SGD: General Analysis and Improved Rates

Robert M. Gower [1] **Nicolas Loizou** [2] **Xun Qian** [3] **Alibek Sailanbayev** [3] **Egor Shulgin** [4] **Peter Richtárik** [3 2 4]

Abstract

We propose a general yet simple theorem describing the convergence of SGD under the *arbitrary sampling* paradigm. Our theorem describes the convergence of an infinite array of variants of SGD, each of which is associated with a specific probability law governing the data selection rule used to form minibatches. This is the first time such an analysis is performed, and most of our variants of SGD were never explicitly considered in the literature before. Our analysis relies on the recently introduced notion of expected smoothness and does not rely on a uniform bound on the variance of the stochastic gradients. By specializing our theorem to different mini-batching strategies, such as sampling with replacement and independent sampling, we derive exact expressions for the stepsize as a function of the mini-batch size. With this we can also determine the mini-batch size that optimizes the total complexity, and show explicitly that as the variance of the stochastic gradient evaluated at the minimum grows, so does the optimal mini-batch size. For zero variance, the optimal mini-batch size is one. Moreover, we prove insightful stepsize-switching rules which describe when one should switch from a constant to a decreasing stepsize regime.

1. Introduction

We consider the optimization problem

$$x^* = \arg\min_{x\in\mathbb{R}^d} \left[f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) \right], \qquad (1)$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is smooth (but not necessarily convex). Further, we assume that f has a unique[1] global minimizer x^* and is μ–strongly quasi-convex (Karimi et al., 2016; Necoara et al., 2018):

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

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

[1]Télécom Paris Tech, France [2]University of Edinburgh, United Kingdom [3]King Abdullah University of Science and Technology, Kingdom of Saudi Arabia [4]Moscow Institute of Physics and Technology, Russian Federation. Correspondence to: Peter Richtárik <peter.richtarik@kaust.edu.sa>.

January 26, 2019

[1]This assumption can be relaxed; but for simplicity of exposition we enforce it.

1.1. Background and contributions

Stochastic gradient descent (SGD) (Robbins & Monro, 1951; Nemirovski & Yudin, 1978; 1983; Shalev-Shwartz et al., 2007; Nemirovski et al., 2009; Hardt et al., 2016), has become the workhorse for training supervised machine learning problems which have the generic form (1).

Linear convergence of SGD. Among the first non-asymptotic analyses of SGD was done by Moulines & Bach (2011) who provided, among other results, a linear convergence analysis for strongly convex f up to a certain noise level. Needell et al. (2016) improved upon these results by removing the quadratic dependency on the condition number in the iteration complexity results, and considered importance sampling. The analysis of Needell et al. (2016) was later extended to a mini-batch variant where the mini-batches are formed by partitioning the data (Needell & Ward, 2017). These works are the main starting point for ours.

Contributions: We further *tighten and generalize these results to virtually all forms of sampling.* We introduce an *expected smoothness* assumption (Assumption 2.1), first introduced in (Gower et al., 2018) in the context of a certain class of variance-reduced methods. This assumption is a joint property of f and the sampling scheme $\mathcal{D}$ utilized by an SGD method, and allows us prove a generic complexity result (Theorem 3.1) that holds for arbitrary sampling schemes $\mathcal{D}$. Our work is the *first time SGD is analysed under this assumption.* We obtain *linear convergence rates without strong convexity*; in particular, assuming strong quasi-convexity (this class includes some non-convex functions as well). Furthermore, we *do not require the functions f_i to be convex.*

Gradient noise assumptions. Shamir & Zhang (2013) extended the analysis of SGD to convex non-smooth optimization (including the strongly convex case). However, their proofs still rely on the assumption that the variance of the stochastic gradient is bounded for all iterates of the algorithm: there exists $c \in \mathbb{R}$ such that $\mathbb{E}_i\|\nabla f_i(x^k)\|^2 \leq c$ for

all k. The same assumption was used in the analysis of several recent papers (Recht et al., 2011; Hazan & Kale, 2014; Rakhlin et al., 2012). Bottou et al. (2018) establish a linear convergence of SGD under a somewhat less restrictive condition, namely $\mathbb{E}_i\|\nabla f_i(x^k)\|^2 \leq c_1 + c_2\mathbb{E}\|\nabla f(x^k)\|^2$ for all k. Recently, Nguyen et al. (2018) turn this assumption into a theorem by establishing formulas c_1 and c_2 under some reasonable conditions, and provide further insights into the workings of SGD and and its parallel asynchronous cousin, Hogwild!. Similar conditions have been also proved and used in the analysis of decentralized variants of SGD (Lian et al., 2017; Assran et al., 2018). Based on a strong growth condition of the stochastic gradients, Cevher & Vu (2017) give sufficient and necessary conditions for the linear convergence of SGD. This strong growth condition holds when using SGD for solving a consistent linear system, but it does not hold for a wide range of problems.

Contributions: Our analysis does not directly assume a growth condition. Instead, we make use of the remarkably weak *expected smoothness* assumption.

Optimal mini-batch size. Recently it was experimentally shown by Goyal et al. (2017) that using larger mini-batches sizes is key to efficient training of large scale non-convex problems, leading to the training of ImageNet in under 1 hour. The authors conjectured that the stepsize should grow linearly with the mini-batch size.

Contributions: We prove (see Section 4) that this is the case, upto a certain optimal mini-batch size, and provide *exact formulas for the dependency of the stepsizes on the mini-batch sizes.*

Learning schedules. Chee & Toulis (2018) develop techniques for detecting the convergence of SGD within a region around the solution.

Contributions: We provide a closed-form formula for when should SGD *switch from a constant stepsize to a decreasing stepsize* (see Theorem 3.2). Further, we clearly show how the *optimal stepsize (learning rate) increases and the iteration complexity decreases as the mini-batch size* increases for both independent sampling and sampling with replacement. We also recover the well known $L/\mu\log(1/\epsilon)$ convergence rate of gradient descent (GD) when the mini-batch size is n; this is the *first time a generic SGD analysis recovers the correct rate of GD.*

Over-parameterized models. There has been some recent work in analysing SGD in the setting where the underlying model being trained has more parameters than there is data available. In this *zero–noise* setting, Ma et al. (2018) showed that SGD converges linearly.

Contributions: In the case of over-parametrized models, we extend the findings of Ma et al. (2018)[2] to independent sampling and sampling with replacement by showing that the optimal mini-batch size is 1. Moreover, we provide results in the more general setting where the model is not necessarily over-parametrized.

Practical performance. We corroborate our theoretical results with extensive experimental testing.

1.2. Stochastic reformulation

In this work we provide a single theorem through which we can analyse all importance sampling and mini-batch variants of SGD. To do this, we need to introduce a *sampling vector* which we will use to re-write our problem (1).

Definition 1.1. We say that a random vector $v \in \mathbb{R}^n$ drawn from some distribution $\mathcal{D}$ is a *sampling vector* if its mean is the vector of all ones:

$$\mathbb{E}_{\mathcal{D}}\left[v_i\right] = 1, \quad \forall i \in [n]. \tag{3}$$

With each distribution $\mathcal{D}$ we now introduce a *stochastic reformulation* of (1) as follows

$$\min_{x\in\mathbb{R}^d} \ \mathbb{E}_{\mathcal{D}}\left[f_v(x) := \frac{1}{n}\sum_{i=1}^n v_i f_i(x)\right]. \tag{4}$$

By the definition of the sampling vector, $f_v(x)$ and $\nabla f_v(x)$ are unbiased estimators of $f(x)$ and $\nabla f(x)$, respectively, and hence probem (4) is indeed equivalent (i.e., a reformulation) of the original problem (1). In the case of the gradient, for instance, we get

$$\mathbb{E}_{\mathcal{D}}\left[\nabla f_v(x)\right] \overset{(4)}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathcal{D}}\left[v_i\right]\nabla f_i(x) \overset{(3)}{=} \nabla f(x). \tag{5}$$

Similar but different stochastic reformulations were recently proposed by Richtárik & Takáč (2017) and further used in (Loizou & Richtárik, 2017) for the more special problem of solving linear systems, and by Gower et al. (2018) in the context of variance-reduced methods. Reformulation (4) can be solved using SGD in a natural way:

$$\boxed{x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k)} \tag{6}$$

where $v^k \sim \mathcal{D}$ is sampled i.i.d at each iteration and $\gamma^k > 0$ is a stepsize. However, for different distributions $\mathcal{D}$, (6) has a *different interpretation* as an SGD method for solving the *original* problem (1). In our main result we will analyse (6) for any $\mathcal{D}$ satisfying (3). By substituting specific choices of $\mathcal{D}$, we obtain specific variants of SGD for solving (1).

[2]Recently, the results of Ma et al. (2018) were extended to the accelerated case by Vaswani et al. (2018); however, we do not study accelerated methods in this work.

2. Expected Smoothness and Gradient Noise

In our analysis of SGD (6) applied to the stochastic reformulation (4) we rely on a generic and remarkably weak assumption of *expected smoothness*, which we now define and relate to existing growth conditions.

2.1. Expected smoothness

Expected smoothness (Gower et al., 2018) is an assumption that combines both the properties of the distribution $\mathcal{D}$ and the smoothness properties of function f.

Assumption 2.1 (Expected Smoothness)**.** We say that f is $\mathcal{L}$–smooth in expectation with respect to distribution $\mathcal{D}$ if there exists $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$ such that

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2\right] \leq 2\mathcal{L}(f(x) - f(x^*)), \tag{7}$$

for all $x \in \mathbb{R}^d$. For simplicity, we will write $(f, \mathcal{D}) \sim ES(\mathcal{L})$ to say that (7) holds. When $\mathcal{D}$ is clear from the context, we will often ignore mentioning it, and simply state that the expected smoothness constant is $\mathcal{L}$.

In Section 3.3 we show how convexity and L_i–smoothness of f_i implies expected smoothness. However, the opposite implication does not hold. Indeed, the expected smoothness assumption can hold even when the f_i's and f are not convex, as we show in the next example.

Example 2.2 (Non-convexity and expected smoothness)**.** Let $f_i = \phi$ for $i = 1, \ldots, n$, where ϕ is a L_ϕ–smooth and non-convex function which has a global minimum $x^* \in \mathbb{R}^d$ (such functions exist[a]). Consequently $f = \phi$ and $f_v = \frac{\sum_i v_i}{n}\phi$. Letting $\theta := \mathbb{E}_{\mathcal{D}}\left[\left(\sum_i v_i\right)^2\right]$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2\right] &= \tfrac{\theta}{n^2}\|\nabla\phi(x) - \nabla\phi(x^*)\|^2 \\ &\leq \tfrac{2\theta L_\phi}{n^2}(f(x) - f(x^*)),\end{aligned}$$

where the last inequality follows from Proposition A.1. So, $(f, \mathcal{D}) \sim ES(\mathcal{L})$ for $\mathcal{L} = \frac{\theta L_\phi}{n^2}$.

[a] There exists invex functions that satisfy these conditions (Karimi et al., 2016). As an example $\phi(x) = x^2 + 3\sin^2(x)$ is smooth, non-convex, and has a unique global minimizer.

