Tighter Analysis for ProxSkip

Zhengmian Hu [1 2] **Heng Huang** [1]

Abstract

In this paper, we provide a tighter analysis for ProxSkip, an algorithm that allows fewer proximal operator computations to solve composite optimization problems. We improve the existing decreasing speed of Lyapunov function from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$, when p, the frequency of the proximal operators is small enough. Our theoretical analysis also reveals the drawbacks of using large step sizes for gradient descent in ProxSkip when the proximal operator part is the bottleneck. Our main motivation comes from the continuous limit in which the original analysis of ProxSkip fails to guarantee convergence when both the step size γ and frequency p tend to zero. We construct a counterexample to demonstrate why such counter-intuitive behavior occurs for the original analysis and then propose a novel Lyapunov function variant to construct a tighter analysis, avoiding the problem of the old one. Such a new Lyapunov function can be directly extended to many other variants of ProxSkip. When applied to stochastic gradient setup, our analysis leads to an improved proximal operator complexity for SProxSkip from $\mathcal{O}(\sqrt{1/\varepsilon\mu^2}\log(1/\varepsilon))$ to $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$.

1. Introduction

Composite optimization is a problem of the form

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

where $f:\mathbb{R}^d\to\mathbb{R}$ is a smooth function and $\psi:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is a proper, closed and convex function. In this paper, we focus on the case where f is strongly convex. This type of problems arises in a wide range of applications in machine learning and statistics modeling (Candès et al., 2011; Candès & Recht, 2012; Lustig et al., 2007; Tibshirani, 2011; Bao et al., 2022).

Subgradient-based optimization algorithms are not generally used because of their slow convergence rates. Proximal gradient descent (PGD) (Combettes & Wajs, 2005; Passty, 1979; Nesterov, 2013) is a canonical method for solving composite optimization problems. It is based on the proximal operator $\operatorname{prox}_\psi(y)=\arg\min_{x\in\mathbb{R}^d}\left\{\frac{1}{2}\|x-y\|^2+\psi(x)\right\}$ and has a gradient complexity of $\mathcal{O}(\kappa\log(1/\varepsilon))$. Accelerated proximal gradient descent (APGD) (Nesterov, 1998; 2013), an accelerated variant of deterministic gradient descent, has a better gradient complexity of $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$.

Most prior work focused on the scenario where the cost of the proximal operator is low and the gradient descent part is the bottleneck. In this case, typically the gradient complexity and proximal operator complexity are equal. On the other hand, in federated learning (FL) one needs to minimize the average of multiple different functions $f(x):=\frac{1}{n}\sum_{i=1}^n f_i(x)$. It can be put in consensus form (Parikh et al., 2014): $f(x_1,\dots,x_n):=\frac{1}{n}\sum_{i=1}^n f_i(x_i)$, $\psi(x_1,\dots,x_n):=\begin{cases}0, & \text{if } x_1=\cdots=x_n,\\ +\infty, & \text{otherwise}\end{cases}$ as a special case of composite optimization. This attracted attention because the proximal operator $\operatorname{prox}_{\gamma\psi}(x_1,\dots,x_n)=(\bar{x},\dots,\bar{x})\in\mathbb{R}^{nd}$ amounts to a single global communication of the parameters from all clients for computing the average. In this context, people sought fewer proximal operator steps and hence fewer communications, rather than focusing on the gradient complexity of optimization.

Local training is a common approach to reduce communication and consists of multiple local gradient descent steps interspersed by a few proximal operator steps. The simplest local training approach, known as Local SGD/FedAvg (Mangasarian & Solodov, 1993; McDonald et al., 2010; McMahan et al., 2016; Zhang et al., 2016; Stich, 2018; Lin et al., 2018), is equivalent to a multi-step GD followed by a single proximal operator step in composite optimization and has been shown to suffer from client drift (Khaled et al., 2019; Karimireddy et al., 2020; Wu et al., 2022). To address this, various methods have been proposed, including Scaffold (Karimireddy et al., 2020), S-Local-GD (Gorbunov

[1]Department of Computer Science, University of Maryland, College Park, MD, USA. [2]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA.. Correspondence to: Zhengmian Hu <huzhengmian@gmail.com>, Heng Huang <henghuanghh@gmail.com>.

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

et al., 2021), and FedLin (Mitra et al., 2021), that leverage control variates to obtain linear convergence. However, the theoretical understanding of such methods is still in its early stages, and the communication complexity remains at $\mathcal{O}(\kappa \log(1/\varepsilon))$, which is no better than that of PGD.

Recently, ProxSkip (Mishchenko et al., 2022) was proposed to address this issue by randomly applying the proximal operator step. With a tighter analysis, it has successfully achieved an acceleration over PGD even when Nesterov momentum is not used, with an optimal proximal operator complexity of $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$. In comparison with Accelerated Proximal Gradient Descent (APGD), it has a worse gradient complexity, but is much simpler in local training and is allowed to have a non-optimal gradient complexity when the bottleneck is only in communication. However, this work does not imply that local training has been thoroughly understood. Specifically, existing analysis is not tight in certain regimes and cannot extrapolate to continuous limits, and only can achieve sublinear convergence with respect to proximal operator complexity when a stochastic gradient is used.

In this paper, we improve upon the above and make contributions as follows:

- We provide a tighter analysis of ProxSkip in Section 3, which improves the decreasing speed of Lyapunov function from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$ when the proximal operator part is the bottleneck.
- We reveal an effect of step size that is not present in the previous analysis, which suggests that large step size actually hinders convergence when the proximal operator part is the bottleneck, as illustrated in Section 3.1.
- Our analysis embodies the continuous limit as a special case of ProxSkip, which demonstrates that gradient flow can also solve composite optimization collaboratively with a proximal operator and achieves $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$ proximal operator complexity, as discussed in Section 4.
- We propose a new Lyapunov function and explain in Section 5 why the old one does not fit and what new proving techniques are needed.
- In Section 6, we extend our analysis to a variety of ProxSkip variants, including SProxSkip, ProxSkip-VR, and GradSkip+. This demonstrates the generality of our analysis methodology.
- For SProxSkip, we further improve the proximal operator complexity from $\mathcal{O}(\sqrt{1/\varepsilon\mu^2}\log(1/\varepsilon))$ to $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$.

We remark that this paper only includes basic numerical experimental results, which are separated in different sections, to verify and illustrate our theoretical analysis results. This is for two reasons: firstly, our main contribution is to derive a tighter theoretical analysis and the algorithms are not new; secondly, ProxSkip and other methods studied in this paper are abstract frameworks and have multiple instantiations suitable for different scenarios. Interested readers should refer to the papers where these algorithms were proposed for evaluation and comparison with other algorithms.

Algorithm 1 ProxSkip

stepsize $\gamma > 0$, probability $p > 0$, initial iterate $x_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, number of iterations $T \geq 1$
for $t = 0, 1, ..., T-1$ **do**
  $\hat{x}_{t+1} = x_t - \gamma(\nabla f(x_t) - h_t)$
  Flip a coin $\theta_t \in \{0, 1\}$ where $\text{Prob}(\theta_t = 1) = p$
  if $\theta_t = 1$ **then**
    $x_{t+1} = \text{prox}_{\frac{\gamma}{p}\psi}(\hat{x}_{t+1} - \frac{\gamma}{p}h_t)$
  else
    $x_{t+1} = \hat{x}_{t+1}$
  end if
  $h_{t+1} = h_t + \frac{p}{\gamma}(x_{t+1} - \hat{x}_{t+1})$
end for

2. Preliminary

We start by introducing ProxSkip, the algorithm we are primarily studying in this paper. Algorithm 1 presents the pseudocode of ProxSkip. Mishchenko et al. (2022) proposed ProxSkip and provides a convergence analysis, based on the following Lyapunov function

$$\Psi_t^{\text{old}} = \|x_t - x_\star\|^2 + \frac{\gamma^2}{p^2}\|h_t - h_\star\|^2. \tag{1}$$

If this Lyapunov function converges to zero, then x_t converges to $x_\star$ and h_t converges to $h_\star = \nabla f(x_\star)$. Further, the original paper guarantees the convergence of Ψ_t^{old} as follows

Theorem 2.1. *Let f be μ-strongly convex with positive $\mu > 0$ and L-smooth, and let $0 < \gamma \leq \frac{1}{L}$ and $0 < p \leq 1$. Then, the iterates of ProxSkip (Algorithm 1) satisfy*

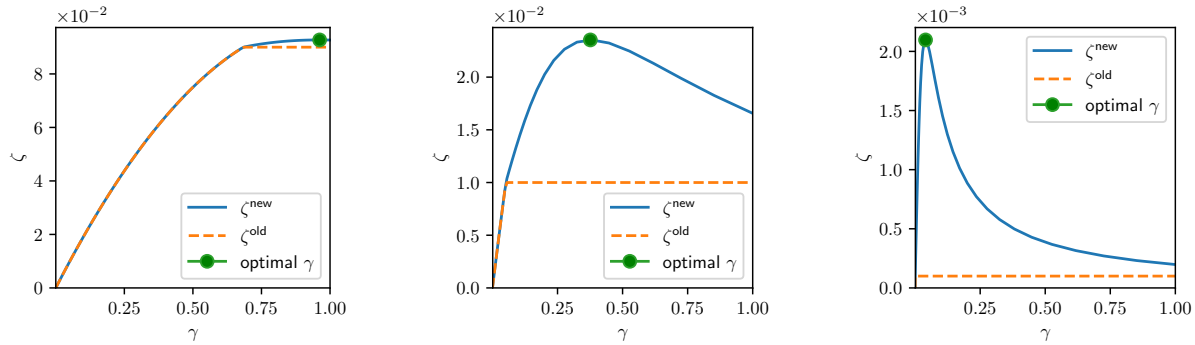
$$\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq (1 - \zeta^{\text{old}})\Psi_t^{\text{old}}, \quad \zeta^{\text{old}} = \min(\gamma\mu(2 - \gamma L), p^2).$$

In other words,

$$\zeta^{\text{old}} = \begin{cases} \gamma\mu(2 - \gamma L) & \text{if } \gamma \leq \gamma_{\text{crit}} \\ p^2 & \text{if } \gamma > \gamma_{\text{crit}}. \end{cases}$$

γ_{crit} is the root of $\gamma\mu(2 - \gamma L) = p^2$ if it exists, otherwise positive infinity.

Remark 2.2. The bound in Theorem 2.1 is slightly tighter than original result $\zeta = \min(\gamma\mu, p^2)$ in (Mishchenko et al., 2022).



(a) $p = 0.3$ (b) $p = 0.1$ (c) $p = 0.01$

Figure 1: Relationship between step size and decreasing speed of Lyapunov function when $\mu = 0.1$ and $L = 1$

Notably, Theorem 2.1 suggests that there are two regimes: gradient descent is the bottleneck when γ is sufficiently small; proximal operator is the bottleneck when γ is sufficiently large or equivalently when p is sufficiently small. In our improved analysis later on, we use a critical value of γ, denoted by γ_{crit}, to separate the two regimes. All our improvements are focused on the second regime, where proximal operator becomes the bottleneck, *i.e.* $\gamma > \gamma_{\text{crit}}$.

3. New Analysis

In this section, we present a new analysis of ProxSkip. The major difference between our analysis and the existing analysis is a novel Lyapunov function. In particular,

$$\begin{aligned}\Psi_t^{\text{new}} =& \|x_t - x_\star\|^2 + \frac{\gamma^2}{p^2}\|h_t - h_\star\|^2 \\ & -2\Delta\frac{\gamma}{p}\langle x_t - x_\star, h_t - h_\star\rangle.\end{aligned} \tag{2}$$

Compared with the old Lyapunov function, there is an additional inner product term. Note that the new Lyapunov function remains a quadratic form, and it is still positive definite when $-1 < \Delta < 1$. Thus, when Ψ_t^{new} converges, we still have x_t converging to $x_\star$ and h_t converging to $h_\star = \nabla f(x_\star)$.

Based on this new Lyapunov function, we derive a new convergence rate as follows.

Theorem 3.1. *Let f be μ-strongly convex with positive $\mu > 0$ and L-smooth with $L > \mu$, and let $0 < \gamma \leq \frac{1}{L}$ and $0 < p \leq 1$. Moreover, if $\gamma \leq \gamma_{\text{crit}}$, let $\Delta = 0$, otherwise, let Δ be the unique solution in $(0, 1)$ of following equation:*

$$\begin{aligned}0 =& \Delta^3 - 2\left(\frac{\mu\gamma}{p} + 1\right)\Delta^2 + \left(\frac{L\mu\gamma^2}{p^2} + 2\frac{\mu\gamma}{p} + 2\right)\Delta \\ & - \frac{\gamma\mu(2-\gamma L) - p^2}{p - p^2}.\end{aligned} \tag{3}$$

Then, the iterates of ProxSkip (Algorithm 1) satisfy

$$\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq (1 - \zeta^{\text{new}})\Psi_t^{\text{new}},$$

$$\zeta^{\text{new}} = \begin{cases}\gamma\mu(2-\gamma L) & \text{if } \gamma \leq \gamma_{\text{crit}} \\ p^2 + \Delta(p - p^2) & \text{if } \gamma > \gamma_{\text{crit}}.\end{cases}$$

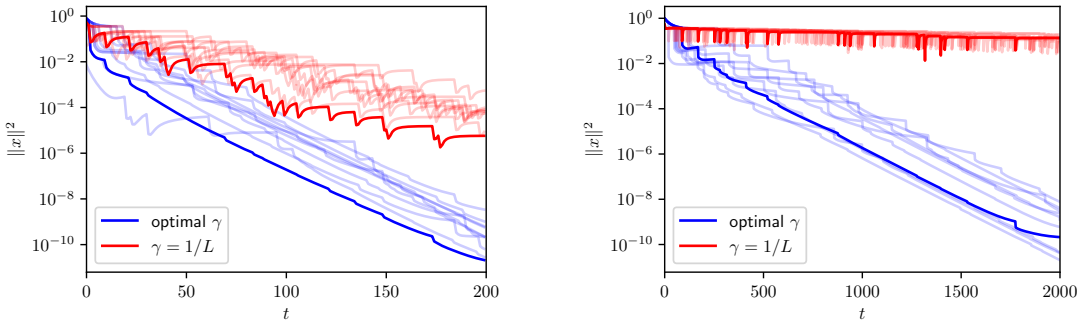
Our improvement depends on the construction of the new Lyapunov function which will be explained in detail in Section 5.

