

**ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ
ՄԱԹԵՄԱՏԻԿԱՅԻ ԵՎ ՄԵԽԱՆԻԿԱՅԻ ՖԱԿՈՒԼՏԵՏ**

**Հավանականության տեսության և վիճակագրության
ամբիոն**

**ԿԻՐԱՌԱԿԱՆ ՎԻՃԱԿԱԳՐՈՒԹՅՈՒՆ ԵՎ ՏՎՅԱԼՆԵՐԻ
ԳԻՏՈՒԹՅՈՒՆ ԿՐԹԱԿԱՆ ԾՐԱԳԻՐ**

ՄԱՐԱՆՁՅԱՆ ԱՐՏԱՎԱԶԴ ԱՐՄԵՆԻ

ՄԱԳԻՍՏՐՈՍԱԿԱՆ ԹԵԶ

ՏԵՂԱՅՆԱԿԱՆ ՈՒՍՈՒՑՄԱՆ ՄԵԹՈԴՆԵՐԻ ՄԱՍԻՆ

*«Վիճակագրություն» մասնագիտությամբ վիճակագրության
մագիստրոսի որակավորման աստիճանի հայցման համար*

ԵՐԵՎԱՆ 2023

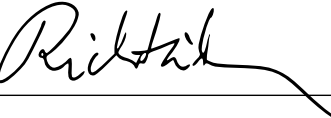
Ուսանող՝

ստորագրություն

Մարանջյան Արտավազդ

ազգանուն, անուն

Գիտական ղեկավար՝



ստորագրություն

Ֆ.մ.գ.դ., պրոֆեսոր Պիտեր Ռիստարիկ

գիտ. աստիճան, կոչում, ազգանուն, անուն

«Թուլյատրել Պաշտպանության»

Ծրագրի ղեկավար՝

ստորագրություն

Ֆ.մ.գ.դ., ասիստենտ Կարեն Բեռյան

գիտ. աստիճան, կոչում, ազգանուն, անուն

« 24 » 05 2023թ.

On local training methods*

Տեղայնական ուսուցման մեթոդների մասին

Abstract

We study a class of distributed optimization algorithms that aim to alleviate high communication costs by allowing clients to perform multiple local gradient-type training steps prior to communication. While methods of this type have been studied for about a decade, the empirically observed acceleration properties of local training eluded all attempts at theoretical understanding. In a recent breakthrough, [Mishchenko et al. \(2022\)](#) proved that local training, when properly executed, leads to provable communication acceleration, which holds in the strongly convex regime without relying on any data similarity assumptions. However, their method [ProxSkip](#) requires all clients to take the same number of local training steps in each communication round. Inspired by a common sense intuition, we start our investigation by conjecturing that clients with “less important” data should be able to get away with fewer local training steps without impacting the overall communication complexity of the method. It turns out that this intuition is correct: we managed to redesign the original [ProxSkip](#) method to achieve this. In particular, we prove that our modified method, for which we coin the name [GradSkip](#), converges linearly under the same assumptions and has the same accelerated communication complexity, while the number of local gradient steps can be reduced relative to a local condition number. We further generalize our method by extending the randomness of probabilistic alternations to arbitrary unbiased compression operators and considering a generic proximable regularizer. This generalization, which we call [GradSkip+](#), recovers several related methods in the literature as special cases. Finally, we present an empirical study on carefully designed toy problems that confirm our theoretical claims.

*This thesis is based on [Maranjyan et al. \(2022\)](#).

Contents

1	Introduction	6
1.1	Communication costs	7
1.2	Statistical heterogeneity	9
1.3	System heterogeneity	10
2	Contributions	11
2.1	GradSkip: efficient gradient skipping algorithm	11
2.2	GradSkip+: general GradSkip method	12
2.3	VR-GradSkip+: variance reduced GradSkip+	12
3	GradSkip	13
3.1	Algorithm structure	13
3.2	Reduced local computation	15
3.3	Convergence theory	17
4	GradSkip+	20
4.1	Algorithm description	22
4.2	Special Cases	22
4.3	Convergence theory	24
5	VR-GradSkip+	26
5.1	Algorithm description	26
5.2	Special Cases	28
5.3	Convergence theory	28
6	Experiments	29
6.1	Number of gradient computations: GradSkip vs ProxSkip	29
6.2	Real Dataset	32

Bibliography	33
A Limitations and Future Work	41
B Proofs for Section 3 (GradSkip)	41
B.1 Proof of Lemma 3.1	41
B.2 Proof of Lemma 3.2	42
B.3 Proof of Theorem 3.5	43
B.4 Proof of Theorem 3.6	48
C Proofs for Section 5 (VR-GradSkip+)	49
D Proofs for Section 4 (GradSkip+)	54
D.1 Proof of Lemma 4.2	54
D.2 Proof of Theorem 4.5	54

1 Introduction

Federated Learning (FL) is an emerging distributed machine learning paradigm where diverse data holders or clients (e.g., smart watches, mobile devices, laptops, hospitals) collectively aim to train a single machine learning model without revealing local data to each other or the orchestrating central server (McMahan et al., 2017; Kairouz et al., 2019; Wang, 2021). Training such models amounts to solving federated optimization problems of the form

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

where d is the (typically large) number of parameters of the model $x \in \mathbb{R}^d$ we aim to train, n is the (potentially large) total number of devices in the federated environment. We denote by $f_i(x)$ the loss or risk associated with the data \mathcal{D}_i stored on client $i \in [n] := \{1, 2, \dots, n\}$. Formally, our goal is to minimize the overall loss/risk denoted by $f(x)$.

Due to their efficiency, *gradient-type methods* with its numerous extensions (Duchi et al., 2011; Zeiler, 2012; Ghadimi and Lan, 2013; Kingma and Ba, 2015; Schmidt et al., 2017; Qian et al., 2019; Gorbunov et al., 2020a) is by far the most dominant method for solving (1) in practice.

The simplest implementation of gradient descent for federated setup requires all workers $i \in [n]$ in each time step $t \geq 0$ to (i) compute local gradient $\nabla f_i(x_t)$ at the current model x_t , (ii) update the current global model x_t using locally computed gradient $\nabla f_i(x_t)$ via (2) with some step size $\gamma > 0$, (iii) average the updated local models

$\hat{x}_{i,t+1}$ via (3) to get the new global model x_{t+1} .

$$\hat{x}_{i,t+1} = x_t - \gamma \nabla f_i(x_t), \quad (2)$$

$$x_{t+1} = \frac{1}{n} \sum_{i=1}^n \hat{x}_{i,t+1}. \quad (3)$$

Distinguishing challenges that characterize FL as a separate distributed training setup and dictate adjustments to the training algorithm include *high communication costs*, *heterogeneous data distribution* and *system heterogeneity* across clients. Next, we briefly discuss these challenges and possible algorithmic workarounds.

1.1 Communication costs

It has been repeatedly observed and advocated that in federated optimization, communication costs dominate and can be a primary bottleneck because of slow and unreliable wireless links between clients and the central server (McMahan et al., 2017). The communication step (3) between clients cannot be taken out entirely as otherwise, all clients would keep training on local data only, resulting in a poor model due to limited local data.

A simple trick to reduce communication costs is to perform the costly synchronization step (3) infrequently, allowing multiple local gradient steps (2) in each communication round (Mangasarjan, 1995). This trick appears in the celebrated FedAvg algorithm of McMahan et al. (2016; 2017) and its further variations (Haddadpour and Mahdavi, 2019; Li et al., 2019a; Khaled et al., 2019a;b; Karimireddy et al., 2020; Horváth et al., 2022) under the name of *local gradient methods*. However, until very recently, theoretical guarantees on the convergence rates of local gradient methods were worse than the rate of classical gradient descent, which synchro-

nizes after every gradient step.

In a recent line of works (Mishchenko et al., 2022; Malinovsky et al., 2022; Condat and Richtárik, 2022; Sadiev et al., 2022), initiated by Mishchenko et al. (2022), a novel local gradient method, called **ProxSkip**, was proposed which performs a *random number* of local gradient steps before each communication (alternation between local training and synchronization is probabilistic) and guarantees strong communication acceleration properties. First, they reformulate the problem (1) into an equivalent regularized consensus problem of the form

$$\min_{x_1, \dots, x_n \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n f_i(x_i) + \psi(x_1, \dots, x_n) \right\}, \quad (4)$$

$$\psi(x_1, \dots, x_n) := \begin{cases} 0, & \text{if } x_1 = \dots = x_n \\ +\infty, & \text{otherwise} \end{cases}, \quad (5)$$

where communication between the clients and averaging local models x_1, \dots, x_n is encoded as taking the proximal step with respect to ψ , i.e., $\text{prox}_\psi([x_1 \dots x_n]^\top) = [\bar{x} \dots \bar{x}]^\top$, where $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$. With this reformulation, **ProxSkip** method of Mishchenko et al. (2022) performs the proximal (equivalently averaging) step with small probability $p = 1/\sqrt{\kappa}$, where κ is the condition number of the problem. Then the key result of the method for smooth and strongly convex setup is $\mathcal{O}(\kappa \log 1/\epsilon)$ iteration complexity with $\mathcal{O}(\sqrt{\kappa} \log 1/\epsilon)$ communication rounds to achieve $\epsilon > 0$ accuracy. Follow-up works extend the method to variance reduced gradient methods (Malinovsky et al., 2022), randomized application of proximal operator (Condat and Richtárik, 2022), and accelerated primal-dual algorithms (Sadiev et al., 2022). Our work was inspired by development of this new generation of local gradient methods, which we detail shortly.

An orthogonal approach utilizes communication compression

strategies on the information that is transferred. Informally, instead of communicating full precision models infrequently we might communicate compressed version of the local model in each iteration via an application of lossy compression operators. Such strategies include sparsification (Alistarh et al., 2018; Mishchenko et al., 2020; Wang et al., 2018), quantization (Alistarh et al., 2017; Sun et al., 2019; Wang et al., 2022), sketching (Hanzely et al., 2018; Safaryan et al., 2021) and low-rank approximation (Vogels et al., 2019).

Our work contributes to the first approach to handling high communication costs that is less understood in theory and, at the same time, immensely popular in the practice of FL.

1.2 Statistical heterogeneity

Because of the decentralized nature of the training data, distributions of local datasets can vary from client to client. This heterogeneity in data distributions poses an additional challenge since allowing multiple local steps would make the local models deviate from each other, an issue widely known as *client drift*. On the other hand, if training datasets are identical across the clients (commonly referred to as homogeneous setup), then the mentioned drifting issue disappears and the training can be done without any communication whatsoever. Now, if we interpolate between these two extremes, then under some data similarity conditions (which typically expressed as gradient similarity conditions) multiple local gradient steps should be useful. In fact, initial theoretical guarantees of local gradient methods utilize such assumptions (Haddadpour and Mahdavi, 2019; Yu et al., 2019; Li et al., 2019b; 2020).

In the fully heterogeneous setup, client drift reduction tech-

niques were designed and analyzed to mitigate the adverse effect of local model deviations ([Karimireddy et al., 2020](#); [Gorbunov et al., 2021](#)). A very close analogy is variance reduction techniques called error feedback mechanisms for the compression noise added to lessen the number of bits required to transfer ([Condat et al., 2022](#)).

1.3 System heterogeneity

Lastly, system heterogeneity refers to the diversity of clients in terms of their computation capabilities or the amount of resources they are willing to use during the training. In a typical FL setup, all participating clients must perform the same amount of local gradient steps before each communication. Consequently, a highly heterogeneous cluster of devices results in significant and unexpected delays due to slow clients or stragglers.

One approach addressing system heterogeneity or dealing with slow clients is client selection strategies ([Luo et al., 2021](#); [Reisizadeh et al., 2020](#); [Wang and Joshi, 2019](#)). Basically, client sampling can be organized in such a way that slow clients do not delay the global synchronization and clients with similar computational capabilities are sampled in each communication round.

In contrast to the above strategy, we propose that clients perform as few or as many local steps as their local resources allow. In other words, we consider the full participation setup where each client decides how much local computation to perform before communication. Informally, slow clients do less local work than fast clients, and during the synchronization of local trained models, the slowdown caused by the stragglers will be minimized.

2 Contributions

We now briefly summarize key contributions of our work.

2.1 GradSkip: efficient gradient skipping algorithm

We design a new local gradient-type method for distributed optimization with communication and computation constraints. The proposed **GradSkip** method (see Algorithm 1) is an extension of recently developed **ProxSkip** method (Mishchenko et al., 2022), which was the first method showing communication acceleration property of performing multiple local steps without any data similarity assumptions. Our **GradSkip** method inherits the same accelerated communication complexity from **ProxSkip**, while further improving computational complexity allowing clients to terminate their local gradient computations independently from each other.

