Sharper Rates and Flexible Framework for Nonconvex SGD with Client and Data Sampling

Alexander Tyurin *alexandertiurin@gmail.com*
KAUST, Saudi Arabia

Lukang Sun *lukang.sun@kaust.edu.sa*
KAUST, Saudi Arabia

Konstantin Burlachenko *burlachenkok@gmail.com*
KAUST, Saudi Arabia

Peter Richtárik *richtarik@gmail.com*
KAUST, Saudi Arabia

Reviewed on OpenReview: *https://openreview.net/forum?id=zKgJ6TWAFE*

Abstract

We revisit the classical problem of finding an approximately stationary point of the average of n smooth and possibly nonconvex functions. The optimal complexity of stochastic first-order methods in terms of the number of gradient evaluations of individual functions is $\mathcal{O}\left(n + n^{1/2}\varepsilon^{-1}\right)$, attained by the optimal SGD methods SPIDER (Fang et al., 2018) and PAGE (Li et al., 2021), for example, where ε is the error tolerance. However, i) the big-$\mathcal{O}$ notation hides crucial dependencies on the smoothness constants associated with the functions, and ii) the rates and theory in these methods assume simplistic sampling mechanisms that do not offer any flexibility. In this work we remedy the situation. First, we generalize the PAGE algorithm so that it can provably work with virtually any (unbiased) sampling mechanism. This is particularly useful in federated learning, as it allows us to construct and better understand the impact of various combinations of client and data sampling strategies. Second, our analysis is sharper as we make explicit use of certain novel inequalities that capture the intricate interplay between the smoothness constants and the sampling procedure. Indeed, our analysis is better even for the simple sampling procedure analyzed in the PAGE paper. However, this already improved bound can be further sharpened by a different sampling scheme which we propose. In summary, we provide the most general and most accurate analysis of optimal SGD in the smooth nonconvex regime. Finally, our theoretical findings are supported by carefully designed experiments.

1 Introduction

In this paper, we consider the minimization of the average of n smooth functions (1) in the nonconvex setting in the regime when the number of functions n is very large. In this regime, calculation of the exact gradient can be infeasible and the classical gradient descent method (GD) (Nesterov, 2018) can not be applied. The structure of the problem is generic, and such problems arise in many applications, including machine learning (Bishop & Nasrabadi, 2006) and computer vision (Goodfellow et al., 2016). Problems of this form are the basis of empirical risk minimization (ERM), which is the prevalent paradigm for training supervised machine learning models.

1.1 Finite-sum optimization in the smooth nonconvex regime

We consider the finite-sum optimization problem

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

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is a smooth (and possibly nonconvex) function for all $i \in [n] := \{1,\dots,n\}$. We are interested in randomized algorithms that find an ε-stationary point of (1) by returning a random point $\widehat{x}$ such that $\mathrm{E}\left[\left\|\nabla f(\widehat{x})\right\|^2\right] \le \varepsilon$. The main efficiency metric of gradient-based algorithms for finding such a point is the (expected) number of gradient evaluations ∇f_i; we will refer to it as the *complexity* of an algorithm.

1.2 Related work

The area of algorithmic research devoted to designing methods for solving the ERM problem (1) in the smooth nonconvex regime is one of the most highly developed and most competitive in optimization.

The path to optimality. Let us provide a lightning-speed overview of recent progress. The complexity of GD for solving (1) is $\mathcal{O}\left(n\varepsilon^{-1}\right)$, but this was subsequently improved by more elaborate stochastic methods, including SAGA, SVRG and SCSG (Defazio et al., 2014; Johnson & Zhang, 2013; Lei et al., 2017; Horváth & Richtárik, 2019), which enjoy the better complexity $\mathcal{O}\left(n + n^{2/3}\varepsilon^{-1}\right)$. Further progress was obtained by methods such as SNVRG and Geom-SARAH (Zhou et al., 2018; Horváth et al., 2020), improving the complexity to $\widetilde{\mathcal{O}}\left(n + n^{1/2}\varepsilon^{-1}\right)$. Finally, the methods SPIDER, SpiderBoost, SARAH and PAGE (Fang et al., 2018; Wang et al., 2019; Nguyen et al., 2017; Li et al., 2021), among others, shaved-off certain logarithmic factors and obtained the *optimal* complexity $\mathcal{O}\left(n + n^{1/2}\varepsilon^{-1}\right)$, matching lower bounds (Li et al., 2021).

Optimal, but hiding a secret. While it may look that this is the end of the road, the starting point of our work is the observation that *the big-$\mathcal{O}$ notation in the above results hides important and typically very large data-dependent constants.* For instance, it is rarely noted that the more precise complexity of GD is $\mathcal{O}\left(L_{-}n\varepsilon^{-1}\right)$, while the complexity of the optimal methods, for instance PAGE, is $\mathcal{O}\left(n + L_{+}n^{1/2}\varepsilon^{-1}\right)$, where $L_{-} \le L_{+}$ are *different* and often *very large* smoothness constants. Moreover, it is easy to generate examples of problems (see Example 1) in which the ratio L_{+}/L_{-} is *as large as one desires.*

Client and data sampling in federated learning. Furthermore, several modern applications, notably federated learning (Konečný et al., 2016; McMahan et al., 2017), depend on elaborate *client* and *data sampling* mechanisms, which are not properly understood. However, optimal SGD methods were considered in combination with very simple mechanisms only, such as sampling a random function f_i several times independently with replacement (Li et al., 2021). We thus believe that an in-depth study of sampling mechanisms for optimal methods will be of interest to the federated learning community. There exists prior work on analyzing non-optimal SGD variants with flexible mechanisms For example, using the "arbitrary sampling" paradigm, originally proposed by Richtárik & Takáč (2016) in the study of randomized coordinate descent methods, Horváth & Richtárik (2019) and Qian et al. (2021) analyzed SVRG, SAGA, and SARAH methods, and showed that it is possible to improve the dependence of these methods on the smoothness constants via carefully crafted sampling strategies. Further, Zhao & Zhang (2014) investigated the stratified sampling, but only provided the analysis for vanilla SGD, and in the convex case.

1.3 Summary of contributions

- Specifically, in the original paper (Li et al., 2021), the optimal (w.r.t. n and ε) optimization method PAGE was analyzed with a simple uniform mini-batch sampling with replacement. We analyze PAGE with virtually any (unbiased) sampling mechanism using a novel Definition 1. Moreover, we show that some samplings can improve the convergence rate $\mathcal{O}\left(n + L_{+}n^{1/2}\varepsilon^{-1}\right)$ of PAGE. *We obtain the new state-of-the-art convergence rates in the considered setting* (see Table 2).
- We improve the analysis of PAGE using a new quantity, the weighted Hessian Variance $L_{\pm}$ (or $L_{\pm,w}$), that is well-defined if the functions f_i are L_i–smooth. It is important that we get strictly better theoretical guarantees without any additional assumptions on the structure of (1). We show that, when the functions

Table 1: The constants A, B, w_i and $|\mathbf{S}|$ that characterize the samplings in Definition 1.

Sampling scheme	A	w_i	B	$\lvert\mathbf{S}\rvert$	Reference
Uniform With Replacement	$1/\tau$	$1/n$	$1/\tau$	$\leq \tau$	Sec. E.3
Importance	$1/\tau$	q_i	$1/\tau$	$\leq \tau$	Sec. E.3
Nice (= *Uniform Without Replacement*)	$\frac{n-\tau}{\tau(n-1)}$	$1/n$	$\frac{n-\tau}{\tau(n-1)}$	τ	Sec. E.1
Independent	$\frac{1}{\sum_{i=1}^n \frac{p_i}{1-p_i}}$	$\frac{\frac{p_i}{1-p_i}}{\sum_{i=1}^n \frac{p_i}{1-p_i}}$	0	$\sum_{i=1}^n p_i$	Sec. E.2
Extended Nice	$\frac{n-\tau}{\tau(n-1)}$	$\frac{l_i}{\sum_{i=1}^n l_i}$	$\frac{n-\tau}{\tau(n-1)}$	$\leq \tau$	Sec. E.4

We now provide a brief explanation of the samplings (more mathematically rigorous explanations are provided in the corresponding sections from "Reference"): 1) *Uniform With Replacement* is a sampling, where we take exactly τ elements from a set with replacement. 2) *Importance* is a sampling, where we take exactly τ elements from a set with replacement, and a new sampled element is taken non-uniformly. 3) *Nice* is a sampling, where we take exactly τ elements from a set without replacement. 4) *Independent* is a sampling, where we take each element independently with a predefined probability. 5) *Extended Nice* is a sampling, where we repeat each i^{th} element l_i times before applying the *Nice* sampling.
Notation: n = # of data points; τ = batch size; q_i = probability to sample i^{th} data point in the multinomial distribution; p_i = probability to sample i^{th} data point in the bernoulli distribution; l_i = # of times to repeat i^{th} data point before apply the *Nice* sampling.

f_i are "similar" in the sense of the weighted Hessian Variance, PAGE enjoys faster convergence rates (see Table 2). Also, unlike (Szlendak et al., 2022), we introduce weights w_i that can play a crucial role in some samplings. Moreover, the experiments in Sec 5 agree with our theoretical results.

• Our framework is flexible and can be generalized to *the composition of samplings*. These samplings naturally emerge in federated learning (Konečný et al., 2016; McMahan et al., 2017), and we show that our framework can be helpful in the analysis of problems from federated learning.

2 Assumptions

We need the following standard assumptions from nonconvex optimization.

Assumption 1. *There exists $f^* \in \mathbb{R}$ such that $f(x) \geq f^*$ for all $x \in \mathbb{R}^d$.*

Assumption 2. *There exists $L_- \geq 0$ such that $\|\nabla f(x) - \nabla f(y)\| \leq L_- \|x - y\|$ for all $x, y \in \mathbb{R}^d$.*

Assumption 3. *For all $i \in [n]$, there exists a constant $L_i > 0$ such that $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$ for all $x, y \in \mathbb{R}^d$.*

Note that Assumption 2 follows from Assumption 3 with $L_- \leq \frac{1}{n}\sum_{i=1}^n L_i$.

2.1 Tight variance control of general sampling estimators

In Algorithm 1 (a generalization of PAGE), we form an estimator of the gradient ∇f via subsampling. In our search for achieving the combined goal of providing a *general* (in terms of the range of sampling techniques we cater for) and *refined* (in terms of the sharpness of our results, even when compared to known results using the *same* sampling technique) analysis of PAGE, we have identified several powerful tools, the first of which is Definition 1. Let

$$\mathcal{S}^n := \left\{ (w_1, ..., w_n) \in \mathbb{R}^n \,\middle|\, w_1, ..., w_n \geq 0, \sum_{i=1}^n w_i = 1 \right\}$$

be the *standard simplex* and $(\Omega, \mathcal{F}, \mathbf{P})$ a probability space.

Definition 1 ($\mathbb{S}(A, B, \{w_i\}_{i=1}^n)$ and Weighted AB Inequality)**.** Consider any random mapping $\mathbf{S} : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, which we will call "sampling", such that $\mathrm{E}\left[\mathbf{S}(a_1, \ldots, a_n; \omega)\right] = \frac{1}{n}\sum_{i=1}^n a_i$ for all

$a_1, \dots, a_n \in \mathbb{R}^d$, and there exist $A, B \geq 0$ and weights $(w_1, \dots, w_n) \in \mathcal{S}^n$ such that

$$\mathrm{E}\left[\left\|\mathbf{S}(a_1, \dots, a_n; \omega) - \frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right] \leq \frac{A}{n}\sum_{i=1}^n \frac{1}{nw_i}\left\|a_i\right\|^2 - B\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2, \tag{2}$$

for all $a_1, \dots, a_n \in \mathbb{R}^d$. The collection of such samplings will denote as $\mathbb{S}(A, B, \{w_i\}_{i=1}^n)$.