We remark that for the regime where proximal operator is the bottleneck, Theorem 3.1 always provides better result than Theorem 2.1, since the decreasing speed of Lyapunov function is improved from p^2 to $p^2 + \Delta(p - p^2)$. This improvement can be more pronounced in certain special limits when $\Delta = \omega(p)$. Here $\omega(p)$ means it decreases strictly and asymptotically slower than p for small p. As an example, we discuss continuous limit in Section 4, where Δ converges to a constant at the limit of $p \to 0$. Another example is SProxSkip, where the gradients are replaced by noisy stochastic gradients. The result for SProxSkip heavily depends on the convergence rate in the continuous limit, which will be discussed in Section 6.1.

3.1. Parameter Selection

In this subsection, we discuss how the step size γ used in gradient descent affects the convergence rate of ProxSkip. The traditional theory of convergence rate has led to a misconception that, within the step size range $\gamma \leq 1/L$, larger step sizes always lead to faster convergence, as ζ^{old} is monotonically increasing with respect to γ. However, our new theory does not support this conjecture.

In our new theory, ζ^{new} is not monotonic with respect to γ. As demonstrated in Figure 1, our new theory yields a much higher decreasing speed of Lyapunov function than the traditional theory when p is small enough, and for this rate to be achieved a proper step size must be chosen.



(a) $p = 0.1$ (b) $p = 0.01$

Figure 2: ProxSkip experiment for comparing of different step sizes.

This raises a natural question: given that the frequency p of the proximal operator is fixed, how do we choose the optimal step size to maximize ζ^{new}? To this end, we propose Algorithm 2 to solve this problem. We also provide theoretical guarantees for this algorithm in Theorem 3.2.

Algorithm 2 Step size selection for ProxSkip

1: probability $p > 0$, smoothness constant L, strongly convex constant μ.
2: **if** $p \geq \sqrt{\frac{1}{\kappa}}$ **then return end if**
3: **while** γ doesn't converge **do**
4: $\quad\Delta \leftarrow$ Solve Equation (3)
5: $\quad\gamma \leftarrow \frac{1}{L} \times \frac{p(1-\Delta(1-\Delta)(1-p))}{(p+(1-p)\Delta)}$
6: **end while**

Theorem 3.2. *The iterations in Algorithm 2 always converge and γ will converge to* $\operatorname{argmax}_\gamma \zeta^{\text{new}}$.

Finally, we conduct an experiment to verify that the "cost" of large step sizes is not just an artifact of theory. We consider Nesterov's "worst function in the world" (Nesterov, 1998) in its strongly convex version as $f(x)$ and an indicator function as $\psi(x)$ (see Appendix A.5 for the details). As shown in Figure 1, for $p = 0.1$, there is about twofold gap between the convergence rate when using the largest step size and when using Algorithm 2 to select step size. For $p = 0.01$, this gap is around 20-fold. Figure 2 provides similar results for each step size by plotting 10 random samples of the numerical process.

Note that we are not claiming that Algorithm 2 can replace step size tuning in practice. One reason is that the algorithm is highly dependent on κ which is prone to under/over-estimation and is only a global characterization of the local behavior. Another reason is that Theorem 3.1 is not tight, therefore the optimal Δ in the sense of Theorem 3.1 is not really optimal for ProxSkip.

Algorithm 3 ODEProx

1: Horizon $\tau > 0$, initial iterate $x_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, total time $T \geq 0$
2: $t \leftarrow 0$
3: **loop**
4: $\quad$Sample random variable $\hat{\tau} \sim \operatorname{Exp}(\frac{1}{\tau})$
5: $\quad t' \leftarrow \min(T, t + \hat{\tau})$
6: $\quad \hat{x}_t \leftarrow x_t$
7: $\quad$Solve $\frac{\mathrm{d}\hat{x}_s}{\mathrm{d}s} = -(\nabla f(\hat{x}_s) - h_t)$ for $s \in [t, t']$
8: $\quad$**if** $t' = T$ **then**
9: $\quad\quad x_{t'} = \hat{x}_{t'}, h_{t'} = h_t$
10: $\quad\quad$**break**
11: $\quad$**end if**
12: $\quad x_{t'} = \operatorname{prox}_{\tau\psi}(\hat{x}_{t'} - \tau h_t)$
13: $\quad h_{t'} = h_t + \frac{1}{\tau}(x_{t'} - \hat{x}_{t'})$
14: $\quad t \leftarrow t'$
15: **end loop**

4. Continuous Limit

In this section, we consider the situation in which $\tau = \gamma/p$ is fixed and both γ and p tend to 0. Under this setting, ProxSkip will evolve into a new algorithm, which we call ODEProx (Algorithm 3), just like Gradient Descent transforms into Gradient Flow in the limit of infinitesimal stepsize.

The Lyapunov function for ODEProx is as follows for $t \geq 0$:

$$\Psi_t = \|x_t - x_\star\|^2 + \tau^2\|h_t - h_\star\|^2 - 2\Delta\tau\langle x_t - x_\star, h_t - h_\star\rangle$$

Theorem 3.1 implies following convergence guarantee.

Corollary 4.1. *Let f be μ-strongly convex and L-Lipschitz smooth. Moreover, let Δ be the unique solution in $(0, 1)$ of following equation:*

$$\Delta^3 - 2(\mu\tau + 1)\Delta^2 + \left(L\mu\tau^2 + 2\mu\tau + 2\right)\Delta - 2\mu\tau = 0. \tag{4}$$

For $T \geq 0$, the iterates of ODEProx (Algorithm 3) satisfy

$$\mathbb{E}[\Psi_T] \leq \left(1 - \frac{\Delta}{\tau}\right)^T \Psi_0.$$

Algorithm 2 transforms into Algorithm 4.

Algorithm 4 Horizon selection for ODEProx

1: Smoothness constant L, strongly convex constant μ.
2: $\tau \leftarrow \frac{1}{L}$
3: **while** τ doesn't converge **do**
4: $\quad \Delta \leftarrow$ Solve Equation (4)
5: $\quad \tau \leftarrow \frac{1}{L} \times (\frac{1}{\Delta} - (1 - \Delta))$
6: **end while**

The following theorem characterizes the proximal operator complexity of ODEProx with τ from Algorithm 4.

Corollary 4.2. *Let f be μ-strongly convex and L-Lipschitz smooth. Under optimal choice of τ, we have*

$$\Delta^2 (\kappa - 1) \left(\Delta^2 - 2\Delta + 2\right) - (1 - \Delta)^2 = 0,$$

and when κ is large, we have $\Delta = \Theta(1/\sqrt{\kappa})$. Furthermore, the expected oracle complexity of proximal operator in order to achieve $\mathbb{E}[\Psi_T] \leq \varepsilon$ is $T/\tau = \tilde{\Theta}(1/\Delta) = \tilde{\Theta}(\sqrt{\kappa})$.

The continuous limit has three implications.

First, it highlights the deficiency of the existing theory as we observe that the convergence rate of the existing theory (Theorem 2.1) reaches 0 in the continuous limit. This is because the decreasing speed of Lyapunov function of the existing theory is $\mathcal{O}(p^2)$, while that of the new theory is $\mathcal{O}(p)$. We will explain in detail why there is such a difference between them in the next section.

Second, it implies that if we can solve a gradient flow with respect to $f(x)$, we can also solve the corresponding composite optimization problem by incorporating the proximal operator, as demonstrated in Figure 3.

Finally, under this continuous limit, the asymptotic proximal operator complexity is identical to that with finite step sizes, which suggests that we always can reduce both γ and p simultaneously without worrying about degenerated convergence rates.

5. Main Idea of Proof

In this section, we will introduce the main idea behind the construction of our new Lyapunov function and illustrate the intuition through a simple example. Moreover, we will discuss the new proof techniques required for the new Lyapunov function.

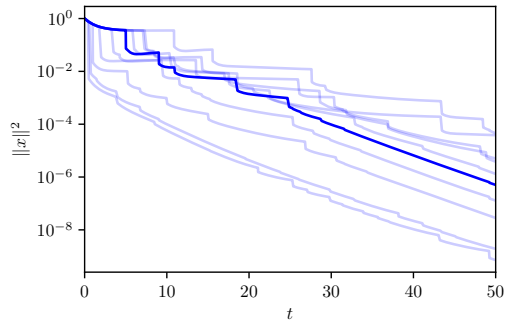


Figure 3: Convergence of ODEProx.

5.1. Intuition

To better explain why the old theory is not tight, we construct a special example in which the old theory is unable to guarantee convergence. As we observe under continuous limit, the lower bound of the decreasing speed of the old theory approaches 0, which indicates that we can find a special point such that the expected decreasing speed of Ψ_t^{old} from this point on is 0.

Our particular construction is as follows. Consider a one-dimensional problem with $\psi(x) = 0$ and $f(x) = \|x\|^2$, with starting point $x_0 = 0$, $h_0 = 1$. The contour plot of the old Lyapunov function is shown in Figure 4.

In ODEProx, there are two operations involved, namely gradient flow and proximal operator. As can be seen from the figure, neither of them can directly lower the Lyapunov function, since for gradient flow, the moving direction is tangent to the contour lines, while for proximal operator, both start and end points are on the same contour line. Therefore in this example, the decrease speed of the old Lyapunov function is 0.

With this example, we can show how our new Lyapunov function bypasses the problem of the old Lyapunov function

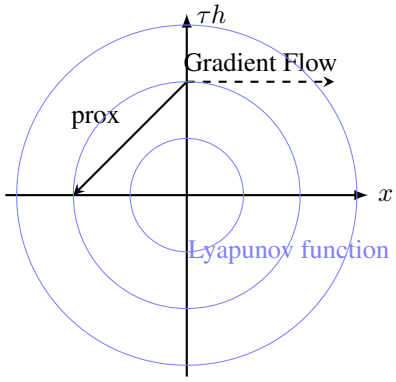


Figure 4: Illustration of old Lyapunov function

in a very straightforward way. We add an inner product term to the Lyapunov function. The contour lines of the new Lyapunov function are no longer circles, but ellipses, as shown in Figure 5, which enables gradient flow to lower the Lyapunov function directly in this example.

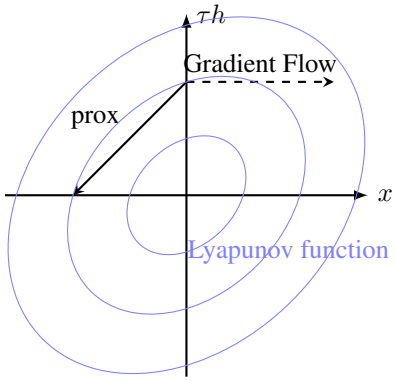


Figure 5: Illustration of new Lyapunov function

Thus we have skirted around the issue of the old Lyapunov function.

5.2. Property of Proximal Operator

Our new Lyapunov function, although resolving the issue of the old Lyapunov function, also makes previous proof techniques unavailable. In previous proofs, a lemma based on firm nonexpansiveness is used to give an upper bound for the Lyapunov function after applying proximal operator.

Lemma 5.1 (Firm nonexpansiveness)**.** *For any $t \geq 0$, if $\theta_t = 1$*

$$\|x_{t+1} - x_\star\|^2 + \frac{\gamma^2}{p^2}\|h_{t+1} - h_\star\|^2 \leq \left\|(\hat{x}_{t+1} - x_\star) - \frac{\gamma}{p}(h_t - h_\star)\right\|^2.$$

Note that the left-hand side of Lemma 5.1 is consistent with Ψ^{old}_{t+1}, thus this lemma is sufficient for proving convergence of the old Lyapunov function. However, this is not the case for the new Lyapunov function.

Therefore, we consider two independent lemmas instead, as shown in Lemmas 5.2 and 5.3.

Lemma 5.2 (Invariance)**.** *For any $t \geq 0$,*

$$(x_{t+1} - x_\star) - \frac{\gamma}{p}(h_{t+1} - h_\star) = (\hat{x}_{t+1} - x_\star) - \frac{\gamma}{p}(h_t - h_\star).$$

Lemma 5.3 (Monotonicity)**.** *For any $t \geq 0$, if $\theta_t = 1$*

$$\langle x_{t+1} - x_\star, h_{t+1} - h_\star \rangle \leq 0.$$

Note that Lemma 5.1 is a simple corollary of Lemmas 5.2 and 5.3, but not vice versa. Hence replacing the firm nonexpansiveness property with an invariance plus monotonicity gives us more freedom in our proof, allowing us to complete the proof of convergence for the new Lyapunov function. The full proof is written in Appendix A.3.

6. Extensions

In this section, we extend the ideas and techniques from the previous sections to other variants of ProxSkip. Specifically, we consider:

- SProxSkip (Mishchenko et al., 2022), which substitutes the gradient with a stochastic gradient;
- ProxSkip-VR (Malinovsky et al., 2022), which allows ProxSkip to be combined with variance reduction techniques;
- GradSkip+ (Maranjyan et al., 2022), which allows an additional unbiased compressor to be added to the gradient descent step and extends the random procedure of proximal operators to a general unbiased compressor.

6.1. SProxSkip

When only a stochastic gradient oracle is available or computing the full gradient is too costly, SProxSkip (Mishchenko et al., 2022) can be used as a stochastic gradient variant of ProxSkip. The only difference between the two algorithms lies in the substitution of $\nabla f(x_t)$ with $g_t(x_t)$. We provide the pseudocode of SProxSkip in Algorithm 5 in the Appendix.

For stochastic gradients, we need to introduce the following assumptions:

Assumption 6.1 (Unbiasedness)**.** For all $t \geq 0$, $g_t(x_t)$ is an unbiased estimator of the gradient $\nabla f(x_t)$. That is,

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

Assumption 6.2 (Expected smoothness)**.** There exist constants $A \geq 0$ and $C \geq 0$ such that for all $t \geq 0$

$$\mathbb{E}[\|g_t(x_t) - \nabla f(x_\star)\|^2 \mid x_t] \leq 2AD f(x_t, x_\star) + C$$

Remark 6.3. The commonly used bounded variance assumption $\mathrm{Var}[g_t(x_t) \mid x_t] \leq \sigma^2$ implies Assumption 6.2 with $A = L$ and $C = \sigma^2$.