The key technical novelty of the proposed algorithm is the construction of auxiliary shifts $\hat{h}_{i,t}$ to handle gradient skipping for each client $i \in [n]$. **GradSkip** also maintains shifts $h_{i,t}$ initially introduced in **ProxSkip** to handle communication skipping across the clients. We prove that **GradSkip** converges linearly in strongly convex and smooth setup, has the same $\mathcal{O}(\sqrt{\kappa_{\max}} \log 1/\epsilon)$ accelerated communication complexity as **ProxSkip**, and requires clients to compute (in expectation) at most $\min(\kappa_i, \sqrt{\kappa_{\max}})$ local gradients in each communication round (see Theorem 3.6), where κ_i is the condition number for client $i \in [n]$ and $\kappa_{\max} = \max_i \kappa_i$. Thus, for **GradSkip**, clients with well-conditioned problems $\kappa_i < \sqrt{\kappa_{\max}}$ perform much less local work to achieve the same convergence rate of **ProxSkip**, which assumes $\sqrt{\kappa_{\max}}$ local steps on average for all clients.

2.2 GradSkip+: general GradSkip method

Next, we generalize the construction and the analysis of [GradSkip](#) by extending it in two directions: handling optimization problems with arbitrary proximable regularizer and incorporating general randomization procedures using unbiased compression operators with custom variance bounds. With such enhancements, we propose our second method, [GradSkip+](#) (see [Algorithm 2](#)), which recovers several methods in the literature as a special case, including the standard proximal gradient descent ([ProxGD](#)), [ProxSkip](#) ([Mishchenko et al., 2022](#)), [RandProx-FB](#) ([Condat and Richtárik, 2022](#)) and [GradSkip](#).

2.3 VR-GradSkip+: variance reduced GradSkip+

Finally, we generalize [GradSkip+](#) by combining it with recently developed [ProxSkip-VR](#) method ([Malinovsky et al., 2022](#)). [ProxSkip-VR](#) reduces computational complexity by allowing computationally cheaper stochastic gradient estimators instead of full batch gradients. This approach of reducing computational complexity is blind to statistical heterogeneity and is entirely orthogonal to our approach of reducing computational complexity in [GradSkip](#). So the generalization is quite natural and we propose our most general method, [VR-GradSkip+](#) (see [Algorithm 3](#)), which is both a generalization of [GradSkip+](#) and [ProxSkip-VR](#).

3 GradSkip

In this section, we present our first algorithm, **GradSkip**, and discuss its benefits in detail. Later we will generalize it, unifying several other methods as special cases. Recall that our target is to address three challenges in FL mentioned in the introductory part, which are (i) reduction in communication cost via infrequent synchronization of local models, (ii) statistical or data heterogeneity, and (iii) reduction in computational cost via limiting local gradient calls based on the local subproblem. We now describe all the steps of the algorithm and how it handles these three challenges.

3.1 Algorithm structure

For the sake of presentation, we describe the progress of the algorithm using two variables $x_{i,t}, \hat{x}_{i,t}$ for the local models and two variables $h_{i,t}, \hat{h}_{i,t}$ for the local gradient shifts. Essentially, we want to maintain two variables for the local models since clients get synchronized infrequently. The shifts $h_{i,t}$ are designed to reduce the client drift caused by the statistical heterogeneity. Finally, we introduce an auxiliary shifts $\hat{h}_{i,t}$ to take care of the different number of local steps. The **GradSkip** method is formally presented in Algorithm 1.

As an initialization step, we choose probability $p > 0$ to control communication rounds, probabilities $q_i > 0$ for each client $i \in [n]$ to control local gradient steps and initial control variates (or shifts) $h_{i,0} \in \mathbb{R}^d$ to control the client drift. Besides, we fix the stepsize $\gamma > 0$ and assume that all clients commence with the same local model, namely $x_{1,0} = \dots = x_{n,0} \in \mathbb{R}^d$. Then, each iteration of the method

Algorithm 1 GradSkip

- 1: **Input:** stepsize $\gamma > 0$, synchronization probability p , probabilities $q_i > 0$ controlling local steps, initial local iterates $x_{1,0} = \dots = x_{n,0} \in \mathbb{R}^d$, initial shifts $h_{1,0}, \dots, h_{n,0} \in \mathbb{R}^d$, total number of iterations $T \geq 1$
 - 2: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 3: **server:** Flip a coin $\theta_t \in \{0, 1\}$ with $\text{Prob}(\theta_t = 1) = p$ \diamond Decide when to skip communication
 - 4: **for all devices** $i \in [n]$ **in parallel do**
 - 5: Flip a coin $\eta_{i,t} \in \{0, 1\}$ with $\text{Prob}(\eta_{i,t} = 1) = q_i$ \diamond Decide when to skip gradient steps (see Lemma 3.1)
 - 6: $\hat{h}_{i,t+1} = \eta_{i,t} h_{i,t} + (1 - \eta_{i,t}) \nabla f_i(x_{i,t})$ \diamond Update the local auxiliary shifts $\hat{h}_{i,t}$
 - 7: $\hat{x}_{i,t+1} = x_{i,t} - \gamma(\nabla f_i(x_{i,t}) - \hat{h}_{i,t+1})$ \diamond Update the local auxiliary iterate $\hat{x}_{i,t}$ via shifted gradient step
 - 8: **if** $\theta_t = 1$ **then**
 - 9: $x_{i,t+1} = \frac{1}{n} \sum_{j=1}^n (\hat{x}_{j,t+1} - \frac{\gamma}{p} \hat{h}_{j,t+1})$ \diamond Average shifted iterates, but only very rarely!
 - 10: **else**
 - 11: $x_{i,t+1} = \hat{x}_{i,t+1}$ \diamond Skip communication!
 - 12: **end if**
 - 13: $h_{i,t+1} = \hat{h}_{i,t+1} + \frac{p}{\gamma}(x_{i,t+1} - \hat{x}_{i,t+1})$ \diamond Update the local shifts $h_{i,t}$
 - 14: **end for**
 - 15: **end for**
-

comprises two stages, the local stage and the communication stage, operating probabilistically. Specifically, the probabilistic nature of these stages is the following. The local stage requires computation only with some predefined probability; otherwise, the stage is void. Similarly, the communication stage requires synchronization between all clients only with probability p ; otherwise, the stage is void.

In the local stage (lines 5–7), all clients $i \in [n]$ in parallel update their local variables $(\hat{x}_{i,t+1}, \hat{h}_{i,t+1})$ using values $(x_{i,t}, h_{i,t})$ from pre-

vious iterate either by computing the local gradient $\nabla f_i(x_{i,t})$ or by just copying the previous values. Afterwards, in the communication stage (lines 8–13), all clients in parallel update their local variables $(x_{i,t+1}, h_{i,t+1})$ from $(\hat{x}_{i,t+1}, \hat{h}_{i,t+1})$ by either averaging across the clients or copying previous values.

3.2 Reduced local computation

Clearly, communication costs are reduced as the averaging step occurs only when $\theta_t = 1$ with probability p of our choice. However, it is not directly apparent how the computational costs are reduced during the local stage. Indeed, both options $\eta_{i,t} = 1$ and $\eta_{i,t} = 0$ involve the expression $\nabla f_i(x_{i,t})$ as if local gradients need to be evaluated in every iteration. As we show in the following lemma, this is not the case.

Lemma 3.1 (Fake local steps). *Suppose that Algorithm 1 does not communicate for $\tau \geq 1$ consecutive iterates, i.e., $\theta_t = \theta_{t+1} = \dots = \theta_{t+\tau-1} = 0$ for some fixed $t \geq 0$. Besides, let for some client $i \in [n]$ we have $\eta_{i,t} = 0$. Then, regardless of the coin tosses $\{\eta_{i,t+j}\}_{j=1}^{\tau}$, client i does fake local steps without any gradient computation in τ iterates. Formally, for all $j = 1, 2, \dots, \tau + 1$, we have*

$$\hat{x}_{i,t+j} = x_{i,t+j} = x_{i,t}, \quad (6)$$

$$\hat{h}_{i,t+j} = h_{i,t+j} = h_{i,t} = \nabla f_i(x_{i,t}). \quad (7)$$

Let us reformulate the above lemma. During the local stage of **GradSkip**, when clients do not communicate with server, i^{th} client terminates its local gradient steps once the local coin toss $\eta_{i,t} = 0$. Thus, smaller probability q_i implies sooner coin toss $\eta_{i,t} = 0$ in expectation, hence, less amount of local computation for client i .

Therefore, we can relax computational requirements of clients by adjusting these probabilities q_i and controlling the amount of local gradient computations.

Next, let us find out how the expected number of local gradient steps depends on probabilities p and q_i . Let Θ and H_i be random variables representing the number of coin tosses (Bernoulli trials) until the first occurrence of $\theta_t = 1$ and $\eta_{i,t} = 0$ respectively. Equivalently, $\Theta \sim \text{Geo}(p)$ is a geometric random variable with parameter p , and $H_i \sim \text{Geo}(1 - q_i)$ are geometric random variables with parameter $1 - q_i$ for $i \in [n]$. Notice that, within one communication round, i^{th} client performs $\min(\Theta, H_i)$ number of local gradient computations, which is again a geometric random variable with parameter $1 - (1 - (1 - q_i))(1 - p) = 1 - q_i(1 - p)$. Therefore, the expected number of local gradient steps is $\mathbb{E}[\min(\Theta, H_i)] = 1/(1 - q_i(1 - p))$. Let us formulate this observation as a separate lemma.

Lemma 3.2 (Expected number of local steps). *The expected number of local gradient computations in each communication round of GradSkip is $\frac{1}{(1 - q_i(1 - p))}$ for all clients $i \in [n]$.*

Notice that, in the special case of $q_i = 1$ for all $i \in [n]$, GradSkip recovers Scaffnew method of Mishchenko et al. (2022). However, as we will show, we can choose probabilities q_i smaller, reducing computational complexity and obtaining the same convergence rate as Scaffnew.

Remark 3.3 (System Heterogeneity). From this discussion we conclude that GradSkip can also address system or device heterogeneity. In particular, probabilities $\{q_i\}_{i=1}^n$ can be assigned to clients in accordance with their local computational resources; slow clients with scarce compute power should get small q_i , while faster clients

with rich resources should get bigger $q_i \leq 1$.

3.3 Convergence theory

Now that we explained the structure and computational benefits of the algorithm, let us proceed to the theoretical guarantees. We consider the same strongly convex and smooth setup as considered by [Mishchenko et al. \(2022\)](#) for the distributed case.

Assumption 3.4. All functions $f_i(x)$ are strongly convex with parameter $\mu > 0$ and have Lipschitz continuous gradients with Lipschitz constants $L_i > 0$, i.e., for all $i \in [n]$ and any $x, y \in \mathbb{R}^d$ we have

$$\frac{\mu}{2}\|x - y\|^2 \leq D_{f_i}(x, y) \leq \frac{L_i}{2}\|x - y\|^2, \quad (8)$$

where $D_{f_i}(x, y) := f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle$ is the Bregman divergence associated with f_i at points $x, y \in \mathbb{R}^d$.

We present Lyapunov-type analysis to prove the convergence, which is a very common approach for iterative algorithms. Consider the Lyapunov function

$$\Psi_t := \sum_{i=1}^n \|x_{i,t} - x_\star\|^2 + \frac{\gamma^2}{p^2} \sum_{i=1}^n \|h_{i,t} - h_{i,\star}\|^2, \quad (9)$$

where $\gamma > 0$ is the stepsize, x_\star is the (necessary) unique minimizer of $f(x)$ and $h_{i,\star} = \nabla f_i(x_\star)$ is the optimal gradient shift. As we show next, Ψ_t decreases at a linear rate.

Theorem 3.5. *Let Assumption 3.4 hold. If the stepsize satisfies $\gamma \leq \min_i \left\{ \frac{1}{L_i} \frac{p^2}{1 - q_i(1 - p^2)} \right\}$ and probabilities are chosen so that $0 < p, q_i \leq 1$, then the iterates of **GradSkip** (Algorithm 1) satisfy*

$$\mathbb{E}[\Psi_t] \leq (1 - \rho)^t \Psi_0, \quad (10)$$

for all $t \geq 1$ with $\rho := \min \{ \gamma\mu, 1 - q_{\max}(1 - p^2) \} > 0$.

Let us comment on this result.

- The first and immediate observation from the above result is that, with a proper stepsize choice, **GradSkip** converges linearly for any choice of probabilities p and q_i from $(0, 1]$.

- Furthermore, by choosing all probabilities $q_i = 1$ we get the same rate of **Scaffnew** with $\rho = \min\{\gamma\mu, p^2\}$ (see Theorem 3.6 in (Mishchenko et al., 2022)). If we further choose the largest admissible stepsize $\gamma = 1/L_{\max}$ and the optimal synchronization probability $p = 1/\sqrt{\kappa_{\max}}$, we get $\mathcal{O}(\kappa_{\max} \log 1/\epsilon)$ iteration complexity, $\mathcal{O}(\sqrt{\kappa_{\max}} \log 1/\epsilon)$ accelerated communication complexity with $1/p = \sqrt{\kappa_{\max}}$ expected number of local steps in each communication round. Here, we used notation $\kappa_{\max} = \max_i \kappa_i$ where $\kappa_i = L_i/\mu$ is the condition number for client $i \in [n]$.