2.2. Gradient noise

Our second key assumption is finiteness of gradient noise, defined next:

Assumption 2.3 (Finite Gradient Noise)**.** The *gradient noise* $\sigma = \sigma(f, \mathcal{D})$, defined by

$$\sigma^2 := \mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x^*)\|^2], \tag{8}$$

is finite.

This is a very weak assumption, and should intuitively be really seen as an assumption on $\mathcal{D}$ rather than on f. For instance, if the sampling vector v is non-negative with probability one and $\mathbb{E}[v_i \sum_j v_j]$ is finite for all i, then σ is finite. When (1) is the training problem of an over-parametrized model, which often occurs in deep neural networks, each individual loss function f_i attains its minimum at x^*, and thus $\nabla f_i(x^*) = 0$. It follows that $\sigma = 0$.

2.3. Key lemma and connection to the weak growth condition

A common assumption used to prove the convergence of SGD is uniform boundedness of the stochastic gradients[3]: there exist $0 < c < \infty$ such that $\mathbb{E}\|\nabla f_v(x)\|^2 \leq c$ for all x. However, this assumption often does not hold, such as in the case when f is strongly convex (Bottou et al., 2018; Nguyen et al., 2018). We do not assume such a bound. Instead, we use the following direct consequence of expected smoothness to bound the expected norm of the stochastic gradients.

Lemma 2.4. If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x)\|^2\right] \leq 4\mathcal{L}(f(x) - f(x^*)) + 2\sigma^2. \tag{9}$$

When the gradient noise is zero ($\sigma = 0$), inequality (9) is known as the *weak growth condition* (Vaswani et al., 2018). We have the following corollary:

Corollary 2.5. If $(f, \mathcal{D}) \sim ES(\mathcal{L})$ and if $\sigma = 0$, then f satisfies the weak growth condition

$$\mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x)\|^2] \leq 2\rho(f(x) - f(x^*)),$$

with $\rho = 2\mathcal{L}$.

This corollary should be contrasted with Proposition 2 in (Vaswani et al., 2018) and Lemma 1 in (Nguyen et al., 2018), where it is shown, by assuming the f_i functions to be smooth and convex, that the weak growth condition holds with $\rho = 2L_{\max}$. However, as we will show in Lemma C.1, $L_{\max} \geq \mathcal{L}$, and hence our bound is often tighter.

3. Convergence Analysis

3.1. Main results

We now present our main theorem, and include its proof to highlight how we make use of expected smoothness and gradient noise.

[3] Or it is assumed that $\mathbb{E}\|\nabla f_v(x^k)\|^2 \leq c$ for all k iterates. But this too has issues since it implicitly assumes that the iterates remain within a compact set, and yet it it used to prove the convergence to within a compact set, raising issues of a circular argument.

Theorem 3.1. Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Choose $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]$ for all k. Then iterates of SGD given by (6) satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1-\gamma\mu)^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}. \quad (10)$$

Hence, given any $\epsilon > 0$, choosing stepsize

$$\gamma = \min\left\{\frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2}\right\}, \quad (11)$$

and

$$k \geq \max\left\{\frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2}\right\} \log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right), \quad (12)$$

implies $\mathbb{E}\|x^k - x^*\|^2 \leq \epsilon$.

Proof. Let $r^k = x^k - x^*$. From (6), we have

$$\begin{aligned}\|r^{k+1}\|^2 &\overset{(6)}{=} \|x^k - x^* - \gamma\nabla f_{v^k}(x^k)\|^2 \\ &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f_{v^k}(x^k)\rangle + \gamma^2\|\nabla f_{v^k}(x^k)\|^2.\end{aligned}$$

Taking expectation conditioned on x^k we obtain:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\|r^{k+1}\|^2 &\overset{(5)}{\leq} \|r^k\|^2 - 2\gamma\langle r^k, \nabla f(x^k)\rangle \\ &\quad + \gamma^2\mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2 \\ &\overset{(2)}{\leq} (1-\gamma\mu)\|r^k\|^2 - 2\gamma[f(x^k) - f(x^*)] \\ &\quad + \gamma^2\mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2.\end{aligned}$$

Taking expectations again and using Lemma 2.4:

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\overset{(9)}{\leq} (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2 \\ &\quad + 2\gamma(2\gamma\mathcal{L} - 1)\mathbb{E}[f(x^k) - f(x^*)] \\ &\leq (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2,\end{aligned}$$

where we used in the last inequality that $2\gamma\mathcal{L} \leq 1$ since $\gamma \leq \frac{1}{2\mathcal{L}}$. Recursively applying the above and summing up the resulting geometric series gives

$$\begin{aligned}\mathbb{E}\|r^k\|^2 &\leq (1-\gamma\mu)^k \|r^0\|^2 + 2\sum_{j=0}^{k-1}(1-\gamma\mu)^j \gamma^2\sigma^2 \\ &\leq (1-\gamma\mu)^k \|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu}. \quad (13)\end{aligned}$$

To obtain an iteration complexity result from the above, we use standard techniques as shown in Section A.1. □

Note that we do not assume f_i nor f to be convex. Theorem 3.1 states that SGD converges linearly up to the additive constant $2\gamma\sigma^2/\mu$ which depends on the gradient noise σ^2 and on the stepsize γ. We obtain a more accurate solution with a smaller stepsize, but then the convergence rate slows down. Since we control $\mathcal{D}$, we also control σ^2 and $\mathcal{L}$ (we compute these parameters for several distributions $\mathcal{D}$ in Section 3.3).

Furthermore, we can control this additive constant by carefully choosing the stepsize, as shown in the next result.

Theorem 3.2 (Decreasing stepsizes)**.** Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Let $\mathcal{K} := \mathcal{L}/\mu$ and

$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & \text{for} \quad k \leq 4\lceil\mathcal{K}\rceil \\ \frac{2k+1}{(k+1)^2\mu} & \text{for} \quad k > 4\lceil\mathcal{K}\rceil. \end{cases} \quad (14)$$

If $k \geq 4\lceil\mathcal{K}\rceil$, then SGD iterates given by (6) satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{k} + \frac{16\lceil\mathcal{K}\rceil^2}{e^2k^2}\|x^0 - x^*\|^2. \quad (15)$$

3.2. Choosing $\mathcal{D}$

For (6) to be efficient, the sampling vector v should be sparse. For this reason we will construct v so that only a (small and random) subset of its entries are non-zero.

Before we formally define v, let us first establish some random set terminology. Let $C \subseteq [n]$ and let $e_C := \sum_{i\in C} e_i$, where $\{e_1, \ldots, e_n\}$ are the standard basis vectors in $\mathbb{R}^n$. These subsets will be selected using a random set valued map S, in the literature referred to by the name *sampling* (Richtárik & Takáč, 2016; Qu & Richtárik, 2016). A sampling is uniquely characterized by choosing subset probabilities $p_C \geq 0$ for all subsets C of $[n]$:

$$\mathbb{P}[S = C] = p_C, \quad \forall C \subset [n], \quad (16)$$

where $\sum_{C\subseteq[n]} p_C = 1$. We will only consider *proper* samplings. A sampling S is called proper if $p_i \overset{\text{def}}{=} \mathbb{P}[i \in S] = \sum_{C: i\in C} p_C$ is positive for all i.

The first analysis of a randomized optimization method with an *arbitrary (proper) sampling* was performed by Richtárik & Takáč (2016) in the context of randomized coordinate descent for strongly convex functions. This arbitrary sampling paradigm was later adopted in many other settings, including accelerated coordinate descent for strongly convex functions (Hanzely & Richtárik, 2018), coordinate and accelerated descent for convex functions (Qu & Richtárik, 2016), primal-dual methods (Qu et al., 2015; Chambolle et al., 2018), variance-reduced methods with convex (Csiba & Richtárik, 2015) and nonconvex (Horváth & Richtárik, 2018) objectives. Arbitrary sampling arises as a special case of our more general analysis by specializing the sampling vector to one dependent on a sampling S. We now define practical sampling vector $v = v(S)$ as follows:

Lemma 3.3. Let S be a proper sampling, and let $\hat{\mathbf{P}} = \mathrm{Diag}(p_1, ..., p_n)$. Then the random vector $v = v(S)$ given

by

$$v = \hat{\mathbf{P}}^{-1} e_S \tag{17}$$

is a sampling vector.

Proof. Note that $v_i = \mathbf{1}_{(i\in S)}/p_i$, where $\mathbf{1}_{(i\in S)}$ is the indicator function of the event $i \in S$. It follows that $\mathbb{E}[v_i] = \mathbb{E}\left[\mathbf{1}_{(i\in S)}\right]/p_i = 1$. □

We can further specialize and define the following commonly used samplings. Each sampling S gives rise to a particular sampling vector $v = v(S)$ (i.e., distribution $\mathcal{D}$), which in turn gives rise to a particular stochastic reformulation (4) and SGD variant (6).

Independent sampling. The sampling S includes every i, independently, with probability $p_i > 0$. This type of sampling was considered in different contexts in (Horváth & Richtárik, 2018; Hanzely & Richtárik, 2018).

Partition sampling. A partition $\mathcal{G}$ of $[n]$ is a set consisting of subsets of $[n]$ such that $\cup_{C\in\mathcal{G}} C = [n]$ and $C_i \cap C_j = \emptyset$ for any $C_i, C_j \in \mathcal{G}$ with $i \neq j$. A partition sampling S is a sampling such that $p_C = \mathbb{P}[S = C] > 0$ for all $C \in \mathcal{G}$ and $\sum_{C\in\mathcal{G}} p_C = 1$.

Single element sampling. Only the singleton sets $\{i\}$ for $i = 1, \ldots, n$ have a non-zero probability of being sampled; that is, $\mathbb{P}[|S| = 1] = 1$. We have $\mathbb{P}[v(S) = e_i/p_i] = p_i$.

τ–nice sampling. We say that S is a τ–nice if S samples from all subsets of $[n]$ of cardinality τ uniformly at random. In this case we have that $p_i = \frac{\tau}{n}$ for all $i \in [n]$. So, $\mathbb{P}\left[v(S) = \frac{n}{\tau} e_C\right] = 1/\binom{n}{\tau}$ for all subsets $C \subseteq \{1, \ldots, n\}$ with τ elements.

3.3. Bounding $\mathcal{L}$ and σ^2

By assuming that the f_i functions are convex and smooth we can calculate closed form expressions for the expected smoothness $\mathcal{L}$ and gradient noise σ^2. In particular we make the following smoothness assumption:

Assumption 3.4. There exists a symmetric positive definite matrix $\mathbf{M}_i \in \mathbb{R}^{d\times d}$ such that

$$f_i(x+h) \leq f_i(x) + \langle \nabla f_i(x), h\rangle + \tfrac{1}{2}\|h\|^2_{\mathbf{M}_i}, \tag{18}$$

for all $x, h \in \mathbb{R}^d$, and $i \in [n]$, where $\|h\|^2_{\mathbf{M}_i} := \langle \mathbf{M}_i h, h\rangle$. In this case we say that f_i is $\mathbf{M}_i$–smooth. Furthermore, we assume that each f_i is convex.

