x), y-x\rangle \\ & - \alpha\|x-y\|^{2-p}L(x)^{p/2}. \end{aligned} \tag{24}$$

Then plug in $G(x) = \frac{\beta}{2}\|x\|^2 - L(x)$, Eq. (24) implies

$$\begin{aligned} \frac{\beta}{2}\|y\|^2 - L(y) \geq & \frac{\beta}{2}\|x\|^2 - L(x) + \langle \nabla G(x), y-x\rangle \\ & - \alpha\|x-y\|^{2-p}L(x)^{p/2} \end{aligned}$$

which is equivalent to

$$\begin{aligned} L(y) \leq & L(x) + \langle \nabla L(x), y-x\rangle \\ & + \frac{\beta}{2}\big(\|y\|^2 - 2\langle x, y\rangle + \|x\|^2\big) \\ & + \alpha\|x-y\|^{2-p}L(x)^{p/2} \\ = & L(x) + \langle \nabla L(x), y-x\rangle + \frac{\beta}{2}\|y-x\|^2 \\ & + \alpha\|x-y\|^{2-p}L(x)^{p/2}. \end{aligned}$$

□

J. From F to L

In this section, we show that if we impose mild conditions on F, it will imply certain key conditions on L, which is critical in proving the convergence of our loss function. We start with a list of assumptions of F.

Assumption J.1. *Let x^* denote the global minimum of $L(x)$ in $\mathbb{R}^m$. We without loss of generality assume $L(x^*) = 0$.*

Table 2: Summary of functions with different properties. We use semi-s. to denote semi-smooth, we use semi-Lg. to denote semi-Lipschitz gradient. We use semi-sc. to denote semi-strongly convex. We use N/A because such function is impossible due to we've proved semi-Lipschitz gradient implies the semi-smoothness in Lemma I.9. †: assume $w^\top x \geq 0$.

Function	assumption	semi-s.	semi-Lg.	semi-sc.	non-critical point
$\|x\|_2^2$		✓	✓	✓	✓
N/A		×	✓	✓	✓
$x^\top Ax - \lambda_{\min}(A)$		✓	✓	✓	×
$\sqrt{w^\top x + b}^\dagger$	$w^\top x \geq 0$	✓	✓	×	×
$\ln(1 + e^x)$	$x \in [-1, 1]^\dagger$	✓	✓	×	✓
$\mathrm{sigmoid}(w^\top x + b)$		✓	✓	×	×
$(w^\top x)^2 \cdot \sin(1/(w^\top x))$	$\|w\|, \|x\| = 1, w^\top x \neq 0$	✓	×	×	✓
$(w^\top x)^2 \cdot \sin(1/(w^\top x))$		✓	×	×	×
$\ln(w^\top x)$	$w^\top x > 0$	✓	×	×	×
N/A		×	✓	✓	×
N/A		×	✓	×	✓
$\cosh(w^\top x)$	$\|w\|, \|x\| = 1$	×	×	✓	✓
N/A		×	✓	×	×
$\cosh(w^\top x)$		×	×	✓	×
$\mathrm{ReLU}(w^\top x)$		×	×	×	×
$1/\|x\|_2$		×	×	×	×
$\tanh(w^\top x)$		×	×	×	×

We give a brief justification of this assumption. Notice that

$$\begin{aligned} L(x) &= \|\nabla_w F(x; w) - g\|^2 \\ &= \|\nabla_w F(x; w) - \nabla_w F(\widetilde{x}; w)\|^2. \end{aligned}$$

Hence $L(x) \geq 0$ and $L(\widetilde{x}) = 0$. So it is reasonable to assume $\min_{x \in \mathbb{R}^m} L(x) = 0$. Even if $L(x)$ has other forms and $\min_{x \in \mathbb{R}^m} L(x) \neq 0$, we can define a dummy objective function $L'(x)$ as

$$L'(x) = L(x) - C,$$

where $C = \min_x L(x)$. Suppose we apply gradient descent with initialization x_0 on $L(x)$ and apply gradient descent with initialization y_0 on $L'(y)$. Based on the fact that $\nabla L(x) = \nabla L'(x)$, we could show if $x_t = y_t$ then

$$y_{t+1} = y_t - \eta \cdot \nabla L'(y_t) = x_t - \eta \cdot \nabla L(x_t) = x_{t+1}.$$

Hence by induction, for all t, $x_t = y_t$ when $x_0 = y_0$. Thus the convergence rate for $L(x)$ and $L'(x)$ are exactly the same as long as the initialization is the same. Since we choose $C = \min_{x \in \mathbb{R}^m} L(x)$, we can easily verify that

$$\min_{x \in \mathbb{R}^m} L'(x) = \min_{x \in \mathbb{R}^m} L(x) - C = 0.$$

Therefore, without loss of generality, we make Assumption J.1.