- Finally, exploiting smaller probabilities q_i , we can optimize computational complexity subject to the same communication complexity as **Scaffnew**. To do that, note that the largest possible stepsize that Theorem 3.5 allows is $\gamma = 1/L_{\max}$ as $\min_i \left\{ \frac{1}{L_i} \frac{p^2}{1-q_i(1-p^2)} \right\} \leq \min_i \frac{1}{L_i} \leq \frac{1}{L_{\max}}$. Hence, taking into account $\rho \leq \gamma\mu$, the best iteration complexity from the rate (10) is $\mathcal{O}(\kappa_{\max} \log 1/\epsilon)$, which can be obtained by choosing the probabilities appropriately as formalized in the following result.

Theorem 3.6 (Optimal parameter choices). *Let Assumption 3.4 hold and choose probabilities $q_i = \frac{1 - \frac{1}{\kappa_i}}{1 - \frac{1}{\kappa_{\max}}} \leq 1$ and $p = \frac{1}{\sqrt{\kappa_{\max}}}$. Then, with the largest admissible stepsize $\gamma = \frac{1}{L_{\max}}$, **GradSkip** enjoys the following properties:*

- (i) $\mathcal{O}(\kappa_{\max} \log \frac{1}{\epsilon})$ iteration complexity,
- (ii) $\mathcal{O}(\sqrt{\kappa_{\max}} \log \frac{1}{\epsilon})$ communication complexity,

(iii) for each client $i \in [n]$, the expected number of local gradient computations per communication round is

$$\frac{1}{1 - q_i(1 - p)} = \frac{\kappa_i(1 + \sqrt{\kappa_{\max}})}{\kappa_i + \sqrt{\kappa_{\max}}} \leq \min(\kappa_i, \sqrt{\kappa_{\max}}). \quad (11)$$

This result clearly quantifies the benefits of using smaller probabilities q_i . In particular, if the condition number κ_i of client i is smaller than $\sqrt{\kappa_{\max}}$, then within each communication round it does only κ_i number of local gradient steps. However, for a client having the maximal condition number (namely, clients $\arg \max_i \{\kappa_i\}$), the number of local gradient steps is $\sqrt{\kappa_{\max}}$, which is the same for **Scaffnew**. From this we conclude that, in terms of computational complexity, **GradSkip** is always better and can be $\mathcal{O}(n)$ times better than **Scaffnew** (Mishchenko et al., 2022).

4 GradSkip+

Here we aim to present a deeper understanding for GradSkip by extending it in two directions and designed our generic GradSkip+ method.

The first direction is the formulation of the optimization problem. As we discussed earlier, distributed optimization (1) with consensus constraints can be transformed into regularized optimization problem (4) in the lifted space. Thus, following Mishchenko et al. (2022), we consider the (lifted) problem¹

$$\min_{x \in \mathbb{R}^d} f(x) + \psi(x), \quad (12)$$

where $f(x)$ is strongly convex and smooth loss, while $\psi(x)$ is closed, proper and convex regularizer (e.g., see (5)). The requirement we impose on the regularizer is that the proximal operator of ψ is a single-valued function that can be computed.

The second extension in GradSkip+ is the generalization of the randomization procedure of probabilistic alternations in GradSkip by allowing arbitrary unbiased compression operators with certain bounds on the variance. Let us formally define the class of compressors we will be working with.

Definition 4.1 (Unbiased Compressors). For any positive semidefinite matrix $\Omega \succeq 0$, denote by $\mathbb{B}^d(\Omega)$ the class of (possibly randomized) unbiased compression operators $\mathcal{C}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[\mathcal{C}(x)] = x, \quad \mathbb{E}[\|(\mathbf{I} + \Omega)^{-1}\mathcal{C}(x)\|^2] \leq \|x\|_{(\mathbf{I} + \Omega)^{-1}}^2.$$

¹To be precise, the lifted problem is in \mathbb{R}^{nd} as we stack all local variables $x_1, \dots, x_n \in \mathbb{R}^d$ into one.

The class $\mathbb{B}^d(\Omega)$ is a generalization of commonly used class $\mathbb{B}^d(\omega)$ of unbiased compressors with variance bound $\mathbb{E} \left[\|\mathcal{C}(x)\|^2 \right] \leq (1 + \omega)\|x\|^2$ for some scalar $\omega \geq 0$. Indeed, when the matrix $\Omega = \omega\mathbf{I}$, then $\mathbb{B}^d(\omega\mathbf{I})$ coincides with $\mathbb{B}^d(\omega)$. Furthermore, the following inclusion holds:

Lemma 4.2. $\mathbb{B}^d(\Omega) \subseteq \mathbb{B}^d((1 + \lambda_{\max}(\Omega))^2 / (1 + \lambda_{\min}(\Omega)) - 1)$.

The purpose of this new variance bound with matrix parameter Ω is to introduce non-uniformity on the level of compression across different directions. For example, in the reformulation (4) each client controls $1/n$ portion of the directions and the level of compression. For example, consider compression operator $\mathcal{C}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ defined as

$$\mathcal{C}(x)_j = \begin{cases} \frac{x_j}{p_j} & \text{with probability } p_j, \\ 0 & \text{with probability } 1 - p_j, \end{cases} \quad (13)$$

for all coordinates $j \in [d]$ and for any $x \in \mathbb{R}^d$, where $p_j \in (0, 1]$ are given probabilities. Then, it is easy to check that $\mathcal{C} \in \mathbb{B}^d(\Omega)$ with diagonal matrix $\Omega = \text{Diag}(1/p_j - 1)$ having diagonal entries $1/p_j - 1 \geq 0$.

Now that we have finer control over the compression operator, we can make use of the granular smoothness information of the loss function f through the so-called smoothness matrices (Qu and Richtárik, 2016a;b).

Definition 4.3 (Matrix Smoothness). A differentiable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is called \mathbf{L} -smooth with some symmetric and positive definite matrix $\mathbf{L} \succ 0$ if

$$D_f(x, y) \leq \frac{1}{2} \|x - y\|_{\mathbf{L}}^2, \quad \forall x, y \in \mathbb{R}^d. \quad (14)$$

The standard L -smoothness condition in (8) with scalar $L > 0$ is obtained as a special case of (14) for matrices of the form $\mathbf{L} = L\mathbf{I}$,

where \mathbf{I} is the identity matrix. The notion of matrix smoothness provides much more information about the function than mere scalar smoothness. In particular, if f is \mathbf{L} -smooth, then it is also $\lambda_{\max}(\mathbf{L})$ -smooth due to the relation $\mathbf{L} \preceq \lambda_{\max}(\mathbf{L})\mathbf{I}$. Smoothness matrices have been used in the literature of randomized coordinate descent (Richtárik and Takáč, 2016; Hanzely and Richtárik, 2019a;b) and distributed optimization (Safaryan et al., 2021; Wang et al., 2022).

4.1 Algorithm description

Similar to **GradSkip**, we maintain two variables x_t, \hat{x}_t for the model, and two variables h_t, \hat{h}_t for the gradient shifts in **GradSkip+**. Initial values $x_0 \in \mathbb{R}^d$ and $h_0 \in \mathbb{R}^d$ can be chosen arbitrarily. In each iteration, **GradSkip+** first updates the auxiliary shift \hat{h}_{t+1} using the previous shift h_t and gradient $\nabla f(x_t)$ (line 4). This shift \hat{h}_{t+1} is then used to update the auxiliary iterate x_t via shifted gradient step (line 5). Then we estimate the proximal gradient \hat{g}_t (line 6) in order to update the main iterate x_{t+1} (line 7). Lastly, we complete the iteration by updating the main shift h_t (line 8). See Algorithm 2 for the formal steps.

4.2 Special Cases

Before we proceed to the theoretical results, let us consider a few special cases of **GradSkip+**.

Case 1 (ProxGD). If \mathcal{C}_ω is the identity compressor (i.e., $\omega = 0$), then Algorithm 2 reduces to the **ProxGD** algorithm as $x_{t+1} = \text{prox}_{\gamma\psi}(\hat{x}_{t+1} - \gamma\hat{h}_{t+1}) = \text{prox}_{\gamma\psi}(x_t - \gamma\nabla f(x_t))$ for any choice of \mathcal{C}_Ω .

Case 2 (ProxSkip). Let \mathcal{C}_Ω be the identity compressor (i.e., $\Omega = \mathbf{I}$) and \mathcal{C}_ω be the Bernoulli compressor \mathcal{C}_p with parameter $p \in (0, 1]$

Algorithm 2 GradSkip+

- 1: **Parameters:** stepsize $\gamma > 0$, compressors $\mathcal{C}_\omega \in \mathbb{B}^d(\omega)$ and $\mathcal{C}_\Omega \in \mathbb{B}^d(\Omega)$.
 - 2: **Input:** initial iterate $x_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, number of iterations $T \geq 1$.
 - 3: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 4: $\hat{h}_{t+1} = \nabla f(x_t) - (\mathbf{I} + \Omega)^{-1} \mathcal{C}_\Omega (\nabla f(x_t) - h_t)$ \diamond Update the shift \hat{h}_t via shifted compression
 - 5: $\hat{x}_{t+1} = x_t - \gamma(\nabla f(x_t) - \hat{h}_{t+1})$ \diamond Update the iterate \hat{x}_t via shifted gradient step
 - 6: $\hat{g}_t = \frac{1}{\gamma(1+\omega)} \mathcal{C}_\omega \left(\hat{x}_{t+1} - \text{prox}_{\gamma(1+\omega)\psi} \left(\hat{x}_{t+1} - \gamma(1+\omega)\hat{h}_{t+1} \right) \right)$ \diamond Estimate the proximal gradient
 - 7: $x_{t+1} = \hat{x}_{t+1} - \gamma\hat{g}_t$ \diamond Update the main iterate x_t
 - 8: $h_{t+1} = \hat{h}_{t+1} + \frac{1}{\gamma(1+\omega)}(x_{t+1} - \hat{x}_{t+1})$ \diamond Update the main shift h_t
 - 9: **end for**
-

(note that here $\omega = 1/p - 1$). In this case, $\hat{h}_{t+1} \equiv h_t$ and x_{t+1} is either $\text{prox}_{\gamma/p\psi}(\hat{x}_{t+1} - \gamma/p h_t)$ (with probability p) or \hat{x}_{t+1} (with probability $1 - p$). Thus, we recover the ProxSkip algorithm.

Case 3 (RandProx-FB). Let \mathcal{C}_Ω be the identity compressor and $\mathcal{C}_\omega = \mathcal{R} \in \mathbb{B}^d(\omega)$. Then, after the following change of notation: $h_t = -u_t$, $\hat{g}_t = d_t/1+\omega_2$, the method is equivalent to RandProx-FB (Condat and Richtárik, 2022), which is a generalization of ProxSkip when additional smoothness information for the regularizer ψ is known².

Case 4 (GradSkip). Finally, we can specialize GradSkip+ to recover GradSkip. Consider the lifted space \mathbb{R}^{nd} where $x \in \mathbb{R}^{nd}$ represents the concatenations of models $x_1, \dots, x_n \in \mathbb{R}^d$ from all client's. The central example of an unbiased compression operator for that would be the probabilistic switching mechanism used in GradSkip, which is sometimes referred to as Bernoulli compressor: for any

²We do not consider smooth regularizers as our primary example of regularizer is the non-smooth consensus constraint (5).

given $p \in [0, 1]$, the compressor $\mathcal{C}_p^{nd}(x)$ outputs x/p (with probability p) or 0 (with probability $1 - p$) for any input vector $x \in \mathbb{R}^{nd}$. **GradSkip** employs one Bernoulli compressor \mathcal{C}_p^{nd} with parameter $p \in (0, 1]$ controlling communication rounds, and one Bernoulli compressor $\mathcal{C}_{q_i}^d$ with parameter $q_i \in (0, 1]$ for each client to control local gradient steps. Therefore, choosing $\mathcal{C}_\omega = \mathcal{C}_p^{nd}$ and $\mathcal{C}_\Omega = \mathcal{C}_{q_1}^d \times \dots \times \mathcal{C}_{q_n}^d$ in the lifted space \mathbb{R}^{nd} , **GradSkip+** reduces to **GradSkip**.

4.3 Convergence theory

We now present the convergence theory for **GradSkip+**, for which we replace the scalar smoothness Assumption 3.4 by matrix smoothness.

Assumption 4.4 (Convexity and smoothness). We assume that the loss function f is μ -strongly convex with positive $\mu > 0$ and \mathbf{L} -smooth with positive definite matrix $\mathbf{L} \succ 0$.