For SProxSkip, we define a Lyapunov function Ψ^{old} and Ψ^{new} identical to ProxSkip's. We quote from (Mishchenko et al., 2022) for existing analysis results for SProxSkip:

Theorem 6.4. *Under Assumptions 6.1 and 6.2, let $0 < \gamma \leq 1/A$ and $0 < p \leq 1$. Then, the iterates of SProxSkip*

Method	Communication Complexity
FedAvg (Li et al., 2020)	$\mathcal{O}\left(\frac{\sigma^2}{\mu N K \varepsilon} + \frac{G^2 K}{\mu^2 \varepsilon}\right)$
Scaffold (Karimireddy et al., 2020)	$\widetilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu S K \varepsilon} + \frac{1}{\mu} + \frac{N}{S}\right)$
SProxSkip (Mishchenko et al., 2022)	$\mathcal{O}\left(\sqrt{1/\varepsilon\mu^2}\log(1/\varepsilon)\right)$
SProxSkip (Our)	$\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$

Table 1: Different stochastic local methods for strongly convex optimization.

(Algorithm 5) satisfy

$$\mathbb{E}\left[\Psi_T^{\text{old}}\right] \leq (1-\zeta^{\text{old}})^T \Psi_0^{\text{old}} + \frac{\gamma^2 C}{\zeta^{\text{old}}}$$

$$\zeta^{\text{old}} = \min(\gamma\mu(2-\gamma A), p^2).$$

γ_{crit} is the root for two branches to be equal if it exists, otherwise positive infinity.

Remark 6.5. Similar to Theorem 2.1, the bound in Theorem 6.4 is slightly improved compared to original result in (Mishchenko et al., 2022)..

By applying similar analysis techniques as Theorem 3.1, we can prove the following convergence result:

Theorem 6.6. *Under the same assumption as Theorem 6.4, if $\gamma \leq \gamma_{\text{crit}}$, let $\Delta = 0$. Otherwise, let Δ be the unique solution in $(0,1)$ of following equation:*

$$0 = \Delta^3 - 2\left(\frac{\mu\gamma}{p}+1\right)\Delta^2 + \left(\frac{L\mu\gamma^2}{p^2} + 2\frac{\mu\gamma}{p} + 2\right)\Delta - \frac{\gamma\mu(2-\gamma A) - p^2}{p-p^2}. \tag{5}$$

Then, the iterates of SProxSkip (Algorithm 5) satisfy

$$\mathbb{E}\left[\Psi_T^{\text{new}}\right] \leq (1-\zeta^{\text{new}})^T \Psi_0^{\text{new}} + \frac{\gamma^2 C}{\zeta^{\text{new}}}$$

$$\zeta^{\text{new}} = \begin{cases} \gamma\mu(2-\gamma A) & \text{if } \gamma \leq \gamma_{\text{crit}} \\ p^2 + \Delta(p-p^2) & \text{if } \gamma > \gamma_{\text{crit}}. \end{cases}$$

The new decreasing speed of Lyapunov function is $\zeta^{\text{new}} = \mathcal{O}(p)$, compared to $\zeta^{\text{old}} = \mathcal{O}(p^2)$, suggesting an asymptotically better result on proximal operator complexity.

We first review the asymptotic complexity of the old analysis: based on (Mishchenko et al., 2022), the stepsize is $\gamma = \min\left\{\frac{1}{A}, \frac{\varepsilon\mu}{2C}\right\}$, and the proximal operator frequency is $p = \sqrt{\gamma\mu}$, and the proximal operator complexity is $\max\left\{\sqrt{\frac{A}{\mu}}, \sqrt{\frac{2C}{\varepsilon\mu^2}}\right\}\log\left(\frac{2\Psi_0}{\varepsilon}\right)$. Note that the choice $p = \mathcal{O}(\sqrt{\gamma})$ is a direct consequence of $\zeta^{\text{old}} = \mathcal{O}(\min(\gamma, p^2))$, in order to make sure ζ^{old} is not too small. The improvement of ζ^{old} directly allows us to use a smaller frequency of p, thus reducing the proximal operator complexity.

Our new theory suggests following asymptotic complexity:

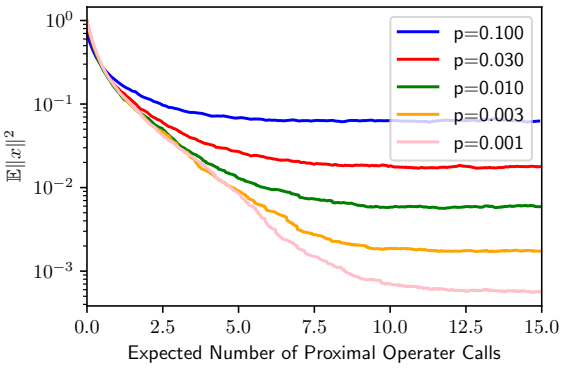


Figure 6: SProxSkip experiment.

Corollary 6.7. *Under the same assumption as Theorem 6.4, we define Δ^{ODE} as the solution of Equation (4) and $\tau = \frac{\gamma}{p}$. If ε is small, we set $p = \frac{\Delta^{\text{ODE}}}{\tau^2 C}\varepsilon$ and $T = \frac{1}{p\Delta^{\text{ODE}}}\log\frac{\Psi_0^{\text{new}}}{\varepsilon}$. Then we have $\mathbb{E}[\Psi_T^{\text{new}}] = 2\varepsilon + \mathcal{O}(\varepsilon^2)$ and the oracle complexity of proximal operator is $pT = \frac{1}{\Delta^{\text{ODE}}}\log\frac{\Psi_0^{\text{new}}}{\varepsilon}$. When κ is large, with good choice of τ obtained from Algorithm 4, the oracle complexity of proximal operator is $pT = \Theta(\sqrt{\kappa}\log(1/\varepsilon))$.*

Comparing the new and old theories, we see that the proximal operator complexity in the new theory only logarithmically depends on ε, which is much better than the old theory.

To verify our new convergence rate, we conducted an experiment by manually adding gradient noise: we fixed $\tau = \gamma/p$ to be the optimal value for ODEProx and tried different values of p and $\gamma = \tau p$. To reduce noise, we take the average of 1000 random runs of the algorithm. As shown in Figure 6, SProxSkip maintains linear convergence until it reaches a flat error, which can be further reduced by decreasing both γ and p.

6.2. ProxSkip-VR

In order to be compatible with various variance reduction techniques, ProxSkip-VR uses a stochastic gradient $g(x_t, y_t, \xi_t)$ similar to SProxSkip instead of the full gradient $\nabla f(x_t)$. It satisfies the following assumption:

Assumption 6.8 (Unbiasedness)**.** For all $t \geq 0, g_t = g(x_t, y_t, \xi_t)$ is an unbiased estimator of the gradient $\nabla f(x_t)$. That is,

$$\mathbb{E}[g_t \mid x_t, y_t] = \nabla f(x_t)$$

Assumption 6.9 (Variance reduction)**.** There exist constants $A \geq 0$ and $C \geq 0$ such that for all $t \geq 0$

$$\mathbb{E}[\|g_t - \nabla f(x_\star)\|^2 \mid x_t, y_t] \leq 2ADf(x_t, x_\star) + B\sigma_t + C$$
$$\mathbb{E}[\sigma_{t+1} \mid x_t, y_t] \leq 2\tilde{A}Df(x_t, x_\star) + \tilde{B}\sigma_t + \tilde{C}$$

The Lyapunov function for ProxSkip-VR is given as:

$$\Psi_t^{\text{old}} = \|x_t - x_\star\|^2 + \frac{\gamma^2}{p^2}\|h_t - h_\star\|^2 + \gamma^2 W\sigma_t$$

and its convergence analysis is as follows:

Theorem 6.10. *Under Assumptions 6.8 and 6.9 with $B \neq 0$ and a number $W > B/(1-\tilde{B})$, let $0 < \gamma \leq 1/(A + W\tilde{A})$ and $0 < p \leq 1$. Then, the iterates of ProxSkip-VR (Algorithm 6) satisfy*

$$\mathbb{E}\left[\Psi_T^{\text{old}}\right] \leq (1 - \widetilde{\zeta^{\text{old}}})^T \Psi_0^{\text{old}} + \frac{\gamma^2(C + W\tilde{C})}{\widetilde{\zeta^{\text{old}}}}$$

$$\zeta^{\text{old}} = \min(\gamma\mu(2 - \gamma(A + W\tilde{A})), p^2).$$
$$\widetilde{\zeta^{\text{old}}} = \min(\zeta^{\text{old}}, 1 - (B + W\tilde{B})/W)$$

γ_{crit} is the root for two branches of ζ^{old} to be equal if it exists, otherwise positive infinity.

Remark 6.11. In Theorem 6.10, we didn't discuss the condition of $B = 0$ as in (Malinovsky et al., 2022), because in that case, Assumption 6.9 for ProxSkip-VR degenerates into Assumption 6.2 for SProxSkip, therefore the Lyapunov function and analysis in Theorems 6.4 and 6.6 applies.

Our new Lyapunov function and convergence analysis are presented as follows:

$$\Psi_t^{\text{new}} = \|x_t - x_\star\|^2 + \frac{\gamma^2}{p^2}\|h_t - h_\star\|^2 - 2\Delta\frac{\gamma}{p}\langle x_t - x_\star, h_t - h_\star\rangle + \gamma^2 W\sigma_t$$

Theorem 6.12. *Under the same assumption as Theorem 6.10, if $\gamma \leq \gamma_{\text{crit}}$, let $\Delta = 0$. Otherwise, let Δ be the unique solution in $(0, 1)$ of following equation:*

$$0 = \Delta^3 - 2\left(\frac{\mu\gamma}{p} + 1\right)\Delta^2 + \left(\frac{L\mu\gamma^2}{p^2} + 2\frac{\mu\gamma}{p} + 2\right)\Delta - \frac{\gamma\mu(2 - \gamma(A + W\tilde{A})) - p^2}{p - p^2}.$$

Then, the iterates of ProxSkip-VR (Algorithm 6) satisfy

$$\mathbb{E}\left[\Psi_T^{\text{new}}\right] \leq (1 - \widetilde{\zeta^{\text{new}}})^T \Psi_0^{\text{new}} + \frac{\gamma^2(C + W\tilde{C})}{\widetilde{\zeta^{\text{new}}}}$$

$$\zeta^{\text{new}} = \begin{cases} \gamma\mu(2 - \gamma(A + W\tilde{A})) & \text{if } \gamma \leq \gamma_{\text{crit}} \\ p^2 + \Delta(p - p^2) & \text{if } \gamma > \gamma_{\text{crit}}. \end{cases}$$

$$\widetilde{\zeta^{\text{new}}} = \min(\zeta^{\text{new}}, 1 - (B + W\tilde{B})/W)$$

ProxSkip-VR encompasses a range of algorithms such as ProxSkip-HUB and ProxSkip-LSVRG as special cases (Malinovsky et al., 2022). We do not elaborate further on each of them individually since their convergence can be derived from the general description in Theorem 6.12.

6.3. GradSkip+

As a general theoretical framework, GradSkip+ is built on the foundation of unbiased compressors which satisfy the following condition, where $\mathbf{I}$ is identity matrix:

Definition 6.13 (Unbiased Compressors)**.** For any positive semidefinite matrix $\mathbf{\Omega} \geq 0$, denote by $\mathbb{B}^d(\mathbf{\Omega})$ the class of (possibly randomized) unbiased compression operators $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$ we have

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

The class $\mathbb{B}^d(\mathbf{\Omega})$ is a generalization of commonly used class $\mathbb{B}^d(\omega)$ of unbiased compressors with variance bound

$$E[\|C(x)\|^2] \leq (1 + \omega)\|x\|^2$$

for some scalar $\omega \geq 0$.

The analysis of GradSkip+ is further based on matrix smoothness, a generalization of Lipschitz-smoothness, defined as:

Definition 6.14 (Matrix Smoothness)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\mathbf{L}$-smooth with some symmetric and positive definite matrix $\mathbf{L} > 0$ if $D_f(x, y) \leq \frac{1}{2}\|x - y\|^2_{\mathbf{L}}, \forall x, y \in \mathbb{R}^d$.

The old analysis has the Lyapunov function:

$$\Psi^{\text{old}} = \|x_t - x_\star\|^2 + \gamma^2(1 + \omega^2)\|h_t - h_\star\|^2,$$

and its convergence is as follows:

Theorem 6.15. *Let f be μ-strongly convex with positive $\mu > 0$ and $\mathbf{L}$-smooth, $\mathcal{C}_\omega \in \mathbb{B}^d(\omega)$ and $\mathcal{C}_{\mathbf{\Omega}} \in \mathbb{B}^d(\mathbf{\Omega})$ be the compression operators, and*

$$\tilde{\mathbf{\Omega}} := \mathbf{I} + \omega(\omega + 2)\mathbf{\Omega}(\mathbf{I} + \mathbf{\Omega})^{-1}.$$

Then, if the stepsize $\gamma \leq \lambda_{\max}^{-1}(\mathbf{L}\tilde{\mathbf{\Omega}})$, the iterates of GradSkip+ (Algorithm 7) satisfy

$$\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq (1 - \zeta^{\text{old}})\Psi_t^{\text{old}},$$

$$\zeta^{\text{old}} = \min\left(\gamma\mu(2 - \gamma\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})), \frac{\lambda_{\min}(\tilde{\mathbf{\Omega}})}{(1 + \omega)^2}\right).$$

γ_{crit} is the root for two branches to be equal if it exists, otherwise positive infinity.

Remark 6.16. Again, the bound in Theorem 6.15 is slightly improved compared to original result in (Maranjyan et al., 2022).

The new analysis with the Lyapunov function and convergence guarantee is presented below:

$$\begin{aligned}\Psi_t^{\text{new}} =&\|x_t - x_\star\|^2 + \gamma^2(1+\omega)^2\|h_t - h_\star\|^2 \\ &-2\Delta\gamma(1+\omega)\langle x_t - x_\star, h_t - h_\star\rangle\end{aligned}$$

Theorem 6.17. *Under the same assumption as Theorem 6.15, if $\gamma \leq \gamma_{\text{crit}}$, let $\Delta = 0$ and Ψ_t^{new} is the same as Ψ_t^{old}.*

When $\gamma > \gamma_{\text{crit}}$, let

$$\begin{aligned}\tilde{\mathbf{\Omega}} &:= (1+\omega\Delta)\mathbf{I} + \omega(\omega+2-\Delta)\mathbf{\Omega}(\mathbf{I}+\mathbf{\Omega})^{-1} \\ \alpha &= \gamma^2\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}) \\ \zeta &= \frac{\lambda_{\min}(\tilde{\mathbf{\Omega}})}{(1+\omega)^2} \\ \beta &= 2\gamma\left(\Delta\left(\zeta\left(\omega+1\right)-1\right)+1\right)-\alpha\end{aligned}$$

There always exists some $\Delta \in (0,1)$ such that

$$(\beta\mu-\zeta)\mathbf{I} - \Delta\frac{\left(\omega\mathbf{I} + \left(\mathbf{\Omega}+\mathbf{I}\right)\left(\omega+1\right)\left(1-\zeta\right)\right)^2}{\omega\left(\mathbf{\Omega}+\mathbf{I}\right)} \succeq 0, \quad (6)$$

and for any Δ that satisfy above condtion, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{new}}] &\leq (1-\zeta^{\text{new}})\Psi_t^{\text{new}}, \\ \zeta^{\text{new}} &= \zeta > \zeta^{\text{old}}.\end{aligned}$$

Remark 6.18. The optimal choice of Δ is the largest value such that inequality (6) holds. We didn't compute the optimal value as it involves an intricate optimization. However, in practice, the optimal Δ can be easily determined with bisect search. It might also be easy to solve in special cases if $\mathbf{L}$ and $\mathbf{\Omega}$ have special structure.