To better relate the above assumption to the standard smoothness assumptions we make the following remark.

Remark 3.5. As a consequence of Assumption 3.4 we also have that each f_i is $L_i := \lambda_{\max}(\mathbf{M}_i)$–smooth and f is $L := \frac{1}{n}\lambda_{\max}(\sum_{i=1}^n \mathbf{M}_i)$–smooth. Let $L_{\max} := \max_{i\in[n]} L_i$.

Using Assumption 3.4 and a sampling we establish the following bounds on $\mathcal{L}$.

Theorem 3.6. Let S be a proper sampling, and $v = v(S)$ (i.e., v is defined by (17). Let f_i be $\mathbf{M}_i$-smooth, and $\mathbf{P} \in \mathbb{R}^{n\times n}$ be defined by $\mathbf{P}_{ij} = \mathbb{P}[i \in S \;\&\; j \in S]$. Then $(f, \mathcal{D}) \sim ES(\mathcal{L})$, where

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\max} := \max_{i\in[n]}\left\{\sum_{C:i\in C} \frac{p_C}{p_i} L_C\right\} \\ &\leq \frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]} \mathbf{P}_{ij}\frac{\lambda_{\max}(\mathbf{M}_j)}{p_i p_j}\right\}, \end{aligned} \tag{19}$$

and $L_C := \frac{1}{n}\lambda_{\max}(\sum_{j\in C}\frac{1}{p_j}\mathbf{M}_j)$. If $|S| \equiv \tau$, then

$$L \;\leq\; \mathcal{L}_{\max} \;\leq\; L_{\max} = \max_{i\in[n]} \lambda_{\max}(\mathbf{M}_i). \tag{20}$$

By applying the above result to specific samplings, we obtain the following practical bounds on $\mathcal{L}$:

Proposition 3.7. (i) For single element sampling S, we have

$$\mathcal{L}_{\max} = \frac{1}{n}\max_{i\in[n]} \frac{\lambda_{\max}(\mathbf{M}_i)}{p_i}. \tag{21}$$

(ii) For independent sampling S, we have

$$\mathcal{L}_{\max} \leq \sum_i \frac{\lambda_{\max}(\mathbf{M}_i)}{n} + \max_{i\in[n]} \frac{1-p_i}{p_i}\frac{\lambda_{\max}(\mathbf{M}_i)}{n}. \tag{22}$$

(iii) For τ-nice sampling S, we have

$$\begin{aligned}\mathcal{L}_{\max} &= \frac{1}{\tau\binom{n-1}{\tau-1}} \max_i \sum_{C:i\in C} \lambda_{\max}\Big(\sum_{j\in C}\mathbf{M}_j\Big) &(23)\\ &\leq \frac{1}{\tau}\frac{\tau-1}{n-1}\sum_{j\in[n]}\lambda_{\max}(\mathbf{M}_j) \\ &\quad + \frac{1}{\tau}\left(1 - \frac{\tau-1}{n-1}\right)\max_i \lambda_{\max}(\mathbf{M}_i). &(24)\end{aligned}$$

(iv) For partition sampling S with partition $\mathcal{G}$, we have

$$\mathcal{L}_{\max} = \frac{1}{n}\max_{C\in\mathcal{G}}\left\{\frac{1}{p_C}\lambda_{\max}\Big(\sum_{j\in C}\mathbf{M}_j\Big)\right\}. \tag{25}$$

For $v = v(S)$, formulas for the gradient noise σ^2 are provided in the next result:

Theorem 3.8. Let $h_i = \nabla f_i(x^*)$. Then

$$\sigma^2 = \frac{1}{n^2}\sum_{i,j\in[n]} \frac{\mathbf{P}_{ij}}{p_i p_j}\langle h_i, h_j\rangle. \tag{26}$$

Specializing the above theorem to specific samplings S gives the following formulas for σ^2:

Proposition 3.9. (i) For single element sampling S, we have

$$\sigma^2 = \frac{1}{n^2}\sum_{i\in[n]}\frac{1}{p_i}\|h_i\|^2. \tag{27}$$

(ii) For independent sampling S with $\mathbb{E}[|S|] = \tau$, we have

$$\sigma^2 = \frac{1}{n^2}\sum_{i\in[n]} \frac{1-p_i}{p_i}\|h_i\|^2. \quad (28)$$

(iii) For τ-nice sampling S, we have

$$\sigma^2 = \frac{1}{n\tau}\cdot\frac{n-\tau}{n-1}\sum_{i\in[n]} \|h_i\|^2. \quad (29)$$

(iv) For partition sampling S with partition $\mathcal{G}$, we have

$$\sigma^2 = \frac{1}{n^2}\sum_{C\in\mathcal{G}} \frac{1}{p_C}\|\sum_{i\in C} h_i\|^2. \quad (30)$$

Generally, we do not know the values of $h_i = \nabla f_i(x^*)$. But if we have prior knowledge that x^* belongs to some set $\mathcal{C}$, we can obtain upper bounds for σ^2 for these samplings from Proposition 3.9 in a straightforward way.

4. Optimal Mini-Batch Size

Here we develop the iteration complexity for different samplings by plugging in the bounds on $\mathcal{L}$ and σ given in Section 3.3 into Theorem 3.1. To keep the notation brief, in this section we drop the logarithmic term $\log\left(2\|x^0 - x^*\|^2/\epsilon\right)$ from the iteration complexity results. Furthermore, for brevity and to better compare our results to others in the literature, we will use $L_i = \lambda_{\max}(\mathbf{M}_i)$, $L_{\max} = \max_{i\in[n]} L_i$ and $\overline{L} = \frac{1}{n}\sum_{i\in[n]} L_i$ (see Remark 3.5). Finally let $\overline{h} = \frac{1}{n}\sum_{i\in[n]} \|h_i\|^2$ for brevity.

Gradient descent. As a first sanity check, we consider the case where $|S| = n$ with probability one. That is, each iteration (6) uses the full batch gradient. Thus $\sigma = 0$ and it is not hard to see that for $\tau = n$ in (23) we have $\mathcal{L}_{\max} = L$. Consequently, the resulting iteration complexity (12) is now $k \geq 2L/\mu$. This is exactly the rate of gradient descent, which is precisely what we would expect since the resulting method *is* gradient descent. Though an obvious sanity check, we believe this is the first convergence theorem of SGD that includes gradient descent as a special case. Clearly, this is a necessary pre-requisite if we are to hope to understand the complexity of mini-batching.

4.1. Nonzero gradient noise

To better appreciate how our iteration complexity evolves with increased mini-batch sizes, we now consider independent sampling with $|S| = \tau$ and τ-nice sampling.

Independent sampling. Inserting the bound on $\mathcal{L}$ (22) and σ (28) into (12) gives the following iteration complexity

$$k \geq \frac{2}{\mu n}\max\left\{\sum_{j\in[n]} L_j + \max_{i\in[n]} \frac{1-p_i}{p_i}L_i\ , \right.$$
$$\left. \frac{2}{\mu\epsilon}\frac{1-p_i}{p_i}\overline{h}\right\}. \quad (31)$$

This is a completely new mini-batch complexity result, which opens up the possibility of optimizing the mini-batch size and probabilities of sampling. For instance, if we fix uniform probabilities with $p_i = \frac{\tau}{n}$ then (31) becomes $k \geq \frac{2}{\mu}\max\{l(\tau), r(\tau)\}$, where

$$l(\tau) := \overline{L} + \left(\frac{1}{\tau} - \frac{1}{n}\right)L_{\max};\ r(\tau) := \frac{2}{\mu\epsilon}\left(\frac{1}{\tau} - \frac{1}{n}\right)\overline{h}. \quad (32)$$

This complexity result corresponds to using the stepsize

$$\gamma = \frac{1}{2}\min\left\{\frac{1}{l(\tau)}, \frac{1}{r(\tau)}\right\} \quad (33)$$

if $\tau < n$, otherwise only the left-hand-side term in the minimization remains. The stepsize (33) is increasing since both $l(\tau)$ and $r(\tau)$ decrease as τ increases.

With such a simple expression for the iteration complexity we can choose a mini-batch size that optimizes the *total complexity*. By defining the *total complexity* $T(\tau)$ as the number of iterations k times the number of gradient evaluations (τ) per iteration gives

$$T(\tau) := \frac{2}{\mu n}\max\left\{\tau n\overline{L} + (n-\tau)L_{\max}, \frac{2(n-\tau)\overline{h}}{\mu\epsilon}\right\}. \quad (34)$$

Minimizing $T(\tau)$ in τ is easy because $T(\tau)$ is a max of a linearly increasing term $\tau \times l(\tau)$ and a linearly decreasing term $\tau \times r(\tau)$ in τ. Furthermore $n \times l(n) \geq 0 = n \times r(n)$. Consequently, if $l(1) \geq r(1)$, then $\tau^* = 1$, otherwise

$$\tau^* = n\frac{\frac{2}{\mu\epsilon}\overline{h} - L_{\max}}{\frac{2}{\mu\epsilon}\overline{h} - L_{\max} + \sum_{j\in[n]} L_j}. \quad (35)$$

Since $l(1)$ is proportional to the noise and $1/\epsilon$ and $r(1)$ is proportional to the smoothness the condition $l(1) \leq r(1)$ holds when there is comparatively a lot of noise or the precision is high. As we will see in Section 4.2 this logic extends to the case where the noise is zero, where the optimal mini-batch size is $\tau^* = 1$.

τ–nice sampling. Inserting the bound on $\mathcal{L}$ (24) and σ (29) into (12) gives the iteration complexity $k \geq \frac{2}{\mu}\max\{l(\tau), r(\tau)\}$, where

$$l(\tau) = \frac{n(\tau-1)}{\tau(n-1)}\overline{L} + \frac{n-\tau}{\tau(n-1)}L_{\max}, \quad (36)$$

$$r(\tau) = \frac{2(n-\tau)}{\epsilon\mu(n-1)}\frac{\overline{h}}{\tau}, \quad (37)$$

which holds for the stepsize

$$\gamma = \frac{1}{2}\min\{\frac{1}{l(\tau)}, \frac{1}{r(\tau)}\} \quad (38)$$

Again, this is an increasing function of τ.

We are now again able to calculate the mini-batch size that optimizes the total complexity $T(\tau)$ given by $T(\tau) = \frac{2\tau}{\mu}\max\{l(\tau), r(\tau)\}$. Once again $T(\tau)$ is a max of a linearly increasing term $\tau \times l(\tau)$ and a linearly decreasing term $\tau \times r(\tau)$ in τ. Furthermore $r(n) = 0 \leq l(n)$. Consequently, if $r(1) \leq l(1)$ then $\tau^* = 1$, otherwise

$$\tau^* = n\frac{\overline{L} - L_{\max} + \frac{2}{\epsilon\mu}\cdot\overline{h}}{n\overline{L} - L_{\max} + \frac{2}{\epsilon\mu}\cdot\overline{h}}. \quad (39)$$

4.2. Zero gradient noise

Consider the case where the gradient noise is zero ($\sigma = 0$). According to Theorem 3.1, the resulting complexity of SGD with constant stepsize $\gamma = \frac{1}{2\mathcal{L}}$ is given by the very simple expression

$$k \geq \frac{2\mathcal{L}}{\mu}, \tag{40}$$

where we have dropped the logarithmic term $\log\left(\|x^0 - x^*\|^2/\epsilon\right)$. In this setting, due to Corollary 2.5, we know that f satisfies the weak growth condition. Thus our results are directly comparable to those developed in (Ma et al., 2018) and in (Vaswani et al., 2018).