Analogous to (9), we analyze **GradSkip+** by consider the following Lyapunov function.

$$\Psi_t := \|x_t - x_\star\|^2 + \gamma^2(1 + \omega)^2 \|h_t - h_\star\|^2,$$

where $h_\star = \nabla f(x_\star)$. In the following theorem, we show the general linear convergence result.

Theorem 4.5. *Let Assumption 4.4 hold, $\mathcal{C}_\omega \in \mathbb{B}^d(\omega)$ and $\mathcal{C}_\Omega \in \mathbb{B}^d(\Omega)$ be the compression operators, and $\tilde{\Omega} := \mathbf{I} + \omega(\omega + 2)\Omega(\mathbf{I} + \Omega)^{-1}$. Then, if the stepsize $\gamma \leq \lambda_{\max}^{-1}(\mathbf{L}\tilde{\Omega})$, the iterates of **GradSkip+** (Algorithm 2) satisfy*

$$\mathbb{E}[\Psi_t] \leq (1 - \min\{\gamma\mu, \delta\})^t \Psi_0, \quad (15)$$

where $\delta = 1 - \frac{1}{1 + \lambda_{\min}(\Omega)} \left(1 - \frac{1}{(1 + \omega)^2}\right) \in [0, 1]$.

Let us discuss some implications from this rate. If we choose \mathcal{C}_Ω to be the identity compression (i.e., $\Omega = 0$), then **GradSkip+** reduces to **RandProx-FB** and we recover asymptotically the same rate with linear factor $(1 - \min\{\gamma\mu, 1/(1+\omega)^2\})$ (see Theorem 3 of [Condat and Richtárik \(2022\)](#)). If we further choose \mathcal{C}_ω to be the Bernoulli compression with parameter $p \in (0, 1]$, then $\omega = 1/p - 1$ and we get the rate of **ProxSkip**.

In order to recover the rate (10) of **GradSkip**, consider the lifted space \mathbb{R}^{nd} with reformulation (4)-(5) and objective function $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x_i)$, where $x_i \in \mathbb{R}^d$ and $x = (x_1, \dots, x_n) \in \mathbb{R}^{nd}$. From μ -strong convexity of each loss function f_i , we conclude that f is also μ -strongly convex. Regarding the smoothness condition, we have $L_i \mathbf{I} \in \mathbb{R}^{d \times d}$ smoothness matrices (equivalent to scalar L_i -smoothness) for each f_i , which implies that the overall loss function f has $\mathbf{L} = \text{Diag}(L_1 \mathbf{I}, \dots, L_n \mathbf{I}) \in \mathbb{R}^{nd \times nd}$ as a smoothness matrix. Furthermore, choosing Bernoulli compression operators $\mathcal{C}_\omega = \mathcal{C}_p^{nd}$ and $\mathcal{C}_\Omega = \mathcal{C}_{q_1}^d \times \dots \times \mathcal{C}_{q_n}^d$ in the lifted space \mathbb{R}^{nd} , we get $\omega = 1/p - 1$ and $\Omega = \text{Diag}(1/q_i - 1)$. It remains to plug all these expressions into Theorem 4.5 and recover Theorem 3.6. Indeed, $\lambda_{\min}(\Omega) = 1/q_{\max} - 1$ and, hence, $\delta = 1 - q_{\max}(1 - p^2)$. Lastly, Theorem 4.5 recovers the same stepsize bound as $\lambda_{\max}^{-1}(\mathbf{L}\tilde{\Omega}) = \min_i (L_i (1 + (1 - q_i) (1/p^2 - 1)))^{-1} = \min_i \left\{ \frac{1}{L_i} \frac{p^2}{1 - q_i(1 - p^2)} \right\}$.

5 VR-GradSkip+

Here we present our most general algorithm, [VR-GradSkip+](#).

Recently developed [ProxSkip-VR](#) method ([Malinovsky et al., 2022](#)) reduces computational complexity by allowing computationally cheaper stochastic gradient estimators instead of full batch gradients. This approach of reducing computational complexity is blind to statistical heterogeneity and is entirely orthogonal to our approach of reducing computational complexity in [GradSkip](#). It is natural to ask the following question.

Is it possible to combine these two methods ([ProxSkip-VR](#) and [GradSkip](#)) to achieve even better computational complexity?

We give an affirmative answer to the question by developing our most general [VR-GradSkip+](#) method.

5.1 Algorithm description

We get [VR-GradSkip+](#) from [GradSkip+](#) by replacing the gradient $\nabla f(x_t)$ by an unbiased estimator $g_t = \text{StochasticGradient}(x_t, f)$, see [Algorithm 3](#).

Our next assumption, initially introduced by [Gorbunov et al. \(2020a\)](#), postulates several parametric inequalities characterizing the behavior and ultimately the quality of a gradient estimator. Similar assumptions appeared later in ([Gorbunov et al., 2021](#); [2020b](#)).

Assumption 5.1. Let $\{x_t\}$ be the iterates produced by [VR-GradSkip+](#). We first assume unbiasedness of the stochastic gradi-

Algorithm 3 VR-GradSkip+

- 1: **Parameters:** stepsize $\gamma > 0$, compressors $\mathcal{C}_\omega \in \mathbb{B}^d(\omega)$ and $\mathcal{C}_\Omega \in \mathbb{B}^d(\Omega)$.
 - 2: **Input:** initial iterate $x_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, number of iterations $T \geq 1$.
 - 3: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 4: $g_t = \text{StochasticGradient}(x_t, f)$ ◇ Construct an unbiased estimator of $\nabla f(x_t)$
 - 5: $\hat{h}_{t+1} = g_t - (\mathbf{I} + \Omega)^{-1} \mathcal{C}_\Omega (g_t - h_t)$ ◇ Update the shift \hat{h}_t via shifted compression
 - 6: $\hat{x}_{t+1} = x_t - \gamma(g_t - \hat{h}_{t+1})$ ◇ Update the iterate \hat{x}_t via shifted stochastic gradient step
 - 7: $\hat{g}_t = \frac{1}{\gamma(1+\omega)} \mathcal{C}_\omega \left(\hat{x}_{t+1} - \text{prox}_{\gamma(1+\omega)\psi} \left(\hat{x}_{t+1} - \gamma(1+\omega)\hat{h}_{t+1} \right) \right)$ ◇ Estimate the proximal gradient
 - 8: $x_{t+1} = \hat{x}_{t+1} - \gamma\hat{g}_t$ ◇ Update the main iterate x_t
 - 9: $h_{t+1} = \hat{h}_{t+1} + \frac{1}{\gamma(1+\omega)}(x_{t+1} - \hat{x}_{t+1})$ ◇ Update the main shift h_t
 - 10: **end for**
-

ents g_t for all iterations $t \geq 0$, i.e.,

$$\mathbb{E}[g_t \mid x_t] = \nabla f(x_t). \quad (16)$$

Next, we assume that for some non-negative constants $A, B, C, \tilde{A}, \tilde{B}, \tilde{C}$, with $\tilde{B} < 1$, and non-negative sequence $\{\sigma_t\}_{t \geq 0}$ the following inequalities hold for all $t \geq 0$:

$$\mathbb{E}[\|g_t - \nabla f(x_\star)\|_{\mathbf{L}^{-1}}^2 \mid x_t] \leq 2AD_f(x_t, x_\star) + B\sigma_t + C, \quad (17)$$

$$\mathbb{E}[\sigma_{t+1} \mid x_t] \leq 2\tilde{A}D_f(x_t, x_\star) + \tilde{B}\sigma_t + \tilde{C}. \quad (18)$$

Assumption 5.1 covers a very large collection of gradient estimators, including an infinite variety of subsampling/minibatch estimators, gradient sparsification and quantization estimators, and their combinations; see (Gorbunov et al., 2020a) for examples. VR estimators are characterized by $C = \tilde{C} = 0$; most non-VR estimators by $\tilde{A} = \tilde{B} = \tilde{C} = B = 0$ and $C > 0$ (Gower et al., 2019).

5.2 Special Cases

Case 1 (GradSkip+). Consider the case when stochastic gradients are full batch gradients, i.e., $g_t = \nabla f(x_t)$ for all $t \geq 0$. Then Algorithm 3 reduces to GradSkip+.

Case 2 (ProxSkip-VR). To recover ProxSkip-VR from VR-GradSkip+, we need the same conditions we had for recovering ProxSkip from GradSkip+. That is, let \mathcal{C}_Ω be the identity compressor (i.e., $\Omega = \mathbf{I}$) and \mathcal{C}_ω be the Bernoulli compressor \mathcal{C}_p with parameter $p \in (0, 1]$ (note that here $\omega = 1/p - 1$). In this case, $\hat{h}_{t+1} \equiv h_t$ and x_{t+1} is either $\text{prox}_{\gamma/p\psi}(\hat{x}_{t+1} - \gamma/p h_t)$ (with probability p) or \hat{x}_{t+1} (with probability $1 - p$). Thus, we recover the ProxSkip-VR algorithm.

5.3 Convergence theory

Consider the Lyapunov function:

$$\Psi_t := \|x_t - x_*\|^2 + \gamma^2(1 + \omega)^2 \|h_t - h_*\|^2 + \gamma^2 W \sigma_t,$$

where $h_* = \nabla f(x_*)$.

Theorem 5.2. *Let Assumption 4.4 hold, and let g_t be a gradient estimator satisfying Assumption 5.1. Let $\mathcal{C}_\omega \in \mathbb{B}^d(\omega)$ and $\mathcal{C}_\Omega \in \mathbb{B}^d(\Omega)$ be the compression operators. If $B > 0$, choose any $W > \frac{\lambda_{\max}(\mathbf{L}\tilde{\Omega})B}{1-\tilde{B}}$ and then $\beta = 1 - \tilde{B} - \frac{\lambda_{\max}(\mathbf{L}\tilde{\Omega})B}{W} > 0$. In case of $B = 0$, set $W = 0$ and $\beta = \tilde{B}$. If the stepsize $\gamma \leq \frac{1}{A\lambda_{\max}(\mathbf{L}\tilde{\Omega}) + W\tilde{A}}$, then the iterates of VR-GradSkip+ (Algorithm 3) satisfy*

$$\mathbb{E}[\Psi_t] \leq (1 - \min(\gamma\mu, \delta, \beta))^t \Psi_0 + \gamma^2 \frac{\lambda_{\max}(\mathbf{L}\tilde{\Omega})C + W\tilde{C}}{\min(\gamma\mu, \delta, \beta)},$$

where

$$\delta = 1 - \frac{1}{1 + \lambda_{\min}(\Omega)} \left(1 - \frac{1}{(1 + \omega)^2} \right), \quad \tilde{\Omega} = \mathbf{I} + \omega(\omega + 2)\Omega(\mathbf{I} + \Omega)^{-1}. \quad (19)$$

6 Experiments

To test the performance of **GradSkip** and illustrate theoretical results, we use classical logistic regression problem. The loss function for this model has the following form:

$$f(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{m_i} \sum_{j=1}^{m_i} \log (1 + \exp (-b_{ij} a_{ij}^\top x)) + \frac{\lambda}{2} \|x\|^2,$$

where n is the number of clients, m_i is the number of data points per worker, $a_{ij} \in \mathbb{R}^d$ and $b_{ij} \in \{-1, +1\}$ are the data samples, and λ is the regularization parameter.

We conduct several experiments on artificially generated data and on the “*australian*” dataset from LibSVM library ([Chang and Lin, 2011](#)). All algorithms are implemented in Python using the package RAY ([Moritz et al., 2018](#)) to utilize parallelization. We run all algorithms using their theoretically optimal hyper-parameters (stepsize, probabilities).