Note that Theorem 6.17 and Theorem 3.1 for ProxSkip are slightly different in their formula since certain simplifications cannot be made for GradSkip+ due to $\mathbf{L}$ being a matrix. GradSkip+ encompasses a range of algorithms such as GradSkip(Maranjyan et al., 2022), ProxSkip and RandProx-FB (Condat & Richtárik, 2022) as special cases.

7. Conclusion

In this paper, we proposed a new way to improve the theoretical analysis of ProxSkip and provided deeper understanding on ProxSkip technique. In the context of federated learning, fewer proximal operator steps mean lower communication complexity. Through our discussion, we provided a lower proximal operator complexity when the proximal operator is the bottleneck. Our analysis revealed that in local training methods like ProxSkip, the step size should not be too large, especially when the frequency of proximal operators is relatively low. We for the first time demonstrated that SProxSkip, a local training method using stochastic gradients, can train to a certain accuracy with communication complexity logarithmically dependent on the accuracy. Finally, we showed that our proof technique can be directly extended to many other variants of ProxSkip.

Acknowledgement

This work was partially supported by NSF IIS 1838627, 1837956, 1956002, 2211492, CNS 2213701, CCF 2217003, DBI 2225775.

References

Bao, R., Wu, X., Xian, W., and Huang, H. Doubly sparse asynchronous learning for stochastic composite optimization. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pp. 1916–1922, 2022.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Communications of the ACM*, 55 (6):111–119, 2012.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200, 2005.

Condat, L. and Richtárik, P. Randprox: Primal-dual optimization algorithms with randomized proximal updates. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Gorbunov, E., Hanzely, F., and Richtárik, P. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 3556–3564. PMLR, 2021.

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

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

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Malinovsky, G., Yi, K., and Richtárik, P. Variance reduced proxskip: Algorithm, theory and application to federated learning. *NeurIPS*, 2022.

Mangasarian, O. L. and Solodov, M. V. Backpropagation convergence via deterministic nonmonotone perturbed minimization. *Advances in Neural Information Processing Systems*, 6, 1993.

Maranjyan, A., Safaryan, M., and Richtárik, P. Gradskip: Communication-accelerated local gradient methods with better computational complexity. *arXiv preprint arXiv:2210.16402*, 2022.

McDonald, R., Hall, K., and Mann, G. Distributed training strategies for the structured perceptron. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 456–464, 2010.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2, 2016.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15750–15769. PMLR, 2022.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Nesterov, Y. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Parikh, N., Boyd, S., et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Passty, G. B. Ergodic convergence to a zero of the sum of monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

Stich, S. U. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

Wu, X., Huang, F., Hu, Z., and Huang, H. Faster adaptive federated learning. *arXiv preprint arXiv:2212.00974*, 2022.

Zhang, J., De Sa, C., Mitliagkas, I., and Ré, C. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.

A. ProxSkip

A.1. Useful Lemmas

We define

$$\tau \coloneqq \frac{\gamma}{p}.$$

Proof of Lemma 5.1. Define

$$P(x) \coloneqq \mathrm{prox}_{\tau\psi}(x),$$
$$Q(x) \coloneqq x - P(x).$$

Due to firm nonexpansiveness, we have for any x, y:

$$\|P(x) - P(y)\|^2 + \|Q(x) - Q(y)\|^2 \leq \|x - y\|^2.$$

Let

$$x = \hat{x}_{t+1} - \tau h_t,$$
$$y = x_\star - \tau h_\star.$$

Notice that $P(y) = x_\star$, we have

$$\|x_{t+1} - x_\star\|^2 + \|\hat{x}_{t+1} - x_{t+1} - \tau h_t + \tau h_\star\|^2 \leq \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2.$$

According to Line 10 in Algorithm 1,

$$\hat{x}_{t+1} - x_{t+1} - \tau h_t = -\tau h_{t+1},$$

Therefore we have

$$\|x_{t+1} - x_\star\|^2 + \tau^2\|h_{t+1} - h_\star\|^2 \leq \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2$$

□

Another proof of Lemma 5.1 based on Lemmas 5.2 and 5.3. Due to Lemma 5.2, we have

$$\begin{aligned}\|\tau(h_t - h_\star) - (\hat{x}_{t+1} - x_\star)\|^2 =& \|\tau(h_{t+1} - h_\star) - (x_{t+1} - x_\star)\|^2 \\ =& \tau^2\|h_{t+1} - h_\star\|^2 + \|x_{t+1} - x_\star\|^2 - 2\tau\langle h_{t+1} - h_\star, x_{t+1} - x_\star\rangle \\ \geq& \tau^2\|h_{t+1} - h_\star\|^2 + \|x_{t+1} - x_\star\|^2.\end{aligned}$$

The last equation is due to Lemma 5.3. □

Proof of Lemma 5.2. According to Line 10 in Algorithm 1,

$$\tau h_{t+1} - x_{t+1} = \tau h_t - \hat{x}_{t+1},$$
$$\tau(h_{t+1} - h_\star) - (x_{t+1} - x_\star) = \tau(h_t - h_\star) - (\hat{x}_{t+1} - x_\star).$$

□

Proof of Lemma 5.3. According to Line 6 in Algorithm 1,

$$(\hat{x}_{t+1} - \tau h_t) - x_{t+1} \in \tau\partial\psi(x_{t+1}).$$

According to Line 10 in Algorithm 1,

$$-h_{t+1} \in \partial\psi(x_{t+1}).$$

Similarly, we have

$$-h_\star = -\nabla f(x_\star) \in \partial\psi(x_\star).$$

Since ψ is closed convex function, its subgradient $\partial\psi$ is a monotone operator, therefore

$$\langle(-h_{t+1}) - (-h_\star), x_{t+1} - x_\star\rangle \geq 0.$$

□

A.2. Old Analysis

Proof of Theorem 2.1.

$$\begin{aligned}\Psi_{t+1}^{\text{old}} =&\|x_{t+1}-x_\star\|^2+\tau^2\|h_{t+1}-h_\star\|^2\\ =&\|(x_{t+1}-x_\star)-\tau(h_{t+1}-h_\star)\|^2+2\tau\langle x_{t+1}-x_\star,h_{t+1}-h_\star\rangle.\end{aligned}$$

Step 1 (expand the proximal operator) If $\theta_t=1$, according to Lemmas 5.2 and 5.3,

$$\Psi_{t+1}^{\text{old}}|_{\theta_t=1}\leq\|(\hat{x}_{t+1}-x_\star)-\tau(h_t-h_\star)\|^2.$$

If $\theta_t=0$,

$$\Psi_{t+1}^{\text{old}}|_{\theta_t=0}=\|(\hat{x}_{t+1}-x_\star)-\tau(h_t-h_\star)\|^2+2\tau\langle\hat{x}_{t+1}-x_\star,h_t-h_\star\rangle.$$

Taking expectation gives

$$\mathbb{E}[\Psi_{t+1}^{\text{old}}]\leq\|(\hat{x}_{t+1}-x_\star)-\tau(h_t-h_\star)\|^2+2\tau(1-p)\langle\hat{x}_{t+1}-x_\star,h_t-h_\star\rangle. \tag{7}$$

Step 2 (expand the gradient descent) Expand $\hat{x}_{t+1}$ according to Line 3 in Algorithm 1 gives

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}]\leq&(1-p^2)\tau^2\|h_t-h_\star\|^2+\|x_t-x_\star\|^2\\ &-2\gamma\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle+\gamma^2\|\nabla f(x_t)-\nabla f(x_\star)\|^2.\end{aligned}$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle\geq\frac{1}{L}\|\nabla f(x_t)-\nabla f(x_\star)\|^2,$$

$$\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle\geq\mu\|x_t-x_\star\|^2.$$

Apply the above two inequalities with additional multipliers α and β, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}]\leq&(1-p^2)\tau^2\|h_t-h_\star\|^2+(1-\beta\mu)\|x_t-x_\star\|^2\\ &+(\gamma^2-\frac{\alpha}{L})\|\nabla f(x_t)-\nabla f(x_\star)\|^2\\ &+(\alpha+\beta-2\gamma)\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle.\end{aligned}$$

We require

$$\begin{aligned}\gamma^2-\frac{\alpha}{L}&=0,\\ \alpha+\beta-2\gamma&=0,\end{aligned}$$

which gives $\alpha=L\gamma^2$ and $\beta=\gamma(2-\gamma L)$. Then we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}]\leq&(1-p^2)\tau^2\|h_t-h_\star\|^2+(1-\gamma\mu(2-\gamma L))\|x_t-x_\star\|^2\\ \leq&(1-\zeta^{\text{old}})\Psi_t^{\text{old}},\end{aligned}$$

$$\zeta^{\text{old}}=\min(\gamma\mu(2-\gamma L),p^2).$$

□

A.3. New Analysis

Proof of Theorem 3.1. When $\gamma\leq\gamma_{\text{crit}}$, we have $\Delta=0$, so the Lyapunov function and the convergence rate is same as Theorem 2.1. Therefore, we only need to focus on $\gamma>\gamma_{\text{crit}}$ case. The following analysis only assumes $|\Delta|<1$, such that Lyapunov function is positive definite.

$$\begin{aligned}\Psi_{t+1}^{\text{new}} =&\|x_{t+1}-x_\star\|^2+\tau^2\|h_{t+1}-h_\star\|^2-2\Delta\tau\langle x_{t+1}-x_\star,h_{t+1}-h_\star\rangle\\ =&\|(x_{t+1}-x_\star)-\tau(h_{t+1}-h_\star)\|^2+2(1-\Delta)\tau\langle x_{t+1}-x_\star,h_{t+1}-h_\star\rangle.\end{aligned}$$

Step 1 (expand the proximal operator) If $\theta_t = 1$, according to Lemmas 5.2 and 5.3,

$$\Psi_{t+1}^{\text{new}}|_{\theta_t=1} \leq \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2.$$

If $\theta_t = 0$,

$$\Psi_{t+1}^{\text{new}}|_{\theta_t=0} = \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2 + 2(1-\Delta)\tau\langle \hat{x}_{t+1} - x_\star, h_t - h_\star\rangle.$$

Taking expectation gives

$$\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2 + 2(1-\Delta)\tau(1-p)\langle \hat{x}_{t+1} - x_\star, h_t - h_\star\rangle. \tag{8}$$

Step 2 (expand the gradient descent) Expand $\hat{x}_{t+1}$ according to Line 3 in Algorithm 1 gives

$$\begin{aligned}
\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \|x_{t+1} - x_\star\|^2 \\
& + (2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \tau^2)\|h_t - h_\star\|^2 \\
& + \gamma^2\|\nabla f(x_t) - \nabla f(x_\star)\|^2 \\
& + 2\Delta\gamma(\tau - \gamma)\langle \nabla f(x_t) - \nabla f(x_\star), h_t - h_\star\rangle \\
& - 2\gamma\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\
& - 2\Delta(\tau - \gamma)\langle h_t - h_\star, x_t - x_\star\rangle.
\end{aligned}$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \frac{1}{L}\|\nabla f(x_t) - \nabla f(x_\star)\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \mu\|x_t - x_\star\|^2.$$

, Additionally, we have

$$\|c_1(h_t - h_\star) - c_2(x_t - x_\star) - c_3(\nabla f(x_t) - \nabla f(x_\star))\|^2 \geq 0.$$

We also have

$$\|x_t - x_\star\|^2 + \tau^2\|h_t - h_\star\|^2 - 2\Delta\tau\langle x_t - x_\star, h_t - h_\star\rangle^2 = \Psi_t^{\text{new}}.$$

Apply the above three inequalities and one equality with additional multipliers α, β, 1 and $-(1-\zeta)$, we have

$$\begin{aligned}
\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \left(\zeta + c_2^2 - \beta\mu\right)\|x_{t+1} - x_\star\|^2 \\
& + \left(2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \zeta\tau^2 + c_1^2\right)\|h_t - h_\star\|^2 \\
& + \left(\gamma^2 + c_3^2 - \frac{\alpha}{L}\right)\|\nabla f(x_t) - \nabla f(x_\star)\|^2 \\
& - 2\left(\Delta\gamma^2 - \Delta\gamma\tau + c_1c_3\right)\langle \nabla f(x_t) - \nabla f(x_\star), h_t - h_\star\rangle \\
& + \left(\alpha + \beta - 2\gamma + 2c_2c_3\right)\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\
& + 2\left(\Delta\gamma - \Delta\zeta\tau - c_1c_2\right)\langle h_t - h_\star, x_t - x_\star\rangle \\
& + (1-\zeta)\Psi_t^{\text{new}}.
\end{aligned}$$

We require

$$\alpha + \beta - 2\gamma + 2c_2c_3 = 0, \tag{9}$$

$$\gamma^2 + c_3^2 - \frac{\alpha}{L} = 0, \tag{10}$$

$$\zeta + c_2^2 - \beta\mu = 0, \tag{11}$$

$$2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \zeta\tau^2 + c_1^2 = 0, \tag{12}$$

$$\Delta\gamma^2 - \Delta\gamma\tau + c_1c_3 = 0, \tag{13}$$

$$\Delta\gamma - \Delta\zeta\tau - c_1c_2 = 0. \tag{14}$$

Solving c_1, c_2, c_3, α and β on Equations (10) to (14) gives

$$\begin{aligned} c_1 &= \pm\sqrt{-2\Delta\gamma^2 + 2\Delta\gamma\tau + \gamma^2 - \zeta\tau^2}, \\ c_2 &= \frac{\Delta\left(\gamma - \zeta\tau\right)}{c_1}, \\ c_3 &= \frac{\Delta\gamma\left(\tau - \gamma\right)}{c_1}, \\ \alpha &= L(\gamma^2 + c_3^2), \\ \beta &= \frac{\zeta + c_2^2}{\mu}. \end{aligned}$$