For simplicity, we denote $\mathbf{S}\left(\{a_i\}_{i=1}^n\right) := \mathbf{S}(a_1, \dots, a_n) := \mathbf{S}(a_1, \dots, a_n; \omega)$. The main purpose of a sampling $\mathbf{S} \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$ is to estimate the mean $\frac{1}{n}\sum_{i=1}^n a_i$ using some random subsets (possibly containing some elements more than once) of the set $\{a_1, \dots, a_n\}$. We refer to Table 1, where we provide examples of samplings that satisfy this definition. One of the simplest examples is the sampling

$$\mathbf{S}(a_1, \dots, a_n; \omega) = a_\chi, \tag{3}$$

where $\chi \in [n]$ is uniformly sampled from the set $[n]$ (the *Uniform With Replacement* sampling with $\tau = 1$ from Table 1).

We now define the cardinality $|\mathbf{S}|$ of a sampling $\mathbf{S} \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$.

Definition 2 (Cardinality of a Sampling)**.** Let us take $\mathbf{S} \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$, and define the function $\mathbf{S}_\omega(a_1, \dots, a_n) : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $\mathbf{S}_\omega(a_1, \dots, a_n) := \mathbf{S}(a_1, \dots, a_n; \omega)$. If the function $\mathbf{S}_\omega(a_1, \dots, a_n)$ depends only on a subset $\mathcal{A}(\omega)$ of the arguments $(a_1, \dots, a_n)$, where $\mathcal{A}(\omega) : \Omega \rightarrow 2^{\{a_1, \dots, a_n\}}$, we define $|\mathbf{S}| := \mathrm{E}\left[|\mathcal{A}(\omega)|\right]$.

In terms of our problem (1), the cardinality of a sampling represents the number of the gradients ∇f_i that a sampling calculates in each draw. Note that it can be easily calculated (see Table 1 for details). In the case of (3), one can see that $|\mathbf{S}| = 1$.

Definition 1 is most closely related to two independent works: (Horváth & Richtárik, 2019) and (Szlendak et al., 2022). Horváth & Richtárik (2019) analyzed several non-optimal SGD methods for "arbitrary samplings"; these are random set-valued mappings S with values being the subsets of $[n]$. The distribution of a such a sampling is uniquely determined by assigning probabilities to all 2^n subsets of $[n]$. In particular, they consider the samplings $\mathbf{S}(a_1, \dots, a_n) = \frac{1}{n}\sum_{i \in S}\frac{a_i}{p_i}$, $p_i := \mathbf{Prob}(i \in S)$, $|\mathbf{S}| = |S|$, some $A \geq 0$, $w_1, \dots, w_n \geq 0$ and $B = 0$. Our new set $\mathbb{S}(A, B, \{w_i\}_{i=1}^n)$ includes these samplings. Recently, Szlendak et al. (2022) studied a similar inequality, but in the context of communication-efficient distributed training with randomized gradient compression operators. They explicitly set out to study *correlated* compressors, and for this reason introduced the second term in the right hand side; i.e., they considered the possibility of B being nonzero, as in this way they obtain a tighter inequality, which they can use in their analysis. However, their inequality only involves uniform weights $\{w_i\}$. See Table 1 for an overview of several samplings and the values A, B and $\{w_i\}$ that belong to $\mathbb{S}(A, B, \{w_i\}_{i=1}^n)$ from Definition 1. *Our Definition 1 offers the tightest known way to control of the variance of the sampling estimator, and our analysis can take advantage of it.*

2.2 Sampling-dependent smoothness constants

We now define two smoothness constants that depend on the weights $\{w_i\}_{i=1}^n$ of a sampling $\mathbf{S}$ and on the functions f_i.

Definition 3. Given a sampling $\mathbf{S} \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$, let $L_{+,w}$ be a constant for which

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i}\left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 \leq L_{+,w}^2\left\|x - y\right\|^2, \tag{4}$$

for all $x, y \in \mathbb{R}^d$. For $(w_1, \dots, w_n) = (1/n, \dots, 1/n)$, we define $L_+ := L_{+,w}$.

Definition 4. Given a sampling $\mathbf{S} \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$, let $L_{\pm,w}$ be a constant for which

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i}\left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 - \left\|\nabla f(x) - \nabla f(y)\right\|^2 \leq L_{\pm,w}^2\left\|x - y\right\|^2, \qquad \forall x, y \in \mathbb{R}^d. \tag{5}$$

For $(w_1, \dots, w_n) = (1/n, \dots, 1/n)$, we define $L_\pm := L_{\pm,w}$.

Table 2: The complexity of methods and samplings from Table 1 and Sec 4.

Sampling scheme	Complexity	Comment
Independent (Horváth & Richtárik, 2019)	$\Theta\left(n + \frac{n^{2/3}\left(\frac{1}{n}\sum_{i=1}^{n} L_i\right)}{\varepsilon}\right)$	SVRG method $p_i \propto L_i$
Uniform With Replacement (Li et al., 2021)	$\Theta\left(n + \frac{\sqrt{n}L_+}{\varepsilon}\right)$	—
Uniform With Replacement (new)	$\Theta\left(n + \frac{\max\{\sqrt{n}L_\pm, L_-\}}{\varepsilon}\right)$	—
Importance	$\Theta\left(n + \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} L_i\right)}{\varepsilon}\right)$	$q_i = \frac{L_i}{\sum_{i=1} L_i}$
Stratified	$\Theta\left(n + \frac{\max\left\{\sqrt{n}\sqrt{\frac{1}{g}\sum_{i=1}^{g} L_{i,\pm}^2}, gL_-\right\}}{\varepsilon}\right)$	The functions f_i are splitted into g groups

Notation: n = # of data points; ε = error tolerance; $L_-, L_i, L_\pm, L_+$ and $L_{i,\pm}$ are smoothness constants such that $L_- \leq \frac{1}{n}\sum_{i=1}^{n} L_i$, $L_- \leq L_+$ and $L_\pm \leq L_+$; g = # of groups in the *Stratified* sampling.

One can interpret Definition 3 as *weighted* mean-squared smoothness property (Arjevani et al., 2019), and Definition 4 as *weighted* Hessian variance (Szlendak et al., 2022) that captures the similarity between the functions f_i. The constants $L_{+,w}$ and $L_{\pm,w}$ help us better understand the structure of the optimization problem (1) *in connection* with a particular choice of a sampling scheme. Note that Definitions 3, 4 and Definition 1 are connected with the weights $\{w_i\}_{i=1}^n$.

The next result states that $L_{+,w}^2$ and $L_{\pm,w}^2$ are well-defined and finite provided the functions f_i are L_i–smooth for all $i \in [n]$.

Theorem 5. *If Assumption 3 holds, then* $L_{+,w}^2 = L_{\pm,w}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{nw_i} L_i^2$ *satisfy Def. 3 and 4.*

Indeed, from Assumption 3 and the inequality $\|\nabla f(x) - \nabla f(y)\|^2 \geq 0$ we get

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{nw_i}\|\nabla f_i(x) - \nabla f_i(y)\|^2 - \|\nabla f(x) - \nabla f(y)\|^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{nw_i} L_i^2\right)\|x - y\|^2,$$

thus we can take $L_{\pm,w}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{nw_i} L_i^2$. The proof for $L_{+,w}^2$ is the same.

From the proof, one can see that we ignore $\|\nabla f(x) - \nabla f(y)\|^2$ when estimating $L_{\pm,w}^2$. However, by doing that, the obtained result is not tight.

2.3 Clarifications and Discussion

In the literature, including (Li et al., 2021), it is usually considered that one of the following inequalities holds. For all $i \in [n]$,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|, \qquad \forall x, y \in \mathbb{R}^d,$$

for some $L_i < \infty$ (see Assumption 3). The inequality

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_+^2 \|x - y\|^2, \qquad \forall x, y \in \mathbb{R}^d,$$

holds for some $L_+ < \infty$. These inequalities are standard smoothness assumptions (the latter follows from the former for some $L_+^2 \leq \frac{1}{n}\sum_{i=1}^{n} L_i^2$). Under the last inequality, Li et al. (2021) get the theoretical bound $\mathcal{O}\left(n + L_+ n^{1/2}\varepsilon^{-1}\right)$ that depend on L_+.

i) We want to emphasize that we also only consider Assumption 3. However, further we show that our theoretical bounds depend on $L_{\pm,w}$ (and $L_{+,w}$) instead of L_+, and we provide the examples when $L_{\pm,w}$ (and $L_{+,w}$) are much smaller than L_+ (see Table 2). We get this improvement without additional assumptions on the problem (1).

ii) The weights w_i depend only on a choice of a sampling, and play an essential role in the *Importance* sampling, for instance. In Sec 3.3, we show that if we take the *Importance* sampling with $q_i = L_i / \sum_{i=1}^n L_i$ for all $i \in [n]$, then $w_i = L_i / \sum_{i=1}^n L_i$ for all $i \in [n]$, and it leads to the fact that $L_{\pm,w} \leq L_{+,w} \leq \frac{1}{n}\sum_{i=1}^n L_i$. One can see that a good choice of a sampling can lead to good weights w_i such that the inequalities in Definitions 3 and 4 are satisfied with small $L_{\pm,w}$ and $L_{+,w}$. It holds without additional assumptions about our problem.

3 A General and Refined Theoretical Analysis of PAGE

In the Algorithm 1, we provide the description of the PAGE method. The description of the method can be found in (Li et al., 2021, Sec. 2.1) The choice of PAGE as the base method is driven by the simplicity of the proof in the original paper. However, we believe that other methods, including SPIDER and SARAH, can also admit samplings from Definition 1.

Algorithm 1 PAGE

```
Input: initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, probability $p \in (0, 1]$
$g^0 = \nabla f(x^0)$
for $t = 0, 1, \ldots, T$ do
    $x^{t+1} = x^t - \gamma g^t$
    Generate a random sampling function $\mathbf{S}^t$
    $g^{t+1} = \begin{cases} \nabla f(x^{t+1}) & \text{with probability } p \\ g^t + \mathbf{S}^t\left(\{\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\}_{i=1}^n\right) & \text{with probability } 1-p \end{cases}$
end for
```

In this section, we provide theoretical results for Algorithm 1. Let us define $\Delta_0 := f(x^0) - f^*$.

Theorem 6. *Suppose that Assumptions 1, 2, 3 hold and the samplings* $\mathbf{S}^t \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$. *Then Algorithm 1 (PAGE) has the convergence rate* $\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \leq \frac{2\Delta_0}{\gamma T}$, *where* $\gamma \leq \left(L_- + \sqrt{\frac{1-p}{p}\left((A-B)\,L_{+,w}^2 + BL_{\pm,w}^2\right)}\right)^{-1}$.

To reach an ε-stationary point, it is enough to do

$$T := \frac{2\Delta_0}{\varepsilon}\left(L_- + \sqrt{\frac{1-p}{p}\left((A-B)\,L_{+,w}^2 + BL_{\pm,w}^2\right)}\right) \tag{6}$$

iterations of Algorithm 1. To deduce the gradient complexity, we provide the following corollary.

Corollary 1. *Suppose that the assumptions of Thm 6 hold. Let us take* $p = \frac{|\mathbf{S}|}{|\mathbf{S}|+n}$. *Then the complexity (the expected number of gradient calculations* ∇f_i*) of Algorithm 1 equals*

$$N := \Theta\left(n + |\mathbf{S}|T\right) = \Theta\left(n + \frac{\Delta_0}{\varepsilon}|\mathbf{S}|\left(L_- + \sqrt{\frac{n}{|\mathbf{S}|}\left((A-B)\,L_{+,w}^2 + BL_{\pm,w}^2\right)}\right)\right).$$

Proof. At each iteration, the expected # gradient calculations equals $pn + (1-p)|\mathbf{S}| \leq 2|\mathbf{S}|$. Thus the total expected number of gradient calculations equals $n + 2|\mathbf{S}|T$ to get an ε-stationary point. □

The original result from (Li et al., 2021) states that the complexity of PAGE with batch size τ is

$$N_{\text{orig}} := \Theta\left(n + \frac{\Delta_0}{\varepsilon}\tau\left(L_- + \frac{\sqrt{n}}{\tau}L_+\right)\right) \geq \Theta\left(n + \frac{\Delta_0\sqrt{n}L_+}{\varepsilon}\right) \tag{7}$$

for all $\tau \in \{1, 2, \ldots, n\}$.