The next assumption is a standard Lipschitz gradient assumption.

Assumption J.2. *$\nabla_w F(x, w)$ is β-Lipschitz with respect to x, i.e., for any $x \in \mathbb{R}^m$ we have*

$$\|\nabla_w F(x_1; w) - \nabla_w F(x_2; w)\| \leq \beta \cdot \|x_1 - x_2\|.$$

The next assumption is necessary to ensure L has non-critical point property.

Assumption J.3. *Let $\theta_2 \geq \theta_1 > 0$. $\forall x \in \mathbb{R}^m$, let $K(x)$ be defined as Definition I.7. $K(x)$'s eigenvalues can be bounded by*

$$\theta_1^2 \leq \lambda_1^2(x) \leq \cdots \leq \lambda_{\min(m,d)}^2(x) \leq \theta_2^2.$$

J.1. What F Implies Semi-smoothness

Lemma J.4. *Let $\Phi(x, w)$ be defined as Def. I.7. Suppose that*

- *(Assumption J.2) $\nabla_w F(x; w)$ is β-Lipschitz with respect to x, $\forall x \in \mathbb{R}^m$;*
- *(Assumption J.3) spectral norm of Hessian matrix is bounded by $\|\Phi(x, w)\| \leq \theta_2$, $\forall x \in \mathbb{R}^m$.*

Then, $L(x) = \|\nabla_w F(x; w) - g\|^2$ is (a, b, p)-semi-smooth, where $b = \beta^2, a = 2(\beta + \theta_2)$, and $p = 1/2$, i.e.

$$\begin{aligned} L(x) \leq\ & L(y) + \langle \nabla L(y), x - y\rangle + b\|x - y\|^2 \\ & + a\|x - y\| L(y)^{1/2}. \end{aligned}$$

Proof. We define $\mathcal{A}_1$ and $\mathcal{A}_2$ as follows:

$$\begin{aligned} \mathcal{A}_1 :=\ & \langle \nabla_w F(y; w) - \nabla_w F(x; w), \\ & 2g - \nabla_w F(x; w) - \nabla_w F(y; w)\rangle, \\ \mathcal{A}_2 :=\ & \langle \Phi(y, w)(\nabla_w F(y; w) - g), y - x\rangle. \end{aligned}$$

Notice that for $\mathcal{A}_1$,

$$\begin{aligned} \mathcal{A}_1 =\ & \langle \nabla_w F(y; w) - \nabla_w F(x; w), \\ & 2g - \nabla_w F(x; w) - \nabla_w F(y; w)\rangle \\ \leq\ & \|\nabla_w F(y; w) - \nabla_w F(x; w)\| \\ & \cdot \|2g - \nabla_w F(x; w) - \nabla_w F(y; w)\| \\ \leq\ & \beta\|y - x\| \cdot \|2g - \nabla_w F(x; w) - \nabla_w F(y; w)\| \\ \leq\ & \beta\|y - x\| \cdot \|\nabla_w F(y; w) - \nabla_w F(x; w)\| \\ & + \beta\|y - x\| \cdot 2\|g - \nabla_w F(y; w)\| \\ \leq\ & \beta\|y - x\| \cdot (\beta\|y - x\| + 2\|g - \nabla_w F(y; w)\|) \\ =\ & \beta^2\|y - x\|^2 + 2\beta\|y - x\| L(y)^{1/2}. \end{aligned}$$

The second step is Cauchy–Schwartz inequality, the third step is derived from the assumption that $F(x, w)$ has β-Lipschitz gradient, and the fourth step is the triangle inequality.