Applying above solution into Equation (9) gives

$$\begin{aligned} E_1 \coloneqq & L\gamma^4\mu - 2L\gamma^3\mu\tau + L\gamma^2\mu\tau^2 - 2\gamma^3\mu + 2\gamma^2\mu\zeta\tau + 2\gamma^2\mu\tau + \gamma^2 - 2\gamma\mu\zeta\tau^2 - 2\gamma\zeta\tau + \zeta^2\tau^2 \\ & + \frac{-2L\gamma^4\mu + 2L\gamma^3\mu\tau + 4\gamma^3\mu - 4\gamma^2\mu\tau - 2\gamma^2\zeta + 2\gamma\zeta\tau}{\Delta} \\ & + \frac{L\gamma^4\mu - L\gamma^2\mu\zeta\tau^2 - 2\gamma^3\mu + \gamma^2\zeta + 2\gamma\mu\zeta\tau^2 - \zeta^2\tau^2}{\Delta^2} \\ = & 0. \end{aligned}$$

Step 4 (optimize free parameter Δ) $E_1(\zeta, \Delta) = 0$ is an implicit function, and we wish the decreasing speed ζ to be optimized. Therefore, we require

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\Delta} = -\frac{\partial_\Delta E_1}{\partial_\zeta E_1} = 0,$$

which further implies

$$\left(L\gamma^2\mu - 2\gamma\mu + \zeta\right)\left(\Delta\gamma^2 - \Delta\gamma\tau - \gamma^2 + \zeta\tau^2\right) = 0.$$

Two roots are

$$\begin{aligned} \zeta &= \gamma\mu(2 - \gamma L), \\ \zeta &= \frac{\gamma(\gamma + \Delta(\tau - \gamma)))}{\tau^2}. \end{aligned}$$

If we pick the first root, $E_1 = 0$ implies

$$\left(L\gamma^2\mu - 2\gamma\mu + 1\right)\left(L\mu\tau^2 - 2\mu\tau + 1\right) = 0,$$

which is impossible because

$$\left(L\gamma^2\mu - 2\gamma\mu + 1\right)\left(L\mu\tau^2 - 2\mu\tau + 1\right) > (1 - \gamma\mu)^2(1 - \mu\tau)^2 \geq 0.$$

Therefore, we proceed with the second root. In that case $E_1 = 0$ implies

$$E_3 \coloneqq \Delta^3 - 2\left(\mu\tau + 1\right)\Delta^2 + \left(L\mu\tau^2 + 2\mu\tau + 2\right)\Delta - \frac{\mu\tau^2(2 - \gamma L) - \gamma}{\tau - \gamma} = 0. \tag{15}$$

The existence and uniqueness of the solution is discussed in Lemma A.1. □

Lemma A.1. *Equation (15) contains exactly one root in interval $(0, 1)$ and doesn't have root in $(-1, 0]$.*

Proof of Lemma A.1. We first prove that E_3 is monotonically increasing in $(-1, 1)$.

$\frac{\mathrm{d}E_3}{\mathrm{d}\Delta}$ is a quadratic function for Δ. Its minimum is achieved at $\Delta = \frac{2}{3}(1 + \mu\tau)$.

If $\frac{2}{3}(1+\mu\tau) < 1$, then the minimum is $E_3|_{\Delta=\frac{2}{3}(1+\mu\tau)} = L\mu\tau^2 - \frac{4\mu^2\tau^2}{3} - \frac{2\mu\tau}{3} + \frac{2}{3} > -\frac{\mu^2\tau^2}{3} - \frac{2\mu\tau}{3} + \frac{2}{3} > \frac{1}{4}$. The first inequality follows $L > \mu$ and second follows $0 < \mu\tau < \frac{1}{2}$.

If $\frac{2}{3}(1+\mu\tau) \geq 1$, then $\inf_{\Delta\in(-1,1)} \frac{\mathrm{d}E_3}{\mathrm{d}\Delta} = \frac{\mathrm{d}E_3}{\mathrm{d}\Delta}|_{\Delta=1} = L\mu\tau^2 - 2\mu\tau + 1 > \mu^2\tau^2 - 2\mu\tau + 1 \geq 0$.

Due to monotonic property, there is at most one root in $(-1, 1)$.

$$E_3|_{\Delta=0} = -\frac{\mu\tau^2(2-\gamma L) - \gamma}{\tau - \gamma} = -\frac{\gamma\mu(2-\gamma L) - p^2}{p - p^2} < 0,$$

$$E_3|_{\Delta=1} = \frac{\tau}{\tau-\gamma}(1 - 2\mu\tau + L\mu\tau^2) > \frac{\tau}{\tau-\gamma}(1 - 2\mu\tau + \mu^2\tau^2) \geq 0.$$

Therefore there is exactly one root in $(0, 1)$. □

A.4. Parameter Selection

Proof of Theorem 3.2. If $p \geq \sqrt{\frac{1}{\kappa}}$, then we can always choose $\gamma = \frac{1}{L}$ to obtain best ζ^{new}.

Otherwise, let Δ_i and γ_i be the result of Lines 4 and 5 in Algorithm 2 after $(i+1)$-th iteration for $i \in \mathbb{N}$. We also use γ_{-1} to denote the initial step size. Additionally, we use $E_3(\Delta, \gamma)$ to denote the left hand side of Equation (3).

Note that according to the discussion in Lemma A.1, $E_3(\Delta, \gamma)$ is a monotonically increasing function for $\Delta \in (0, 1)$. Moreover, $E_3(\Delta, \gamma)$ is a quadratic and convex function for γ.

The iterate in Algorithm 2 can be rewritten as

$$\begin{aligned}\Delta_{i+1} &= \operatorname*{Solve}_{\Delta} E_3(\Delta, \gamma_i) = 0,\\ \gamma_{i+1} &= \operatorname*{argmin}_{\gamma} E_3(\Delta_{i+1}, \gamma).\end{aligned}$$

Since Δ_i is obtained by finding a root, we have $E_3(\Delta_i, \gamma_{i-1}) = 0$. Since γ_i is obtained by minimization, we have $E_3(\Delta_i, \gamma_i) \leq 0$. We also have $E_3(\Delta_{i+1}, \gamma_i) = 0$. According to the fact that $E_3(\Delta, \gamma)$ is monotonically increasing function for $\Delta \in (0, 1)$, we have $\Delta_{i+1} > \Delta_i$.

Since $\Delta_i < 1$ all the time, we can establish the convergence of Δ.

Next, the function $E_3(\Delta, \cdot)$ is always strongly convex and converge uniformly, therefore $\gamma_i = \operatorname{argmin}_\gamma E_3(\Delta_i, \gamma)$ also converge.

Finally, we have Δ converge to the maximum possible value, therefore $\zeta^{\text{new}} = p^2 + \Delta(p - p^2)$ will also converge to its maximum. □

A.5. Toy Model Setup

The following definitions are used for producing Figures 2, 3 and 6.

$$f(x) = \frac{\mu(\kappa-1)}{4}\left\{\frac{1}{2}\left[x^{(1)} + \sum_{i=1}^{d-1}(x^{(i)} - x^{(i+1)})^2 + x^{(d)}\right] - x^{(1)}\right\} + \frac{\mu}{2}\|x\|^2,$$

$$\psi(x) = \begin{cases} 0 & \text{if } x^{(1)} = 0 \\ +\infty & \text{otherwise} \end{cases},$$

$$d = 10, \kappa = 10, \mu = 0.1.$$

The global optimum is $x_\star = 0$. The initial point is $x_0 = [1, 0, 0, \dots]$ and $h_0 = 0$.

When SProxSkip is used, we set $g_t(x_t) = \nabla f(x_t) + 0.1 \times e$ where $e \sim \mathcal{N}(0, I)$.

A.6. Continuous Limit

Proof of Corollary 4.1. By substituting $p = \gamma/\tau$, we have

$$\Delta^3 - 2\left(\mu\tau + 1\right)\Delta^2 + \left(L\mu\tau^2 + 2\mu\tau + 2\right)\Delta - \frac{\mu\tau^2(2-\gamma L) - \gamma}{\tau - \gamma} = 0.$$

Taking the limit of $\gamma \to 0$, we get the condition Equation (4). □

Proof of Corollary 4.2. By substituting $\tau = \frac{1}{L} \times (\frac{1}{\Delta} - (1-\Delta))$, we have

$$\Delta^2\left(\kappa - 1\right)\left(\Delta^2 - 2\Delta + 2\right) - \left(1-\Delta\right)^2 = 0.$$

□

B. SProxSkip

B.1. Algorithm

Algorithm 5 SProxSkip

Stepsize $\gamma > 0$, probability $p > 0$, initial iterate $x_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, number of iterations $T \geq 1$
for $t = 0, 1, ..., T-1$ **do**
  $\hat{x}_{t+1} = x_t - \gamma(g_t(x_t) - h_t)$ ◇ Take a gradient-type step adjusted via the control variate h_t
  Flip a coin $\theta_t \in \{0, 1\}$ where Prob$(\theta_t = 1) = p$ ◇ Flip a coin that decides whether to skip the prox or not
  if $\theta_t = 1$ **then**
    $x_{t+1} = \mathrm{prox}_{\frac{\gamma}{p}\psi}(\hat{x}_{t+1} - \frac{\gamma}{p}h_t)$ ◇ Apply prox, but only very rarely! (with small probability p)
  else
    $x_{t+1} = \hat{x}_{t+1}$ ◇ Skip the prox!
  end if
  $h_{t+1} = h_t + \frac{p}{\gamma}(x_{t+1} - \hat{x}_{t+1})$ ◇ Update the control variate h_t
end for

B.2. Old Analysis

Proof of Theorem 6.4. Based on the same argument as step 1 in Appendix A.2, we have following middle result same as Equation (7):

$$\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2 + 2\tau(1-p)\langle \hat{x}_{t+1} - x_\star, h_t - h_\star\rangle. \tag{16}$$

Step 2 (expand the gradient descent) Expand $\hat{x}_{t+1}$ according to Line 3 in Algorithm 5 gives

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq &(1-p^2)\tau^2\|h_t - h_\star\|^2 + \|x_t - x_\star\|^2 \\ &- 2\gamma\langle x_t - x_\star, \mathbb{E}[g_t(x_t)] - \nabla f(x_\star)\rangle + \gamma^2\mathbb{E}\|g_t(x_t) - \nabla f(x_\star)\|^2 \\ \leq &(1-p^2)\tau^2\|h_t - h_\star\|^2 + \|x_t - x_\star\|^2 \\ &- 2\gamma\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle + 2\gamma^2 A D_f(x_t, x_\star) + \gamma^2 C.\end{aligned}$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq D_f(x_t, x_\star) + \frac{\mu}{2}\|x_t - x_\star\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \mu\|x_t - x_\star\|^2.$$

Apply the above two inequalities with additional multipliers β_1 and β_2, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq &(1-p^2)\tau^2\|h_t - h_\star\|^2 + (1 - (\beta_1/2 + \beta_2)\mu)\|x_t - x_\star\|^2 \\ &+ (\beta_1 + \beta_2 - 2\gamma)\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \\ &+ (2\gamma^2 A - \beta_1) D_f(x_t, x_\star) + \gamma^2 C.\end{aligned}$$

We require

$$\begin{aligned} 2\gamma^2 A - \beta_1 &= 0, \\ \beta_1 + \beta_2 - 2\gamma &= 0, \end{aligned}$$

which gives $\beta_1 = 2\gamma^2 A$ and $\beta_2 = 2\gamma(1-\gamma A)$. Then we have

$$\begin{aligned} \mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq & (1-p^2)\tau^2\|h_t - h_\star\|^2 + (1-\gamma\mu(2-\gamma A))\|x_t - x_\star\|^2 + \gamma^2 C \\ \leq & (1-\zeta^{\text{old}})\Psi_t^{\text{old}} + \gamma^2 C, \end{aligned}$$

$$\zeta^{\text{old}} = \min(\gamma\mu(2-\gamma A), p^2).$$

Apply Gronwall's inequality, we have

$$\mathbb{E}\left[\Psi_T^{\text{old}}\right] \leq (1-\zeta^{\text{old}})^T \Psi_0^{\text{old}} + \frac{\gamma^2 C}{\zeta^{\text{old}}},$$

□

B.3. New Analysis

Proof of Theorem 6.6. Based on the same argument as step 1 in Appendix A.3, we have following middle result same as Equation (8):

$$\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2 + 2(1-\Delta)\tau(1-p)\langle \hat{x}_{t+1} - x_\star, h_t - h_\star\rangle. \tag{17}$$

Step 2 (expand the gradient descent) Expand $\hat{x}_{t+1}$ according to Line 3 in Algorithm 5 gives

$$\begin{aligned} \mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \gamma^2 \mathbb{E}\|g_t(x_t) - \nabla f(x_\star)\|^2 - 2\gamma\langle x_t - x_\star, \mathbb{E}[g_t(x_t)] - \nabla f(x_\star)\rangle \\ & + 2\gamma\Delta(\tau-\gamma)\langle h_t - h_\star, \mathbb{E}[g_t(x_t)] - \nabla f(x_\star)\rangle \\ & + 2\Delta(\gamma-\tau)\langle h_t - h_\star, x_t - x_\star\rangle \\ & + (\gamma-\tau)\left(\gamma\left(2\Delta-1\right)-\tau\right)\|h_t - h_*\|^2 + \|x_t - x_\star\|^2 \\ \leq & 2\gamma\Delta\left(\tau-\gamma\right)\langle \nabla f(x_t) - \nabla f(x_\star), h_t - h_\star\rangle \\ & - 2\gamma\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\ & + (\gamma-\tau)\left(\gamma\left(2\Delta-1\right)-\tau\right)\|h_t - h_\star\|^2 \\ & + 2\Delta\left(\gamma-\tau\right)\langle h_t - h_\star, x_t - x_\star\rangle \\ & + (x_t - x_\star)^2 + 2A\gamma^2 D_f(x, x_*) + C\gamma^2. \end{aligned}$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \frac{1}{L}\|\nabla f(x_t) - \nabla f(x_\star)\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq D_f(x_t, x_\star) + \frac{\mu}{2}\|x_t - x_\star\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \mu\|x_t - x_\star\|^2.$$

Additionally, we have

$$\|c_1(h_t - h_\star) - c_2(x_t - x_\star) - c_3(\nabla f(x_t) - \nabla f(x_\star))\|^2 \geq 0.$$

We also have

$$\|x_t - x_\star\|^2 + \tau^2\|h_t - h_\star\|^2 - 2\Delta\tau\langle x_t - x_\star, h_t - h_\star\rangle^2 = \Psi_t^{\text{new}}.$$