In particular, Theorem 1 in (Ma et al., 2018) states that when running SGD with mini-batches based on sampling with replacement, the resulting iteration complexity is

$$k \geq \frac{L_{\max}}{\mu}\frac{1}{\tau} + \frac{L}{\mu}\frac{\tau-1}{\tau}, \tag{41}$$

again dropping the logarithmic term. Now gaining insight into the complexity (40) is a matter of studying the expected smoothness parameter $\mathcal{L}$ for different sampling strategies.

Independent sampling. Setting $\sigma = 0$ (thus $\overline{h} = 0$) and using uniform probabilities with $p_i = \frac{\tau}{n}$ in (31) gives

$$k \quad \geq \quad \frac{2\overline{L}}{\mu} + \left(\frac{1}{\tau} - \frac{1}{n}\right)\frac{2L_{\max}}{\mu}. \tag{42}$$

τ –nice sampling. If we use a uniform sampling and $\sigma = 0$ then the resulting iteration complexity is given by

$$k \quad \geq \quad \frac{n(\tau-1)}{\tau(n-1)}\frac{2\overline{L}}{\mu} + \frac{n-\tau}{\tau(n-1)}\frac{2L_{\max}}{\mu}. \tag{43}$$

Iteration complexities (41), (42) and (43) tell essentially the same story. Namely, the complexity improves as τ increases to n, but this improvement is not enough when considering the total complexity (multiplying by τ). Indeed, for total complexity, these results all say that $\tau = 1$ is optimal.

5. Importance Sampling

In this section we propose importance sampling for single element sampling and independent sampling with $\mathbb{E}[|S|] = \tau$, respectively. Due to lack of space, the details of this section are in the appendix, Section I. Again we drop the log term in (12) and adopt the notation in Remark 3.5.

5.1. Single element sampling

For single element sampling, plugging (21) and (27) into (12) gives the following iteration complexity

$$\frac{2}{\epsilon\mu^2}\max\left\{\frac{\epsilon\mu}{n}\max_{i\in[n]}\frac{L_i}{p_i}, \frac{2}{n^2}\sum_{i\in[n]}\frac{1}{p_i}\|h_i\|^2\right\},$$

where $0 < p_i \leq 1$ and $\sum_{i\in[n]} p_i = 1$. In order to optimize this iteration complexity over p_i, we need to solve a n dimensional linearly constrained nonsmooth convex minimization problem, which could be harder than the original problem (1). So instead, we will focus on minimizing $\mathcal{L}_{\max}$ and σ^2 over p_i seperately. We will then use these two resulting (sub)optimal probabilities to construct a sampling.

In particular, for single element sampling we can recover the *partially biased sampling* developed in (Needell et al., 2016). First, from (21) it is easy to see that the probabilities that minimize $\mathcal{L}_{\max}$ are $p_i^{\mathcal{L}} = L_i/\sum_{j\in[n]} L_j$, for all i. Using these suboptimal probabilities we can construct a partially biased sampling by letting $\hat{p}_i := \frac{1}{2}p_i^{\mathcal{L}} + \frac{1}{2n}$. Plugging this sampling in (21) gives $\mathcal{L}_{\max} \leq 2\overline{L}$, and from (27), we have $\sigma^2 \leq \frac{2}{n}\sum_{i\in[n]}\|h_i\|^2 := 2\overline{h}$. This sampling is the same as the partially biased sampling in (Needell et al., 2016). From (31) in Theorem 3.1, we get that the total complexity is now given by

$$k \geq \max\left\{\frac{4\overline{L}}{\alpha\mu}, \frac{8\overline{h}}{\epsilon\mu^2}\right\}. \tag{44}$$

For uniform sampling, $\mathcal{L}_{\max} = \max_{i\in[n]} L_i \geq \overline{L}$ and $\sigma^2 = \frac{1}{n}\sum_{i\in[n]}\|h_i\|^2$. Hence, compared to uniform sampling, the iteration complexity of partially biased sampling is at most two times larger, but could be $n/2$ smaller in the extreme case where $L_{\max} = n\,\overline{L}$.

5.2. Minibatches

Importance sampling for minibatches was first considered in (Csiba & Richtárik, 2018); but not in the context of SGD. Here we propose the first importance sampling for mini-batch SGD. In Section I.2 in the appendix we introduce the use of partially biased sampling together with independent sampling with $|S| = \tau$ and show that we can achieve a total complexity of (by Proposition I.3)

$$k \geq \max\left\{\left(1 - \frac{2}{\tau}\right)\frac{2\overline{L}}{\alpha\mu}, \left(\frac{2}{\tau} - \frac{1}{n}\right)\frac{8\overline{h}}{\epsilon\mu^2}\right\}, \tag{45}$$

which not only eliminates the dependence on $L_{\max}$, but also improves as the mini-batch size τ increases.

6. Experiments

In this section, we empirically validate our theoretical results. We perform three experiments in each of which we highlight a different aspect of our contributions.

In the first two experiments we focus on ridge regression and regularized logistic regression problems (problems with strongly convex objective f and components f_i) and we evaluate the performance of SGD on both synthetic and real data. In particular, in the first experiment (Section 6.1) we numerically verify the performance of SGD (in the case of uniform single element sampling) as predicted from Theorems 3.1 and 3.2 for both constant and decreasing step-sizes. In the second experiment (Section 6.2) we compare the convergence of SGD for several choices of the distribution $\mathcal{D}$

(different sampling strategies) as described in the previous sections. In the last experiment (Section 6.3) we focus on the problem of principal component analysis (PCA) which by construction can be seen as a problem with a strongly convex objective f but with non-convex functions f_i (Allen-Zhu & Yuan, 2016; Garber & Hazan, 2015; Shalev-Shwartz, 2016).

In all experiments, to evaluate SGD we use the relative error measure $\frac{\|x^k-x^*\|^2}{\|x^0-x^*\|^2}$. For all implementations, the starting point x^0 is standard Gaussian. We run each method until $\|x^k - x^*\|^2 \leq 10^{-3}$ or until a pre-specified maximum number of epochs is achieved. For the horizontal axis we always use the number of epochs. The code for the experiments is written in Python 3

For more experiments we refer the interested reader to Section J of the Appendix.

Regularized Regression Problems: In the case of the *ridge regression* problem we solve:

$$\min_x f(x) = \frac{1}{2n}\sum_{i=1}^n (\mathbf{A}[i,:]x - y_i)^2 + \frac{\lambda}{2}\|x\|^2,$$

while for the *L2-regularized logistic regression* problem we solve:

$$\min_x f(x) = \frac{1}{2n}\sum_{i=1}^n \log\left(1+\exp(-y_i\mathbf{A}[i,:]x)\right)+\frac{\lambda}{2}\|x\|^2.$$

In both problems $\mathbf{A} \in \mathbb{R}^{n\times d}, y \in \mathbb{R}^n$ are the given data and $\lambda > 0$ is the regularization parameter. For the generation of the synthetic data in both problems, the rows of matrix $\mathbf{A}$ ($\mathbf{A}[i,:]$) were sampled from the standard Gaussian distribution $\mathcal{N}(0,1)$. For the synthetic data in the case of ridge regression we choose vector y to be Gaussian vector while in the case of logistic regression $y \in \{-1,1\}^n$ where $\mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = \frac{1}{2}$ The regularization parameter λ varies depending on the experiment. For our experiments on real data we choose several LIBSVM (Chang & Lin, 2011) datasets.

6.1. Constant vs decreasing step size

We now compare the performance of SGD in the constant and decreasing stepsize regimes considered in Theorems 3.1 (see (11)) and 3.2 (see (14)), respectively. As expected from theory, we see in Figure 1 that the decreasing stepsize regime is vastly superior at reaching a higher precision than the constant step-size variant. In our plots, the vertical red line denotes the value of $4\lceil \mathcal{L}/\mu \rceil$ predicted from Theorem 3.2 and highlights the point where SGD needs to change its update rule from constant to decreasing step-size.

6.2. Minibatches

In Figures 2 and 5 we compare the single element sampling (uniform and importance), τ independent sampling

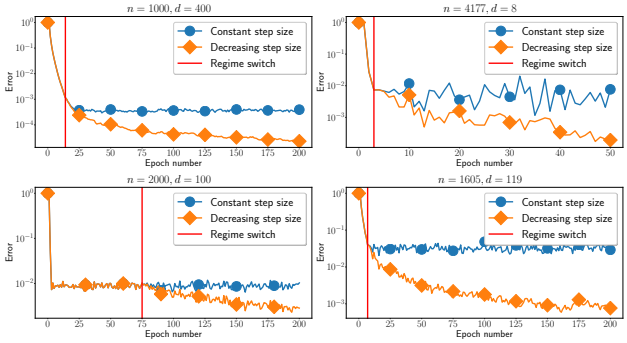

Figure 1. Comparison between constant and decreasing step size regimes of SGD. Ridge regression problem (first row): on left - synthetic data, on right - real dataset: abalone from LIBSVM. Logistic regression problem(second row): on left - synthetic data, on right - real data-set: a1a from LIBSVM. In all experiments $\lambda = 1/n$.

(uniform, uniform with optimal batch size and importance) and τ nice sampling (with some τ and with optimal τ^*). The probabilities of importance samplings in the single element sampling and τ independent sampling are calculated by formulas (65) and (73) (see the Appendix). Formulas for optimal minibatch size τ^* in independent sampling and τ-nice samplings are given in (35) and (39), respectively. Observe that minibatching with optimal τ^* gives the best convergence. In addition, note that for constant step size, importance sampling variants depend on the accuracy ϵ. It is clear that before the error approaches required accuracy, importance sampling is comparable or better than their uniform sampling.