In the following examples, we will see that Cor 1 can guarantee better complexities than (7). These improvements we get using a refined and tighter analysis of the gradient estimator g^{t+1} in Sec D. In particular, we provide a better upper bound to the variance $\mathrm{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right]$, which makes our dependencies on smoothness constants tighter.

3.1 *Uniform With Replacement* sampling

Let us do a sanity check and substitute the parameters of the sampling that the original paper uses. We take the *Uniform With Replacement* sampling (see Sec E.3) with batch size τ (note that $\tau \geq |\mathbf{S}|$), $A = B = 1/\tau$ and $w_i = 1/n$ for all $i \in [n]$ (see Table 1) and get the complexity $N_{\text{uniform}} = \Theta\left(n + \frac{\Delta_0}{\varepsilon}\tau\left(L_- + \frac{\sqrt{n}}{\tau}L_\pm\right)\right)$ for all $\tau \in \{1, 2, \dots, n\}$. Next, let us fix $\tau \leq \max\left\{\frac{\sqrt{n}L_\pm}{L_-}, 1\right\}$, and, finally, obtain that $N_{\text{uniform}} = \Theta\left(n + \frac{\Delta_0 \max\{\sqrt{n}L_\pm, L_-\}}{\varepsilon}\right)$. Let us compare it with (7). With the **same sampling**, our analysis provides better complexity; indeed, note that $\max\{\sqrt{n}L_\pm, L_-\} \leq \sqrt{n}L_+$ (see Lemma 2 in Szlendak et al. (2022)). Moreover, Szlendak et al. (2022) provides examples of the optimization problems when $L_\pm$ is small and L_+ is large, so the difference can be arbitrary large.

Note that the original analysis of PAGE (Li et al., 2021) is also based on Assumptions 1, 2 and 3, and we do not have additional assumptions on the structure of the problem (1). We have better theoretical guarantees in the same setting.

3.2 *Nice* sampling (= *Uniform Without Replacement*) sampling

Next, we consider the *Nice* sampling (see Sec E.1) and get that the complexity $N_{\text{nice}} = \Theta\left(n + \frac{\Delta_0}{\varepsilon}\tau\left(L_- + \frac{1}{\tau}\sqrt{\frac{n(n-\tau)}{(n-1)}}L_\pm\right)\right)$. Unlike the *Uniform With Replacement* sampling, for ε small enough, the *Nice* sampling recovers the complexity of GD for $\tau = n$, which is equal to $\Theta\left(\frac{\Delta_0 n L_-}{\varepsilon}\right)$.

3.3 *Importance* sampling

Let us consider the *Importance* sampling (see Sec E.3) that justifies the introduction of the weights w_i. We can get the complexity

$$N_{\text{importance}} = \Theta\left(n + \frac{\Delta_0}{\varepsilon}\tau\left(L_- + \frac{\sqrt{n}}{\tau}L_{\pm,w}\right)\right) \leq \Theta\left(n + \frac{\Delta_0 \max\{\sqrt{n}L_{\pm,w}, L_-\}}{\varepsilon}\right)$$

for $\tau \leq \max\left\{\frac{\sqrt{n}L_{\pm,w}}{L_-}, 1\right\}$. Now, we take $q_i = w_i = \frac{L_i}{\sum_{i=1}^n L_i}$ and use the results from Sec F to obtain $N_{\text{importance}} = \Theta\left(n + \frac{\Delta_0\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n L_i\right)}{\varepsilon}\right)$ (See Sec G). In Example 2, we consider the optimization task where $\frac{1}{n}\sum_{i=1}^n L_i$ is $\sqrt{n}$ times smaller than L_+. Thus the complexity $N_{\text{importance}}$ can be at least $\sqrt{n}$ times smaller that the complexity N_{orig}. However, we do not guarantee that $N_{\text{importance}} \leq N_{\text{orig}}$ in general.

3.4 The power of $B > 0$

This is the first analysis of optimal SGD, which uses $B > 0$. In all previous examples, the constant $A = B > 0$. If $A = B$, then the complexity

$$N = \Theta\left(n + \frac{\Delta_0}{\varepsilon}|\mathbf{S}|\left(L_- + \sqrt{\frac{n}{|\mathbf{S}|}BL_{\pm,w}^2}\right)\right),$$

thus the complexity N does not depend on $L_{+,w}^2$, which greater of equal to $L_{\pm,w}^2$. It can even be possible that $L_{\pm,w}^2$ is close or equal to zero, while $L_{+,w}^2$ is arbitrary large. Unlike Definition 3 of $L_{+,w}^2$,, Definition 4 of $L_{\pm,w}^2$ captures *the variance of gradient differences*. Imagine that $(w_1, \dots, w_n) = (1/n, \dots, 1/n)$ and $f_i = f$ for all $i \in [n]$, then the the left-hand side of (5) equals zero, and we can take $L_{\pm,w}^2 = 0$. In general, we *can not* take $L_{+,w}^2 = 0$ in (4) if $f_i = f$ for all $i \in [n]$.

3.5 Analysis under PŁ Condition

The previous results can be extended to the optimization problems that satisfy the Polyak-Łojasiewicz condition. Under this assumption, Algorithm 1 enjoys a linear convergence rate.

Assumption 4. *There exists $\mu > 0$ such that the function f satisfy (Polyak-Łojasiewicz) PŁ-condition:*

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*) \quad \forall x \in \mathbb{R},$$

where $f^* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.

Using Assumption 4, we can improve the convergence rate of PAGE.

Theorem 7. *Suppose that Assumptions 1, 2, 3, 4 and the samplings $\mathbf{S}^t \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$. Then Algorithm 1 (PAGE) has the convergence rate $\mathrm{E}\left[f(x^T)\right] - f^* \leq (1 - \gamma\mu)^T \Delta_0$, where $\gamma \leq \min\left\{\left(L_- + \sqrt{\frac{2(1-p)}{p}\left((A-B)\, L_{+,w}^2 + BL_{\pm,w}^2\right)}\right)^{-1}, \frac{p}{2\mu}\right\}$.*

Algorithm 2 PAGE with composition of samplings

```
Input: initial point $x^0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, probability $p \in (0,1]$, $g^0 = \nabla f(x^0)$
for $t = 0, 1, \ldots, T$ do
    $x^{t+1} = x^t - \gamma g^t$
    $c^{t+1} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$
    if $c^{t+1} = 1$ then
        $g^{t+1} = \nabla f(x^{t+1})$
        /* FL Interpretation: With the small probability, calculate the full gradients $\nabla f_i$ on the nodes and collect them */
    else
        Generate samplings $\mathbf{S}_i^t$ for all $i \in [n]$
        $h_i^{t+1} = \mathbf{S}_i^t\left(\{\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\}_{j=1}^{m_i}\right)$ for all $i \in [n]$
        /* FL Interpretation: Calculate the mini-batches $h_i^{t+1}$ on the nodes */
        Generate a sampling $\mathbf{S}^t$ and set $g^{t+1} = g^t + \mathbf{S}^t\left(\{h_i^{t+1}\}_{i=1}^n\right)$
        /* FL Interpretation: Collect $h_i^{t+1}$ only from the sampled nodes */
    end if
end for
```

4 Composition of Samplings: Application to Federated Learning

In Sec 3, we analyze the PAGE method with samplings that belong to Definition 1. Now, let us assume that the functions f_i have the finite-sum form, i.e., $f_i(x) := \frac{1}{m_i}\sum_{j=1}^{m_i} f_{ij}(x)$, thus we an optimization problem

$$\min_{x \in \mathbb{R}^d}\left\{f(x) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} f_{ij}(x)\right\}, \tag{8}$$

Another way to get the problem is to assume that we split the functions f_i into groups of sizes m_i. Let us consider (8) instead of (1).

The problem (8) occurs in many applications, including distributed optimization and federated learning (Konečný et al., 2016; McMahan et al., 2017). In federated learning, many devices and machines (nodes) store local datasets that they do not share with other nodes. The local datasets are represented by functions f_i, and all nodes solve the common optimization problem (8). Due to privacy reasons and communication bottlenecks (Kairouz et al., 2021), it is infeasible to store and compute the functions f_i locally in one machine.

In general, when we solve (1) in one machine, we have the freedom of choosing a sampling $\mathbf{S}$ for the functions f_i, which we have shown in Sec 3. However, in federated learning, a sampling of nodes or the functions f_i is dictated by hardware limits or network quality (Kairouz et al., 2021). Still, each i^{th} node can choose sampling $\mathbf{S}_i$ to sample the functions f_{ij}. As a result, we have a composition of the sampling $\mathbf{S}$ and the samplings $\mathbf{S}_i$ (see Algorithm 2).

Assumption 5. *For all $j \in [m_i], i \in [n]$, there exists a Lipschitz constant L_{ij} such that $\left\|\nabla f_{ij}(x) - \nabla f_{ij}(y)\right\| \le L_{ij}\left\|x - y\right\|$ for all $x, y \in \mathbb{R}^d$.*

We now introduce the counterpart of Definitions 3 and 4.

Definition 8. For all $i \in [n]$ and any sampling $\mathbf{S}_i \in \mathbb{S}(A_i, B_i, \{w_{ij}\}_{j=1}^{m_i})$, define constant $L_{i,+,w_i}$ such that

$$\tfrac{1}{m_i}\sum_{j=1}^{m_i} \tfrac{1}{m_i w_{ij}}\left\|\nabla f_{ij}(x) - \nabla f_{ij}(y)\right\|^2 \le L^2_{i,+,w_i}\left\|x - y\right\|^2$$

for all $x, y \in \mathbb{R}^d$.

Definition 9. For all $i \in [n]$ and any sampling $\mathbf{S}_i \in \mathbb{S}(A_i, B_i, \{w_{ij}\}_{j=1}^{m_i})$, define constant $L_{i,\pm,w_i}$ such that

$$\frac{1}{m_i}\sum_{j=1}^{m_i} \frac{1}{m_i w_{ij}}\left\|\nabla f_{ij}(x) - \nabla f_{ij}(y)\right\|^2 - \left\|\nabla f_i(x) - \nabla f_i(y)\right\|^2 \le L^2_{i,\pm,w_i}\left\|x - y\right\|^2$$

for all $x, y \in \mathbb{R}^d$.

Let us provide the counterpart of Thm 6 for Algorithm 2.

Theorem 10. *Suppose that Assumptions 1, 2, 3, 5 hold and the samplings $\mathbf{S}^t \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$ and the samplings $\mathbf{S}_i^t \in \mathbb{S}(A_i, B_i, \{w_{ij}\}_{j=1}^{m_i})$ for all $i \in [n]$. Moreover, $B \le 1$. Then Algorithm 2 has the convergence rate $\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \le \frac{2\Delta_0}{\gamma T}$, where $\gamma \le \left(L_- + \sqrt{\frac{1-p}{p}\left(\mathcal{L}_l + \mathcal{L}_g\right)}\right)^{-1}$,*

$$\mathcal{L}_l := \frac{1}{n}\sum_{i=1}^n \left(\frac{A}{nw_i} + \frac{(1-B)}{n}\right)((A_i - B_i)L^2_{i,+,w_i} + B_i L^2_{i,\pm,w_i}),$$

and

$$\mathcal{L}_g := (A - B)L^2_{+,w} + BL^2_{\pm,w}.$$

The obtained theorem provides a general framework that helps analyze the convergence rates of the composition of samplings that belong to Definition 1. We discuss the obtained result in different contexts.