Apply the above four inequalities and one equality with additional multipliers α, β_1, β_2, 1 and $-(1-\zeta)$, we have

$$\begin{aligned}
\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \left(\zeta + c_2^2 - (\beta_1/2 + \beta_2)\mu\right) \|x_{t+1} - x_\star\|^2 \\
& + \left(2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \zeta\tau^2 + c_1^2\right) \|h_t - h_\star\|^2 \\
& + \left(c_3^2 - \frac{\alpha}{L}\right) \|\nabla f(x_t) - \nabla f(x_\star)\|^2 \\
& - 2\left(\Delta\gamma^2 - \Delta\gamma\tau + c_1c_3\right) \langle \nabla f(x_t) - \nabla f(x_\star), h_t - h_\star \rangle \\
& + (\alpha + \beta_1 + \beta_2 - 2\gamma + 2c_2c_3) \langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star \rangle \\
& + 2\left(\Delta\gamma - \Delta\zeta\tau - c_1c_2\right) \langle h_t - h_\star, x_t - x_\star \rangle \\
& + (2\gamma^2 A - \beta_1) D_f(x_t - x_\star) \\
& + (1-\zeta)\Psi_t^{\text{new}} + \gamma^2 C.
\end{aligned}$$

We require

$$\alpha + \beta_1 + \beta_2 - 2\gamma + 2c_2c_3 = 0, \tag{18}$$

$$2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \zeta\tau^2 + c_1^2 = 0, \tag{19}$$

$$c_3^2 - \frac{\alpha}{L} = 0, \tag{20}$$

$$\zeta + c_2^2 - (\beta_1/2 + \beta_2)\mu = 0, \tag{21}$$

$$\Delta\gamma^2 - \Delta\gamma\tau + c_1c_3 = 0, \tag{22}$$

$$\Delta\gamma - \Delta\zeta\tau - c_1c_2 = 0. \tag{23}$$

$$2\gamma^2 A - \beta_1 = 0, \tag{24}$$

Solving $c_1, c_2, c_3, \alpha, \beta_1$ and β_2 on Equations (19) to (24) gives

$$\begin{aligned}
c_1 &= \pm\sqrt{-2\Delta\gamma^2 + 2\Delta\gamma\tau + \gamma^2 - \zeta\tau^2}, \\
c_2 &= \frac{\Delta(\gamma - \zeta\tau)}{c_1}, \\
c_3 &= \frac{\Delta\gamma(\tau - \gamma)}{c_1}, \\
\alpha &= Lc_3^2, \\
\beta_1 &= 2\gamma^2 A, \\
\beta_2 &= \frac{-A\gamma^2\mu + \zeta + c_2^2}{\mu}.
\end{aligned}$$

Applying above solution into Equation (18) gives

$$\begin{aligned}
E_1 \coloneqq & L\gamma^4\mu - 2L\gamma^3\mu\tau + L\gamma^2\mu\tau^2 - 2\gamma^3\mu + 2\gamma^2\mu\zeta\tau + 2\gamma^2\mu\tau + \gamma^2 \\
& - 2\gamma\mu\zeta\tau^2 - 2\gamma\zeta\tau + \zeta^2\tau^2 \\
& + \frac{-2A\gamma^4\mu + 2A\gamma^3\mu\tau + 4\gamma^3\mu - 4\gamma^2\mu\tau - 2\gamma^2\zeta + 2\gamma\zeta\tau}{\Delta} \\
& + \frac{A\gamma^4\mu - A\gamma^2\mu\zeta\tau^2 - 2\gamma^3\mu + \gamma^2\zeta + 2\gamma\mu\zeta\tau^2 - \zeta^2\tau^2}{\Delta^2} \\
= & 0.
\end{aligned}$$

Step 4 (optimize free parameter Δ) $E_1(\zeta, \Delta) = 0$ is an implicit function, and we wish the decreasing speed ζ to be optimized. Therefore, we require

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\Delta} = -\frac{\partial_\Delta E_1}{\partial_\zeta E_1} = 0,$$

which further implies

$$\left(A\gamma^2\mu - 2\gamma\mu + \zeta\right)\left(\Delta\gamma^2 - \Delta\gamma\tau - \gamma^2 + \zeta\tau^2\right) = 0.$$

According to the analysis for special case in Appendix A.3, the first root $\gamma\mu(2-\gamma A)$ is discarded. And we proceed with second root $\zeta = \frac{\gamma(\gamma+\Delta(\tau-\gamma)))}{\tau^2}$. Then, $E_1 = 0$ implies

$$\Delta^3 - 2\left(\mu\tau + 1\right)\Delta^2 + \left(L\mu\tau^2 + 2\mu\tau + 2\right)\Delta - \frac{\mu\tau^2(2-\gamma A) - \gamma}{\tau - \gamma} = 0. \tag{25}$$

Finally, we have

$$\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq (1-\zeta)\Psi_t^{\text{new}} + \gamma^2 C.$$

Apply Gronwall's inequality, we have

$$\mathbb{E}\left[\Psi_T^{\text{new}}\right] \leq (1-\zeta^{\text{new}})^T \Psi_0^{\text{new}} + \frac{\gamma^2 C}{\zeta^{\text{new}}}.$$

□

B.4. Asymptotic Convergence Rate

Proof of Corollary 6.7. Let E_3 be the left hand side of optimality condition Equation (5) for SProxSkip, and E_3^{ODE} be left hand side of optimality condition Equation (4) for ODEProx, we have $E_3 = E_3^{\text{ODE}} + \frac{1-2\mu\tau+A\mu\tau^2}{(1-p)^2}p + \mathcal{O}\left(p^2\right) = E_3^{\text{ODE}} + \mathcal{O}(p)$.

Let Δ and Δ^{ODE} be the unique solution in $(0,1)$ for optimality condition $E_3 = 0$ and $E_3^{\text{ODE}} = 0$ separately and $\zeta = \zeta^{\text{new}}$ for abbreviation, then we have $\Delta = \Delta^{\text{ODE}} + \mathcal{O}(p)$ and $\zeta = \Delta^{\text{ODE}}p + \mathcal{O}\left(p^2\right)$.

For each terms in upper bound of $\mathbb{E}[\Psi_T^{\text{new}}]$ in Theorem 6.6, we have

$$\begin{aligned}
(1-\zeta^{\text{new}})^T\Psi_0^{\text{new}} &= \left[\left(1 - p\Delta^{\text{ODE}} + \mathcal{O}\left(p^2\right)\right)^{\frac{1}{p\Delta^{\text{ODE}}}}\right]^{\log\frac{\Psi_0}{\varepsilon}}\Psi_0 \\
&= \left(\frac{1}{e} + \mathcal{O}(p)\right)^{\log\frac{\Psi_0}{\varepsilon}}\Psi_0 \\
&= \varepsilon(1 + \mathcal{O}(p)) \\
&= \varepsilon + \mathcal{O}\left(\varepsilon^2\right),
\end{aligned}$$

$$\frac{\gamma^2 C}{\zeta^{\text{new}}} = \frac{\tau^2 p^2 C}{p\Delta^{\text{ODE}} + \mathcal{O}(p^2)} = \frac{\tau^2 C}{\Delta^{\text{ODE}}}p + \mathcal{O}\left(p^2\right) = \varepsilon + \mathcal{O}\left(\varepsilon^2\right).$$

Finally we have

$$\mathbb{E}[\Psi_T^{\text{new}}] \leq 2\varepsilon + \mathcal{O}\left(\varepsilon^2\right). \tag{26}$$

The oracle complexity of proximal operator is $pT = \frac{1}{\Delta^{\text{ODE}}}\log\frac{\Psi_0^{\text{new}}}{\varepsilon} = \tilde{\Theta}(1/\Delta^{\text{ODE}})$. According the analysis of ODEProx, when κ is large, we have $\Delta^{\text{ODE}} = \Theta(1/\sqrt{\kappa})$. □

C. ProxSkip-VR

C.1. Algorithm

Algorithm 6 ProxSkip-VR

1: Stepsize $\gamma > 0$, probability $p > 0$, initial iterate $x_0 \in \mathbb{R}^d$, initial control vector $y_0 \in \mathbb{R}^d$, initial control variate $h_0 \in \mathbb{R}^d$, number of iterations $T \geq 1$
2: **for** $t = 0, 1, ..., T-1$ **do**
3: $g_t = g(x_t, y_t, \xi_t)$
4: $\hat{x}_{t+1} = x_t - \gamma(g_t - h_t)$ ◇ Take a gradient-type step adjusted via the control variate h_t
5: Construct new control vector y_{t+1}
6: Flip a coin $\theta_t \in \{0, 1\}$ where $\text{Prob}(\theta_t = 1) = p$ ◇ Flip a coin that decides whether to skip the prox or not
7: **if** $\theta_t = 1$ **then** ◇ Apply prox, but only very rarely! (with small probability p)
8: $x_{t+1} = \text{prox}_{\frac{\gamma}{p}\psi}(\hat{x}_{t+1} - \frac{\gamma}{p} h_t)$
9: **else**
10: $x_{t+1} = \hat{x}_{t+1}$ ◇ Skip the prox!
11: **end if**
12: $h_{t+1} = h_t + \frac{p}{\gamma}(x_{t+1} - \hat{x}_{t+1})$ ◇ Update the control variate h_t
13: **end for**

C.2. Old Analysis

Proof of Theorem 6.10. Based on the same argument as step 1 in Appendix A.2, we have following middle result similar Equation (7):

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq & \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2 + 2\tau(1-p)\langle \hat{x}_{t+1} - x_\star, h_t - h_\star\rangle \\ & + \gamma^2 W \mathbb{E}[\sigma_{t+1}].\end{aligned}$$

Step 2 (expand the gradient descent) Expand $\hat{x}_{t+1}$ according to Line 4 in Algorithm 6 gives

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq & (1-p^2)\tau^2\|h_t - h_\star\|^2 + \|x_t - x_\star\|^2 \\ & - 2\gamma\langle x_t - x_\star, \mathbb{E}[g_t] - \nabla f(x_\star)\rangle + \gamma^2 \mathbb{E}\|g_t - \nabla f(x_\star)\|^2 \\ & + \gamma^2 W \mathbb{E}[\sigma_{t+1}] \\ \leq & (1-p^2)\tau^2\|h_t - h_\star\|^2 + \|x_t - x_\star\|^2 \\ & - 2\gamma\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle + 2\gamma^2(A + W\tilde{A})D_f(x_t, x_\star) \\ & + \gamma^2(C + W\tilde{C}) + \gamma^2(B + W\tilde{B})\sigma_t.\end{aligned}$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq D_f(x_t, x_\star) + \frac{\mu}{2}\|x_t - x_\star\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \mu\|x_t - x_\star\|^2.$$

Apply the above two inequalities with additional multipliers β_1 and β_2, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq & (1-p^2)\tau^2\|h_t - h_\star\|^2 + (1 - (\beta_1/2 + \beta_2)\mu)\|x_t - x_\star\|^2 \\ & + (\beta_1 + \beta_2 - 2\gamma)\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \\ & + (2\gamma^2(A + W\tilde{A}) - \beta_1)D_f(x_t, x_\star) + \gamma^2(C + W\tilde{C}) \\ & + \gamma^2(B + W\tilde{B})\sigma_t.\end{aligned}$$

We require

$$\begin{aligned} 2\gamma^2 A - \beta_1 &= 0, \\ \beta_1 + \beta_2 - 2\gamma &= 0, \end{aligned}$$

which gives $\beta_1 = 2\gamma^2(A + W\tilde{A})$ and $\beta_2 = 2\gamma(1 - \gamma(A + W\tilde{A}))$. Then we have

$$\begin{aligned} \mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq & (1-p^2)\tau^2\|h_t - h_\star\|^2 + (1 - \gamma\mu(2 - \gamma(A + W\tilde{A})))\|x_t - x_\star\|^2 \\ & + \gamma^2(B + W\tilde{B})\sigma_t + \gamma^2(C + W\tilde{C}) \\ \leq & (1 - \widetilde{\zeta^{\text{old}}})\Psi_t^{\text{old}} + \gamma^2(C + W\tilde{C}), \end{aligned}$$

$$\widetilde{\zeta^{\text{old}}} = \min(\gamma\mu(2 - \gamma(A + W\tilde{A})), p^2, 1 - (B + W\tilde{B})/W).$$

Apply Gronwall's inequality, we have

$$\mathbb{E}\left[\Psi_T^{\text{old}}\right] \leq (1 - \widetilde{\zeta^{\text{old}}})^T \Psi_0^{\text{old}} + \frac{\gamma^2(C + W\tilde{C})}{\widetilde{\zeta^{\text{old}}}}.$$

□

C.3. New Analysis

Proof of Theorem 6.12. Based on the same argument as step 1 in Appendix A.3, we have following middle result similar to Equation (8):

$$\begin{aligned} \mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \|(\hat{x}_{t+1} - x_\star) - \tau(h_t - h_\star)\|^2 + 2(1-\Delta)\tau(1-p)\langle \hat{x}_{t+1} - x_\star, h_t - h_\star\rangle \\ & + \gamma^2 W \mathbb{E}[\sigma_{t+1}]. \end{aligned}$$

Step 2 (expand the gradient descent) Expand $\hat{x}_{t+1}$ according to Line 4 in Algorithm 6 gives

$$\begin{aligned} \mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \gamma^2 \mathbb{E}\|g_t(x_t) - \nabla f(x_\star)\|^2 - 2\gamma\langle x_t - x_\star, \mathbb{E}[g_t] - \nabla f(x_\star)\rangle \\ & + 2\gamma\Delta(\tau - \gamma)\langle h_t - h_\star, \mathbb{E}[g_t] - \nabla f(x_\star)\rangle \\ & + 2\Delta(\gamma - \tau)\langle h_t - h_\star, x_t - x_\star\rangle \\ & + (\gamma - \tau)\left(\gamma\left(2\Delta - 1\right) - \tau\right)\|h_t - h_*\|^2 + \|x_t - x_\star\|^2 \\ & + \gamma^2 W \mathbb{E}[\sigma_{t+1}] \\ \leq & 2\gamma\Delta\left(\tau - \gamma\right)\langle \nabla f(x_t) - \nabla f(x_\star), h_t - h_\star\rangle \\ & - 2\gamma\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\ & + (\gamma - \tau)\left(\gamma\left(2\Delta - 1\right) - \tau\right)\|h_t - h_\star\|^2 \\ & + 2\Delta\left(\gamma - \tau\right)\langle h_t - h_\star, x_t - x_\star\rangle \\ & + (x_t - x_\star)^2 + 2(A + W\tilde{A})\gamma^2 D_f(x, x_*) + \gamma^2(C + W\tilde{C}) \\ & + \gamma^2(B + W\tilde{B})\sigma_t. \end{aligned}$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \frac{1}{L}\|\nabla f(x_t) - \nabla f(x_\star)\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq D_f(x_t, x_\star) + \frac{\mu}{2}\|x_t - x_\star\|^2,$$

$$\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \mu\|x_t - x_\star\|^2.$$