6.1 Number of gradient computations: GradSkip vs ProxSkip

We compare **GradSkip** only to **ProxSkip** since **ProxSkip** has SOTA accelerated communication complexity. Although **ProxSkip-VR** has better computational complexity. The difference (in terms of computational complexity) between **VR-GradSkip+** over **ProxSkip-VR** and **GradSkip** over **ProxSkip** will be the same.

$$\sum_{i=1}^n \frac{\kappa_i (1 + \sqrt{\kappa_{\max}})}{\kappa_i + \sqrt{\kappa_{\max}}} \leq \sum_{i=1}^n \min \{ \kappa_i, \sqrt{\kappa_{\max}} \},$$

(see (11)), while for **ProxSkip** we have $n\sqrt{\kappa_{\max}}$. That is the gradient computation ratio of **ProxSkip** over **GradSkip** depends on the number

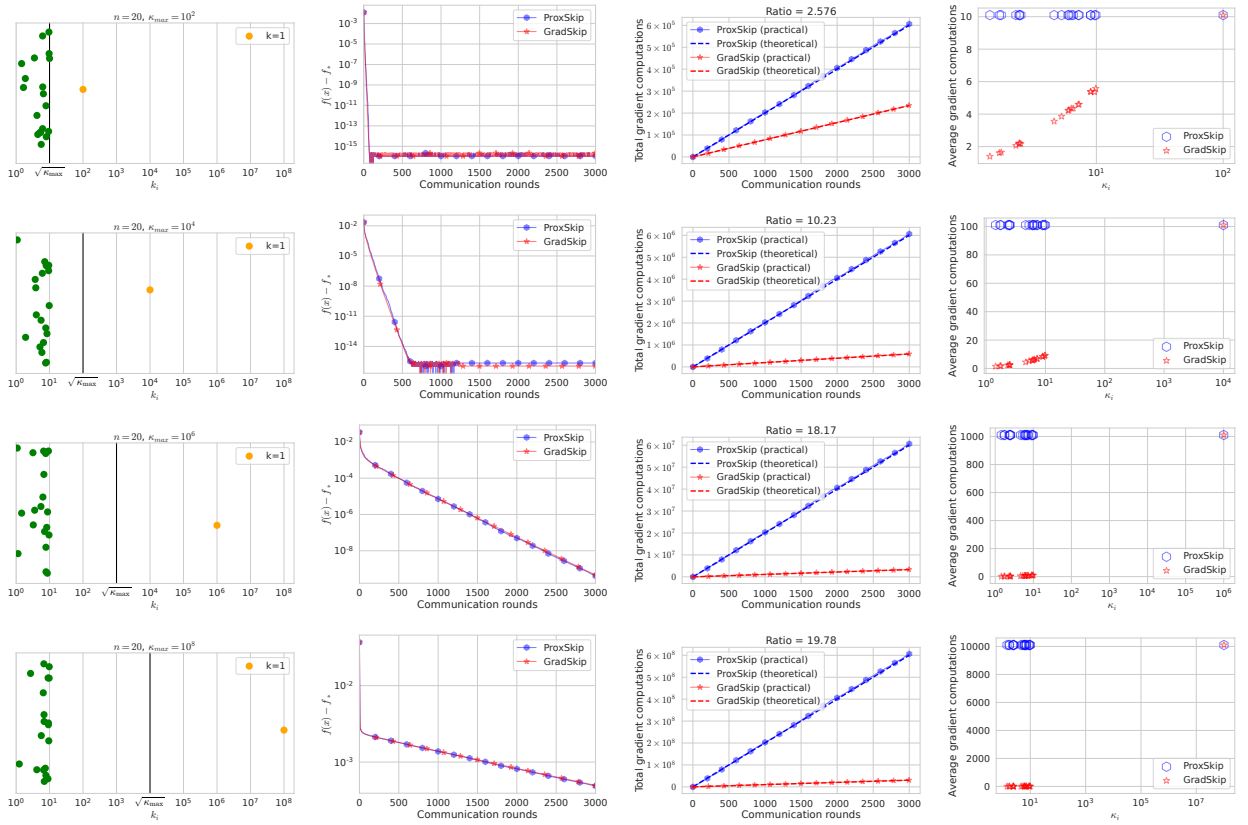


Figure 1: In the first column we show the condition numbers for devices. In the second column we show the convergence per communication rounds. In the third column we show theoretical and practical difference between number of gradient computations. In the last column we have the average gradient computations for each device having κ_i condition number, we see that for **GradSkip** the device with $\kappa_i = \kappa_{max}$ does the same number of gradient computations as all devices in **ProxSkip**.

of devices having $\kappa_i \geq \sqrt{\kappa_{max}}$ condition number. If there are $k \leq n$ such devices, then the gradient computation ratio of **ProxSkip** over **GradSkip** converges to $n/k \geq 1$ when $\kappa_{max} \rightarrow \infty$.

In our experiments we have only one device with ill-conditioned local problem ($k = 1$). And to show this convergence we artificially generate data for having control over the smoothness constants. We also set the regularization parameter $\lambda = 10^{-1} = \mu$.

In Figure 1, we have $n = 20$ devices. We set large $L_i = L_{max}$ for one device and for the rest we have $L_i \sim \text{Uniform}(0.1, 1)$. We can see that the convergence is the same for **GradSkip** and **ProxSkip**. Next,

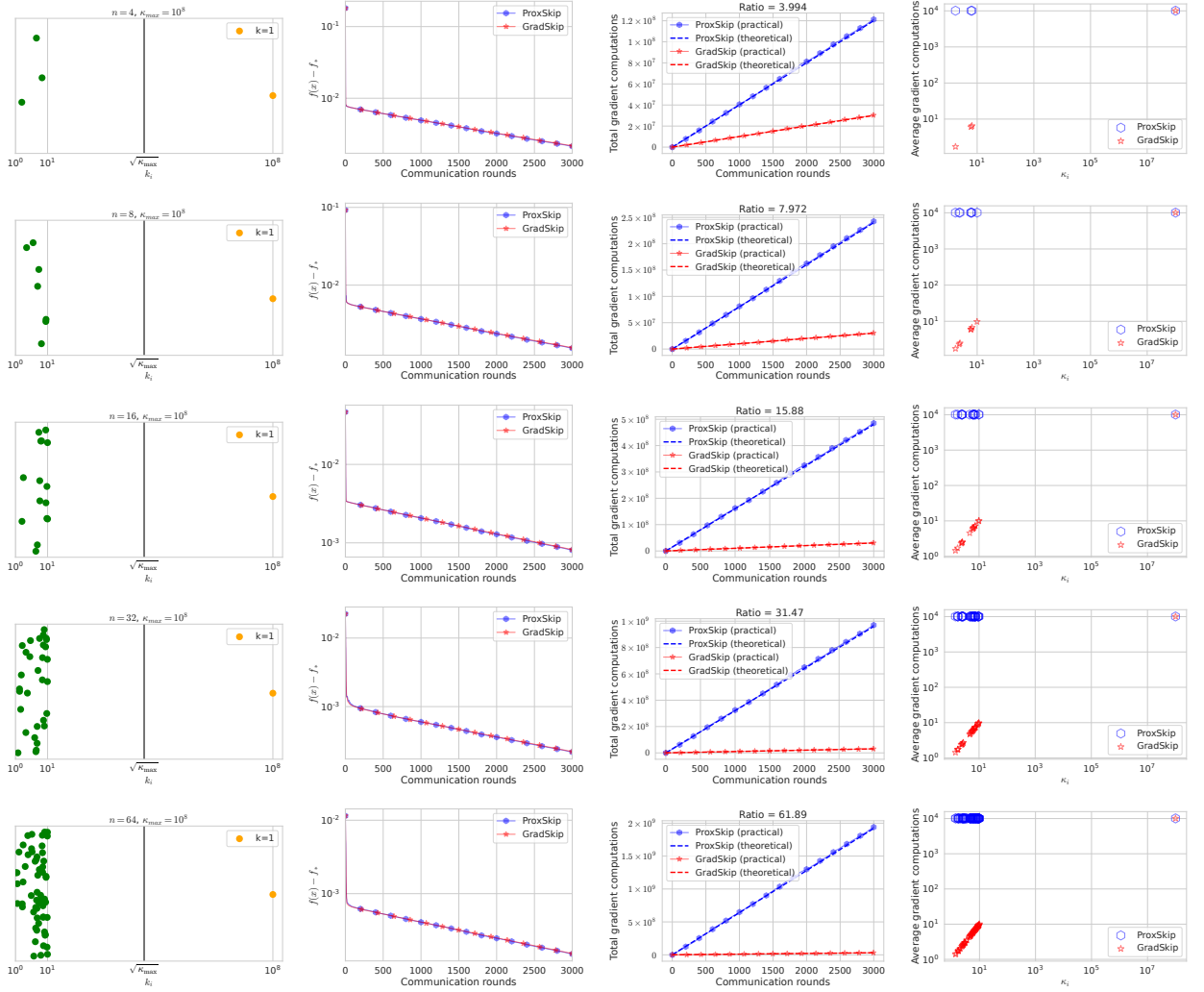


Figure 2: The columns have the same meaning as in Figure 1.

we increase L_{\max} in each row to show that the ratio indeed converges to $n = 20$.

In Figure 2 we show that this ratio can be made arbitrarily large by increasing the number of clients (n). We set large $L_i = L_{\max} = 10^7$ for one device and for the rest we again have $L_i \sim \text{Uniform}(0.1, 1)$, and we increase n in each row.

Remarkably, the experimental results follows the same pattern as our theoretical prediction.

6.2 Real Dataset

We also do the same experiment using the “*australian*” dataset from LibSVM library (Chang and Lin, 2011). We set the regularization parameter $\lambda = 10^{-4}L_{\max}$. We split the dataset equally into $n = 20$ devices. In this case we get $k = 8$ devices with ill-conditioned local problems, so the gradient computation ratio of ProxSkip over GradSkip should be close to $n/k = 2.5$. It can be seen in Figure 3.

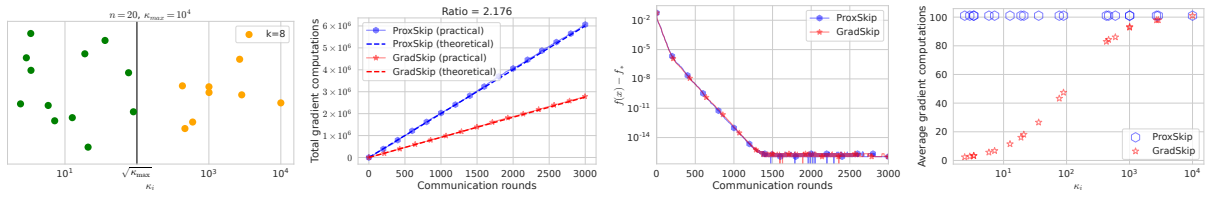


Figure 3: The plots have the same meaning as in Figure 1.

Bibliography

- Artavazd Maranjyan, Mher Safaryan, and Peter Richtárik. Gradskip: Communication-accelerated local gradient methods with better computational complexity. *arXiv preprint arXiv:2210.16402*, 2022.
- Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! *39th International Conference on Machine Learning (ICML)*, 2022.
- H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- Peter Kairouz et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1-2):1-210., 2019.
- Jianguo Wang, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.
- John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Journal of Machine Learning Research*, page 2121-2159, 2011.
- Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. In *arXiv e-prints*, *arXiv:1212.5701*, 2012.

- Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. In *SIAM Journal on Optimization*, volume 23(4), page 2341–2368, 2013.
- Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. In *Mathematical Programming*, volume 162(1-2), page 83–112, 2017.
- Xun Qian, Peter Richtárik, Robert Mansel Gower, Alibek Sailanbayev, Nicolas Loizou, and Egor Shulgin. SGD with arbitrary sampling: General analysis and improved rates. In *International Conference on Machine Learning*, 2019.
- Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020a.
- O. L. Mangasarian. Parallel gradient distribution in unconstrained optimization. In *SIAM Journal on Control and Optimization*, 33(6):1916–1925, 1995.
- H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. In *arXiv preprint arXiv:1602.05629*, 2016.
- F. Haddadpour and M. Mahdavi. On the convergence of local descent methods in federated learning. In *arXiv preprint arXiv:1910.14425*, 2019.

- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, , and V. Smith. Federated optimization for heterogeneous networks. In *Proceedings of the 1st Adaptive Multitask Learning Workshop*, 2019a.
- A. Khaled, K. Mishchenko, and P. Richtárik. First analysis of local gd on heterogeneous data. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, pages 1–11, 2019a.
- A. Khaled, K. Mishchenko, and P. Richtárik. Better communication complexity for local sgd. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, pages 1–11, 2019b.
- Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAF-FOLD: Stochastic controlled averaging for on-device federated learning. In *International Conference on Machine Learning (ICML)*, 2020.
- S. Horváth, M. Sanjabi, L. Xiao, P. Richtárik, and M. Rabbat. FedShuffle: Recipes for better use of local work in federated learning. In *arXiv preprint arXiv:2204.13169*, 2022.
- Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance Reduced ProxSkip: Algorithm, Theory and Application to Federated Learning. In *arXiv:2207.04338*, 2022.
- Laurent Condat and Peter Richtárik. RandProx: Primal-Dual Optimization Algorithms with Randomized Proximal Updates. In *arXiv:2207.12891*, 2022.
- Abdurakhmon Sadiev, Dmitry Kovalev, and Peter Richtárik. Communication Acceleration of Local Gradient Methods via an

Accelerated Primal-Dual Algorithm with Inexact Prox. In *arXiv:2207.03957*, 2022.

Dan Alistarh, Torsten Hoefer, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The convergence of sparsified gradient methods. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5977–5987. Curran Associates, Inc., 2018.

Konstantin Mishchenko, Filip Hanzely, and Peter Richtárik. 99% of worker-master communication in distributed optimization is not needed. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 979–988. PMLR, 03–06 Aug 2020.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, 2018.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

J. Sun, T. Chen, G. Giannakis, and Z. Yang. Communication-efficient distributed learning via lazily aggregated quantized gradients. *Advances in Neural Information Processing Systems*, 32:3370–3380, 2019.