4.1 Federated Learning

For simplicity, let us assume that the samplings $\mathbf{S}^t$ and $\mathbf{S}_i^t$ are *Uniform With Replacement* samplings with batch sizes τ and τ_i for all $i \in n$, accordingly, then to get ε-stationary point, it is enough to do

$$T := \Theta\left(\tfrac{\Delta_0}{\varepsilon}\left(L_- + \sqrt{\tfrac{1-p}{p\tau}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{1}{\tau_i}L^2_{i,\pm} + L^2_{\pm}\right)}\right)\right)$$

iterations. Note that $T \ge \Theta\left(\frac{\Delta_0}{\varepsilon}\left(L_- + \sqrt{\frac{1-p}{p\tau}L^2_{\pm}}\right)\right)$ for all $\tau_i \ge 1$ for all $i \in [n]$. It means that after some point, there is no benefit in increasing batch sizes τ_i. In order to balance $\frac{1}{n}\sum_{i=1}^n \frac{1}{\tau_i}L^2_{i,\pm}$ and $L^2_{\pm}$, one can take $\tau_i = \Theta\left({L^2_{i,\pm}}/{L^2_{\pm}}\right)$. The constant $L^2_{i,\pm}$ captures the *intra-variance* inside i^{th} node, while $L^2_{\pm}$ captures the *inter-variance* between nodes. If the *intra-variance* is small with respect to the *inter-variance*, then our theory suggests taking small batch sizes and vice versa.

4.2 *Stratified* sampling

Let us provide another example that is closely related to (Zhao & Zhang, 2014). Let us consider (1) and use a variation of the *Stratified* sampling (Zhao & Zhang, 2014): we split the functions f_i into $g = {n}/{m}$ groups, where m is the number of functions in each group. Thus we get the problem (8) with $f(x) = \frac{1}{g}\sum_{i=1}^g \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$. Let us assume that we always sample *all* groups, thus $A = B = 0$, and the sampling $\mathbf{S}_i^t$ are *Nice* samplings with batch sizes τ_1 for all $i \in [n]$. Applying Thm 10, we get the convergence rate

$$T_{\text{group}} := \Theta\left(\tfrac{\Delta_0}{\varepsilon}\left(L_- + \sqrt{\tfrac{1-p}{pg\tau_1}\left(\tfrac{1}{g}\textstyle\sum_{i=1}^g L^2_{i,\pm}\right)}\right)\right).$$

At each iteration, the algorithm calculates $g\tau_1$ gradients, thus we should take $p = \frac{g\tau_1}{g\tau_1 + n}$ to get the complexity $N_{\text{group}} := \Theta\left(n + g\tau_1 T\right) = \Theta\left(n + \frac{\Delta_0}{\varepsilon}\left(g\tau_1 L_- + \sqrt{n}\sqrt{\frac{1}{g}\sum_{i=1}^{g} L_{i,\pm}^2}\right)\right)$. Let us take $\tau_1 \leq \max\left\{\frac{\sqrt{n}\sqrt{\frac{1}{g}\sum_{i=1}^{g} L_{i,\pm}^2}}{gL_-}, 1\right\}$ to obtain the complexity $N_{\text{group}} = \Theta\left(n + \frac{\Delta_0 \max\left\{\sqrt{n}\sqrt{\frac{1}{g}\sum_{i=1}^{g} L_{i,\pm}^2}, gL_-\right\}}{\varepsilon}\right)$. Comparing the complexity N_{group} with the complexity N_{uniform} from Sec 3, one can see that if split the functions f_i in a "right way", such that $L_{i,\pm}$ is small for $i \in [n]$ (see Example 3), then we can get at least $\sqrt{n}/\sqrt{g}$ times improvement with the *Stratified* sampling. However, we can be "unlucky" with a splitting of the functions; thus, we do not guarantee improvement in general.

5 Experiments

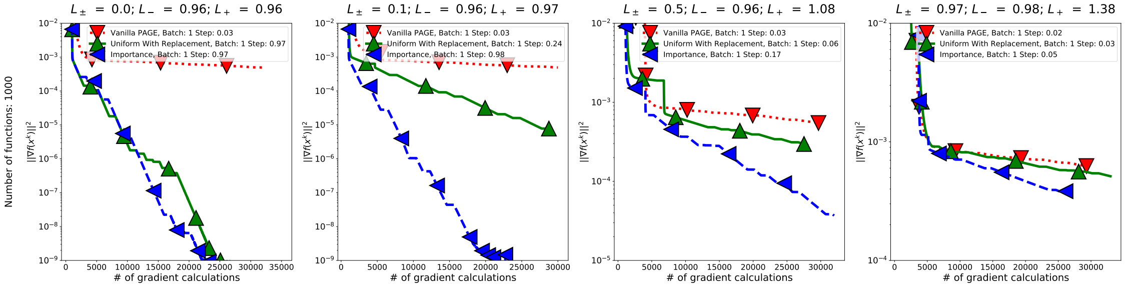


Figure 1: Comparison of samplings and methods on quadratic optimization tasks with various $L_\pm$.

We now provide experiments[1] with synthetic quadratic optimization tasks, where the functions f_i, in general, are nonconvex quadratic functions. Note that our goal here is to check whether the dependencies that our theory predicts are correct for the problem (1). The procedures that generate synthetic quadratic optimization tasks give us control over the choice of smoothness constants. All parameters, including the step sizes, are chosen as suggested by the corresponding theory. In the plots, we represent the relation between the norm of gradients and the number of gradient calculations ∇f_i.

5.1 Quadratic optimization tasks with various Hessian Variances $L_\pm$

Using Algorithm 3 (see Appendix), we generated various quadratic optimization tasks with different smoothness constants $L_\pm \in [0, 1.0]$ and fixed $L_- \approx 1.0$ (see Fig. 1). We choose $d = 10$, $n = 1000$, regularization $\lambda = 0.001$, and the noise scale $s \in \{0, 0.1, 0.5, 1\}$. According to Sec 3 and Table 2, the gradient complexity of original PAGE method ("Vanilla PAGE" in Fig. 1) is proportional to L_+. While the gradient complexity of the new analysis with the *Uniform With Replacement* sampling ("Uniform With Replacement" in Fig. 1) is proportional to $L_\pm$, which is always less or equal L_+. In Fig. 1, one can see that the smaller $L_\pm$ with respect to L_+, the better the performance of "Uniform With Replacement." Moreover, we provide experiments with the *Importance* sampling ("Importance" in Fig. 1) with $q_i = \frac{L_i}{\sum_{i=1}^n L_i}$ for all $i \in [n]$. This sampling has the best performance in all regimes.

5.2 Quadratic optimization tasks with various local Lipschitz constatns L_i

Using Algorithm 4 (see Appendix), we synthesized various quadratic optimization tasks with different smoothness constants L_i (see Fig. 2). We choose $d = 10$, $n = 1000$, the regularization $\lambda = 0.001$, and the noise scale $s \in \{0, 0.1, 0.5, 10.0\}$. We generated tasks in such way that the difference between $\max_i L_i$ and $\min_i L_i$ increases. First, one can see that the *Uniform With Replacement* sampling with the new analysis ("Uniform With Replacement" in Fig. 2) has better performance even in the cases of significant variations of L_i. Next, we see the stability of the *Importance* sampling ("Importance" in Fig. 2) with respect to this variations.

[1] Our code: https://github.com/mysteryresearcher/sampling-in-optimal-sgd

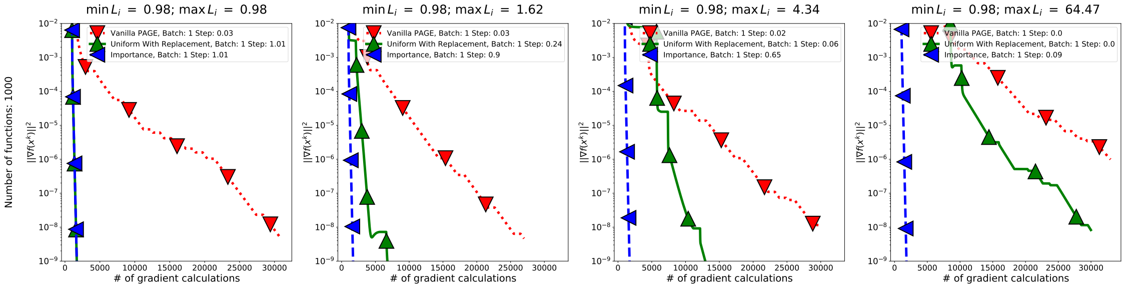
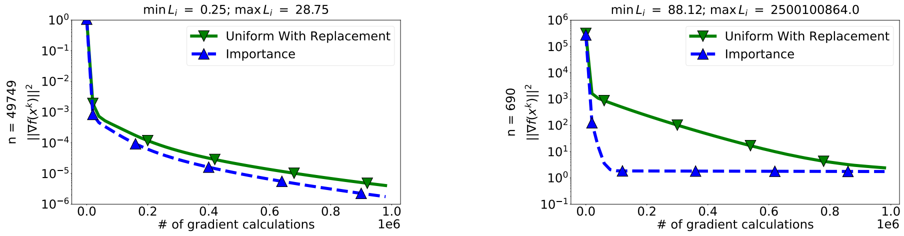


Figure 2: Comparison of samplings and methods on quadratic optimization tasks with various L_i.



w8a dataset *australian* dataset

Figure 3: Comparison of samplings on nonconvex machine learning tasks with LIBSVM datasets.

5.3 Nonconvex classification problem with LIBSVM datasets

We now solve nonconvex machine learning tasks and compare samplings on LIBSVM datasets (Chang & Lin, 2011) (see details in Sec A.2). As in previous sections, PAGE with the *Importance* sampling performs better in Fig. 3, especially in the *australian* dataset where the variation of L_i is large.

References

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Konstantin Burlachenko, Samuel Horváth, and Peter Richtárik. Fl_pytorch: optimization research simulator for federated learning. In *Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning*, pp. 1–7, 2021.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *NeurIPS Information Processing Systems*, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Samuel Horváth and Peter Richtárik. Nonconvex variance reduced optimization with arbitrary sampling. In *International Conference on Machine Learning*, pp. 2781–2789. PMLR, 2019.

Samuel Horváth, Lihua Lei, Peter Richtárik, and Michael I Jordan. Adaptivity of stochastic gradient methods for nonconvex optimization. *arXiv preprint arXiv:2002.05359*, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

Lam Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *The 34th International Conference on Machine Learning*, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Xun Qian, Zheng Qu, and Peter Richtárik. L-SVRG and L-Katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22:1–49, 2021.

Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1):433–484, 2016.

Rafał Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. *International Conference on Learning Representations (ICLR)*, 2022.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Contents

A Extra Experiments and Details

A.1 Quadratic optimization tasks with various batch sizes.

In this section, we consider the same setup as in Sec 5.1. In Figure 4, we fix $L_\pm$, and show that the *Importance* sampling has better convergence rates with different batch sizes. Note that with large batches, the competitors reduce to the GD method, and the difference is not significant.

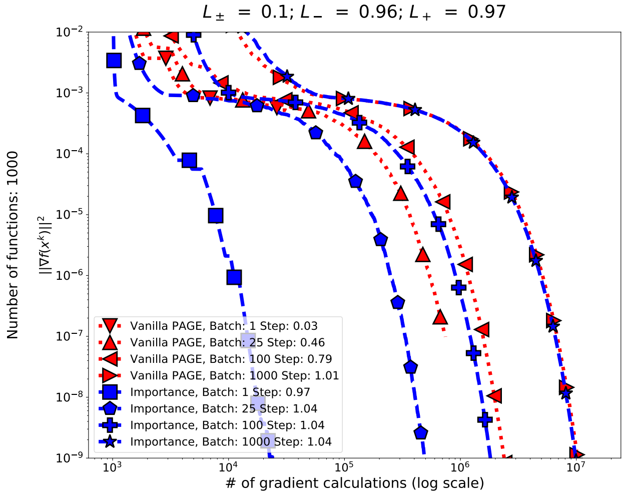


Figure 4: Comparison of samplings and methods with various batch sizes.