Additionally, we have

$$\|c_1(h_t - h_\star) - c_2(x_t - x_\star) - c_3(\nabla f(x_t) - \nabla f(x_\star))\|^2 \geq 0.$$

We also have

$$\|x_t - x_\star\|^2 + \tau^2\|h_t - h_\star\|^2 - 2\Delta\tau\langle x_t - x_\star, h_t - h_\star\rangle^2 = \Psi_t^{\text{new}}.$$

Apply the above four inequalities and one equality with additional multipliers α, β_1, β_2, 1 and $-(1-\zeta)$, we have

$$\begin{aligned}
\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq & \left(\zeta + c_2^2 - (\beta_1/2 + \beta_2)\mu\right)\|x_{t+1} - x_\star\|^2 \\
& + \left(2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \zeta\tau^2 + c_1^2\right)\|h_t - h_\star\|^2 \\
& + \left(c_3^2 - \frac{\alpha}{L}\right)\|\nabla f(x_t) - \nabla f(x_\star)\|^2 \\
& - 2\left(\Delta\gamma^2 - \Delta\gamma\tau + c_1c_3\right)\langle \nabla f(x_t) - \nabla f(x_\star), h_t - h_\star\rangle \\
& + (\alpha + \beta_1 + \beta_2 - 2\gamma + 2c_2c_3)\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\
& + 2\left(\Delta\gamma - \Delta\zeta\tau - c_1c_2\right)\langle h_t - h_\star, x_t - x_\star\rangle \\
& + (2\gamma^2(A + W\tilde{A}) - \beta_1)D_f(x_t - x_\star) \\
& + (1-\zeta)\Psi_t^{\text{new}} + \gamma^2(C + W\tilde{C}) \\
& + \gamma^2(B + W\tilde{B})\sigma_t.
\end{aligned}$$

We require

$$\alpha + \beta_1 + \beta_2 - 2\gamma + 2c_2c_3 = 0, \tag{27}$$

$$2\Delta\gamma^2 - 2\Delta\gamma\tau - \gamma^2 + \zeta\tau^2 + c_1^2 = 0, \tag{28}$$

$$c_3^2 - \frac{\alpha}{L} = 0, \tag{29}$$

$$\zeta + c_2^2 - (\beta_1/2 + \beta_2)\mu = 0, \tag{30}$$

$$\Delta\gamma^2 - \Delta\gamma\tau + c_1c_3 = 0, \tag{31}$$

$$\Delta\gamma - \Delta\zeta\tau - c_1c_2 = 0. \tag{32}$$

$$2\gamma^2(A + W\tilde{A}) - \beta_1 = 0, \tag{33}$$

Solving $c_1, c_2, c_3, \alpha, \beta_1$ and β_2 on Equations (28) to (33) gives

$$\begin{aligned}
c_1 &= \pm\sqrt{-2\Delta\gamma^2 + 2\Delta\gamma\tau + \gamma^2 - \zeta\tau^2}, \\
c_2 &= \frac{\Delta(\gamma - \zeta\tau)}{c_1}, \\
c_3 &= \frac{\Delta\gamma(\tau - \gamma)}{c_1}, \\
\alpha &= Lc_3^2, \\
\beta_1 &= 2\gamma^2(A + W\tilde{A}), \\
\beta_2 &= \frac{-(A + W\tilde{A})\gamma^2\mu + \zeta + c_2^2}{\mu}.
\end{aligned}$$

Applying above solution into Equation (27) gives

$$\begin{aligned}
E_1 :=& L\gamma^4\mu - 2L\gamma^3\mu\tau + L\gamma^2\mu\tau^2 - 2\gamma^3\mu + 2\gamma^2\mu\zeta\tau + 2\gamma^2\mu\tau + \gamma^2 \\
& - 2\gamma\mu\zeta\tau^2 - 2\gamma\zeta\tau + \zeta^2\tau^2 \\
& + \frac{-2A'\gamma^4\mu + 2A'\gamma^3\mu\tau + 4\gamma^3\mu - 4\gamma^2\mu\tau - 2\gamma^2\zeta + 2\gamma\zeta\tau}{\Delta} \\
& + \frac{A'\gamma^4\mu - A'\gamma^2\mu\zeta\tau^2 - 2\gamma^3\mu + \gamma^2\zeta + 2\gamma\mu\zeta\tau^2 - \zeta^2\tau^2}{\Delta^2} \\
=& 0.
\end{aligned}$$

For simplicity, in above formula, we use $A' = A + W\tilde{A}$.

Step 4 (optimize free parameter Δ) $E_1(\zeta, \Delta) = 0$ is an implicit function, and we wish the decreasing speed ζ to be optimized. Therefore, we require

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\Delta} = -\frac{\partial_\Delta E_1}{\partial_\zeta E_1} = 0,$$

which further implies

$$\left(A'\gamma^2\mu - 2\gamma\mu + \zeta\right)\left(\Delta\gamma^2 - \Delta\gamma\tau - \gamma^2 + \zeta\tau^2\right) = 0$$

According to the analysis for special case in Appendix A.3, the first root $\gamma\mu(2-\gamma A')$ is discarded. And we proceed with second root $\zeta = \frac{\gamma(\gamma+\Delta(\tau-\gamma)))}{\tau^2}$. Then, $E_1 = 0$ implies

$$\Delta^3 - 2\left(\mu\tau + 1\right)\Delta^2 + \left(L\mu\tau^2 + 2\mu\tau + 2\right)\Delta - \frac{\mu\tau^2(2-\gamma A') - \gamma}{\tau - \gamma} = 0. \tag{34}$$

Finally, we have

$$\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq \max((1-\zeta), (B + W\tilde{B})/W)\Psi_t^{\text{new}} + \gamma^2(C + W\tilde{C}).$$

We define

$$\widetilde{\zeta^{\text{new}}} = \min(\zeta, 1 - (B + W\tilde{B})/W).$$

Apply Gronwall's inequality, we have

$$\mathbb{E}\left[\Psi_T^{\text{new}}\right] \leq (1 - \widetilde{\zeta^{\text{new}}})^T \Psi_0^{\text{new}} + \frac{\gamma^2(C + W\tilde{C})}{\widetilde{\zeta^{\text{new}}}}.$$

□

D. GradSkip+

D.1. Algorithm

Algorithm 7 GradSkip+

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

D.2. Useful Lemmas

We define

$$p \coloneqq \frac{1}{1+\omega} \in (0, 1],$$
$$\tau \coloneqq \gamma(1+\omega).$$

Lemma D.1. *For any $t \geq 0$,*

$$\|\tau(h_{t+1} - h_\star) - (x_{t+1} - x_\star)\|^2 = \left\|\tau(\hat{h}_{t+1} - h_\star) - (\hat{x}_{t+1} - x_\star)\right\|^2.$$

Proof of Lemma D.1. According to Line 8 in Algorithm 7,

$$\tau h_{t+1} - x_{t+1} = \tau \hat{h}_{t+1} - \hat{x}_{t+1},$$
$$\tau(h_{t+1} - h_\star) - (x_{t+1} - x_\star) = \tau(\hat{h}_{t+1} - h_\star) - (\hat{x}_{t+1} - x_\star).$$

□

Lemma D.2. *For any $t \geq 0$, let $\mathbb{E}_{\mathcal{C}_\omega,t}[\cdot]$ be the expectation over the randomness from unbiased compressor $\mathcal{C}_\omega$ at t-th step,*

$$\mathbb{E}_{\mathcal{C}_\omega,t}\langle x_{t+1} - x_\star, h_{t+1} - h_\star\rangle \leq \left(1 - \frac{1}{1+\omega}\right)\langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star\rangle.$$

Proof of Lemma D.2. We define

$$x_{t+1}^+ = \mathrm{prox}_{\tau\psi}\left(\hat{x}_{t+1} - \tau\hat{h}_{t+1})\right),$$
$$h_{t+1}^+ = \hat{h}_{t+1} + \frac{1}{\tau}(x_{t+1}^+ - \hat{x}_{t+1}).$$

According to Lemma 5.3, $\langle x_{t+1}^+ - x_\star, h_{t+1}^+ - h_\star\rangle \leq 0$.

Then we let $s = \frac{1}{1+\omega}\mathcal{C}_\omega(\hat{x}_{t+1} - x_{t+1}^+)$, and we have

$$x_{t+1} = \hat{x}_{t+1} - s,$$
$$h_{t+1} = \hat{h}_{t+1} - \frac{1}{\tau}s.$$

Note that

$$\mathbb{E}_{\mathcal{C}_\omega,t}[s] = \frac{1}{1+\omega}(\hat{x}_{t+1} - x_{t+1}^+),$$
$$\mathbb{E}_{\mathcal{C}_\omega,t}\|s\|^2 \leq \frac{1}{1+\omega}\left\|\hat{x}_{t+1} - x_{t+1}^+\right\|^2.$$

Then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}_\omega,t}\langle x_{t+1} - x_\star, h_{t+1} - h_\star\rangle \\
=&\mathbb{E}_{\mathcal{C}_\omega,t}\left[\frac{1}{\tau}\langle \hat{x}_{t+1} - s - x_\star, \tau\hat{h}_{t+1} - s - h_\star\rangle\right] \\
=&\frac{1}{\tau}\langle \hat{x}_{t+1} - \mathbb{E}_{\mathcal{C}_\omega,t}[s] - x_\star, \tau\hat{h}_{t+1} - \mathbb{E}_{\mathcal{C}_\omega,t}[s] - h_\star\rangle \\
&+ \frac{1}{\tau}\left(\mathbb{E}_{\mathcal{C}_\omega,t}\|s\|^2 - \|\mathbb{E}_{\mathcal{C}_\omega,t}[s]\|^2\right) \\
\leq&\frac{1}{\tau}\langle \hat{x}_{t+1} - p(\hat{x}_{t+1} - x_{t+1}^+) - x_\star, \tau\hat{h}_{t+1} - p(\hat{x}_{t+1} - x_{t+1}^+) - h_\star\rangle \\
&+ \frac{1}{\tau}p(1-p)\left\|\hat{x}_{t+1} - x_{t+1}^+\right\|^2 \\
=&(1-p)\langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star\rangle + p\langle x_{t+1}^+ - x_\star, h_{t+1}^+ - h_\star\rangle \\
\leq&(1-p)\langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star\rangle.
\end{aligned}$$

□

D.3. Old Analysis

Proof of Theorem 2.1.

$$\begin{aligned}
\Psi_{t+1}^{\text{old}} =&\|x_{t+1} - x_\star\|^2 + \tau^2\|h_{t+1} - h_\star\|^2 \\
=&\|(x_{t+1} - x_\star) - \tau(h_{t+1} - h_\star)\|^2 + 2\tau\langle x_{t+1} - x_\star, h_{t+1} - h_\star\rangle.
\end{aligned}$$

Step 1 (expand the proximal operator) According to Lemmas D.1 and D.2,

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}_\omega,t}[\Psi_{t+1}^{\text{old}}]\\ \le&\left\|(\hat{x}_{t+1}-x_\star)-\tau(\hat{h}_{t+1}-h_\star)\right\|^2+2\tau(1-p)\langle\hat{x}_{t+1}-x_\star,\hat{h}_{t+1}-h_\star\rangle.\end{aligned}$$

Step 2 (expand the gradient descent) Let

$$s=(\mathbf{I}+\mathbf{\Omega})^{-1}\mathcal{C}_{\mathbf{\Omega}}(\nabla f(x_t)-h_t),$$

then

$$\begin{aligned}\hat{h}_{t+1}&=\nabla f(x_t)-s,\\ \hat{x}_{t+1}&=\nabla x_t-\gamma s.\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}_\omega,t}[\Psi_{t+1}^{\text{old}}]\\ \le&\left\|(\hat{x}_{t+1}-x_\star)-\tau(\hat{h}_{t+1}-h_\star)\right\|^2+2\tau(1-p)\langle\hat{x}_{t+1}-x_\star,\hat{h}_{t+1}-h_\star\rangle\\ =&\tau^2\|\nabla f(x_t)-\nabla f(x_\star)\|^2-2\gamma\langle\nabla f(x_t)-\nabla f(x_\star),x_{t+1}-x_\star\rangle\\ &+\|x_{t+1}-x_\star\|^2-2(\tau^2-\gamma^2)\langle\nabla f(x_t)-\nabla f(x_\star),s\rangle+(\tau^2-\gamma^2)\|s\|^2,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}[\Psi_{t+1}^{\text{old}}]\\ \le&\tau^2\|\nabla f(x_t)-\nabla f(x_\star)\|^2-2\gamma\langle\nabla f(x_t)-\nabla f(x_\star),x_{t+1}-x_\star\rangle\\ &+\|x_{t+1}-x_\star\|^2-2(\tau^2-\gamma^2)\langle\nabla f(x_t)-\nabla f(x_\star),\mathbb{E}[s]\rangle+(\tau^2-\gamma^2)\mathbb{E}\|s\|^2\\ \le&-(\tau^2-\gamma^2)\|\nabla f(x_t)-\nabla f(x_\star)\|^2_{(\mathbf{I}+\mathbf{\Omega})^{-1}}+\tau^2\|\nabla f(x_t)-\nabla f(x_\star)\|^2\\ &-2\gamma\langle\nabla f(x_t)-\nabla f(x_\star),x_t-x_\star\rangle+(\tau^2-\gamma^2)\|h_t-h_\star\|^2_{(\mathbf{I}+\mathbf{\Omega})^{-1}}+\|x_t-x_\star\|^2\\ =&\langle\nabla f(x_t)-\nabla f(x_\star),(\tau^2\mathbf{I}-(\tau^2-\gamma^2)(\mathbf{I}+\mathbf{\Omega})^{-1})(\nabla f(x_t)-\nabla f(x_\star))\rangle\\ &-2\gamma\langle\nabla f(x_t)-\nabla f(x_\star),x_t-x_\star\rangle+(\tau^2-\gamma^2)\|h_t-h_\star\|^2_{(\mathbf{I}+\mathbf{\Omega})^{-1}}+\|x_t-x_\star\|^2\\ =&\langle\nabla f(x_t)-\nabla f(x_\star),(\gamma^2\tilde{\mathbf{\Omega}})(\nabla f(x_t)-\nabla f(x_\star))\rangle\\ &-2\gamma\langle\nabla f(x_t)-\nabla f(x_\star),x_t-x_\star\rangle+(\tau^2-\gamma^2)\|h_t-h_\star\|^2_{(\mathbf{I}+\mathbf{\Omega})^{-1}}+\|x_t-x_\star\|^2.\end{aligned}$$