- B. Wang, M. Safaryan, and P. Richtárik. Theoretically Better and Numerically Faster Distributed Optimization with Smoothness-Aware Quantization Techniques. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: variance reduction via gradient sketching. *Advances in Neural Information Processing Systems 31*, pages 2082–2093, 2018.
- Mher Safaryan, Filip Hanzely, and Peter Richtárik. Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.
- H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. *International Conference on Machine Learning (ICML)*, 2019.
- X. Li, W. Yang, S. Wang, and Z. Zhang. Communication-efficient local decentralized SGD methods. *arXiv preprint arXiv:1910.09126*, 2019b.
- X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of FedAvg on non-IID data. *International Conference on Learning Representations*, 2020.

- Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local SGD: unified theory and new efficient methods. *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.
- Laurent Condat, Kai Yi, and Peter Richtárik. EF-BV: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. *arXiv:2205.04180*, 2022.
- Bing Luo, Wenli Xiao, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Tackling System and Statistical Heterogeneity for Federated Learning with Adaptive Client Sampling . In *arXiv:2112.11256*, 2021.
- Amirhossein Reisizadeh, Isidoros Tziotis, Hamed Hassani, Aryan Mokhtari, and Ramtin Pedarsani. Straggler-Resilient Federated Learning: Leveraging the Interplay Between Statistical Accuracy and System Heterogeneity . In *arXiv:2012.14453*, 2020.
- Jianyu Wang and Gauri Joshi. Adaptive Communication Strategies to Achieve the Best Error-Runtime Trade-off in Local-Update SGD . In *arXiv:1810.08313*, 2019.
- Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling i: algorithms and complexity. *Optimization Methods and Software*, 31:829–857, 2016a.
- Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling ii: expected separable overapproximation. *Optimization Methods and Software*, 31:858–884, 2016b. doi: 10.1080/10556788.2016.1190361.

- Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. *Optim Lett*, 10:1233–1243, 2016. doi: <https://doi.org/10.1007/s11590-015-0916-1>.
- Filip Hanzely and Peter Richtárik. One method to rule them all: Variance reduction for data, parameters and many new methods. preprint arXiv:1905.11266, 2019a.
- Filip Hanzely and Peter Richtárik. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 304–312. PMLR, 16–18 Apr 2019b. URL <http://proceedings.mlr.press/v89/hanzely19a.html>.
- Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated sgd. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc., 2020b. URL <https://proceedings.neurips.cc/paper/2020/file/ef9280fbc5317f17d480e4d4f61b3751-Paper.pdf>.
- Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/qian19b.html>.
- Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support

vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 561–577, 2018.

Appendix

A Limitations and Future Work

In this part, we outline some limitations and future research directions related to our work.

- As the previous works on local gradient methods with communication acceleration, our theory does not cover general convex or non-convex objective functions.
- Another key component for designing efficient distributed and federated learning algorithm is partial device participation. This extension seems rather tricky, and we leave this as future work.
- Finally, one can combine the local gradient methods with communication compression to achieve even better communication complexity. Moreover, our proposed gradient skipping approach can be decoupled to address computational complexity too.

B Proofs for Section 3 (GradSkip)

B.1 Proof of Lemma 3.1

Proof. The proof is rather straightforward and follows by following the corresponding lines of the algorithm. Note that $\eta_{i,t} = \theta_t = 0$ implies (see lines 6 and 7 in Algorithm 1) that

$$\hat{x}_{i,t+1} = x_{i,t+1} = x_{i,t}, \quad (20)$$

$$\hat{h}_{i,t+1} = h_{i,t+1} = h_{i,t} = \nabla f_i(x_{i,t}), \quad (21)$$

which proves (6)-(7) when $j = 1$. Consider the two possible cases for $\eta_{i,t+1}$ coupled with $\theta_{t+1} = 0$. If $\eta_{i,t+1} = 1$, then

$$\begin{aligned}\hat{x}_{i,t+2} &= x_{i,t+1} - \gamma(\nabla f_i(x_{i,t+1}) - h_{i,t+1}) \\ &\stackrel{(20)}{=} x_{i,t+1} - \gamma(\nabla f_i(x_{i,t}) - h_{i,t+1}) \\ &\stackrel{(21)}{=} x_{i,t+1} \\ &\stackrel{(20)}{=} x_{i,t},\end{aligned}$$

and

$$\hat{h}_{i,t+2} = h_{i,t+1} \stackrel{(21)}{=} h_{i,t} = \nabla f_i(x_{i,t}).$$

In case of $\eta_{i,t+1} = 0$, we have

$$\hat{x}_{i,t+2} = x_{i,t+1} \stackrel{(20)}{=} x_{i,t}$$

and

$$\hat{h}_{i,t+2} = \nabla f_i(x_{i,t+1}) \stackrel{(20)}{=} \nabla f_i(x_{i,t}) \stackrel{(20)}{=} h_{i,t}.$$

Hence, in both cases we get

$$\hat{x}_{i,t+2} = x_{i,t+1} = x_{i,t}, \tag{22}$$

$$\hat{h}_{i,t+2} = h_{i,t} = \nabla f_i(x_{i,t}). \tag{23}$$

It remains to combine (22)-(23) with the condition that $\theta_{t+1} = 0$, which implies $x_{i,t+2} = \hat{x}_{i,t+2}$, $h_{i,t+2} = \hat{h}_{i,t+2}$. Thus, we proved (6)-(7) when $j = 2$. The proof can be completed by applying induction on j . □

B.2 Proof of Lemma 3.2

As mentioned in the text preceding the lemma, the proof follows from the fact that for two geometric random variables $\Theta \sim \text{Geo}(p)$ and $H \sim \text{Geo}(q)$, their minimum $\min(\Theta, H)$ is also a geometric random variable with parameter $1 - (1 - p)(1 - q)$. To see this, consider

the corresponding Bernoulli trials with success probability p and q for each geometric random variable. Notice that the probability that both trials fail is $(1-p)(1-q)$. Hence, $\min(\Theta, H)$ is the number of joint trials of the two Bernoulli variables until one of them succeeds with probability $1 - (1-p)(1-q)$. Therefore, $\min(\Theta, H)$ is also a geometric random variable with success probability $1 - (1-p)(1-q)$.

B.3 Proof of Theorem 3.5

Denote $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid x_{1,t}, \dots, x_{n,t}]$ the conditional expectation with respect to the randomness of all local models $x_{1,t}, \dots, x_{n,t}$ at t^{th} iterate.

Lemma B.1. *If $\gamma > 0$ and $0 \leq p, q_i \leq 1$, then*

$$\mathbb{E}_t[\Psi_{t+1}] = \sum_{i=1}^n \left[\|w_{i,t} - w_{i,\star}\|^2 + (1 - q_i)(1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 + q_i(1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right],$$

where the expectation is taken over θ_t and $\eta_{i,t}$ in Algorithm 1.

Proof of Lemma B.1. In order to simplify notation, denote

$$x_i := \hat{x}_{i,t+1} - \frac{\gamma}{p} \hat{h}_{i,t+1}, \quad y_i := x_\star - \frac{\gamma}{p} h_{i,\star}. \quad (24)$$

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} := \frac{1}{n} \sum_{i=1}^n y_i = x_\star. \quad (25)$$

STEP 1 (Recalling the steps of the method). Recall that

$$x_{i,t+1} = \begin{cases} \bar{x} & \text{with probability } p \\ \hat{x}_{i,t+1} & \text{with probability } 1 - p \end{cases}, \quad (26)$$

and

$$h_{i,t+1} = \begin{cases} \hat{h}_{i,t+1} + \frac{p}{\gamma} (\bar{x} - \hat{x}_{i,t+1}) & \text{with probability } p \\ \hat{h}_{i,t+1} & \text{with probability } 1 - p \end{cases}. \quad (27)$$

STEP 2 (One-step expectation w.r.t. the global coin toss

θ_t).

The expected value of the Lyapunov function

$$\Psi_t := \sum_{i=1}^n \|x_{i,t} - x_\star\|^2 + \frac{\gamma^2}{p^2} \sum_{i=1}^n \|h_{i,t} - h_{i,\star}\|^2 \quad (28)$$

at $(t+1)^{th}$ iterate with respect to the coin toss θ_t is

$$\begin{aligned} & \mathbb{E}_t [\Psi_{t+1} \mid \eta_{1,t}, \dots, \eta_{n,t}] \\ \stackrel{(26)-(28)}{=} & p \sum_{i=1}^n \left(\|\bar{x} - x_\star\|^2 + \frac{\gamma^2}{p^2} \left\| \hat{h}_{i,t+1} + \frac{p}{\gamma} (\bar{x} - \hat{x}_{i,t+1}) - h_{i,\star} \right\|^2 \right) \\ & + (1-p) \sum_{i=1}^n \left(\|\hat{x}_{i,t+1} - x_\star\|^2 + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right) \\ \stackrel{(25)}{=} & p \sum_{i=1}^n \left(\|\bar{x} - \bar{y}\|^2 + \|\bar{x} - x_i + y_i - \bar{y}\|^2 \right) \\ & + (1-p) \sum_{i=1}^n \left(\|\hat{x}_{i,t+1} - x_\star\|^2 + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right) \\ = & p \sum_{i=1}^n \|x_i - y_i\|^2 + (1-p) \sum_{i=1}^n \left(\|\hat{x}_{i,t+1} - x_\star\|^2 + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right) \\ = & \sum_{i=1}^n \left[p \left\| \hat{x}_{i,t+1} - \frac{\gamma}{p} \hat{h}_{i,t+1} - \left(x_\star - \frac{\gamma}{p} h_{i,\star} \right) \right\|^2 \right. \\ & \left. + (1-p) \left(\|\hat{x}_{i,t+1} - x_\star\|^2 + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right) \right]. \end{aligned}$$

STEP 3 (Simple algebra). Next, we expand the squared norm

and collect the terms, obtaining

$$\begin{aligned}
& \mathbb{E}_t [\Psi_{t+1} \mid \eta_{1,t}, \dots, \eta_{n,t}] \\
&= \sum_{i=1}^n \left[p \|\hat{x}_{i,t+1} - x_\star\|^2 + p \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 - 2\gamma \langle \hat{x}_{i,t+1} - x_\star, \hat{h}_{i,t+1} - h_{i,\star} \rangle \right. \\
&\quad \left. + (1-p) \left(\|\hat{x}_{i,t+1} - x_\star\|^2 + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right) \right] \\
&= \sum_{i=1}^n \left[\|\hat{x}_{i,t+1} - x_\star\|^2 - 2\gamma \langle \hat{x}_{i,t+1} - x_\star, \hat{h}_{i,t+1} - h_{i,\star} \rangle + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right] \\
&= \sum_{i=1}^n \left[\left\| \hat{x}_{i,t+1} - x_\star - \gamma (\hat{h}_{i,t+1} - h_{i,\star}) \right\|^2 - \gamma^2 \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right. \\
&\quad \left. + \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right] \\
&= \sum_{i=1}^n \left[\left\| \hat{x}_{i,t+1} - x_\star - \gamma (\hat{h}_{i,t+1} - h_{i,\star}) \right\|^2 + (1-p^2) \frac{\gamma^2}{p^2} \|\hat{h}_{i,t+1} - h_{i,\star}\|^2 \right].
\end{aligned}$$

STEP 4 (One-step expectation w.r.t. local coin tosses $\eta_{i,t}$).

Applying the expectation with respect to (independent) coin tosses $\eta_{i,t}$ and using the tower property we get

$$\begin{aligned}
& \mathbb{E}_t [\Psi_{t+1}] \\
&= \sum_{i=1}^n \left[q_i \left(\|x_{i,t} - \gamma(\nabla f_i(x_{i,t}) - h_{i,t}) - x_\star - \gamma(h_{i,t} - h_{i,\star})\|^2 \right. \right. \\
&\quad \left. \left. + (1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right) \right. \\
&\quad \left. + (1 - q_i) \left(\|x_{i,t} - x_\star - \gamma(\nabla f(x_{i,t}) - h_{i,\star})\|^2 \right. \right. \\
&\quad \left. \left. + (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \right) \right] \\
&= \sum_{i=1}^n \left[q_i \left(\|x_{i,t} - x_\star - \gamma(\nabla f_i(x_{i,t}) - h_{i,\star})\|^2 + (1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right) \right. \\
&\quad \left. + (1 - q_i) \left(\|x_{i,t} - x_\star - \gamma(\nabla f(x_{i,t}) - h_{i,\star})\|^2 \right. \right. \\
&\quad \left. \left. + (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \right) \right] \\
&= \sum_{i=1}^n \left[\|x_{i,t} - x_\star - \gamma(\nabla f_i(x_{i,t}) - h_{i,\star})\|^2 \right. \\
&\quad \left. + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 + q_i (1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right] \\
&= \sum_{i=1}^n \left[\|w_{i,t} - w_{i,\star}\|^2 + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \right. \\
&\quad \left. + q_i (1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right].
\end{aligned}$$

□

Next, we upper bound the first two terms of the above equality by adjusting the stepsize.