A.2 Details on experiments with LIBSVM datasets

We compare the samplings on practical machine learnings with LIBSVM datasets (Chang & Lin, 2011) (under the 3-clause BSD license). Parameters of Algorithm 1 are chosen as suggested in Thm 6 and Cor 1. We take the parameters for *Uniform With Replacement* and *Importance* samplings from Table 1 with $q_i = \frac{L_i}{\sum_{i=1}^n L_i}$. We consider the logistic regression task with a nonconvex regularization (Wang et al., 2019)

$$f(x_1, x_2) := \frac{1}{n}\sum_{i=1}^{n}\left[-\log\left(\frac{\exp\left(a_i^\top x_{y_i}\right)}{\sum_{y\in\{1,2\}}\exp\left(a_i^\top x_y\right)}\right) + \lambda\sum_{y\in\{1,2\}}\sum_{k=1}^{d}\frac{\{x_y\}_k^2}{1+\{x_y\}_k^2}\right] \to \min_{x_1,x_2\in\mathbb{R}^d},$$

where $\{\cdot\}_k$ is an indexing operation, $a_i \in \mathbb{R}^d$ is the feature of a i^{th} sample, $y_i \in \{1,2\}$ is the label of a i^{th} sample, constant $\lambda = 0.001$. We fix batch size $\tau = 1$ and take *w8a* dataset (dimension $d = 300$, number of samples $n = 49{,}749$) and *australian* dataset (dimension $d = 14$, number of samples $n = 690$) from LIBSVM.

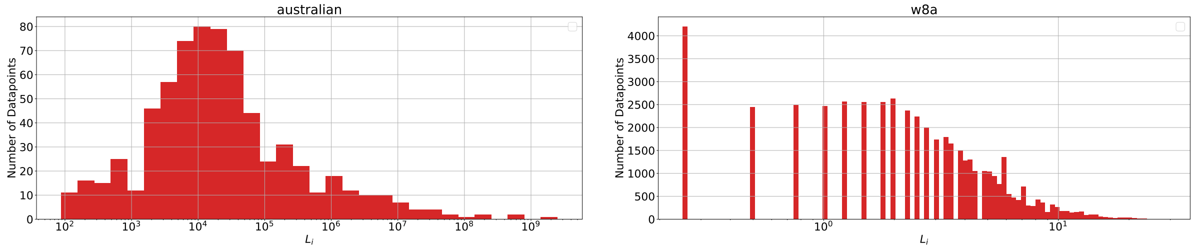
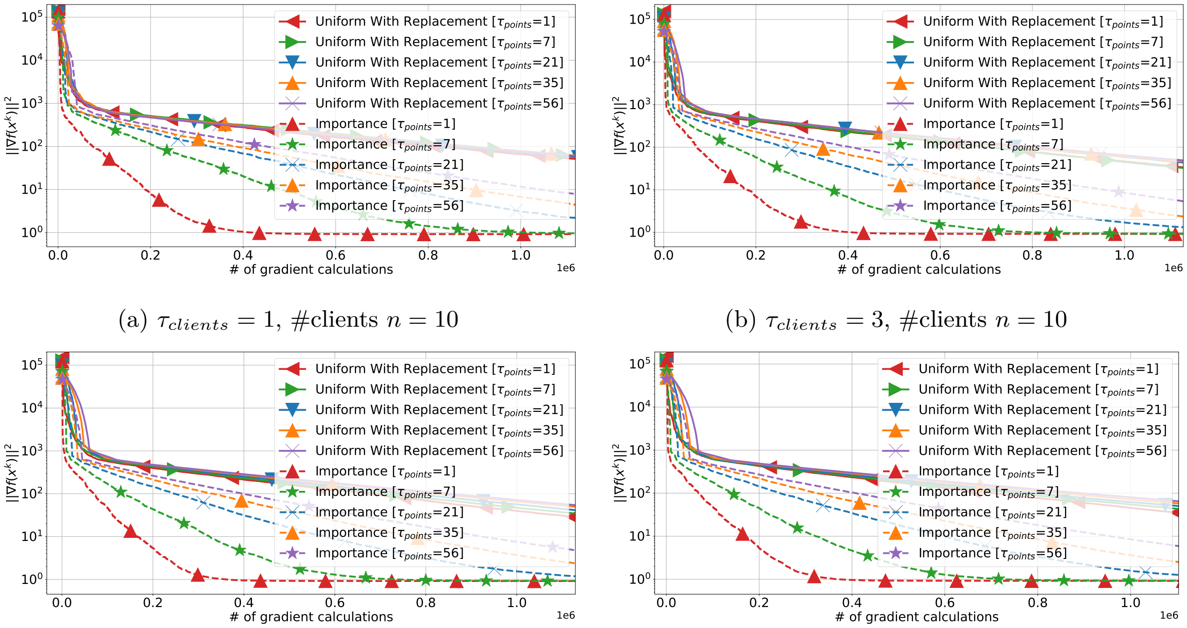


Figure 5: The distribution of Lipschitz constants L_i



(a) $\tau_{clients} = 1$, #clients $n = 10$

(b) $\tau_{clients} = 3$, #clients $n = 10$

(c) $\tau_{clients} = 6$, #clients $n = 10$

(d) $\tau_{clients} = 9$, #clients $n = 10$

Figure 6: Comparison of methods on *australian* dataset from LIBSVM

For the logistic regression, the Lipschitz constants L_i can be estimated. The distribution of Lipschitz constants L_i across datapoints for that two datasets is presented in Fig. 5. We use Thm 5 to obtain $L^2_{+,w}$ and $L^2_{\pm,w}$.

A.3 Federated learning experiments with LIBSVM dataset

In this experiment[2], we compare the *Uniform With Replacement* sampling and the *Importance* sampling on the logistic regression task from Sec. A.2 in a distributed environment. The training of the models is carried on *australian* dataset from LIBSVM. The dataset is reshuffled with uniform distribution, and then it is split across $n = 10$ clients. In all experiments, we use Algorithm 2 with theoretical stepsizes according to Theorem 10. We take the parameters of the *Uniform With Replacement* and *Importance* samplings from Table 1 with $q_i = \frac{L_i}{\sum_{i=1}^n L_i}$.

According to Algorithm 2, we have the samplings $\mathbf{S}^t$ that sample clients, and the samplings $\mathbf{S}^t_i$ that sample data from the local datasets of clients. Algorithm 2 allows mixed sampling strategies that satisfy Assumption 1. For simplicity, we consider that the samplings $\mathbf{S}^t$ and $\mathbf{S}^t_i$ are of the same type.

For the logistic regression, the Lipschitz constants L_i and L_{ij} of the gradients of functions $f_i(x)$ and $f_{ij}(x)$ can be estimated. As in Sec A.2, we use Thm 5 to obtain the constants $L^2_{i,+,w}$, $L^2_{i,\pm,w}$, $L^2_{+,w}$ and $L^2_{\pm,w}$. The results of experiments are provided in Fig. 6. We denote by τ_{points} the batch size of the samplings $\mathbf{S}^t_i$ for all $i \in [n]$, and by $\tau_{clients}$ the batch size of the sampling $\mathbf{S}^t$. The number of gradient calculations in Fig. 6 stands for the total number of gradient calculations in all clients.

We demonstrate results for different values of the batch sizes $\tau_{clients}$ and τ_{points}. As in previous experiments, the *Importance* sampling has better empirical performance than the *Uniform With Replacement* sampling. In addition to it, we observe that plots with small batch sizes τ_{points} converge faster.

[2] Our code: https://github.com/mysteryresearcher/page_ab_fl_experiment_a3

A.4 Computing environment

The code was written in Python 3.6.8 using PyTorch 1.9 (Paszke et al., 2019) and optimization research simulator FL_PyTorch (Burlachenko et al., 2021). The distributed environment was emulated on a machine with Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz and 64 cores.

B Auxiliary fact: Variance Decomposition

We use the following auxiliary fact in our proofs: Let us take a *random vector* $\xi \in \mathbb{R}^d$, then

$$\mathrm{E}\left[\|\xi\|^2\right] = \mathrm{E}\left[\|\xi - \mathrm{E}\left[\xi\right]\|^2\right] + \|\mathrm{E}\left[\xi\right]\|^2. \tag{9}$$

C Examples of Optimization Problems

Example 1. For simplicity, let us assume that n is even. Let us consider the optimization problem (1) with $f_i(x) = \frac{a}{2}x^2 + \frac{b}{2}x^2$ for $i \in \{1, \cdots, n/2\}$ and $f_i(x) = -\frac{a}{2}x^2 + \frac{b}{2}x^2$ for $i \in \{n/2+1, \cdots, n\}$, where $x \in \mathbb{R}$ and $b \geq 0$. Then $f(x) = \frac{b}{2}x^2$ and

$$L_-^2 = \sup_{x \neq y} \frac{\|\nabla f(x) - \nabla f(y)\|^2}{\|x - y\|^2} = b^2.$$

Moreover,

$$L_+^2 = \sup_{x \neq y} \frac{\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2}{\|x - y\|^2} = \frac{1}{2}\left((a+b)^2 + (a-b)^2\right),$$

and we can take a arbitrary large.

Example 2. Let us assume that $n \geq 2$ and consider the optimization problem (1) with $f_1(x) = \frac{b}{2}x^2$ and $f_i(x) = 0$ for $i \in \{2, \cdots, n\}$, where $x \in \mathbb{R}$ and $b \geq 0$. Then $f(x) = \frac{b}{2n}x^2$,

$$L_- = \sup_{x \neq y} \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} = \frac{b}{n},$$

$$\frac{1}{n}\sum_{i=1}^n L_i = \frac{1}{n}\sup_{x \neq y} \frac{\|\nabla f_1(x) - \nabla f_1(y)\|^2}{\|x - y\|^2} = \frac{b}{n},$$

and

$$L_+ = \sqrt{\sup_{x \neq y} \frac{\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2}{\|x - y\|^2}} = \frac{b}{\sqrt{n}}.$$

Example 3. Let us consider the optimization problem (8) with $f(x) = \frac{1}{g}\sum_{i=1}^g \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$ and $f_{ij}(x) = \frac{b_i}{2}x^2$ for all $i \in [g]$ and $j \in [m]$, where $x \in \mathbb{R}$ and $b_1 \geq 0$ and $b_i = 0$ for all $i \in \{2, \ldots, g\}$. Then $f(x) = \frac{b_1}{2g}x^2$,

$$L_- = \sup_{x \neq y} \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} = \frac{b_1}{g},$$

$$L_\pm^2 = \sup_{x \neq y} \frac{\frac{1}{gm}\sum_{i=1}^g \sum_{j=1}^m \|\nabla f_{ij}(x) - \nabla f_{ij}(y)\|^2 - \|\nabla f(x) - \nabla f(y)\|^2}{\|x - y\|^2} = \left(\frac{1}{g} - \frac{1}{g^2}\right) b_1^2,$$

and

$$L_{i,\pm}^2 = \sup_{x \neq y} \frac{\frac{1}{m}\sum_{j=m}^n \|\nabla f_{ij}(x) - \nabla f_{ij}(y)\|^2 - \|\nabla f_i(x) - \nabla f_i(y)\|^2}{\|x - y\|^2} = 0 \quad \forall i \in [n].$$

Substituting the smoothness constants to the complexity N_{uniform} from Sec 3 and N_{uniform} from Sec 4, one can show that

$$N_{\text{uniform}} = \Theta\left(n + \frac{\Delta_0 \max\{\sqrt{n}L_\pm, L_-\}}{\varepsilon}\right) = \Theta\left(n + \frac{\Delta_0 \sqrt{n} b_1}{\varepsilon\sqrt{g}}\right)$$

and

$$N_{\text{group}} = \Theta\left(n + \frac{\Delta_0 \max\left\{\sqrt{n}\sqrt{\frac{1}{g}\sum_{i=1}^g L_{i,\pm}^2}, gL_-\right\}}{\varepsilon}\right) = \Theta\left(n + \frac{\Delta_0 b_1}{\varepsilon}\right).$$

The complexity N_{group} is ${\sqrt{n}}/{\sqrt{g}}$ times better than the complexity N_{uniform}.