6.3. Sum-of-non-convex functions

In Figure 3, our goal is to illustrate that Theorem 3.1 holds even if the functions f_i are non convex. The scheme of the experiment is similar to the one from (Allen-Zhu & Yuan, 2016). In particular, we first generate random vectors $a_1, \ldots, a_n, b \in \mathbb{R}^d$ from $\mathcal{U}(0,10)$ $\mathbf{A} := \frac{1}{n}\sum_{i=1}^n a_i a_i^\top$. Then we consider the minimization problem:

$$\min_x f(x) = \frac{1}{2n}\sum_{i=1}^n x^\top(a_i a_i^\top + D_i)x + b^\top x,$$

where D_i, $i \in [n]$ are diagonal matrices satisfying $D := D_1 + \cdots + D_n = 0$. In particular, to guarantee that $D = 0$, we randomly select half of the matrices and assign their j-th diagonal value $(D_i)_{jj}$ equal to 11; for the other half we assign $(D_i)_{jj}$ to be -11. We repeat that for all diagonal values. Note that under this construction, each f_i is non-convex function. Once again, in the first plot we observe that while both are equally fast in the beginning, the decreasing stepsize variant is better at reaching higher accuracy than the fixed stepsize variant. In the second plot we see, as expected, that all four minibatch versions of SGD outperform single

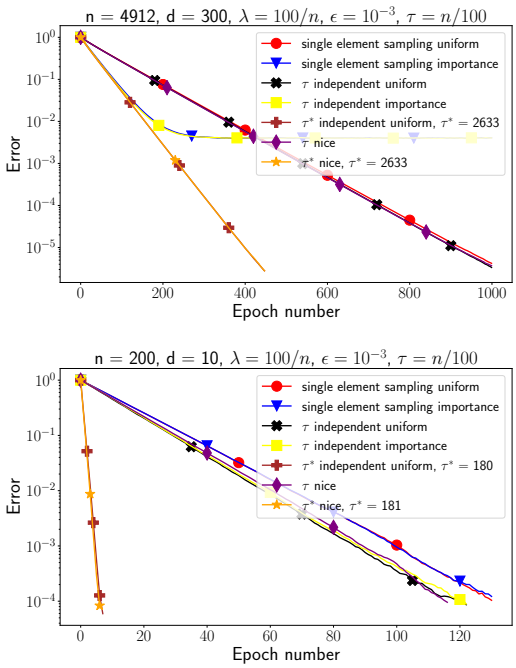

Figure 2. Performance of SGD with several minibatch strategies for logistic regression. On the left: the real data-set w3a from LIBSVM. On the right: synthetic data.

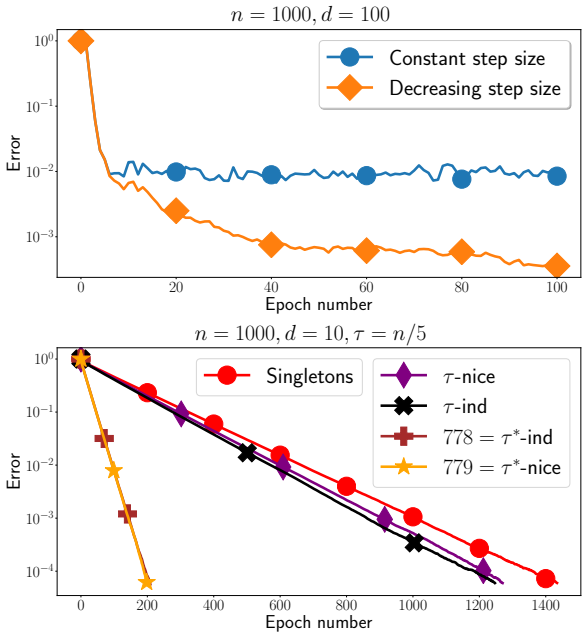

Figure 3. On the left: Comparison between constant and decreasing step size regimes of SGD for PCA. On the right: comparison of different sampling strategies of SGD for PCA.

element SGD. However, while the τ-nice and τ-independent samplings with $\tau = n/5$ lead to a slight improvement only, the theoretically optimal choice $\tau = \tau^*$ leads to a vast improvement.

References

Allen-Zhu, Z. and Yuan, Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International Conference on Machine Learning*, pp. 1080–1089, 2016.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. *arXiv preprint arXiv:1811.10792*, 2018.

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

Cevher, V. and Vu, B. C. On the linear convergence of the stochastic gradient method with constant step-size. *arXiv:1712.01906*, pp. 1–9, 2017.

Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schöenlieb, C.-B. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Chee, J. and Toulis, P. Convergence diagnostics for stochastic gradient descent with constant learning rate. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1476–1485. PMLR, 09–11 Apr 2018.

Csiba, D. and Richtárik, P. Primal method for ERM with flexible mini-batching schemes and non-convex losses. *arXiv:1506.02227*, 2015.

Csiba, D. and Richtárik, P. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27):1–21, 2018.

Garber, D. and Hazan, E. Fast and simple PCA via convex optimization. *arXiv preprint arXiv:1509.05647*, 2015.

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arxiv:1805.02632*, 2018.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

Hanzely, F. and Richtárik, P. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. *arXiv Preprint arXiv: 1809.09354*, 2018.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: stability of stochastic gradient descent. In *33rd International Conference on Machine Learning*, 2016.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

Horváth, S. and Richtárik, P. Nonconvex variance reduced optimization with arbitrary sampling. *arXiv:1809.04146*, 2018.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 5330–5340, 2017.

Loizou, N. and Richtárik, P. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pp. 3331–3340, 2018.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pp. 1–39, 2018. doi: https://doi.org/10.1007/s10107-018-1232-1.

Needell, D. and Ward, R. Batched stochastic gradient descent with weighted sampling. In *Approximation Theory XV, Springer*, volume 204 of *Springer Proceedings in Mathematics & Statistics,*, pp. 279 – 306, 2017.

Needell, D., Srebro, N., and Ward, R. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming, Series A*, 155(1):549–573, 2016.

Nemirovski, A. and Yudin, D. B. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathetmatics Doklady*, 19, 1978.

Nemirovski, A. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley Interscience, 1983.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. SGD and hogwild! Convergence without the bounded gradients assumption. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3750–3758. PMLR, 2018.

Qu, Z. and Richtárik, P. Coordinate descent with arbitrary sampling I: Algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016.

Qu, Z. and Richtárik, P. Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884, 2016.

Qu, Z., Richtárik, P., and Zhang, T. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Advances in Neural Information Processing Systems*, pp. 865–873, 2015.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *29th International Conference on Machine Learning*, volume 12, pp. 1571–1578, 2012.

Recht, B., Re, C., Wright, S., and Niu, F. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 693–701, 2011.

Richtárik, P. and Takáč, M. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, 2016.

Richtárik, P. and Takáč, M. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156 (1-2):433–484, 2016.

Richtárik, P. and Takáč, M. Stochastic reformulations of linear systems: algorithms and convergence theory. *arXiv:1706.01108*, 2017.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Shalev-Shwartz, S. SDCA without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pp. 747–754, 2016.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: primal estimated subgradient solver for SVM. In *24th International Conference on Machine Learning*, pp. 807–814, 2007.

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 71–79, 2013.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

APPENDIX
SGD: General Analysis and Improved Rates

A. Elementary Results

In this section we collect some elementary results; some of them we use repeatedly.

Proposition A.1. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be L_ϕ–smooth, and assume it has a minimizer x^* on $\mathbb{R}^d$. Then

$$\|\nabla\phi(x) - \nabla\phi(x^*)\|^2 \leq 2L_\phi(\phi(x) - \phi(x^*)).$$

Proof. Lipschitz continuity of the gradient implies that

$$\phi(x+h) \leq \phi(x) + \langle\nabla\phi(x), h\rangle + \frac{L_\phi}{2}\|h\|^2.$$

Now plugging $h = -\frac{1}{L_\phi}\nabla\phi(x)$ into the above inequality, we get $\frac{1}{2L_\phi}\|\nabla\phi(x)\|^2 \leq \phi(x) - \phi(x+h) \leq \phi(x) - \phi(x^*)$. It remains to note that $\nabla\phi(x^*) = 0$. □

In this section we summarize some elementary results which we use often in our proofs. We do not claim novelty; we but we include them for completeness and clarity.

Lemma A.2 (Double counting)**.** Let $a_{i,C} \in \mathbb{R}$ for $i = 1, \ldots, n$ and $C \in \mathcal{C}$, where $\mathcal{C}$ is some collection of subsets of $[n]$. Then

$$\sum_{C\in\mathcal{C}}\sum_{i\in C} a_{i,C} \quad = \quad \sum_{i=1}^{n}\sum_{C\in\mathcal{C}\,:\,i\in C} a_{i,C}. \tag{46}$$

Lemma A.3 (Complexity bounds)**.** Let $E > 0$, $0 < \rho \leq 1$ and $0 \leq c < 1$. If $k \in \mathbb{N}$ satisfies

$$k \geq \frac{1}{1-\rho}\log\left(\frac{E}{1-c}\right), \tag{47}$$

then

$$\rho^k \leq (1-c)E. \tag{48}$$

Proof. Taking logarithms and rearranging (48) gives

$$\log\left(\frac{E}{1-c}\right) \leq k\log\left(\frac{1}{\rho}\right). \tag{49}$$

Now using that $\log\left(\frac{1}{\rho}\right) \geq 1 - \rho$, for $0 < \rho \leq 1$ gives (47). □

A.1. The iteration complexity (12) of Theorem 3.1

To analyse the iteration complexity, let $\epsilon > 0$ and choosing the stepsize so that $\frac{2\gamma\sigma^2}{\mu} \leq \frac{1}{2}\epsilon$, gives (11). Next we choose k so that

$$(1-\gamma\mu)^k\|r^0\|^2 \leq \frac{1}{2}\epsilon.$$

Taking logarithms and re-arranging the above gives

$$\log\left(\frac{2\|r^0\|^2}{\epsilon}\right) \leq k\log\left(\frac{1}{1-\gamma\mu}\right). \tag{50}$$

Now using that $\log\left(\frac{1}{\rho}\right) \geq 1-\rho$, for $0<\rho\leq 1$ gives

$$\begin{aligned} k &\geq \frac{1}{\gamma\mu}\log\left(\frac{2\|r^0\|^2}{\epsilon}\right) \\ &\overset{(11)}{=} \frac{1}{\mu}\max\left\{2\mathcal{L},\ \frac{4\sigma^2}{\epsilon\mu}\right\}\log\left(\frac{2\|r^0\|^2}{\epsilon}\right). \end{aligned} \tag{51}$$

Which concludes the proof.

B. Proof of Lemma 2.4

For brevity, let us write $\mathbb{E}[\cdot]$ instead of $\mathbb{E}_{\mathcal{D}}[\cdot]$. Then

$$\begin{aligned} \mathbb{E}\|\nabla f_v(x)\|^2 &= \mathbb{E}\|\nabla f_v(x)-\nabla f_v(x^*)+\nabla f_v(x^*)\|^2 \\ &\leq 2\mathbb{E}\|\nabla f_v(x)-\nabla f_v(x^*)\|^2+2\mathbb{E}\|\nabla f_v(x^*)\|^2 \\ &\leq 4\mathcal{L}[f(x)-f(x^*)]+2\mathbb{E}\|\nabla f_v(x^*)\|^2. \end{aligned}$$

The first inequality follows from the estimate $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$, and the second inequality follows from (7).

C. Bounds on the Expected Smoothness Constant $\mathcal{L}$

Below we establish some lower and upper bounds on the expected smoothness constant $\mathcal{L}=\mathcal{L}_{\max}$. These bounds were referred to in the main paper in Section 2.3.