Lemma B.2. *If $0 < \gamma \leq \min_i \left\{ \frac{1}{L_i} \frac{p^2}{1 - q_i(1 - p^2)} \right\}$, then*

$$\|w_{i,t} - w_{i,\star}\|^2 + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \leq (1 - \gamma\mu) \|x_{i,t} - x_\star\|^2.$$

Proof of Lemma B.2. After some algebraic transformations we get

$$\begin{aligned}
& \|w_{i,t} - w_{i,\star}\|^2 + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \\
= & \|x_{i,t} - x_\star - \gamma (\nabla f_i(x_{i,t}) - h_{i,\star})\|^2 + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \\
= & \|x_{i,t} - x_\star\|^2 - 2\gamma \langle x_{i,t} - x_\star, \nabla f_i(x_{i,t}) - h_{i,\star} \rangle \\
& + \gamma^2 \|\nabla f_i(x_{i,t}) - h_{i,\star}\|^2 + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \\
\leq & (1 - \gamma\mu) \|x_{i,t} - x_\star\|^2 - 2\gamma D_{f_i}(x_{i,t}, x_\star) \\
& + \gamma^2 \left(1 + \frac{(1 - q_i) (1 - p^2)}{p^2} \right) \|\nabla f_i(x_{i,t}) - h_{i,\star}\|^2 \\
\leq & (1 - \gamma\mu) \|x_{i,t} - x_\star\|^2 - 2\gamma D_{f_i}(x_{i,t}, x_\star) \left(1 - \gamma L_i \left(\frac{p^2 + (1 - q_i) (1 - p^2)}{p^2} \right) \right) \\
\leq & (1 - \gamma\mu) \|x_{i,t} - x_\star\|^2,
\end{aligned}$$

where we used the bound $\|\nabla f_i(x_{i,t}) - h_{i,\star}\|^2 \leq 2L_i D_{f_i}(x_{i,t}, x_\star)$ and the last inequality holds since $\gamma \leq \frac{1}{L_i} \frac{p^2}{1 - q_i(1 - p^2)}$. \square

Proof of Theorem 3.5. The proof of the theorem is direct combination of the above proved lemmas.

$$\begin{aligned}
\mathbb{E}_t[\Psi_{t+1}] &= \sum_{i=1}^n \left[\|x_{i,t} - x_\star - \gamma (\nabla f_i(x_{i,t}) - h_{i,\star})\|^2 \right. \\
&\quad + (1 - q_i) (1 - p^2) \frac{\gamma^2}{p^2} \|\nabla f(x_{i,t}) - h_{i,\star}\|^2 \\
&\quad \left. + q_i (1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right] \\
&\leq \sum_{i=1}^n \left[(1 - \gamma\mu) \|x_{i,t} - x_\star\|^2 + q_i (1 - p^2) \frac{\gamma^2}{p^2} \|h_{i,t} - h_{i,\star}\|^2 \right] \\
&\leq (1 - \gamma\mu) \sum_{i=1}^n \|x_{i,t} - x_\star\|^2 + q_{max} (1 - p^2) \frac{\gamma^2}{p^2} \sum_{i=1}^n \|h_{i,t} - h_{i,\star}\|^2 \\
&\leq \max \{ 1 - \gamma\mu, q_{max} (1 - p^2) \} \Psi_t \\
&= (1 - \min \{ \gamma\mu, 1 - q_{max} (1 - p^2) \}) \Psi_t.
\end{aligned}$$

\square

B.4 Proof of Theorem 3.6

From the choice of $q_i = \frac{1-1/\kappa_i}{1-1/\kappa_{\max}}$, we immediately imply $q_{\max} = 1$. Furthermore, choosing the optimal $p = \frac{1}{\sqrt{\kappa_{\max}}}$, we get

$$\gamma = \min_i \left\{ \frac{1}{L_i} \frac{p^2}{1 - q_i(1 - p^2)} \right\} = \min_i \left\{ \frac{L_i p^2}{L_i \mu_{\min}} \right\} = \frac{1}{L_{\max}}.$$

Now, if we plug these values back to the rate (10), we get the best rate of ProxSkip as

$$1 - \min \left\{ \gamma \mu, 1 - q_{\max} (1 - p^2) \right\} = 1 - \min \left\{ \frac{\mu}{L_{\max}}, p^2 \right\} = 1 - \frac{\mu}{L_{\max}} = 1 - \frac{1}{\kappa_{\max}}.$$

This implies $\mathcal{O}(\kappa_{\max} \log \frac{1}{\varepsilon})$ total iteration complexity of the method. Due to the choice $p = \frac{1}{\sqrt{\kappa_{\max}}}$, the method enjoys $\mathcal{O}(\sqrt{\kappa_{\max}} \log \frac{1}{\varepsilon})$ accelerated communication complexity.

We have two geometric random variables, $\Theta \sim \text{Geom}(p)$ and $H_i \sim \text{Geom}(1 - q_i)$, for each client describing local training. From the algorithm description, we see that the number of local steps for client i is $\min(\Theta, H_i)$ which is still a Geometric random variable with parameter $1 - q_i(1 - p)$. Therefore, the expected number of local steps for client i is the inverse of that parameter, i.e., $\frac{1}{1 - q_i(1 - p)}$. If we plug in the values for p and q_i , we have

$$\begin{aligned} \mathbb{E}[\min(\Theta, H_i)] &= \frac{1}{1 - q_i(1 - p)} = \frac{1}{1 - \left(1 - \frac{1}{\sqrt{\kappa_{\max}}}\right) \frac{1-1/\kappa_i}{1-1/\kappa_{\max}}} \\ &= \frac{1}{1 - \frac{1-1/\kappa_i}{1+1/\sqrt{\kappa_{\max}}}} = \frac{1 + 1/\sqrt{\kappa_{\max}}}{1/\kappa_i + 1/\sqrt{\kappa_{\max}}} = \frac{\kappa_i(1 + \sqrt{\kappa_{\max}})}{\kappa_i + \sqrt{\kappa_{\max}}} \\ &\leq \min(\kappa_i, \sqrt{\kappa_{\max}}), \end{aligned}$$

where the last inequality can be verified with simple algebraic steps.

C Proofs for Section 5 (VR-GradSkip+)

Here we start proving convergence of Algorithm 3 by first proving some auxiliary lemmas. Let

$$w_t := x_t - \gamma g_t, \quad \text{and} \quad w_* := x_* - \gamma \nabla f(x_*).$$

Lemma C.1. *If $\gamma > 0$ and $\mathcal{C}_\omega \in \mathbb{B}^d(\omega)$, $\mathcal{C}_\Omega \in \mathbb{B}^d(\Omega)$, then*

$$\begin{aligned} \mathbb{E}_t [\Psi_{t+1} - \gamma^2 W \sigma_{t+1} \mid g_t] &\leq \|w_t - w_*\|^2 \\ &\quad + \left(1 - \frac{1}{(1+\omega)^2}\right) \gamma^2 (1+\omega)^2 \|g_t - h_*\|_{\mathbf{I} - (\mathbf{I} + \Omega)^{-1}}^2 \\ &\quad + \left(1 - \frac{1}{(1+\omega)^2}\right) \gamma^2 (1+\omega)^2 \|h_t - h_*\|_{(\mathbf{I} + \Omega)^{-1}}^2, \end{aligned}$$

where the expectation is with respect to the randomness from \mathcal{C}_ω and \mathcal{C}_Ω .

Proof of Lemma C.1. In order to simplify notation, let $P(\cdot) := \text{prox}_{\gamma(1+\omega)\psi}(\cdot)$, and

$$x := \hat{x}_{t+1} - \gamma(1+\omega)\hat{h}_{t+1}, \quad y := x_* - \gamma(1+\omega)h_*. \quad (29)$$

STEP 1 (Optimality conditions). Using the first-order optimality conditions for $f + \psi$ and using $h_* := \nabla f(x_*)$, we obtain the following fixed-point identity for x_* :

$$x_* = \text{prox}_{\gamma(1+\omega)\psi}(x_* - \gamma(1+\omega)h_*) \stackrel{(29)}{=} P(y). \quad (30)$$

STEP 2 (Recalling the steps of the method). Recall that the vectors x_{t+1} and h_{t+1} are in Algorithm 3 updated as follows:

$$x_{t+1} = \hat{x}_{t+1} - \gamma \hat{g}_t = \hat{x}_{t+1} - \frac{1}{1+\omega} \mathcal{C}_\omega(\hat{x}_{t+1} - P(x)), \quad (31)$$

and

$$h_{t+1} = \hat{h}_{t+1} + \frac{1}{\gamma(1+\omega)}(x_{t+1} - \hat{x}_{t+1}) = \hat{h}_{t+1} - \frac{1}{\gamma(1+\omega)^2} \mathcal{C}_\omega(\hat{x}_{t+1} - P(x)). \quad (32)$$

STEP 3 (One-step expectation of the Lyapunov function).

The expected value of the Lyapunov function

$$\Psi_t := \|x_t - x_\star\|^2 + \gamma^2(1 + \omega)^2 \|h_t - h_\star\|^2 + \gamma^2 W \sigma_t \quad (33)$$

at time $t + 1$, with respect to the randomness of \mathcal{C}_ω , is

$$\begin{aligned} & \mathbb{E}_t [\Psi_{t+1} - \gamma^2 W \sigma_{t+1} \mid \mathcal{C}_\Omega, g_t] \\ &= \mathbb{E}_t \left[\left\| \hat{x}_{t+1} - \frac{1}{1 + \omega} \mathcal{C}_\omega (\hat{x}_{t+1} - P(x)) - x_\star \right\|^2 \mid \mathcal{C}_\Omega, g_t \right] \\ & \quad + \mathbb{E}_t \left[\gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - \frac{1}{\gamma(1 + \omega)^2} \mathcal{C}_\omega (\hat{x}_{t+1} - P(x)) - h_\star \right\|^2 \mid \mathcal{C}_\Omega, g_t \right] \\ &= \mathbb{E}_t \left[\left\| \hat{x}_{t+1} - x_\star \right\|^2 - \frac{2}{1 + \omega} \langle \mathcal{C}_\omega (\hat{x}_{t+1} - P(x)), \hat{x}_{t+1} - x_\star \rangle \right. \\ & \quad \left. + \frac{1}{(1 + \omega)^2} \left\| \mathcal{C}_\omega (\hat{x}_{t+1} - P(x)) \right\|^2 \mid \mathcal{C}_\Omega, g_t \right] \\ & \quad + \mathbb{E}_t \left[\gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - h_\star \right\|^2 - 2\gamma \langle \mathcal{C}_\omega (\hat{x}_{t+1} - P(x)), \hat{h}_{t+1} - h_\star \rangle \right. \\ & \quad \left. + \frac{1}{(1 + \omega)^2} \left\| \mathcal{C}_\omega (\hat{x}_{t+1} - P(x)) \right\|^2 \mid \mathcal{C}_\Omega, g_t \right] \\ &\leq \left\| \hat{x}_{t+1} - x_\star \right\|^2 + \frac{2}{1 + \omega} \langle P(x) - \hat{x}_{t+1}, \hat{x}_{t+1} - x_\star \rangle + \frac{1}{1 + \omega} \left\| P(x) - \hat{x}_{t+1} \right\|^2 \\ & \quad + \gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - h_\star \right\|^2 + \frac{2}{1 + \omega} \langle P(x) - \hat{x}_{t+1}, \gamma(1 + \omega)(\hat{h}_{t+1} - h_\star) \rangle \\ & \quad + \frac{1}{1 + \omega} \left\| P(x) - \hat{x}_{t+1} \right\|^2 \\ &= \left\| \hat{x}_{t+1} - x_\star \right\|^2 + \frac{1}{1 + \omega} \left(\left\| P(x) - x_\star \right\|^2 - \left\| \hat{x}_{t+1} - x_\star \right\|^2 \right) \\ & \quad + \gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - h_\star \right\|^2 \\ & \quad + \frac{1}{1 + \omega} \left(\left\| P(x) - \hat{x}_{t+1} + \gamma(1 + \omega)(\hat{h}_{t+1} - h_\star) \right\|^2 - \gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - h_\star \right\|^2 \right) \\ &= \left(1 - \frac{1}{1 + \omega} \right) \left(\left\| \hat{x}_{t+1} - x_\star \right\|^2 + \gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - h_\star \right\|^2 \right) \\ & \quad + \frac{1}{1 + \omega} \left(\left\| P(x) - x_\star \right\|^2 + \left\| P(x) - \hat{x}_{t+1} + \gamma(1 + \omega)(\hat{h}_{t+1} - h_\star) \right\|^2 \right) \\ &= \left(1 - \frac{1}{1 + \omega} \right) \left(\left\| \hat{x}_{t+1} - x_\star \right\|^2 + \gamma^2(1 + \omega)^2 \left\| \hat{h}_{t+1} - h_\star \right\|^2 \right) \\ & \quad + \frac{1}{1 + \omega} \left(\left\| P(x) - P(y) \right\|^2 + \left\| P(x) - x + y - P(y) \right\|^2 \right). \end{aligned}$$