D Missing Proofs

Lemma 1. *Suppose that Assumption 2 holds and let $x^{t+1} = x^t - \gamma g^t$. Then for any $g^t \in \mathbb{R}^d$ and $\gamma > 0$, we have*

$$f(x^{t+1}) \le f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - \nabla f(x^t)\right\|^2. \tag{10}$$

Proof. Using Assumption 2, we have

$$\begin{aligned} f(x^{t+1}) &\le f(x^t) + \left\langle \nabla f(x^t), x^{t+1} - x^t \right\rangle + \frac{L_-}{2}\left\|x^{t+1} - x^t\right\|^2 \\ &= f(x^t) - \gamma\left\langle \nabla f(x^t), g^t \right\rangle + \frac{L_-}{2}\left\|x^{t+1} - x^t\right\|^2. \end{aligned}$$

Next, due to $-\langle x, y\rangle = \frac{1}{2}\|x - y\|^2 - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2$, we obtain

$$f(x^{t+1}) \le f(x^t) - \frac{\gamma}{2}\left\|\nabla f(x^t)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - \nabla f(x^t)\right\|^2.$$

□

Theorem 6. *Suppose that Assumptions 1, 2, 3 hold and the samplings $\mathbf{S}^t \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$. Then Algorithm 1 (PAGE) has the convergence rate $\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \le \frac{2\Delta_0}{\gamma T}$, where $\gamma \le \left(L_- + \sqrt{\frac{1-p}{p}\left(\left(A - B\right)L_{+,w}^2 + BL_{\pm,w}^2\right)}\right)^{-1}$.*

Proof. We start with the estimation of the variance of the noise:

$$\begin{aligned} &\mathrm{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\ &= (1-p)\mathrm{E}\left[\left\|g^t + \mathbf{S}^t\left(\{\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\}_{i=1}^n\right) - \nabla f(x^{t+1})\right\|^2\right] \\ &= (1-p)\left\|\mathbf{S}^t\left(\{\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\}_{i=1}^n\right) - \left(\nabla f(x^{t+1}) - \nabla f(x^t)\right)\right\|^2 + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2, \end{aligned}$$

where we used the unbiasedness of the sampling. Using Definition 1, we have

$$\begin{aligned} &\mathrm{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\ &\le (1-p)\left(A\sum_{i=1}^n \frac{1}{n^2 w_i}\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 - B\left\|\nabla f(x^{t+1}) - \nabla f(x^t)\right\|^2\right) \\ &\quad + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2. \end{aligned}$$

Using the definition of $L_{+,w}$ and $L_{\pm,w}$, we get

$$\begin{aligned} &\mathrm{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\ &\le (1-p)\left(A\sum_{i=1}^n \frac{1}{n^2 w_i}\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 - B\left\|\nabla f(x^{t+1}) - \nabla f(x^t)\right\|^2\right) \\ &\quad + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2 \\ &= (1-p)\Bigg(\left(A - B\right)\left(\sum_{i=1}^n \frac{1}{n^2 w_i}\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2\right) \\ &\quad + B\left(\sum_{i=1}^n \frac{1}{n^2 w_i}\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2 - \left\|\nabla f(x^{t+1}) - \nabla f(x^t)\right\|^2\right)\Bigg) \\ &\quad + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2 \\ &\le (1-p)\left(\left(A - B\right)L_{+,w}^2 + BL_{\pm,w}^2\right)\left\|x^{t+1} - x^t\right\|^2 + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2. \end{aligned} \tag{11}$$

We now continue the proof using Lemma 1. We add (10) with $\frac{\gamma}{2p}\times$ (11), and take expectation to get

$$\begin{aligned}
&\mathrm{E}\left[f(x^{t+1})-f^*+\frac{\gamma}{2p}\left\|g^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\\
&\leq \mathrm{E}\left[f\left(x^t\right)-f^*-\frac{\gamma}{2}\left\|\nabla f\left(x^t\right)\right\|^2-\left(\frac{1}{2\gamma}-\frac{L_-}{2}\right)\left\|x^{t+1}-x^t\right\|^2+\frac{\gamma}{2}\left\|g^t-\nabla f\left(x^t\right)\right\|^2\right]\\
&\quad+\frac{\gamma}{2p}\mathrm{E}\left[(1-p)\left\|g^t-\nabla f\left(x^t\right)\right\|^2+(1-p)\left(\left(A-B\right)L_{+,w}^2+BL_{\pm,w}^2\right)\left\|x^{t+1}-x^t\right\|^2\right]\\
&=\mathrm{E}\left[f\left(x^t\right)-f^*+\frac{\gamma}{2p}\left\|g^t-\nabla f\left(x^t\right)\right\|^2-\frac{\gamma}{2}\left\|\nabla f\left(x^t\right)\right\|^2\right.\\
&\quad\left.-\left(\frac{1}{2\gamma}-\frac{L_-}{2}-\frac{(1-p)\gamma}{2p}\left(\left(A-B\right)L_{+,w}^2+BL_{\pm,w}^2\right)\right)\left\|x^{t+1}-x^t\right\|^2\right]\\
&\leq \mathrm{E}\left[f\left(x^t\right)-f^*+\frac{\gamma}{2p}\left\|g^t-\nabla f\left(x^t\right)\right\|^2-\frac{\gamma}{2}\left\|\nabla f\left(x^t\right)\right\|^2\right],
\end{aligned} \tag{12}$$

where the last inequality holds due to $\frac{1}{2\gamma}-\frac{L_-}{2}-\frac{(1-p)\gamma}{2p}\left(\left(A-B\right)L_{+,w}^2+BL_{\pm,w}^2\right)\geq 0$ by choosing stepsize

$$\gamma\leq\left(L_-+\sqrt{\frac{1-p}{p}\left(\left(A-B\right)L_{+,w}^2+BL_{\pm,w}^2\right)}\right)^{-1}.$$

Now, if we define $\Phi_t:=f\left(x^t\right)-f^*+\frac{\gamma}{2p}\left\|g^t-\nabla f\left(x^t\right)\right\|^2$, then (12) can be written in the form

$$\mathrm{E}\left[\Phi_{t+1}\right]\leq\mathrm{E}\left[\Phi_t\right]-\frac{\gamma}{2}\mathrm{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right].$$

Summing up from $t=0$ to $T-1$, we get

$$\mathrm{E}\left[\Phi_T\right]\leq\mathrm{E}\left[\Phi_0\right]-\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathrm{E}\left[\left\|\nabla f\left(x^t\right)\right\|^2\right].$$

Then according to the output of the algorithm, i.e., $\widehat{x}_T$ is randomly chosen from $\{x^t\}_{t\in[T]}$ and $\Phi_0=f\left(x^0\right)-f^*+\frac{\gamma}{2p}\|g^0-\nabla f\left(x^0\right)\|^2=f\left(x^0\right)-f^*\overset{\text{def}}{=}\Delta_0$, we have

$$\mathrm{E}\left[\left\|\nabla f\left(\widehat{x}_T\right)\right\|^2\right]\leq\frac{2\Delta_0}{\gamma T}.$$

□

Theorem 7. *Suppose that Assumptions 1, 2, 3, 4 and the samplings* $\mathbf{S}^t\in\mathbb{S}(A,B,\{w_i\}_{i=1}^n)$. *Then Algorithm 1 (PAGE) has the convergence rate* $\mathrm{E}\left[f(x^T)\right]-f^*\leq(1-\gamma\mu)^T\Delta_0$, *where* $\gamma\leq\min\left\{\left(L_-+\sqrt{\frac{2(1-p)}{p}\left(\left(A-B\right)L_{+,w}^2+BL_{\pm,w}^2\right)}\right)^{-1},\frac{p}{2\mu}\right\}$.

Proof. From the proof of Thm 6, we know that

$$\begin{aligned}
&\mathrm{E}\left[\left\|g^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\\
&\leq(1-p)\left(\left(A-B\right)L_{+,w}^2+BL_{\pm,w}^2\right)\left\|x^{t+1}-x^t\right\|^2+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2.
\end{aligned}\tag{13}$$

Using Lemma 1, we add (10) with $\frac{\gamma}{p}\times$ (13), and take expectation to get

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1}) - f^* + \frac{\gamma}{p}\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
&\leq \mathrm{E}\left[f\left(x^t\right) - f^* - \frac{\gamma}{2}\left\|\nabla f\left(x^t\right)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L_-}{2}\right)\left\|x^{t+1} - x^t\right\|^2 + \frac{\gamma}{2}\left\|g^t - \nabla f\left(x^t\right)\right\|^2\right] \\
&\quad + \frac{\gamma}{p}\mathrm{E}\left[(1-p)\left\|g^t - \nabla f\left(x^t\right)\right\|^2 + (1-p)\left(\left(A-B\right)L_{+,w}^2 + BL_{\pm,w}^2\right)\left\|x^{t+1} - x^t\right\|^2\right] \\
&= \mathrm{E}\left[f\left(x^t\right) - f^* + \left(1 - \frac{p}{2}\right)\frac{\gamma}{p}\left\|g^t - \nabla f\left(x^t\right)\right\|^2 - \frac{\gamma}{2}\left\|\nabla f\left(x^t\right)\right\|^2\right. \\
&\quad \left. - \left(\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{(1-p)\gamma}{p}\left(\left(A-B\right)L_{+,w}^2 + BL_{\pm,w}^2\right)\right)\left\|x^{t+1} - x^t\right\|^2\right] \\
&\leq \mathrm{E}\left[f\left(x^t\right) - f^* + \left(1 - \frac{p}{2}\right)\frac{\gamma}{p}\left\|g^t - \nabla f\left(x^t\right)\right\|^2 - \frac{\gamma}{2}\left\|\nabla f\left(x^t\right)\right\|^2\right],
\end{aligned}
$$

where the last inequality holds due to $\frac{1}{2\gamma} - \frac{L_-}{2} - \frac{(1-p)\gamma}{p}\left(\left(A-B\right)L_{+,w}^2 + BL_{\pm,w}^2\right) \geq 0$ by choosing stepsize

$$
\gamma \leq \left(L_- + \sqrt{\frac{2(1-p)}{p}\left(\left(A-B\right)L_{+,w}^2 + BL_{\pm,w}^2\right)}\right)^{-1}.
$$

Next, using Assumption 4 and $\gamma \leq \frac{p}{2\mu}$, we have

$$
\begin{aligned}
&\mathrm{E}\left[f(x^{t+1}) - f^* + \frac{\gamma}{p}\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
&\leq (1 - \gamma\mu)\,\mathrm{E}\left[f\left(x^t\right) - f^* + \frac{\gamma}{p}\left\|g^t - \nabla f\left(x^t\right)\right\|^2\right].
\end{aligned}
$$

Unrolling the recursion and considering that $g^0 = \nabla f(x^0)$, we can complete the proof of theorem. □

E Derivations of the Parameters for the Samplings

E.1 *Nice* sampling

Let S be a random subset uniformly chosen from $[n]$ with a fixed cardinality τ. Let us fix $a_1, \ldots, a_n \in \mathbb{R}^d$. A sampling $\mathbf{S}(a_1, \ldots, a_n) := \frac{1}{n}\sum_{i\in S}\frac{a_i}{p_i}$ is called the *Nice* sampling, where $p_i := \mathbf{Prob}(i \in S)$.

Let us bound $\mathrm{E}\left[\left\|\mathbf{S}(a_1, \ldots, a_n) - \frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right]$ and find parameters from Definition 1. Note that $|\mathbf{S}| = |S| = \tau$. We introduce auxiliary random variables

$$
\chi_i := \begin{cases} 1 & i \in S \\ 0 & \text{otherwise.} \end{cases}.
$$

Due to $p_i = \mathbf{Prob}\left(i \in S\right) = \frac{\tau}{n}$, we have

$$\begin{aligned}
\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i\in S}\frac{a_i}{p_i}\right\|^2\right] &= \mathrm{E}\left[\left\|\frac{1}{\tau}\sum_{i=1}^{n}\chi_i a_i\right\|^2\right] \\
&= \frac{1}{\tau^2}\sum_{i=1}^{n}\mathrm{E}\left[\left\|\chi_i a_i\right\|^2\right] + \frac{1}{\tau^2}\sum_{i\neq j}\mathrm{E}\left[\langle\chi_i a_i, \chi_j a_j\rangle\right] \\
&= \frac{1}{\tau^2}\sum_{i=1}^{n}\mathrm{E}\left[\chi_i\right]\left\|a_i\right\|^2 + \frac{1}{\tau^2}\sum_{i\neq j}\mathrm{E}\left[\langle\chi_i, \chi_j\rangle\right]\langle a_i, a_j\rangle \\
&= \frac{1}{n\tau}\sum_{i=1}^{n}\left\|a_i\right\|^2 + \frac{\tau-1}{n(n-1)\tau}\sum_{i\neq j}\langle a_i, a_j\rangle \\
&= \frac{1}{n\tau}\sum_{i=1}^{n}\left\|a_i\right\|^2 + \frac{\tau-1}{n(n-1)\tau}\left(\left\|\sum_{i=1}^{n}a_i\right\|^2 - \sum_{i=1}^{n}\left\|a_i\right\|^2\right) \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n}\left\|a_i\right\|^2 + \frac{\tau-1}{n(n-1)\tau}\left\|\sum_{i=1}^{n}a_i\right\|^2,
\end{aligned}$$

where we use $\mathrm{E}\left[\chi_i^2\right] = \mathrm{E}\left[\chi_i\right] = \frac{\tau}{n}$ and $\mathrm{E}\left[\chi_i\chi_j\right] = \frac{\tau(\tau-1)}{n(n-1)}$, when $i \neq j$.