For $\mathcal{A}_2$, it can be bounded as

$$\begin{aligned} \mathcal{A}_2 =\ & \langle \Phi(y, w)(\nabla_w F(y; w) - g), y - x\rangle \\ \leq\ & \|\Phi(y, w)(\nabla_w F(y; w) - g)\| \cdot \|y - x\| \\ \leq\ & \|\Phi(y, w)\| \cdot \|\nabla_w F(y; w) - g\| \cdot \|y - x\| \\ \leq\ & \theta_2 \cdot L(y)^{1/2} \cdot \|y - x\|. \end{aligned}$$

The second step is Cauchy-Schwartz inequality and the third step is the assumption on spectral norm.

Let a, b, R be defined as

$$\begin{aligned} b &:= \beta \\ a &:= 2(\beta + \theta_2) \\ R &:= b\|y - x\|^2 + a\|y - x\| L(y)^{1/2} \end{aligned}$$

Combining the bound for $\mathcal{A}_1$ and $\mathcal{A}_2$, we have the bound $\mathcal{A}_1 + 2\mathcal{A}_2 \leq R$. Therefore

$$\begin{aligned} R \geq\ & \mathcal{A}_1 + 2\mathcal{A}_2 \\ =\ & \langle \nabla_w F(y, w) - \nabla_w F(x, w), 2g - \nabla_w F(x, w) \\ & - \nabla_w F(y, w)\rangle + 2\langle \Phi(y, w)(\nabla_w F(y, w) - g), y - x\rangle \end{aligned}$$

$$\begin{aligned}
&= \langle \nabla_w F(y,w) - \nabla_w F(x,w), 2g - \nabla_w F(x,w) \\
&\quad - \nabla_w F(y,w)\rangle - 2\langle \Phi(y,w)(\nabla_w F(y,w) - g), x - y\rangle \\
&= \|\nabla_w F(x,w)\|^2 - \|\nabla_w F(y,w)\|^2 \\
&\quad + 2\langle \nabla_w F(y,w) - \nabla_w F(x,w), g\rangle \\
&\quad - 2\langle \nabla_y \|\nabla_w F(y,w) - g\|^2, x - y\rangle \\
&= \|\nabla_w F(x,w) - g\|^2 - \|\nabla_w F(y,w) - g\|^2 \\
&\quad - \langle \nabla L(y), x - y\rangle \\
&= L(x) - L(y) - \langle \nabla L(y), x - y\rangle.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& L(x) - L(y) \\
\leq & \langle \nabla L(y), x - y\rangle + R \\
= & \langle \nabla L(y), x - y\rangle + b\|y - x\|^2 + a\|y - x\|L(y)^{1/2},
\end{aligned}$$

which is equivalent to the statement that $L(x)$ is $(2(\beta + \theta_2), \beta^2, 1/2)$-semi-smooth. □

J.2. What F Implies Non-critical Point

Lemma J.5. *Let K be defined as in Def. I.7. Denote the eigenvalues of K by $\lambda_1^2(x) \leq \lambda_2^2(x) \leq \cdots \leq \lambda_m^2(x)$. If Assumption J.3 holds, i.e., for all $x \in \mathbb{R}^m$*

- $\theta_1^2 \leq \lambda_1^2(x)$,
- $\theta_2^2 \geq \lambda_{\min(m,d)}^2(x)$.

then, L satisfies (θ_1, θ_2)-non-critical point condition., i.e.

$$\theta_1^2 \cdot L(x) \leq \|\nabla L(x)\|^2 \leq \theta_2^2 \cdot L(x).$$

Proof. Notice that

$$\begin{aligned}
\|\nabla_x L(x)\|^2 &= \|\Phi(x,w)(\nabla_w F(x;w) - g)\|^2 \\
&= (\nabla_w F(x;w) - g)^\top K(x)(\nabla_w F(x;w) - g)
\end{aligned}$$

Given conforming positive definite matrix A and vector y, it is well-known that

$$\lambda_{\min}(A) \leq \frac{y^\top A y}{\|y\|^2} \leq \lambda_{\max}(A),$$

hence,

$$\begin{aligned}
\|\nabla_x L(x)\|^2 &\geq \theta_1^2 \cdot \|\nabla_w F(x;w) - g\|^2 \\
\|\nabla_x L(x)\|^2 &\leq \theta_2^2 \cdot \|\nabla_w F(x;w) - g\|^2,
\end{aligned}$$

which is equivalent to

$$\theta_1^2 \cdot L(x) \leq \|\nabla_x L(x)\|^2 \leq \theta_2^2 \cdot L(x).$$

□

K. Converge to Optimal Solution

One of the important conditions we need to impose on L if we want to converge to the optimal solution is L has a unique minimum. In order to achieve this property, we introduce the notion of semi-strongly convex:

Definition K.1 (semi-strongly convex)**.** *For any $p \in [0,1]$, we say function $L : \mathbb{R}^m \to \mathbb{R}$ is (c,d,p)-semi-strongly-convex if for any $x, y \in \mathbb{R}^m$, we have*

$$\begin{aligned} L(x) \geq & \, L(y) + \langle \nabla L(y), x - y \rangle + d\|x-y\|^2 \\ & - c\|x-y\|^{2-2p} \cdot L(y)^p. \end{aligned}$$

K.1. Conditions for Unique Minimum

Theorem K.2 (Unique Local Minimum)**.** *If $L(x)$ satisfies (θ_1,θ_2)-non-critical point condition ($\theta_1 > 0$), and (c,d,p)-semi-strongly convex ($d > 0, p \neq 1$), then there exists a unique local minima x^* such that $\nabla L(x^*) = 0$.*

Proof. Suppose $x_1^* \in \mathbb{R}^m$ and $x_2^* \in \mathbb{R}^m$ are two local minima such that

$$\nabla L(x_1^*) = \nabla L(x_2^*) = 0.$$

Since $L(x)$ satisfies (θ_1,θ_2)-non-critical point condition,

$$\theta_1^2 \cdot L(x_1^*) \leq \|\nabla L(x_1^*)\|^2 \leq \theta_2^2 \cdot L(x_1^*).$$

Therefore $L(x_1^*) = 0$ holds. Similarly $L(x_2^*) = 0$ also holds. By (c,d,p)-semi-strongly convexity of $L(x)$,

$$\begin{aligned} L(x_1^*) \geq & \, L(x_2^*) + \langle \nabla L(x_2^*), x_1^* - x_2^* \rangle \\ & + d\|x_2^* - x_1^*\|^2 \\ & - c\|x_2^* - x_1^*\|^{2-2p} \cdot L(x_2^*)^p. \end{aligned} \tag{25}$$

Combining with $L(x_1^*) = L(x_2^*) = 0$ and $\nabla L(x_2^*) = 0$, Eq. (25) implies

$$0 \geq d\|x_2^* - x_1^*\|^2.$$

Hence $\|x_2^* - x_1^*\|^2 = 0$ and $x_1^* = x_2^*$. □

K.2. Conditions for Convergence of x_t

Theorem K.3. *Suppose we run gradient descent algorithm to update x_{t+1} in each iteration as follows:*

$$x_{t+1} = x_t - \eta \cdot \nabla L(x)|_{x=x_t}$$

Assume that $\nabla L(x^*) = 0$. If function L is

- *(c,d,p)-semi-strongly convex (Def. K.1)*
- *(α,β,p)-semi-Lipschitiz gradient (Def. I.5)*
- *(θ_1,θ_2)-non-critical point (Def. I.6)*
- $d > \frac{c^{1/2p}}{\theta_1(\theta_1-\alpha)^{1/p}} \left(\beta^2 + (\alpha/\theta_1^p)^{1/(1-p)}\right) + c^{1/(2-2p)}$
- $\theta_1 > \alpha^{1/p}$

by choosing

$$\eta \leq \xi/(2\zeta)$$

where

$$\zeta := \frac{\theta_1}{\theta_1 - \alpha^{1/p}} \cdot \left(\beta^2 + (\alpha/\theta_1^p)^{1/(1-p)}\right)$$

and

$$\xi := 2(d - c^{1/2p}\theta_1^{-2}\zeta - c^{1/(2-2p)}).$$

we have

$$\|x_{t+1} - x^*\| \leq (1-\gamma) \cdot \|x_t - x^*\|,$$

where $\gamma = 1 - \xi\eta/2$.