STEP 4 (Applying firm non-expansiveness). Applying firm

non-expansiveness of prox operator P , this leads to the inequality

$$\begin{aligned}
& \mathbb{E}_t [\Psi_{t+1} - \gamma^2 W \sigma_{t+1} \mid \mathcal{C}_\Omega, g_t] \\
& \leq \left(1 - \frac{1}{1+\omega}\right) \left(\|\hat{x}_{t+1} - x_\star\|^2 + \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2 \right) \\
& \quad + \frac{1}{1+\omega} \|x - y\|^2 \\
& = \left(1 - \frac{1}{1+\omega}\right) \left(\|\hat{x}_{t+1} - x_\star\|^2 + \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2 \right) \\
& \quad + \frac{1}{1+\omega} \left\| \hat{x}_{t+1} - \gamma(1+\omega)\hat{h}_{t+1} - (x_\star - \gamma(1+\omega)h_\star) \right\|^2 \\
& = \left(1 - \frac{1}{1+\omega}\right) \left(\|\hat{x}_{t+1} - x_\star\|^2 + \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2 \right) \\
& \quad + \frac{1}{1+\omega} \left\| \hat{x}_{t+1} - x_\star - \gamma(1+\omega) (\hat{h}_{t+1} - h_\star) \right\|^2.
\end{aligned}$$

STEP 5 (Simple algebra). Next, we expand the squared norm and collect the terms, obtaining

$$\begin{aligned}
& \mathbb{E}_t [\Psi_{t+1} - \gamma^2 W \sigma_{t+1} \mid \mathcal{C}_\Omega, g_t] \\
& \leq \left(1 - \frac{1}{1+\omega}\right) \left(\|\hat{x}_{t+1} - x_\star\|^2 + \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2 \right) \\
& \quad + \frac{1}{1+\omega} \|\hat{x}_{t+1} - x_\star\|^2 - 2\gamma \langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star \rangle + \gamma^2(1+\omega) \|\hat{h}_{t+1} - h_\star\|^2 \\
& = \|\hat{x}_{t+1} - x_\star\|^2 - 2\gamma \langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star \rangle + \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2 \\
& = \|\hat{x}_{t+1} - x_\star - \gamma(\hat{h}_{t+1} - h_\star)\|^2 - \gamma^2 \|\hat{h}_{t+1} - h_\star\|^2 + \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2 \\
& = \|\hat{x}_{t+1} - x_\star - \gamma(\hat{h}_{t+1} - h_\star)\|^2 + \left(1 - \frac{1}{(1+\omega)^2}\right) \gamma^2(1+\omega)^2 \|\hat{h}_{t+1} - h_\star\|^2.
\end{aligned}$$

STEP 6 (Tower property). Applying the expectation with respect to the randomness of \mathcal{C}_Ω and using the tower property, we

get

$$\begin{aligned}
& \mathbb{E}_t [\Psi_{t+1} - \gamma^2 W \sigma_{t+1} \mid g_t] \\
= & \mathbb{E}_t \left[\left\| x_t - \gamma(g_t - \hat{h}_{t+1}) - x_\star - \gamma(\hat{h}_{t+1} - h_\star) \right\|^2 \mid g_t \right] \\
& + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \mathbb{E}_t \left[\left\| g_t - (\mathbf{I} + \mathbf{\Omega})^{-1} \mathbf{C}_\Omega (g_t - h_t) - h_\star \right\|^2 \mid g_t \right] \\
= & \left\| x_t - \gamma g_t - (x_\star - \gamma h_\star) \right\|^2 \\
& + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \mathbb{E}_t \left[\left\| g_t - h_\star - (\mathbf{I} + \mathbf{\Omega})^{-1} \mathbf{C}_\Omega (g_t - h_t) \right\|^2 \mid g_t \right] \\
\leq & \left\| w_t - w_\star \right\|^2 + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \left\| g_t - h_\star \right\|^2 \\
& + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \left(2 \langle g_t - h_\star, h_t - g_t \rangle_{(\mathbf{I} + \mathbf{\Omega})^{-1}} + \left\| g_t - h_t \right\|_{(\mathbf{I} + \mathbf{\Omega})^{-1}}^2 \right) \\
= & \left\| w_t - w_\star \right\|^2 + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega_2)^2 \left\| g_t - h_\star \right\|^2 \\
& + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega_2)^2 \left(\left\| h_t - h_\star \right\|_{(\mathbf{I} + \mathbf{\Omega})^{-1}}^2 - \left\| g_t - h_\star \right\|_{(\mathbf{I} + \mathbf{\Omega})^{-1}}^2 \right) \\
= & \left\| w_t - w_\star \right\|^2 + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \left\| g_t - h_\star \right\|_{\mathbf{I} - (\mathbf{I} + \mathbf{\Omega})^{-1}}^2 \\
& + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \left\| h_t - h_\star \right\|_{(\mathbf{I} + \mathbf{\Omega})^{-1}}^2.
\end{aligned}$$

□

Next, we upper bound the first two terms.

Lemma C.2. Denote $\tilde{\mathbf{\Omega}} = \mathbf{I} + \omega(\omega + 2)\mathbf{\Omega}(\mathbf{I} + \mathbf{\Omega})^{-1}$. Then

$$\begin{aligned}
& \mathbb{E}_t \left[\left\| w_t - w_\star \right\|^2 \right] + \left(1 - \frac{1}{(1+\omega)^2} \right) (1+\omega)^2 \gamma^2 \mathbb{E}_t \left[\left\| g_t - h_\star \right\|_{\mathbf{I} - (\mathbf{I} + \mathbf{\Omega})^{-1}}^2 \right] \\
& \leq (1 - \gamma\mu) \left\| x_t - x_\star \right\|^2 - 2\gamma \left(1 - \gamma A \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) \right) D_f(x_t, x_\star) \\
& \quad + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) B \sigma_t + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) C.
\end{aligned}$$

Proof of Lemma C.2. Expanding the first term and rearranging

terms, we get

$$\begin{aligned}
& \mathbb{E}_t \left[\|w_t - w_\star\|^2 \right] + \left(1 - \frac{1}{(1+\omega)^2} \right) (1+\omega)^2 \gamma^2 \mathbb{E}_t \left[\|g_t - h_\star\|_{\mathbf{I}-(\mathbf{I}+\mathbf{\Omega})^{-1}}^2 \right] \\
&= \mathbb{E}_t \left[\|x_t - x_\star - \gamma(g_t - \nabla f(x_\star))\|^2 \right] \\
&\quad + \omega(\omega+2)\gamma^2 \mathbb{E}_t \left[\|g_t - \nabla f(x_\star)\|_{\mathbf{\Omega}(\mathbf{I}+\mathbf{\Omega})^{-1}}^2 \right] \\
&= \|x_t - x_\star\|^2 - 2\gamma \langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star) \rangle \\
&\quad + \gamma^2 \mathbb{E}_t \left[\|g_t - \nabla f(x_\star)\|^2 \right] + \omega(\omega+2)\gamma^2 \mathbb{E}_t \left[\|g_t - \nabla f(x_\star)\|_{\mathbf{\Omega}(\mathbf{I}+\mathbf{\Omega})^{-1}}^2 \right] \\
&\leq (1-\gamma\mu) \|x_t - x_\star\|^2 - 2\gamma D_f(x_t, x_\star) + \gamma^2 \mathbb{E}_t \left[\|g_t - \nabla f(x_\star)\|_{\tilde{\mathbf{\Omega}}}^2 \right] \\
&\leq (1-\gamma\mu) \|x_t - x_\star\|^2 - 2\gamma D_f(x_t, x_\star) + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) \mathbb{E}_t \left[\|g_t - \nabla f(x_\star)\|_{\mathbf{L}^{-1}}^2 \right] \\
&\leq (1-\gamma\mu) \|x_t - x_\star\|^2 - 2\gamma D_f(x_t, x_\star) \\
&\quad + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) (2AD_f(x_t, x_\star) + B\sigma_t + C) \\
&= (1-\gamma\mu) \|x_t - x_\star\|^2 - 2\gamma \left(1 - \gamma A \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) \right) D_f(x_t, x_\star) \\
&\quad + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) B\sigma_t + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) C.
\end{aligned}$$

□

Proof of Theorem 5.2. The proof is a direct combination of the two lemmas we have proved.

$$\begin{aligned}
\mathbb{E} [\Psi_{t+1}] &\leq (1-\gamma\mu) \|x_t - x_\star\|^2 - 2\gamma \left(1 - \gamma A \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) \right) D_f(x_t, x_\star) \\
&\quad + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) B\sigma_t + \gamma^2 \lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) C \\
&\quad + \left(1 - \frac{1}{(1+\omega)^2} \right) \gamma^2 (1+\omega)^2 \|h_t - h_\star\|_{(\mathbf{I}+\mathbf{\Omega})^{-1}}^2 \\
&\quad + \gamma^2 W \left(2\tilde{A}D_f(x_t, x_\star) + \tilde{B}\sigma_t + \tilde{C} \right) \\
&= (1-\gamma\mu) \|x_t - x_\star\|^2 - 2\gamma \left(1 - \gamma(A\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) + W\tilde{A}) \right) D_f(x_t, x_\star) \\
&\quad + \frac{\omega(\omega+2)}{(1+\lambda_{\min}(\mathbf{\Omega}))(1+\omega)^2} \gamma^2 (1+\omega)^2 \|h_t - h_\star\|^2 \\
&\quad + \left(\frac{\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})B}{W} + \tilde{B} \right) \gamma^2 W\sigma_t + \gamma^2 (\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})C + W\tilde{C}).
\end{aligned}$$

Next we choose the stepsize $\gamma \leq \frac{1}{A\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) + W\tilde{A}}$ so that the term with $D_f(x_t, x_\star)$ is non-negative and can be suppressed for further

steps. Let $\delta = 1 - \frac{\omega(\omega+2)}{(1+\lambda_{\min}(\mathbf{\Omega}))(\omega+1)^2} = 1 - \frac{1}{1+\lambda_{\min}(\mathbf{\Omega})} \left(1 - \frac{1}{(1+\omega)^2}\right) \in [0, 1]$, $\beta = 1 - \tilde{B} - \frac{\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})B}{W} > 0$, provided that $W > \frac{\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})B}{1-\tilde{B}}$, and continue the above derivation

$$\begin{aligned} \mathbb{E}[\Psi_{t+1}] &\leq \max(1 - \gamma\mu, 1 - \delta, 1 - \beta) \Psi_t + \gamma^2(\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})C + W\tilde{C}) \\ &= (1 - \min(\gamma\mu, \delta, \beta)) \Psi_t + \gamma^2(\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})C + W\tilde{C}) \\ &\leq (1 - \min(\gamma\mu, \delta, \beta))^{t+1} \Psi_0 + \gamma^2 \frac{\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})C + W\tilde{C}}{\min(\gamma\mu, \delta, \beta)}. \end{aligned}$$

□

D Proofs for Section 4 (GradSkip+)

D.1 Proof of Lemma 4.2

The proof follows from the following simple inequalities:

$$\begin{aligned} \|x\|_{(\mathbf{I}+\mathbf{\Omega})^{-1}}^2 &\leq \lambda_{\max}((\mathbf{I}+\mathbf{\Omega})^{-1}) \|x\|^2 = \frac{1}{1+\lambda_{\min}(\mathbf{\Omega})} \|x\|^2, \\ \|(\mathbf{I}+\mathbf{\Omega})^{-1}\mathcal{C}(x)\|^2 &\geq \lambda_{\min}((\mathbf{I}+\mathbf{\Omega})^{-1})^2 \|\mathcal{C}(x)\|^2 = \frac{1}{(1+\lambda_{\max}(\mathbf{\Omega}))^2} \|\mathcal{C}(x)\|^2. \end{aligned}$$

D.2 Proof of Theorem 4.5

Since **GradSkip+** is a special case of a **VR-GradSkip+**, Theorem 4.5 is a corollary of Theorem 5.2. To see this, first, let us prove the following lemma.

Lemma D.1. *Let Assumption 4.4 hold. Then for the gradient estimator $g_t = \nabla f(x_t)$, Assumption 5.1 holds with the following parameters:*

$$A = 1, \quad B = 0, \quad C = 0, \quad \tilde{A} = 0, \quad \tilde{B} = 0, \quad \tilde{C} = 0, \quad \sigma_t \equiv 0.$$

Proof. The proof is rather trivial and follows from the \mathbf{L} -smoothness of f ,

$$\mathbb{E} \left[\|g_t - \nabla f(x_\star)\|_{\mathbf{L}^{-1}}^2 \right] = \|\nabla f(x_t) - \nabla f(x_\star)\|_{\mathbf{L}^{-1}}^2 \leq 2D_f(x_t, x_\star).$$

□

Having this, Theorem [5.2](#) reduces to Theorem [4.5](#).