Lemma C.1. Assume that there exists $\tau\in[n]$ such that $|S|=\tau$ with probability 1. Let

$$\bar{\mathcal{L}}_S := \frac{1}{|S|}\sum_{i\in S}\mathcal{L}_i.$$

Then $\mathbb{E}\left[\bar{\mathcal{L}}_S\right]=\mathbb{E}\left[L_S\right]$. Moreover,

$$L\leq \mathbb{E}\left[\bar{\mathcal{L}}_S\right]\leq \mathcal{L}_{\max}\leq L_{\max}. \tag{52}$$

Proof. Define $\mathbf{M}_S := \frac{1}{n}\sum_{i\in S}\frac{\mathbf{M}_i}{p_i}$ and note that f is $\frac{1}{n}\sum_{i\in[n]}\mathbf{M}_i$–smooth Furthermore

$$\mathbb{E}\left[\mathbf{M}_S\right]=\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n\frac{\mathbf{M}_i}{p_i}\mathbf{1}_{(i\in S)}\right]=\frac{1}{n}\sum_{i=1}^n\frac{\mathbf{M}_i}{p_i}\mathbb{E}\left[\mathbf{1}_{(i\in S)}\right]=\frac{1}{n}\sum_{i\in[n]}\mathbf{M}_i.$$

We will now establish the inequalities in (52) starting from left to the right.

(Part I $L\leq\mathbb{E}\left[L_S\right]$). Recalling that $L_S=\lambda_{\max}(\mathbf{M}_S)$ and by Jensen's inequality,

$$L=\lambda_{\max}\left(\mathbb{E}\left[\mathbf{M}_S\right]\right)\leq\mathbb{E}\left[\lambda_{\max}(\mathbf{M}_S)\right]=\mathbb{E}\left[L_S\right].$$

Furthermore

$$\begin{aligned} \mathbb{E}\left[\bar{\mathcal{L}}_S\right] &= \mathbb{E}\left[\frac{1}{\tau}\sum_{i\in S}\mathcal{L}_i\right]=\frac{1}{\tau}\sum_i p_i\mathcal{L}_i \\ &\overset{(60)}{=} \frac{1}{\tau}\sum_i\sum_{C\,:\,i\in C}p_C L_i \overset{\text{Lemma } A.2}{=} \frac{1}{\tau}\sum_C\sum_{i\in C}p_C L_C \\ &= \frac{1}{\tau}\sum_C|C|p_C L_C=\sum_C p_C L_C=\mathbb{E}\left[L_S\right] \end{aligned}$$

(Part II $\mathbb{E}\left[\bar{L}_S\right] \leq \mathcal{L}_{\max}$). We have that

$$\bar{L}_S = \frac{1}{|S|}\sum_{i\in S}\mathcal{L}_i \leq \frac{1}{|S|}\sum_{i\in S}\max_{i\in[n]}\mathcal{L}_i = \mathcal{L}_{\max}.$$

(Part III $\mathcal{L}_{\max} \leq L_{\max}$). Finally, since

$$L_C \leq \frac{1}{\tau}\sum_{j\in C} L_j \leq L_{\max}, \tag{53}$$

we have that

$$\mathcal{L}_i \overset{(60)+(53)}{\leq} \sum_{C\,:\,i\in C}\frac{p_C}{p_i}\frac{1}{\tau}\sum_{j\in C}L_j \overset{(53)}{\leq} \sum_{C\,:\,i\in C}\frac{p_C}{p_i}L_{\max} = L_{\max}.$$

Consequently taking the maximum over $i\in[n]$ in the above gives $\mathcal{L}_{\max}\leq L_{\max}$. □

D. Proof of Theorem 3.2

Proof. Let $\gamma_k := \frac{2k+1}{(k+1)^2\mu}$ and let k^* be an integer that satisfies $\gamma_{k^*} \leq \frac{1}{2\mathcal{L}}$. In particular this holds for

$$k^* \geq \lceil 4\mathcal{K} - 1\rceil.$$

Note that γ_k is decreasing in k and consequently $\gamma_k \leq \frac{1}{2\mathcal{L}}$ for all $k\geq k^*$. This in turn guarantees that (13) holds for all $k\geq k^*$ with γ_k in place of γ, that is

$$\mathbb{E}\|r^{k+1}\|^2 \leq \frac{k^2}{(k+1)^2}\mathbb{E}\|r^k\|^2 + \frac{2\sigma^2}{\mu^2}\frac{(2k+1)^2}{(k+1)^4}. \tag{54}$$

Multiplying both sides by $(k+1)^2$ we obtain

$$\begin{aligned}(k+1)^2\mathbb{E}\|r^{k+1}\|^2 &\leq k^2\mathbb{E}\|r^k\|^2 + \frac{2\sigma^2}{\mu^2}\left(\frac{2k+1}{k+1}\right)^2\\ &\leq k^2\mathbb{E}\|r^k\|^2 + \frac{8\sigma^2}{\mu^2},\end{aligned}$$

where the second inequality holds because $\frac{2k+1}{k+1} < 2$. Rearranging and summing from $t = k^*\ldots k$ we obtain:

$$\sum_{t=k^*}^{k}\left[(t+1)^2\mathbb{E}\|r^{t+1}\|^2 - t^2\mathbb{E}\|r^t\|^2\right] \leq \sum_{t=k^*}^{k}\frac{8\sigma^2}{\mu^2}. \tag{55}$$

Using telescopic cancellation gives

$$(k+1)^2\mathbb{E}\|r^{k+1}\|^2 \leq (k^*)^2\mathbb{E}\|r^{k^*}\|^2 + \frac{8\sigma^2(k-k^*)}{\mu^2}.$$

Dividing the above by $(k+1)^2$ gives

$$\mathbb{E}\|r^{k+1}\|^2 \leq \frac{(k^*)^2}{(k+1)^2}\mathbb{E}\|r^{k^*}\|^2 + \frac{8\sigma^2(k-k^*)}{\mu^2(k+1)^2}. \tag{56}$$

For $k\leq k^*$ we have that (13) holds, which combined with (56), gives

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\leq \frac{(k^*)^2}{(k+1)^2}\left(1-\frac{\mu}{2\mathcal{L}}\right)^{k^*}\|r^0\|^2\\ &+ \frac{\sigma^2}{\mu^2(k+1)^2}\left(8(k-k^*) + \frac{(k^*)^2}{\mathcal{K}}\right).\end{aligned} \tag{57}$$

Choosing k^* that minimizes the second line of the above gives $k^* = 4\lceil\mathcal{K}\rceil$, which when inserted into (57) becomes

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\leq \frac{16\lceil\mathcal{K}\rceil^2}{(k+1)^2}\left(1-\frac{1}{2\mathcal{K}}\right)^{4\lceil\mathcal{K}\rceil}\|r^0\|^2 \\ &\quad + \frac{\sigma^2}{\mu^2}\frac{8(k-2\lceil\mathcal{K}\rceil)}{(k+1)^2} \\ &\leq \frac{16\lceil\mathcal{K}\rceil^2}{e^2(k+1)^2}\|r^0\|^2 + \frac{\sigma^2}{\mu^2}\frac{8}{k+1}, \end{aligned} \tag{58}$$

where we have used that $\left(1-\frac{1}{2x}\right)^{4x} \leq e^{-2}$ for all $x \geq 1$.

□

E. Proof of Theorem 3.6

Proof. Since $v_i = v_i(S) = \mathbf{1}_{(i\in S)}\frac{1}{p_i}$. and since f_i is $\mathbf{M}_i$-smooth, the function

$$f_v(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)v_i = \frac{1}{n}\sum_{i\in S}\frac{f_i(x)}{p_i}, \tag{59}$$

is L_S–smooth where

$$L_S := \frac{1}{n}\lambda_{\max}\left(\sum_{i\in S}\frac{\mathbf{M}_i}{p_i}\right).$$

We also define the following smoothness related quantities

$$\mathcal{L}_i := \sum_{C\,:\,i\in C}\frac{p_C}{p_i}L_C, \quad \mathcal{L}_{\max} := \max_i \mathcal{L}_i, \text{ and;} \quad L_{\max} = \max_{i\in[n]}\lambda_{\max}(\mathbf{M}_i). \tag{60}$$

Since the f_i's are convex and the sampling vector $v \in \mathbb{R}^d_+$ has positive elements, each realization of f_v is convex, and it follows from Lemma 8.1 in (Needell et al., 2016) that

$$\|\nabla f_v(x) - \nabla f_v(y)\|^2 \leq 2L_S\left(f_v(x) - f_v(y) - \langle\nabla f_v(y), x-y\rangle\right). \tag{61}$$

Taking expectation in (61) gives

$$\begin{aligned}\mathbb{E}[\|\nabla f_v(x) - \nabla f_v(y)\|^2] &\leq 2\sum_C p_C L_C\left(f_{v(C)}(x) - f_{v(C)}(y) - \langle\nabla f_{v(C)}(y), x-y\rangle\right) \\ &\overset{(59)}{=} 2\sum_C p_C L_C\sum_{i\in C}\frac{1}{np_i}\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x-y\rangle\right) \\ &\overset{\text{Lemma A.2}}{=} \frac{2}{n}\sum_{i=1}^n\sum_{C:i\in C} p_C\frac{1}{p_i}L_C\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x-y\rangle\right) \\ &\overset{(19)}{\leq} \frac{2}{n}\sum_{i=1}^n\mathcal{L}_{\max}\left(f_i(x) - f_i(y) - \langle\nabla f_i(y), x-y\rangle\right) \\ &= 2\mathcal{L}_{\max}\left(f(x) - f(y) - \langle\nabla f(y), x-y\rangle\right). \end{aligned}$$

Furthermore, for each i,

$$\begin{aligned}\mathcal{L}_i = \sum_{C:i\in C}\frac{p_C}{p_i}L_C \quad &= \quad \frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\lambda_{\max}\left(\sum_{j\in C}\frac{\mathbf{M}_j}{p_j}\right) \\ &\leq \quad \frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\sum_{j\in C}\frac{\lambda_{\max}(\mathbf{M}_j)}{p_j} \\ &\overset{\text{Lemma } A.2}{=} \quad \frac{1}{n}\sum_{j=1}^{n}\sum_{C:i\in C \;\&\; j\in C}\frac{p_C}{p_ip_j}\lambda_{\max}(\mathbf{M}_j) \\ &= \quad \frac{1}{n}\sum_{j=1}^{n}\frac{\mathbf{P}_{ij}}{p_ip_j}\lambda_{\max}(\mathbf{M}_j).\end{aligned} \tag{62}$$

Hence,

$$\mathcal{L}_{\max} \leq \frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\mathbf{P}_{ij}\frac{\lambda_{\max}(\mathbf{M}_j)}{p_ip_j}\right\}. \tag{63}$$

Let $y = x^*$ and notice that $\nabla f(x^*) = 0$, which gives (19). We prove (20) in the following slightly more comprehensive Lemma. □

F. Proof of Proposition 3.7

Proof. First note that by combining (19) and (62) we have that

$$\begin{aligned}\mathcal{L}_{\max} \quad &\overset{(19)}{=} \quad \max_{i\in[n]}\left\{\sum_{C:i\in C}\frac{p_C}{p_i}L_C\right\} \\ &\overset{(62)}{=} \quad \max_{i\in[n]}\left\{\frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\lambda_{\max}\left(\sum_{j\in C}\frac{\mathbf{M}_j}{p_j}\right)\right\}.\end{aligned} \tag{64}$$

(i) By straight forward calculation from (64) and using that each set C is a singleton.