Finally, we have

$$\begin{aligned}
\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i\in S}\frac{a_i}{p_i} - \frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] &= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i\in S}\frac{a_i}{p_i}\right\|^2\right] - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n}\left\|a_i\right\|^2 + \frac{\tau-1}{n(n-1)\tau}\left\|\sum_{i=1}^{n}a_i\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\left\|a_i\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right).
\end{aligned}$$

Thus we have $A = B = \frac{n-\tau}{\tau(n-1)}$ and $w_i = \frac{1}{n}$ for all $i \in [n]$.

E.2 *Independent* sampling

Let us define i.i.d. random variables

$$\chi_i = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i, \end{cases}.$$

for all $i \in [n]$ and take $S := \{i \in [n] \mid \chi_i = 1\}$. We now fix $a_1, \ldots, a_n \in \mathbb{R}^d$. A sampling $\mathbf{S}(a_1, \ldots, a_n) := \frac{1}{n}\sum_{i\in S}\frac{a_i}{p_i}$ is called the *Independent* sampling, where $p_i := \mathbf{Prob}(i \in S)$.

We get

$$\begin{aligned}
\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i\in S}\frac{a_i}{p_i}-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] &= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\chi_i a_i\right\|^2\right]-\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\\
&= \sum_{i=1}^{n}\frac{\mathrm{E}\left[\chi_i\right]}{n^2p_i^2}\left\|a_i\right\|^2+\sum_{i\neq j}\frac{\mathrm{E}\left[\chi_i\right]\mathrm{E}\left[\chi_j\right]}{n^2p_ip_j}\left\langle a_i,a_j\right\rangle-\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\\
&= \sum_{i=1}^{n}\frac{1}{n^2p_i}\left\|a_i\right\|^2+\frac{1}{n^2}\left(\left\|\sum_{i=1}^{n}a_i\right\|^2-\sum_{i=1}^{n}\left\|a_i\right\|^2\right)-\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{p_i}-1\right)\left\|a_i\right\|^2.
\end{aligned}$$

Thus we have $A=\frac{1}{\sum_{i=1}^{n}\frac{p_i}{1-p_i}}$, $B=0$ and $w_i=\frac{\frac{p_i}{1-p_i}}{\sum_{i=1}^{n}\frac{p_i}{1-p_i}}$ for all $i\in[n]$.

E.3 *Importance* and *Uniform With Replacement* sampling

Let us fix $\tau>0$. For all $k\in[\tau]$, we define i.i.d. random variables

$$\chi_k=\begin{cases}1 & \text{with probability } q_1\\ 2 & \text{with probability } q_2\\ & \vdots\\ n & \text{with probability } q_n,\end{cases}$$

where $(q_1,\dots,q_n)\in\mathcal{S}^n$ (simple simplex). A sampling

$$\mathbf{S}(a_1,\dots,a_n):=\frac{1}{\tau}\sum_{k=1}^{\tau}\frac{a_{\chi_k}}{nq_{\chi_k}}$$

is called the *Importance* sampling. The *Importance* sampling reduces to the *Uniform With Replacement* sampling when $q_i={}^1\!/\!_n$ for all $i\in[n]$. Note that $|\mathbf{S}|\leq\tau$.

Let us bound the variance

$$\begin{aligned}
&\mathrm{E}\left[\left\|\frac{1}{\tau}\sum_{k=1}^{\tau}\frac{a_{\chi_k}}{nq_{\chi_k}}-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]\\
&=\frac{1}{\tau^2}\sum_{k=1}^{\tau}\mathrm{E}\left[\left\|\frac{a_{\chi_k}}{nq_{\chi_k}}-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]+\frac{1}{\tau^2}\sum_{k\neq k'}\mathrm{E}\left[\left\langle\frac{a_{\chi_k}}{nq_{\chi_k}}-\frac{1}{n}\sum_{i=1}^{n}a_i,\frac{a_{\chi_{k'}}}{nq_{\chi_{k'}}}-\frac{1}{n}\sum_{i=1}^{n}a_i\right\rangle\right].
\end{aligned}$$

Using the independents and unbiasedness of the random variables, the last term vanishes and we get

$$\begin{aligned}
\mathrm{E}\left[\left\|\frac{1}{\tau}\sum_{k=1}^{\tau}\frac{a_{\chi_k}}{nq_{\chi_k}}-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] &= \frac{1}{\tau^2}\sum_{k=1}^{\tau}\mathrm{E}\left[\left\|\frac{a_{\chi_k}}{nq_{\chi_k}}-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]\\
&\overset{(9)}{=}\frac{1}{\tau^2}\sum_{k=1}^{\tau}\mathrm{E}\left[\left\|\frac{a_{\chi_k}}{nq_{\chi_k}}\right\|^2\right]-\frac{1}{\tau}\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\\
&= \frac{1}{\tau}\sum_{i=1}^{n}q_i\left\|\frac{a_i}{nq_i}\right\|^2-\frac{1}{\tau}\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2
\end{aligned}$$

$$= \frac{1}{\tau}\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{nq_i}\left\|a_i\right\|^2 - \left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2\right).$$

Thus we have $A = B = \frac{1}{\tau}$, and $w_i = q_i$ for all $i \in [n]$.

E.4 *Extended Nice* sampling

In this section, we analyze the extension of *Nice* sampling. First, we l_i times repeat each vector a_i, then we use the *Nice* sampling. We define

$$\tilde{a}_i := \begin{cases} \frac{\sum_{j=1}^n l_j}{nl_1}a_1 & 1 \le i \le l_1 \\ \frac{\sum_{j=1}^n l_j}{nl_2}a_2 & l_1+1 \le i \le l_1+l_2 \\ & \vdots \\ \frac{\sum_{j=1}^n l_j}{nl_n}a_n & \sum_{j=1}^{n-1} l_j \le i \le \sum_{j=1}^n l_j, \end{cases},$$

where $a_i \in \mathbb{R}^d$ and $l_i \ge 1$ for all $i \in [n]$. Then we have

$$\frac{1}{n}\sum_{i=1}^n a_i(x) = \frac{1}{N}\sum_{i=1}^N \tilde{a}_i(x),$$

where $N := \sum_{j=1}^n l_j$. Also, we denote $N_k := \sum_{j=1}^k l_j$.

For some $\tau > 0$, we apply the *Nice* sampling method:

$$\mathbf{S}(a_1,\ldots,a_n) := \frac{1}{N}\sum_{i\in S}\frac{\tilde{a}_i}{p_i} = \sum_{i=1}^N \frac{1}{\tau}\chi_i\tilde{a}_i,$$

where

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & \text{otherwise} \end{cases}, \quad p_i = \mathbf{Prob}\left(i \in S\right),$$

and S is a random set with cardinality τ from $[N]$. The sampling $\mathbf{S}(a_1,\ldots,a_n)$ is called the *Extended Nice* sampling.

We now ready to bound the variance. Using the results for the *Nice* sampling, we obtain

$$\begin{aligned} &\mathrm{E}\left[\left\|\mathbf{S}(a_1,\ldots,a_n) - \frac{1}{n}\sum_{i=1}^n a_i(x)\right\|^2\right] \\ &= \mathrm{E}\left[\left\|\mathbf{S}(a_1,\ldots,a_n) - \frac{1}{N}\sum_{i=1}^N \tilde{a}_i(x)\right\|^2\right] \\ &= \frac{n-\tau}{\tau(n-1)}\frac{1}{N}\sum_{i=1}^N \left\|\tilde{a}_i\right\|^2 - \frac{n-\tau}{\tau(n-1)}\left\|\frac{1}{N}\sum_{i=1}^N \tilde{a}_i\right\|^2 \\ &= \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{N}\left(\frac{N}{nl_1}\right)^2\sum_{i=1}^{N_1}\left\|a_1\right\|^2 + \frac{1}{N}\left(\frac{N}{nl_2}\right)^2\sum_{i=N_1+1}^{N_2}\left\|a_2\right\|^2\right. \\ &\qquad \left. + \cdots + \frac{1}{N}\left(\frac{N}{nl_n}\right)^2\sum_{i=N_{n-1}+1}^{N}\left\|a_n\right\|^2\right) - \frac{n-\tau}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2 \end{aligned}$$

$$= \frac{n-\tau}{\tau(n-1)}\left(\frac{N}{nl_1}\frac{1}{n}\|a_1\|^2 + \frac{N}{nl_2}\frac{1}{n}\|a_2\|^2 + \cdots + \frac{N}{nl_n}\frac{1}{n}\|a_n\|^2\right) - \frac{n-\tau}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2$$

$$= \frac{n-\tau}{\tau(n-1)}\left(\sum_{i=1}^n \frac{1}{n^2 w_i}\|a_i\|^2\right) - \frac{n-\tau}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|^2$$

where $w_i = \frac{l_i}{N}$. Thus we have $A = B = \frac{n-\tau}{\tau(n-1)}$ and $w_i = \frac{l_i}{N}$ for $i \in [n]$.

F The Optimal Choice of w_i

Let us consider $L^2_{+,w}$ and $L^2_{\pm,w}$. In Sec 2, we show that one can take $L^2_{+,w} = L^2_{\pm,w} = \frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i}L_i^2$. Let us minimize $\frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i}L_i^2$ with respect to the weights w_i such that $w_1, \ldots, w_n \geq 0$ and $\sum_{i=1}^n w_i = 1$. Using the method of Lagrange multipliers, we can construct a Lagrangian

$$\mathcal{L}(w, \lambda) := \frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i}L_i^2 - \lambda\left(\sum_{i=1}^n w_i - 1\right).$$

Next, we calculate partial derivatives

$$\frac{\partial \mathcal{L}}{\partial w_i} = -\frac{1}{n^2 w_i^2}L_i^2 - \lambda = 0 \forall i \in [n]$$

and get

$$w_i^2 = -\frac{L_i^2}{n^2\lambda}.$$

Using $\sum_{i=1}^n w_i = 1$, we can show that the weights $w_i^* = \frac{L_i}{\sum_{i=1}^n L_i}$ are the solutions of the minimization problem. Moreover,

$$L^2_{\pm,w^*} = \frac{1}{n}\sum_{i=1}^n \frac{1}{nw_i^*}L_i^2 = \left(\frac{1}{n}\sum_{i=1}^n L_i\right)^2.$$

G The Complexity of Algorithm 1 with the *Importance* sampling

The expected number of gradient calculations ∇f_i of Algorithm 1 with the *Importance* sampling, the optimal ${w_i}^*$ from Sec. F, and $\tau \leq \max\left\{\frac{\sqrt{n}L_{\pm,w}}{L_-}, 1\right\}$ equals

$$\begin{aligned} N_{\text{importance}} &= \mathcal{O}\left(n + \frac{\Delta_0}{\varepsilon}\tau\left(L_- + \frac{\sqrt{n}}{\tau}\sqrt{\frac{1}{n}\sum_{i=1}^n \frac{1}{n{w_i}^*}L_i^2}\right)\right) \\ &= \mathcal{O}\left(n + \frac{\Delta_0}{\varepsilon}\tau\left(L_- + \frac{\sqrt{n}}{\tau}\frac{1}{n}\sum_{i=1}^n L_i\right)\right) \\ &= \mathcal{O}\left(n + \frac{\Delta_0}{\varepsilon}\sqrt{n}L_{\pm,w^*} + \frac{\Delta_0\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n L_i\right)}{\varepsilon}\right) \\ &= \mathcal{O}\left(n + \frac{\Delta_0\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n L_i\right)}{\varepsilon}\right). \end{aligned}$$