Proof. We have

$$\begin{aligned} & \|x_{t+1} - x^*\|^2 \\ = {} & \|x_{t+1} - x_t + x_t - x^*\|^2 \\ = {} & \underbrace{\|x_{t+1} - x_t\|^2}_{\mathcal{A}_1} + 2 \underbrace{\langle x_{t+1} - x_t, x_t - x^* \rangle}_{\mathcal{A}_2} + \|x_t - x^*\|^2. \end{aligned} \tag{26}$$

For the first term in Eq. (26), we have

$$\mathcal{A}_1 = \eta^2 \|\nabla L(x_t)\|^2.$$

Consider x_t, x^*, using (α, β, p)-semi-Lipschitz gradient and $\nabla L(x^*) = 0$, we have

$$\begin{aligned} & \|\nabla L(x_t)\|^2 \\ \leq {} & \beta^2 \|x_t - x^*\|^2 + \alpha^2 \|x_t - x^*\|^{2-2p} \cdot L(x_t)^p \\ \leq {} & \beta^2 \|x_t - x^*\|^2 + \alpha^2 \|x_t - x^*\|^{2-2p} \cdot \|\nabla L(x_t)\|^{2p} / (\theta_1^{2p}), \end{aligned} \tag{27}$$

where the second step follows from non-critical point (Definition I.6). For the last term of the above equation, we have

$$\begin{aligned} & \alpha^2 \|x_t - x^*\|^{2-2p} \cdot \|\nabla L(x_t)\|^{2p} / (\theta_1^{2p}) \\ \leq {} & (\alpha/\theta_1^p)^{2/(2-2p)} \cdot \|x_t - x^*\|^2 + (\alpha/\theta_1^p)^{1/p} \|\nabla L(x_t)\|^2, \end{aligned} \tag{28}$$

where the step follows from $a^{2-2p} b^{2p} \leq a^2 + b^2$.

Thus, Eq. (27) and (28) imply

$$\begin{aligned} \|\nabla L(x_t)\|^2 \leq {} & \frac{\theta_1}{\theta_1 - \alpha^{1/p}} \cdot \Big(\beta^2 + (\alpha/\theta_1^p)^{1/(1-p)} \Big) \\ & \cdot \|x_t - x^*\|^2. \end{aligned}$$

For the second term in Eq. (26), we have

$$\begin{aligned} \mathcal{A}_2 = {} & 2\eta \langle \nabla L(x_t), x_t - x^* \rangle \\ \leq {} & 2\eta \big(\underbrace{L(x^*) - L(x_t)}_{\leq 0} - d\|x_t - x^*\|^2 \\ & + c\|x_t - x^*\|^{2-2p} \cdot L(x_t)^p \big) \\ \leq {} & 2\eta(-d\|x_t - x^*\|^2 + c\|x_t - x^*\|^{2-2p} \cdot L(x_t)^p) \\ \leq {} & (-2\eta d + 2\eta c^{1/(2-2p)}) \|x_t - x^*\|^2 + 2\eta c^{1/(2p)} L(x_t) \\ \leq {} & (-2\eta d + 2\eta c^{1/(2-2p)}) \|x_t - x^*\|^2 \\ & + 2\eta \frac{c^{1/2p}}{\theta_1^2} \|\nabla L(x_t)\|^2, \end{aligned}$$

where the second step follows from (c, d, p)-semi-strongly convex, the third step follows from $L(x^*) - L(x_t) \leq 0$, the fourth step follows from $a^{2-2p} b^{2p} \leq a^2 + b^2$, and the last step is $L(x_t) \leq (1/\theta_1^2) \|\nabla L(x_t)\|^2$.

Putting it to the Eq. (26), we have

$$\begin{aligned} & \|x_{t+1} - x^*\|^2 \\ = {} & \mathcal{A}_1 + \mathcal{A}_2 + \|x_t - x^*\|^2 \\ \leq {} & \left(\eta^2 + 2\eta \frac{c^{1/2p}}{\theta_1^2} \right) \|\nabla L(x_t)\|^2 \end{aligned}$$

$$
\begin{aligned}
&+ (1 - 2\eta d + 2\eta c^{1/(2-2p)})\|x_t - x^*\|^2 \\
\le &\|x_t - x^*\|^2 \cdot \left[\eta^2 \cdot \frac{\theta_1}{\theta_1 - \alpha^{1/p}} \left(\beta^2 + \left(\frac{\alpha}{\theta_1^p}\right)^{1/(1-p)}\right)\right. \\
&-2\eta \left(d - \frac{c^{1/2p}}{\theta_1(\theta_1 - \alpha)^{1/p}} \left(\beta^2 + \left(\frac{\alpha}{\theta_1^p}\right)^{1/(1-p)}\right)\right. \\
&\left.\left. - \ c^{1/(2-2p)}\right) + 1\right].
\end{aligned}
$$