(ii) By the upper bound in (19), and noticing $\mathbf{P}_{ij} = p_ip_j$ for $i \neq j$, we have

$$\begin{aligned}\mathcal{L}_{\max} \quad &\leq \quad \frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\mathbf{P}_{ij}\frac{\lambda_{\max}(\mathbf{M}_j)}{p_ip_j}\right\} \\ &= \quad \frac{1}{n}\max_{i\in[n]}\left\{\sum_{j\in[n]}\lambda_{\max}(\mathbf{M}_j) + \left(\frac{1}{p_i}-1\right)\lambda_{\max}(\mathbf{M}_i)\right\} \\ &= \quad \frac{1}{n}\left(\sum_{j\in[n]}\lambda_{\max}(\mathbf{M}_j) + \max_{i\in[n]}\left(\frac{1}{p_i}-1\right)\lambda_{\max}(\mathbf{M}_i)\right).\end{aligned}$$

(iii) Noticing that for τ-nice sampling S, $p_C = \frac{1}{\binom{n}{\tau}}$ for all $C \in \text{supp}\,(S)$ and $p_i = \frac{\tau}{n}$, we have

$$\begin{aligned}
\mathcal{L}_{\max} &\overset{(64)}{=} \max_{i\in[n]}\left\{\frac{1}{n}\sum_{C:i\in C}\frac{p_C}{p_i}\lambda_{\max}\left(\sum_{j\in C}\frac{\mathbf{M}_j}{p_j}\right)\right\} \\
&= \max_{i\in[n]}\left\{\frac{1}{n}\sum_{C:i\in C}\frac{n^2}{\tau^2\binom{n}{\tau}}\lambda_{\max}\left(\sum_{j\in C}\mathbf{M}_j\right)\right\} \\
&= \frac{1}{\tau\binom{n-1}{\tau-1}}\max_{i\in[n]}\left\{\sum_{C:i\in C}\lambda_{\max}\left(\sum_{j\in C}\mathbf{M}_j\right)\right\},
\end{aligned}$$

where we used the identity $\frac{\tau}{n}\binom{n}{\tau} = \binom{n-1}{\tau-1}$. Furthermore, for $\tau > 1$ since $\mathbf{P}_{ij} = \frac{\tau(\tau-1)}{n(n-1)}$ if $i \neq j$ and $\mathbf{P}_{ii} = \frac{\tau}{n}$ from (63) we have that

$$\begin{aligned}
\mathcal{L}_{\max} &\overset{(63)}{\leq} \frac{1}{n}\frac{n^2}{\tau^2}\max_{i\in[n]}\left\{\sum_{j\neq i}\mathbf{P}_{ij}\lambda_{\max}(\mathbf{M}_j) + \mathbf{P}_{ii}\lambda_{\max}(\mathbf{M}_i)\right\} \\
&= \frac{1}{n}\frac{n^2}{\tau^2}\max_{i\in[n]}\left\{\sum_{j\neq i}\frac{\tau(\tau-1)}{n(n-1)}\lambda_{\max}(\mathbf{M}_j) + \frac{\tau}{n}\lambda_{\max}(\mathbf{M}_i)\right\} \\
&= \frac{1}{\tau}\max_{i\in[n]}\left\{\sum_{j\neq i}\frac{\tau-1}{n-1}\lambda_{\max}(\mathbf{M}_j) + \lambda_{\max}(\mathbf{M}_i)\right\} \\
&= \frac{1}{\tau}\max_{i\in[n]}\left\{\sum_{j\in[n]}\frac{\tau-1}{n-1}\lambda_{\max}(\mathbf{M}_j) + \left(1-\frac{\tau-1}{n-1}\right)\lambda_{\max}(\mathbf{M}_i)\right\} \\
&= \frac{1}{\tau}\frac{\tau-1}{n-1}\sum_{j\in[n]}\lambda_{\max}(\mathbf{M}_j) + \frac{1}{\tau}\left(1-\frac{\tau-1}{n-1}\right)\max_{i\in[n]}\lambda_{\max}(\mathbf{M}_i).
\end{aligned}$$

(iv) For every partition sampling we have that $p_i = p_C$ if $i \in C$, hence

$$\begin{aligned}
\mathcal{L}_{\max} &\overset{(64)}{=} \max_{i\in[n]}\left\{\frac{1}{n}\sum_{C:i\in C}\frac{p_i}{p_i}\lambda_{\max}\left(\sum_{j\in C}\frac{\mathbf{M}_j}{p_C}\right)\right\} \\
&\overset{(62)}{=} \frac{1}{n}\max_{i\in[n]}\left\{\sum_{C:i\in C}\frac{1}{p_C}\lambda_{\max}\left(\sum_{j\in C}\mathbf{M}_j\right)\right\} \\
&= \frac{1}{n}\max_{C\in\mathcal{G}}\left\{\frac{1}{p_C}\lambda_{\max}\left(\sum_{j\in C}\mathbf{M}_j\right)\right\}.
\end{aligned}$$

□

G. Proof of Theorem 3.8

Proof.

$$\begin{aligned}\sigma^2 = \mathbb{E}[\|\nabla f_v(x^*)\|^2] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*)v_i\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n \nabla f_i(x^*)v_i\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i\in S}\frac{1}{p_i}h_i\right\|^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[\left\|\sum_{i=1}^n 1_{i\in S}\frac{1}{p_i}h_i\right\|^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n 1_{i\in S}1_{j\in S}\langle\frac{1}{p_i}h_i,\frac{1}{p_j}h_j\rangle\right] \\ &= \frac{1}{n^2}\sum_{i,j}\frac{\mathbf{P}_{ij}}{p_ip_j}\langle h_i,h_j\rangle.\end{aligned}$$

□

H. Proof of Proposition 3.9

Proof. (i) By straight calculation from (26).

(ii) For independent sampling S, $\mathbf{P}_{ij} = p_ip_j$ for $i \neq j$, hence,

$$\begin{aligned}\sigma^2 &= \frac{1}{n^2}\sum_{i,j\in[n]}\frac{\mathbf{P}_{ij}}{p_ip_j}\langle h_i,h_j\rangle = \frac{1}{n^2}\sum_{i,j\in[n]}\langle h_i,h_j\rangle + \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2 \\ &= \frac{1}{n^2}\|\nabla f(x^*)\|^2 + \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2 = \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i}-1\right)\|h_i\|^2.\end{aligned}$$

(iii) For τ-nice sampling S, if $\tau = 1$, it is obvious. If $\tau \geq 1$, then $\mathbf{P}_{ij} = \frac{C_{n-2}^{\tau-2}}{C_n^\tau}$ for $i \neq j$, and $p_i = \frac{\tau}{n}$ for all i. Hence,

$$\begin{aligned}\sigma^2 &= \frac{1}{n^2}\sum_{i,j\in[n]}\frac{\mathbf{P}_{ij}}{p_ip_j}\langle h_i,h_j\rangle \\ &= \frac{1}{n^2}\sum_{i\neq j}\frac{\tau(\tau-1)}{n(n-1)}\cdot\frac{n^2}{\tau^2}\langle h_i,h_j\rangle + \frac{1}{n^2}\sum_{i\in[n]}\frac{n}{\tau}\|h_i\|^2 \\ &= \frac{1}{n\tau}\left(\sum_{i\neq j}\frac{\tau-1}{n-1}\langle h_i,h_j\rangle + \sum_{i\in[n]}\|h_i\|^2\right) \\ &= \frac{1}{n\tau}\left(\sum_{i,j\in[n]}\frac{\tau-1}{n-1}\langle h_i,h_j\rangle + \sum_{i\in[n]}\frac{n-\tau}{n-1}\|h_i\|^2\right) \\ &= \frac{1}{n\tau}\cdot\frac{n-\tau}{n-1}\sum_{i\in[n]}\|h_i\|^2.\end{aligned}$$

(iv) For partition sampling, $\mathbf{P}_{ij} = p_C$ if $i,j \in C$, and $\mathbf{P}_{ij} = 0$ otherwise. Hence,

$$\sigma^2 = \frac{1}{n^2}\sum_{i,j\in[n]}\frac{\mathbf{P}_{ij}}{p_ip_j}\langle h_i,h_j\rangle = \frac{1}{n^2}\sum_{C\in\mathcal{G}}\sum_{i,j\in C}\frac{1}{p_C}\langle h_i,h_j\rangle = \frac{1}{n^2}\sum_{C\in\mathcal{G}}\frac{1}{p_C}\|\sum_{i\in C}h_i\|^2.$$

□

I. Importance sampling

I.1. Single element sampling

From (21) it is easy to see that the probabilities that minimize $\mathcal{L}_{\max}$ are $p_i^{\mathcal{L}} = L_i / \sum_{j\in[n]} L_j$, for all i, and consequently $\mathcal{L}_{\max} = \overline{L}$. On the other hand the probabilities that minimize (27) are given by $p_i^{\sigma^2} = \|h_i\| / \sum_{j\in[n]} \|h_j\|$, for all i, with $\sigma^2 = (\sum_{i\in[n]} \|h_i\|/n)^2 := \sigma_{opt}^2$.

Importance sampling. From $p_i^{\mathcal{L}}$ and $p_i^{\sigma^2}$, we construct interpolated probabilities p_i as follows:

$$p_i = p_i(\alpha) = \alpha p_i^{\mathcal{L}} + (1-\alpha) p_i^{\sigma^2}, \tag{65}$$

where $\alpha \in (0,1)$. Then $0 < p_i < 1$ and from (21) we have

$$\mathcal{L}_{\max} \leq \frac{1}{\alpha} \cdot \frac{1}{n} \max_{i\in[n]} \frac{L_i}{p_i^{\mathcal{L}}(\tau)} = \frac{1}{\alpha}\overline{L}.$$

Similarly, from (27) we have that $\sigma^2 \leq \frac{1}{1-\alpha}\sigma_{opt}^2$. Now by letting $p_i = p_i(\alpha)$, from (31) in Theorem 3.1, we get an upper bound of the right hand side of (12):

$$\max\left\{\frac{2\overline{L}}{\alpha\mu}, \frac{4\sigma_{opt}^2}{(1-\alpha)\epsilon\mu^2}\right\}. \tag{66}$$

By minimizing this bound in α we can get

$$\alpha = \frac{\overline{L}}{2\sigma_{opt}^2/\epsilon\mu + \overline{L}}, \tag{67}$$

and then the upper bound (66) becomes

$$\frac{4\sigma_{opt}^2}{\epsilon\mu^2} + \frac{2\overline{L}}{\mu} \leq 2\max\left\{\frac{2\overline{L}}{\mu}, \frac{4\sigma_{opt}^2}{\epsilon\mu^2}\right\}, \tag{68}$$

where the right hand side comes by setting $\alpha = 1/2$. Notice that the minimum of the iteration complexity in (12) is not less than $\max\left\{\frac{2\overline{L}}{\mu}, \frac{4\sigma_{opt}^2}{\epsilon\mu^2}\right\}$. Hence, the iteration complexity of this importance sampling(left hand side of (68)) is at most two times larger than the minimum of the iteration complexity in (12) over p_i.