The last step is due to the following equivalent formula of $\tilde{\mathbf{\Omega}}$:

$$\tilde{\mathbf{\Omega}}:=(1+\omega)^2\mathbf{I}-\omega(2+\omega)(\mathbf{I}+\mathbf{\Omega})^{-1}.$$

Step 3 (apply strongly convexity and smoothness) We have

$$\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle\ge\|\nabla f(x_t)-\nabla f(x_\star)\|^2_{\mathbf{L}^{-1}}$$

$$\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle\ge\mu\|x_t-x_\star\|^2.$$

Apply the above two inequalities with additional multipliers α and β, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}^{\text{old}}]\le&(1-p^2)\tau^2\|h_t-h_\star\|^2_{(\mathbf{I}+\mathbf{\Omega})^{-1}}+(1-\beta\mu)\|x_t-x_\star\|^2\\ &+\langle\nabla f(x_t)-\nabla f(x_\star),(\gamma^2\tilde{\mathbf{\Omega}}-\alpha\mathbf{L}^{-1})(\nabla f(x_t)-\nabla f(x_\star))\rangle\\ &+(\alpha+\beta-2\gamma)\langle x_t-x_\star,\nabla f(x_t)-\nabla f(x_\star)\rangle.\end{aligned}$$

We require

$$\lambda_{\max}(\gamma^2\tilde{\mathbf{\Omega}} - \alpha\mathbf{L}^{-1}) \leq 0,$$

The smallest α we that satisfy above condition is $\alpha = \gamma^2\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})$. We also require $\alpha + \beta - 2\gamma = 0$, so we have

$$\begin{aligned}
\mathbb{E}[\Psi_{t+1}^{\text{old}}] \leq& (1-p^2)\tau^2\|h_t - h_\star\|^2_{(\mathbf{I}+\mathbf{\Omega})^{-1}} + (1-\gamma\mu(2-\gamma\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})))\|x_t - x_\star\|^2 \\
\leq& (1-p^2)\tau^2\lambda_{\max}((\mathbf{I}+\mathbf{\Omega})^{-1})\|h_t - h_\star\|^2 \\
&+ (1-\gamma\mu(2-\gamma\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})))\|x_t - x_\star\|^2 \\
\leq& (1-\zeta^{\text{old}})\Psi_t^{\text{old}}
\end{aligned}$$

$$\zeta^{\text{old}} = \min\left(\gamma\mu(2-\gamma\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}})), 1 - \frac{1-p^2}{1+\lambda_{\min}(\mathbf{\Omega})}\right).$$

□

D.4. New Analysis

Proof of Theorem 3.1. We only need to focus on $\gamma > \gamma_{\text{crit}}$ case.

$$\begin{aligned}
\Psi_{t+1}^{\text{new}} =& \|x_{t+1} - x_\star\|^2 + \tau^2\|h_{t+1} - h_\star\|^2 - 2\Delta\tau\langle x_{t+1} - x_\star, h_{t+1} - h_\star\rangle \\
=& \|(x_{t+1} - x_\star) - \tau(h_{t+1} - h_\star)\|^2 + 2(1-\Delta)\tau\langle x_{t+1} - x_\star, h_{t+1} - h_\star\rangle.
\end{aligned}$$

Step 1 (expand the proximal operator) According to Lemmas D.1 and D.2,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}_\omega,t}[\Psi_{t+1}^{\text{new}}] \\
\leq& \left\|(\hat{x}_{t+1} - x_\star) - \tau(\hat{h}_{t+1} - h_\star)\right\|^2 + 2(1-\Delta)\tau(1-p)\langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star\rangle.
\end{aligned}$$

Step 2 (expand the gradient descent) Let

$$s = (\mathbf{I}+\mathbf{\Omega})^{-1}\mathcal{C}_\omega(\nabla f(x_t) - h_t),$$

then

$$\begin{aligned}
\hat{h}_{t+1} &= \nabla f(x_t) - s, \\
\hat{x}_{t+1} &= \nabla x_t - \gamma s,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}_\omega,t}[\Psi_{t+1}^{\text{new}}] \\
\leq& \left\|(\hat{x}_{t+1} - x_\star) - \tau(\hat{h}_{t+1} - h_\star)\right\|^2 + 2(1-\Delta)\tau(1-p)\langle \hat{x}_{t+1} - x_\star, \hat{h}_{t+1} - h_\star\rangle \\
=& \tau^2\|\nabla f(x_t) - \nabla f(x_\star)\|^2 - 2(\Delta(\tau-\gamma)+\gamma)\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\
&- 2(\tau-\gamma)((1-\Delta)\gamma+\tau)\langle \nabla f(x_t) - \nabla f(x_\star), s\rangle \\
&+ \|x_t - x_\star\|^2 + 2\Delta(\tau-\gamma)\langle x_t - x_\star, s\rangle + (\tau-\gamma)((1-2\Delta)\gamma+\tau)\|s\|^2,
\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}[\Psi_{t+1}^{\text{new}}] \\
\leq& \tau^2\|\nabla f(x_t) - \nabla f(x_\star)\|^2 - 2(\Delta(\tau-\gamma)+\gamma)\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\
&- 2(\tau-\gamma)((1-\Delta)\gamma+\tau)\langle \nabla f(x_t) - \nabla f(x_\star), \mathbb{E}[s]\rangle \\
&+ \|x_t - x_\star\|^2 + 2\Delta(\tau-\gamma)\langle x_t - x_\star, \mathbb{E}[s]\rangle + (\tau-\gamma)((1-2\Delta)\gamma+\tau)\mathbb{E}\|s\|^2 \\
\leq& \langle \nabla f(x_t) - \nabla f(x_\star), (\tau^2 - (\tau^2-\gamma^2)(\mathbf{I}+\mathbf{\Omega})^{-1})(\nabla f(x_t) - \nabla f(x_\star))\rangle \\
&+ 2\Delta\gamma(\tau-\gamma)\langle \nabla f(x_t) - \nabla f(x_\star), (\mathbf{I}+\mathbf{\Omega})^{-1}(h_t - h_\star)\rangle \\
&+ \langle \nabla f(x_t) - \nabla f(x_\star), (-2\gamma\mathbf{I} - 2\Delta(\tau-\gamma)(\mathbf{I} - (\mathbf{I}-\mathbf{\Omega})^{-1}))(x_t - x_\star)\rangle \\
&+ (\gamma-\tau)(2\Delta\gamma-\gamma-\tau)\langle h_t - h_\star, (\mathbf{I}+\mathbf{\Omega})^{-1}(h_t - h_\star)\rangle \\
&+ 2\Delta(\gamma-\tau)\langle h_t - h_\star, (\mathbf{I}+\mathbf{\Omega})^{-1}(x_t - x_\star)\rangle + \|x_t - x_\star\|^2.
\end{aligned}
$$

Step 3 (apply strongly convexity and smoothness) We have

$$
\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \|\nabla f(x_t) - \nabla f(x_\star)\|_{\mathbf{L}^{-1}}^2,
$$

$$
\langle x_t - x_\star, \nabla f(x_t) - \nabla f(x_\star)\rangle \geq \mu\|x_t - x_\star\|^2.
$$

Additionally, for any positive semidefinite matrices c_1, c_2, c_3, we have

$$
\|c_1(h_t - h_\star) - c_2(x_t - x_\star) - c_3(\nabla f(x_t) - \nabla f(x_\star))\|^2 \geq 0.
$$

We also have

$$
\|x_t - x_\star\|^2 + \tau^2\|h_t - h_\star\|^2 - 2\Delta\tau\langle x_t - x_\star, h_t - h_\star\rangle^2 = \Psi_t^{\text{new}}.
$$

Apply the above three inequalities and one equality with additional multipliers α, β, 1 and $-(1-\zeta)$, we have

$$
\begin{aligned}
&\mathbb{E}[\Psi_{t+1}^{\text{new}}] \\
\leq& \tau^2\|\nabla f(x_t) - \nabla f(x_\star)\|^2 - 2(\Delta(\tau-\gamma)+\gamma)\langle \nabla f(x_t) - \nabla f(x_\star), x_t - x_\star\rangle \\
&- 2(\tau-\gamma)((1-\Delta)\gamma+\tau)\langle \nabla f(x_t) - \nabla f(x_\star), \mathbb{E}[s]\rangle \\
&+ \|x_t - x_\star\|^2 + 2\Delta(\tau-\gamma)\langle x_t - x_\star, \mathbb{E}[s]\rangle + (\tau-\gamma)((1-2\Delta)\gamma+\tau)\mathbb{E}\|s\|^2 \\
\leq& \langle \nabla f(x_t) - \nabla f(x_\star), (c_3^2 + \tau^2\mathbf{I} - (\tau^2-\gamma^2)(\mathbf{I}+\mathbf{\Omega})^{-1} - \mathbf{L}^{-1}\alpha)(\nabla f(x_t) - \nabla f(x_\star))\rangle \\
&+ \langle \nabla f(x_t) - \nabla f(x_\star), (-c_3c_1 + 2\Delta\gamma(\tau-\gamma)(\mathbf{I}+\mathbf{\Omega})^{-1})(h_t - h_\star)\rangle \\
&+ \langle \nabla f(x_t) - \nabla f(x_\star), ((\alpha+\beta-2\gamma)\mathbf{I} + 2c_3c_2 - 2\Delta(\tau-\gamma)(\mathbf{I} - (\mathbf{I}-\mathbf{\Omega})^{-1}))(x_t - x_\star)\rangle \\
&+ \langle h_t - h_\star, (-\tau^2(1-\zeta)\mathbf{I} + c_1^2 + (\gamma-\tau)(2\Delta\gamma-\gamma-\tau)(\mathbf{I}+\mathbf{\Omega})^{-1})(h_t - h_\star)\rangle \\
&+ 2\langle h_t - h_\star, (\Delta\tau(1-\zeta)\mathbf{I} - c_1c_2 + \Delta(\gamma-\tau)(\mathbf{I}+\mathbf{\Omega})^{-1})(x_t - x_\star)\rangle \\
&+ \langle x_t - x_\star, ((1-\beta\mu+\zeta)\mathbf{I} + c_2^2)(x_t - x_\star)\rangle \\
&+ (1-\zeta)\Psi_t^{\text{new}}.
\end{aligned}
$$

We require

$$
c_3^2 + \tau^2\mathbf{I} - (\tau^2-\gamma^2)(\mathbf{I}+\mathbf{\Omega})^{-1} - \mathbf{L}^{-1}\alpha \preceq 0, \tag{35}
$$

$$
-\tau^2(1-\zeta)\mathbf{I} + c_1^2 + (\gamma-\tau)(2\Delta\gamma-\gamma-\tau)(\mathbf{I}+\mathbf{\Omega})^{-1} \preceq 0, \tag{36}
$$

$$
(1-\beta\mu+\zeta)\mathbf{I} + c_2^2 \preceq 0, \tag{37}
$$

$$
-c_3c_1 + 2\Delta\gamma(\tau-\gamma)(\mathbf{I}+\mathbf{\Omega})^{-1} = 0, \tag{38}
$$

$$
(\alpha+\beta-2\gamma)\mathbf{I} + 2c_3c_2 - 2\Delta(\tau-\gamma)(\mathbf{I} - (\mathbf{I}-\mathbf{\Omega})^{-1}) = 0, \tag{39}
$$

$$
\Delta\tau(1-\zeta)\mathbf{I} - c_1c_2 + \Delta(\gamma-\tau)(\mathbf{I}+\mathbf{\Omega})^{-1} = 0. \tag{40}
$$

Then we have

$$\mathbb{E}[\Psi_{t+1}^{\text{new}}] \leq (1-\zeta)\Psi_t^{\text{new}}.$$

Inspired by an analysis on the special case where $\mathbf{L}$ and $\mathbf{\Omega}$ are isotropic, we manually choose following parameters:

$$\begin{aligned} c_1 &= \Delta\gamma(\tau-\gamma)(\mathbf{I}-\mathbf{\Omega})^{-1}, \\ c_2 &= \Delta(\tau(1-\zeta)-(\tau-\gamma)(\mathbf{I}-\mathbf{\Omega})^{-1})c_1^{-1}, \\ c_3 &= c_1. \end{aligned}$$

With above parameter, Equations (38) to (40) holds when $\alpha+\beta = 2\gamma - 2\Delta(\gamma-\zeta\tau)$. And Equations (35) to (37) becomes

$$\gamma^2\tilde{\mathbf{\Omega}} - \mathbf{L}^{-1}\alpha \preceq 0, \tag{41}$$

$$\zeta\mathbf{I} - \frac{\gamma^2}{\tau^2}\tilde{\mathbf{\Omega}} \preceq 0, \tag{42}$$

$$E_1 \coloneqq (-\beta\mu+\zeta)\mathbf{I} + \frac{\Delta\left(\tau(1-\zeta)-(\mathbf{I}-\mathbf{\Omega})^{-1}(\tau-\gamma)\right)^2}{(\mathbf{I}-\mathbf{\Omega})^{-1}\gamma\left(\tau-\gamma\right)} \preceq 0. \tag{43}$$

We select

$$\begin{aligned} \alpha &= \gamma^2\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}), \\ \zeta &= \frac{\lambda_{\min}(\tilde{\mathbf{\Omega}})}{(1+\omega)^2}, \\ \beta &= 2\gamma\left(\Delta\left(\zeta\left(\omega+1\right)-1\right)+1\right)-\alpha. \end{aligned}$$

which satisfy Equations (41) and (42).
Therefore we only need to make sure Equation (43) holds. It is guaranteed some feasible solution with $\Delta \in (0,1)$ exist, because at the left end point, we have

$$\lambda_{\max}(E_1)|_{\Delta=0} = \left(-\gamma\mu(2-\gamma\lambda_{\max}(\mathbf{L}\tilde{\mathbf{\Omega}}))+1-\frac{1-p^2}{1+\lambda_{\min}(\mathbf{\Omega})}\right),$$

and due to $\gamma > \gamma_{\text{crit}}$, we have $E_1|_{\Delta=0} \prec 0$. □