H Missing Proofs: The Composition of Samplings

Lemma 2. *Let us assume that a random sampling function* $\mathbf{S}$ *belongs to Definition 1 with some* A, B *and weights* w_i*, and a random sampling function* $\mathbf{S}_i$ *belongs to Definition 1 with some* A_i, B_i *and weights* w_{ij} *for* *all $i \in [n]$. Moreover, $B \leq 1$. Then*

$$\mathrm{E}\left[\left\|\mathbf{S}\left(\mathbf{S}_1\left(a_{11},\ldots,a_{1m_1}\right),\ldots,\mathbf{S}_n\left(a_{n1},\ldots,a_{nm_n}\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}a_{ij}\right)\right\|^2\right]$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i}+\frac{(1-B)}{n}\right)\left(\frac{A_i}{m_i}\sum_{j=1}^{m_i}\frac{1}{m_iw_{ij}}\left\|a_{ij}\right\|^2-B_i\left\|\frac{1}{m_i}\sum_{j=1}^{m_i}a_{ij}\right\|^2\right)$$
$$+\frac{A}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|\frac{1}{m_i}\sum_{j=1}^{m_i}a_{ij}\right\|^2-B\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i}\sum_{j=1}^{m_i}a_{ij}\right)\right\|^2,$$

where $a_{ij} \in \mathbb{R}^d$ for all $j \in [m_i]$ and $i \in [n]$.

Proof. We denote $\widehat{a}_i := \mathbf{S}_i\left(a_{i1},\ldots,a_{im_i}\right)$ and $a_i := \frac{1}{m_i}\sum_{j=1}^{m_i}a_{ij}$. Using (9), we have

$$\mathrm{E}\left[\left\|\mathbf{S}\left(\widehat{a}_1,\ldots,\widehat{a}_n\right)-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]$$
$$=\mathrm{E}\left[\mathrm{E}_S\left[\left\|\mathbf{S}\left(\widehat{a}_1,\ldots,\widehat{a}_n\right)-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]\right]$$
$$=\mathrm{E}\left[\mathrm{E}_S\left[\left\|\mathbf{S}\left(\widehat{a}_1,\ldots,\widehat{a}_n\right)-\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i\right\|^2\right]\right]+\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right].$$

Next, using Definition 1 for the sampling $\mathbf{S}$, we get

$$\mathrm{E}\left[\left\|\mathbf{S}\left(\widehat{a}_1,\ldots,\widehat{a}_n\right)-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]$$
$$\leq A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\mathrm{E}\left[\left\|\widehat{a}_i\right\|^2\right]-B\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i\right\|^2\right]$$
$$+\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right].$$

Due to (9), we obtain

$$\mathrm{E}\left[\left\|\mathbf{S}\left(\widehat{a}_1,\ldots,\widehat{a}_n\right)-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]$$
$$\leq A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\mathrm{E}\left[\left\|\widehat{a}_i-a_i\right\|^2\right]+A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|a_i\right\|^2$$
$$-B\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]-B\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2$$
$$+\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i-\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right]$$

$$
\begin{aligned}
&= A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\mathrm{E}\left[\left\|\widehat{a}_i - a_i\right\|^2\right] + A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|a_i\right\|^2 \\
&\quad + (1-B)\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i - \frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] - B\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2 \\
&= A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\mathrm{E}\left[\left\|\widehat{a}_i - a_i\right\|^2\right] + A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|a_i\right\|^2 \\
&\quad + \frac{(1-B)}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\left\|\widehat{a}_i - a_i\right\|^2\right] - B\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i} + \frac{(1-B)}{n}\right)\mathrm{E}\left[\left\|\widehat{a}_i - a_i\right\|^2\right] + A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|a_i\right\|^2 - B\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2.
\end{aligned}
$$

Using Definition 1 for the samplings $\mathbf{S}_i$, we have

$$
\begin{aligned}
&\mathrm{E}\left[\left\|\mathbf{S}\left(\widehat{a}_1, \dots, \widehat{a}_n\right) - \frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2\right] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i} + \frac{(1-B)}{n}\right)\left(A_i\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{m_iw_{ij}}\left\|a_{ij}\right\|^2 - B_i\left\|a_i\right\|^2\right) \\
&\quad + A\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|a_i\right\|^2 - B\left\|\frac{1}{n}\sum_{i=1}^{n}a_i\right\|^2.
\end{aligned}
$$

□

Theorem 10. *Suppose that Assumptions 1, 2, 3, 5 hold and the samplings $\mathbf{S}^t \in \mathbb{S}(A, B, \{w_i\}_{i=1}^n)$ and the samplings $\mathbf{S}_i^t \in \mathbb{S}(A_i, B_i, \{w_{ij}\}_{j=1}^{m_i})$ for all $i \in [n]$. Moreover, $B \le 1$. Then Algorithm 2 has the convergence rate $\mathrm{E}\left[\left\|\nabla f(\widehat{x}^T)\right\|^2\right] \le \frac{2\Delta_0}{\gamma T}$, where $\gamma \le \left(L_- + \sqrt{\frac{1-p}{p}\left(\mathcal{L}_l + \mathcal{L}_g\right)}\right)^{-1}$,*

$$
\mathcal{L}_l := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i} + \frac{(1-B)}{n}\right)((A_i - B_i)L_{i,+,w_i}^2 + B_iL_{i,\pm,w_i}^2),
$$

and

$$
\mathcal{L}_g := (A-B)L_{+,w}^2 + BL_{\pm,w}^2.
$$

Proof. We start with the estimation of the variance of the noise:

$$
\begin{aligned}
&\mathrm{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
&= (1-p)\mathrm{E}\left[\left\|g^t + \mathbf{S}\left(\left\{\mathbf{S}_i\left(\{\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\}_{j=1}^{m_i}\right)\right\}_{i=1}^n\right) - \nabla f(x^{t+1})\right\|^2\right] \\
&= (1-p)\left\|\mathbf{S}\left(\left\{\mathbf{S}_i\left(\{\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\}_{j=1}^{m_i}\right)\right\}_{i=1}^n\right) - \left(\nabla f(x^{t+1}) - \nabla f(x^t)\right)\right\|^2 + (1-p)\left\|g^t - \nabla f(x^t)\right\|^2,
\end{aligned}
$$

where we used the unbiasedness of the composition of samplings. Using Lemma 2, we have

$$
\begin{aligned}
&\mathrm{E}\left[\left\|g^{t+1} - \nabla f(x^{t+1})\right\|^2\right] \\
&\le (1-p)\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i} + \frac{(1-B)}{n}\right)\left(\frac{A_i}{m_i}\sum_{j=1}^{m_i}\frac{1}{m_iw_{ij}}\left\|\nabla f_{ij}(x^{t+1}) - \nabla f_{ij}(x^t)\right\|^2 - B_i\left\|\nabla f_i(x^{t+1}) - \nabla f_i(x^t)\right\|^2\right)\right.
\end{aligned}
$$

$$+\frac{A}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2-B\left\|\nabla f(x^{t+1})-\nabla f(x^t)\right\|^2\Bigg)$$
$$+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2.$$

Using Definitions 3, 4, 8 and 9, we get

$$\begin{aligned}
&\mathrm{E}\left[\left\|g^{t+1}-\nabla f(x^{t+1})\right\|^2\right]\\
&\leq (1-p)\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i}+\frac{(1-B)}{n}\right)\left(\frac{A_i-B_i}{m_i}\sum_{j=1}^{m_i}\frac{1}{m_iw_{ij}}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right\|^2\right.\right.\\
&+\quad B_i\left.\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{1}{m_iw_{ij}}\left\|\nabla f_{ij}(x^{t+1})-\nabla f_{ij}(x^t)\right\|^2-\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2\right)\right)\\
&\quad+\frac{A-B}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2\\
&\quad\left.+B\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{nw_i}\left\|\nabla f_i(x^{t+1})-\nabla f_i(x^t)\right\|^2-\left\|\nabla f(x^{t+1})-\nabla f(x^t)\right\|^2\right)\right)\\
&\quad+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2\\
&\leq (1-p)\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i}+\frac{(1-B)}{n}\right)\left((A_i-B_i)L_{i,+,w_i}^2+B_iL_{i,\pm,w_i}^2\right)+(A-B)L_{+,w}^2+BL_{\pm,w}^2\right)\left\|x^{t+1}-x^t\right\|^2\\
&\quad+(1-p)\left\|g^t-\nabla f(x^t)\right\|^2.
\end{aligned}$$

From this point the proof of theorem repeats the proof of Thm 6 with

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A}{nw_i}+\frac{(1-B)}{n}\right)\left((A_i-B_i)L_{i,+,w_i}^2+B_iL_{i,\pm,w_i}^2\right)+(A-B)L_{+,w}^2+BL_{\pm,w}^2$$

instead of

$$(A-B)\,L_{+,w}^2+BL_{\pm,w}^2.$$

□

I Artificial Quadratic Optimization Tasks

In this section, we provide algorithms that we use to generate artificial optimization tasks for experiments. Algorithm 3 and Algorithm 4 allow us to control the smoothness constants $L_\pm$ and L_i, accordingly, via the noise scales.

Algorithm 3 Generate quadratic optimization task with controlled $L_\pm$ (homogeneity)

```
Parameters: number nodes $n$, dimension $d$, regularizer $\lambda$, and noise scale $s$.
for $i = 1, \dots, n$ do
    Generate random noises $\nu_i^s = 1 + s\xi_i^s$ and $\nu_i^b = s\xi_i^b$, i.i.d. $\xi_i^s, \xi_i^b \sim \mathrm{NormalDistribution}(0, 1)$
    Take vector $b_i = \frac{\nu_i^s}{4}(-1 + \nu_i^b, 0, \cdots, 0) \in \mathbb{R}^d$
    Take the initial tridiagonal matrix
        $\mathbf{A}_i = \frac{\nu_i^s}{4} \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{d \times d}$
end for
Take the mean of matrices $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i$
Find the minimum eigenvalue $\lambda_{\min}(\mathbf{A})$
for $i = 1, \dots, n$ do
    Update matrix $\mathbf{A}_i = \mathbf{A}_i + (\lambda - \lambda_{\min}(\mathbf{A}))\mathbf{I}$
end for
Take starting point $x^0 = (\sqrt{d}, 0, \cdots, 0)$
Output: matrices $\mathbf{A}_1, \cdots, \mathbf{A}_n$, vectors $b_1, \cdots, b_n$, starting point $x^0$
```

Algorithm 4 Generate quadratic optimization task with controlled L_i

```
Parameters: number nodes $n$, dimension $d$, regularizer $\lambda$, and noise scale $s$.
for $i = 1, \dots, n$ do
    Generate random noises $\nu_i^s = 1 + s\xi_i^s$, where i.i.d. $\xi_i^s \sim \mathrm{ExponentialfDistribution}(1.0)$
    Generate random noises $\nu_i^b = s\xi_i^b$, i.i.d. $\xi_i^b \sim \mathrm{NormalDistribution}(0, 1)$
    Take vector $b_i = (-\frac{1}{4} + \nu_i^b, 0, \cdots, 0) \in \mathbb{R}^d$
    Take the initial tridiagonal matrix
        $\mathbf{A}_i = \frac{\nu_i^s}{4} \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{d \times d}$
end for
Take starting point $x^0 = (\sqrt{d}, 0, \cdots, 0)$
Output: matrices $\mathbf{A}_1, \cdots, \mathbf{A}_n$, vectors $b_1, \cdots, b_n$, starting point $x^0$
```