Let

$$\zeta = \frac{\theta_1}{\theta_1 - \alpha^{1/p}} \left(\beta^2 + (\alpha/\theta_1^p)^{1/(1-p)}\right)$$

and

$$\xi = 2(d - c^{1/2p}\theta_1^{-2}\zeta - c^{1/(2-2p)}).$$

Then we have

$$
\begin{aligned}
&\|x_{t+1} - x^*\|^2 \\
\le &(\zeta\eta^2 - \xi\eta + 1)\|x_t - x^*\|^2 \\
\le &(-\xi\eta/2 + 1)\|x_t - x^*\|^2 \\
\le &(1 - \gamma)\|x_t - x^*\|^2,
\end{aligned}
$$

where $\gamma = \xi\eta/2$. The second step holds because we choose $\eta \le \xi/(2\zeta)$ and hence

$$\zeta\eta^2 - \xi\eta \ \le \ (\xi/2)\eta - \xi\eta \ = \ -\xi\eta/2.$$

This concludes our proof. □

L. Converge to Optimal Cost

In this section, we provide the formal proof that if L is semi-smooth and non-critical point, then the loss converges linearly.

L.1. Conditions for Convergence of $L(x_t)$

Theorem L.1. *Suppose we run gradient descent algorithm to update x_{t+1} in each iteration as follows:*

$$x_{t+1} = x_t - \eta \cdot \nabla L(x)|_{x=x_t}$$

If we assume

- *L is (a, b, p)-semi-smooth (Def. I.4),*
- *L is (θ_1, θ_2)-non-critical point (Def. I.6),*
- *$\theta_1^2 > a\theta_2^{2-2p}$,*

using the choice

$$\eta \le (\theta_1^2 - a\theta_2^{2-2p})/(2b\theta_2^2),$$

then we have

$$L(x_{t+1}) - L(x^*) \le (1 - \gamma)(L(x_t) - L(x^*)),$$

where $\gamma = \eta(\theta_1^2 - a\theta_2^{2-2p})/2$.

Proof. We start by bounding the consecutive gap between $L(x_{t+1})$ and $L(x_t)$:

$$\begin{aligned}
& L(x_{t+1}) - L(x_t) \\
\leq & \langle \nabla L(x_t), x_{t+1} - x_t \rangle + b\|x_{t+1} - x_t\|^2 \\
& + a\|x_{t+1} - x_t\|^{2-2p} \cdot L(x_t)^p \\
= & -\eta\|\nabla L(x_t)\|^2 + b\eta^2\|\nabla L(x_t)\|^2 \\
& + a\eta\|\nabla L(x_t)\|^{2-2p} \cdot L(x_t)^p \\
= & -\eta\|\Phi(x_t, w)(\nabla_w F(x_t; w) - g)\|^2 \\
& + b\eta^2\|\Phi(x_t, w)(\nabla_w F(x_t; w) - g)\|^2 \\
& + a\eta\|\Phi(x_t, w)(\nabla_w F(x_t; w) - g)\|^{2-2p} \cdot L(x_t) \\
\leq & -\eta\theta_1^2 L(x_t) + b\eta^2\theta_2^2 L(x_t) + a\eta\theta_2^{2-2p} L(x_t) \\
= & (-\eta\theta_1^2 + b\eta^2\theta_2^2 + a\eta\theta_2^{2-2p})L(x_t),
\end{aligned}$$

where the first step follows from (a, b, p)-semi-smoothness, the third step is due to the identity $\nabla L(x) = \Phi(x, w)(\nabla F_w(x; w) - g)$, the fourth step uses minimum and maximum eigenvalue to give a bound.

This implies that

$$L(x_{t+1}) \leq (1 - \eta\theta_1^2 + b\eta^2\theta_2^2 + a\eta\theta_2^{2-2p})L(x_t).$$

It remains to compute $L(x_{t+1}) - L(x^*)$:

$$\begin{aligned}
& L(x_{t+1}) - L(x^*) \\
\leq & (1 - \eta\theta_1^2 + b\eta^2\theta_2^2 + a\eta\theta_2^{2-2p})L(x_t) - L(x^*) \\
\leq & (1 - (\theta_1^2 - a\theta_2^{2-2p})\eta/2) \cdot L(x_t) - L(x^*) \\
\leq & (1 - \eta) \cdot (L(x_{t+1}) - L(x^*)).
\end{aligned}$$

This completes the proof. □

M. Attack Sketched Gradient

In this section, we consider the setting where the gradient is sketched, i.e., we can only observe a sketched gradient $\mathcal{S}(g)$ where $\mathcal{S} : \mathbb{R}^d \to \mathbb{R}^{b_{\mathrm{sketch}}}$ is a sketching matrix. We can also observe the sketching matrix $\mathcal{S}$, hence, our strategy will be solving the new sketched objective $L_{\mathcal{S}}(x) = \|\mathcal{S}(\nabla_w F(x, w)) - \mathcal{S}(g)\|^2$ and optimize over the sketched objective. We remark this is similar to the classical sketch-and-solve paradigm (Clarkson & Woodruff, 2013; Woodruff, 2014). Let $\mathcal{S} \in \mathbb{R}^{b_{\mathrm{sketch}} \times d}$ be a sketching matrix, popular sketching matrices are random Gaussian, Count Sketch (Charikar et al., 2002), subsampled randomized Hadamard transform (Lu et al., 2013). We impose following assumptions on $\mathcal{S}$.

Assumption M.1. *Let $\tau > 0$, for any in $u, v \in \mathbb{R}^d$,*

$$\|\mathcal{S}(u) - \mathcal{S}(v)\| \leq \tau\|u - v\|.$$

The above assumption is a standard guarantee given by so-called subspace embedding property (Sarlós, 2006).

Assumption M.2. *For any sketching matrix $\mathcal{S} \in \mathbb{R}^{b_{\mathrm{sketch}} \times d}$, we have*

$$0 < \gamma_1 \leq \sigma_1(\mathcal{S}^\top) \leq \ldots \leq \sigma_s(\mathcal{S}^\top) \leq \gamma_2.$$

For typical sketching matrices, the spectral norm is 1 and it is full rank almost-surely, hence, $\gamma_1 > 0$ is a reasonable assumption.

M.1. What Sketching Implies Semi-smoothness

Lemma M.3. *If the sketching mapping $\mathcal{S}$ satisfies $\|\mathcal{S}(u)-\mathcal{S}(v)\| \leq \tau\|u-v\|$ and $\|\mathcal{S}\| \leq \gamma_2$, and F satisfies the conditions of Lemma J.4, then $L(x) := \|\mathcal{S}(\nabla_w F(x;w)) - \mathcal{S}(g)\|^2$ is $(A, B, 1/2)$-semi-smooth where $A = 2\tau\beta + 2\theta_2\gamma_2$ and $B = \tau^2\beta$.*

Proof. For simplicity, let $G(x) := \nabla_w F(x, w)$. Then the objective function $L(x)$ can be represented in the form of $L(x) = \|\mathcal{S}(G(x)) - \mathcal{S}(g)\|^2$. The statement that $L(x)$ is $(A, B, 1/2)$-semi-smooth is equivalent to

$$L(y) \leq L(x) + \langle \nabla L(x), y - x\rangle + B\|y - x\|^2 + A\|y - x\|L(x)^{1/2}.$$

Define

$$\begin{aligned}
\mathcal{A}_1 :=\ & \|\mathcal{S}(G(y))\|^2 - \|\mathcal{S}(G(x))\|^2 \\
& + 2\langle \mathcal{S}(G(x)) - \mathcal{S}(G(y)), \mathcal{S}(g)\rangle, \\
\mathcal{A}_2 :=\ & \langle \nabla L(x), x - y\rangle.
\end{aligned}$$

$\mathcal{A}_1$ can be bounded as

$$\begin{aligned}
\mathcal{A}_1 =\ & \langle \mathcal{S}(G(y)) - \mathcal{S}(G(x)), \\
& \quad \mathcal{S}(G(y)) + \mathcal{S}(G(x)) - 2\mathcal{S}(g)\rangle \\
\leq\ & \|\mathcal{S}(G(y)) - \mathcal{S}(G(x))\| \\
& \cdot \|\mathcal{S}(G(y)) + \mathcal{S}(G(x)) - 2\mathcal{S}(g)\| \\
\leq\ & \|\mathcal{S}(G(y)) - \mathcal{S}(G(x))\| \cdot \|\mathcal{S}(G(y)) - \mathcal{S}(G(x))\| \\
& + \|\mathcal{S}(G(y)) - \mathcal{S}(G(x))\| \cdot 2\|\mathcal{S}(G(x)) - \mathcal{S}(g)\| \\
\leq\ & \tau\|G(y) - G(x)\| \cdot (\tau\|G(y) - G(x)\| + 2L(x)^{1/2}) \\
=\ & \tau^2\|G(y) - G(x)\|^2 + 2\tau\|G(y) - G(x)\| \cdot L(x)^{1/2} \\
\leq\ & \tau^2\beta\|y - x\|^2 + 2\tau\beta\|y - x\|L(x)^{1/2}.
\end{aligned}$$

$\mathcal{A}_2$ can be bounded as

$$\begin{aligned}
\mathcal{A}_2 \leq\ & \|\nabla L(x)\| \cdot \|x - y\| \\
=\ & \|2(\nabla_x G(x))^\top \cdot (\nabla_u \mathcal{S}(u)\big|_{u=G(x)})^\top \\
& \cdot (\mathcal{S}(G(x)) - \mathcal{S}(g))\| \cdot \|x - y\| \\
\leq\ & 2\|\Phi(x)\| \cdot \big\|\nabla_u \mathcal{S}(u)\big|_{u=G(x)}\big\| \cdot \|\mathcal{S}(G(x)) - \mathcal{S}(g)\| \\
& \cdot \|x - y\| \\
\leq\ & 2 \cdot \theta_2 \cdot \gamma_{\mathcal{S}} \cdot \|\mathcal{S}(G(x)) - \mathcal{S}(g)\| \cdot \|x - y\| \\
=\ & 2 \cdot \theta_2 \cdot \gamma_{\mathcal{S}} \cdot L(x)^{1/2} \cdot \|x - y\|.
\end{aligned}$$

Let $A = 2\tau\beta + 2\theta_2\gamma_{\mathcal{S}}$, $B = \tau^2\beta$, and $R = B\|y - x\|^2 + A\|y - x\|L(x)^{1/2}$. Combining the upper bound for $\mathcal{A}_1$ and $\mathcal{A}_2$, we conclude that

$$\begin{aligned}
R \geq\ & \mathcal{A}_1 + \mathcal{A}_2 \\
=\ & \|\mathcal{S}(G(y))\|^2 - \|\mathcal{S}(G(x))\|^2 \\
& + 2\langle \mathcal{S}(G(x)) - \mathcal{S}(G(y)), \mathcal{S}(g)\rangle + \langle \nabla L(x), x - y\rangle \\
=\ & \|\mathcal{S}(G(y)) - \mathcal{S}(g)\|^2 - \|\mathcal{S}(G(x)) - \mathcal{S}(g)\|^2 \\
& + \langle \nabla L(x), x - y\rangle \\
=\ & L(y) - L(x) - \langle \nabla L(x), y - x\rangle.
\end{aligned}$$

Hence,

$$\begin{aligned}L(y) \leq & \; L(x) + \langle \nabla L(x), y - x\rangle + B\|y-x\|^2 \\ & + A\|y-x\|L(x)^{1/2}.\end{aligned}$$

□

M.2. What Sketching Implies Non-critical Point

Lemma M.4. *If the sketching mapping $\mathcal{S}$ satisfies Assumption M.2 and F satisfies Assumption J.3, then $L(x) := \|\mathcal{S}(\nabla_w F(x,w)) - \mathcal{S}(g)\|^2$ is $(2\theta_1\gamma_1, 2\theta_2\gamma_2)$-non-critical-point.*

Proof. Let $G(x) := \nabla_w F(x; w)$. Notice that the norm of $\nabla L(x)$ can be bounded as

$$\begin{aligned} & \|\nabla L(x)\| \\ = & \; \Big\|\nabla_x \|\mathcal{S}(G(x)) - \mathcal{S}(g)\|^2\Big\| \\ = & \; \Big\|2(\nabla_x G(x))^\top \cdot (\nabla_u \mathcal{S}(u)\big|_{u=G(x)})^\top \cdot (\mathcal{S}(G(x)) - \mathcal{S}(g))\Big\| \\ = & \; \big\|2\Phi(x,w) \cdot \mathcal{S}^\top \cdot (\mathcal{S}(G(x)) - \mathcal{S}(g))\big\| \end{aligned}$$

Hence we conclude that

$$(2\theta_1\gamma_1)^2 L(x) \leq \|\nabla L(x)\|^2 \leq (2\theta_2\gamma_2)^2 L(x).$$

□