I.2. Independent sampling

Calculating $p_i^{\mathcal{L}}(\tau)$. Minimizing the upper bound of $\mathcal{L}_{\max}$ in (22) boils down to minimizing $\max_{i\in[n]}(\frac{1}{p_i}-1)\lambda_{\max}(M_i)$, which is not easy generally. Instead, as a proxy we obtain the probabilities p_i by solving

$$\begin{array}{ll} \min & \max_{i\in[n]} \frac{L_i}{p_i} \\ \text{s.t.} & \sum_{i\in[n]} p_i = \tau,\ 0 < p_i \leq 1, \forall i. \end{array} \tag{69}$$

Let $q_i = \frac{L_i}{\sum_{j\in[n]} L_j} \cdot \tau$ for all i, and $T = \{i | q_i > 1\}$. If $T = \emptyset$, it is easy to see $p_i = p_i^{\mathcal{L}}(\tau) = q_i$ solves (69). Otherwise, in order to solve (69), we can choose $p_i = p_i^{\mathcal{L}}(\tau) = 1$ for $i \in T$, and $q_i \leq p_i = p_i^{\mathcal{L}}(\tau) \leq 1$ for $i \notin T$ such that $\sum_{i\in[n]} p_i^{\mathcal{L}}(\tau) = \tau$. By letting $p_i = p_i^{\mathcal{L}}(\tau)$, we have

$$\mathcal{L}_{\max} \leq \frac{1}{n}\left(\left(1+\frac{1}{\tau}\right)\sum_{j\in[n]} L_j - \min_{j\in[n]} L_j\right) \leq \left(1+\frac{1}{\tau}\right)\overline{L}.$$

Calculating $p_i^{\sigma^2}(\tau)$. For σ^2, from (28), we need to solve

$$\begin{array}{ll} \min & \sum_{i\in[n]} \frac{\|h_i\|^2}{p_i} \\ \text{s.t.} & \sum_{i\in[n]} p_i = \tau,\ 0 < p_i \leq 1, \forall i. \end{array} \tag{70}$$

Let $q_i = \frac{\|h_i\|}{\sum_{j\in[n]}\|h_j\|}\cdot\tau$ for all i, and let $T = \{i|q_i > 1\}$. If $T = \emptyset$, it is easy to see that $p_i = p_i^{\sigma^2}(\tau) = q_i$ solve (70). Otherwise, it is a little complicated to find the optimal solution. For simplicity, if $T \neq \emptyset$, we choose $p_i = p_i^{\sigma^2}(\tau) = 1$ for $i \in T$, and $q_i \leq p_i = p_i^{\sigma^2}(\tau) \leq 1$ for $i \notin T$ such that $\sum_{i\in[n]} p_i^{\sigma^2}(\tau) = \tau$. By letting $p_i = p_i^{\sigma^2}(\tau)$, from (28), we have

$$\begin{aligned}\sigma^2 &\leq \frac{1}{n^2}\sum_{i\notin T}\left(\frac{\|h_i\|\sum_{j\in[n]}\|h_j\|}{\tau} - \|h_i\|^2\right)\\ &\leq \frac{1}{\tau}\left(\frac{\sum_{i\in[n]}\|h_i\|}{n}\right)^2 := \sigma^2_{opt}(\tau).\end{aligned}$$

Importance sampling. Since $\mathcal{L} = \left(1+\frac{1}{\tau}\right)\overline{L}$ and $\sigma = \sigma^2_{opt}(\tau)$ are obtained by using the upper bounds in (22) and (28), and the upper bounds are nonincreasing as p_i increases, we get the following property.

Proposition I.1. If $p_i \geq p_i^{\mathcal{L}}(\tau)$ for all i, then $\mathcal{L}_{\max} \leq (1+\frac{1}{\tau})\overline{L}$, and if $p_i \geq p_i^{\sigma^2}(\tau)$, then $\sigma^2 \leq \sigma^2_{opt}(\tau)$.

From Proposition I.1, we can get the following result.

Proposition I.2. For $0 < \alpha < 1$, let $p_i(\alpha)$ satisfy

$$\begin{cases} 1 \geq p_i(\alpha) \geq \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\}, \quad \forall i,\\ \sum_{i\in[n]} p_i(\alpha) = \tau. \end{cases} \tag{71}$$

If $p_i = p_i(\alpha)$ where $p_i(\alpha)$ satisfies (71), then we have

$$\mathcal{L}_{\max} \leq \left(1+\frac{1}{\alpha\tau}\right)\overline{L},$$

and

$$\sigma^2 \leq \sigma^2_{opt}((1-\alpha)\tau) = \frac{1}{(1-\alpha)\tau}\left(\frac{\sum_{i\in[n]}\|h_i\|}{n}\right)^2.$$

Proof. First , we claim that $p_i(\alpha)$ can be constructed to satisfy (71). Since $0 < p_i^{\mathcal{L}}(\alpha) \leq 1$ and $0 < p_i^{\sigma^2}((1-\alpha)\tau) \leq 1$, we know

$$0 < \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\} \leq 1,$$

for all i. Hence, we can first construct $\tilde{q}_i$ such that

$$1 \geq \tilde{q}_i \geq \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\},$$

for all i. Furthermore, since $\sum_{i\in[n]} p_i^{\mathcal{L}}(\alpha\tau) = \alpha\tau$ and $\sum_{i\in[n]} p_i^{\sigma^2}((1-\alpha)\tau) = (1-\alpha)\tau$, we know $\sum_{i\in[n]}\tilde{q}_i \leq \tau$. At last, we increase some $\tilde{q}_i$ which is less than one to make the sum equal to τ, and hence, by letting $p_i(\alpha) = \tilde{q}_i$, $p_i(\alpha)$ satisfies (71).

From (71), we have $p_i = p_i(\alpha) \geq p_i^{\mathcal{L}}(\alpha\tau)$. Then by Proposition I.1, we have

$$\mathcal{L}_{\max} \leq \left(1+\frac{1}{\alpha\tau}\right)\overline{L}.$$

We also have $p_i(\alpha) \geq p_i^{\sigma^2}((1-\alpha)\tau)$, hence, by Proposition I.1, we get

$$\sigma^2 \leq \sigma^2_{opt}((1-\alpha)\tau) = \frac{1}{(1-\alpha)\tau}\left(\frac{\sum_{i\in[n]}\|h_i\|}{n}\right)^2.$$

□

From (12) in Theorem 3.1, by letting $p_i = p_i(\alpha)$ in Proposition I.2, we get an upper bound of the right hand side of (12):

$$\max\left\{\frac{2(1+\frac{1}{\alpha\tau})\overline{L}}{\mu}, \frac{4\sigma_{opt}^2((1-\alpha)\tau)}{\epsilon\mu^2}\right\}.$$

By minimizing this upper bound, we get

$$\alpha = \frac{\tau - a - 1 + \sqrt{4\tau + (\tau - a - 1)^2}}{2\tau}, \tag{72}$$

and the upper bound becomes

$$\frac{2(1+\frac{1}{\alpha\tau})}{\mu}\cdot\overline{L}$$

where $a = 2(\frac{\sum_{i\in[n]}\|h_i\|}{n})^2/(\epsilon\mu\overline{L})$. So suboptimal probabilities

$$p_i = \min\{1, p_i^{\mathcal{L}}(\alpha\tau) + p_i^{\sigma^2}((1-\alpha)\tau)\}, \tag{73}$$

where α is given in Equation (72).

Partially biased sampling. In practice, we do not know $\|h_i\|$ generally. But we can use $p_i^{\mathcal{L}}(\tau)$ and the uniform probability $\frac{\tau}{n}$ to construct a new probability just as that in Proposition I.2. More specific, we have the following result.

Proposition I.3. Let p_i satisfy

$$\begin{cases} 1 \geq p_i \geq \min\{1, p_i^{\mathcal{L}}(\frac{\tau}{2}) + \frac{1}{2}\cdot\frac{\tau}{n}\}, \quad \forall i, \\ \sum_{i\in[n]} p_i = \tau. \end{cases} \tag{74}$$

Then we have

$$\mathcal{L}_{\max} \leq \left(1 + \frac{2}{\tau}\right)\overline{L},$$

and

$$\sigma^2 \leq \left(\frac{2}{\tau} - \frac{1}{n}\right)\cdot\frac{1}{n}\sum_{i\in[n]}\|h_i\|^2.$$

Proof. The proof for $\mathcal{L}_{\max}$ is the same as Proposition I.2. For σ^2, from (28), since $p_i \geq \tau/2n$, we have

$$\sigma^2 = \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{1}{p_i} - 1\right)\|h_i\|^2 \leq \frac{1}{n^2}\sum_{i\in[n]}\left(\frac{2n}{\tau} - 1\right)\|h_i\|^2 = \left(\frac{2}{\tau} - \frac{1}{n}\right)\cdot\frac{1}{n}\sum_{i\in[n]}\|h_i\|^2.$$

□

This sampling is very nice in the sense that it can maintain $\mathcal{L}_{\max}$ at least close to $\overline{L}$, and meanwhile, can acheive nearly linear speedup in σ^2 by increasing τ. We can compare the upper bounds of $\mathcal{L}_{\max}$ and σ^2 for this sampling, τ-nice sampling, and τ-uniform independent sampling when $1 < \tau = \mathcal{O}(1)$ in the following table.

From Table 1, compared to τ-nice sampling and τ-uniform independent sampling, the iteration complexity of this τ-partially biased independent sampling is at most two times larger, but could be about $\frac{2\tau}{n}$ smaller in some extremely case where $L_{\max} \approx n\bar{L}$ and $2\mathcal{L}/\mu$ dominates in (12).

J. Additional Experiments

J.1. From fixed to decreasing stepsizes: analysis of the switching time

Here we evaluate the choice of the switching moment from a constant to a decreasing step size according to (14) from Theorem 3.2. We are using synthetic data that was generated in the same way as it had been in the Section 6 for the ridge

Table 1. Comparison of the upper bounds of $\mathcal{L}_{\max}$ and σ^2 for τ-nice sampling, τ-partially biased independent sampling, and τ-uniform independent sampling.

	$\mathcal{L}_{\max}$	σ^2
τ-NICE SAMPLING	$\frac{n}{\tau}\cdot\frac{\tau-1}{n-1}\bar{L}+\frac{1}{\tau}(1-\frac{\tau-1}{n-1})L_{\max}$	$\frac{1}{\tau}\cdot\frac{n-\tau}{n-1}\bar{h}$
τ-UNIFORM IS	$\bar{L}+(\frac{1}{\tau}-\frac{1}{n})L_{\max}$	$(\frac{1}{\tau}-\frac{1}{n})\bar{h}$
τ-PBA-IS	$(1+\frac{2}{\tau})\bar{L}$	$(\frac{2}{\tau}-\frac{1}{n})\bar{h}$

regression problem. In particular we evaluate 4 different cases: (i) the theoretical moment of regime switch at moment k as predicted from the Theorem, (ii) early switch at $0.3\times k$, (iii) late switch at $0.7\times k$ and (iv) the optimal k for switch, where the optimal k is obtained using one-dimensional numerical minimization of (57) as a strongly convex function of k^*.

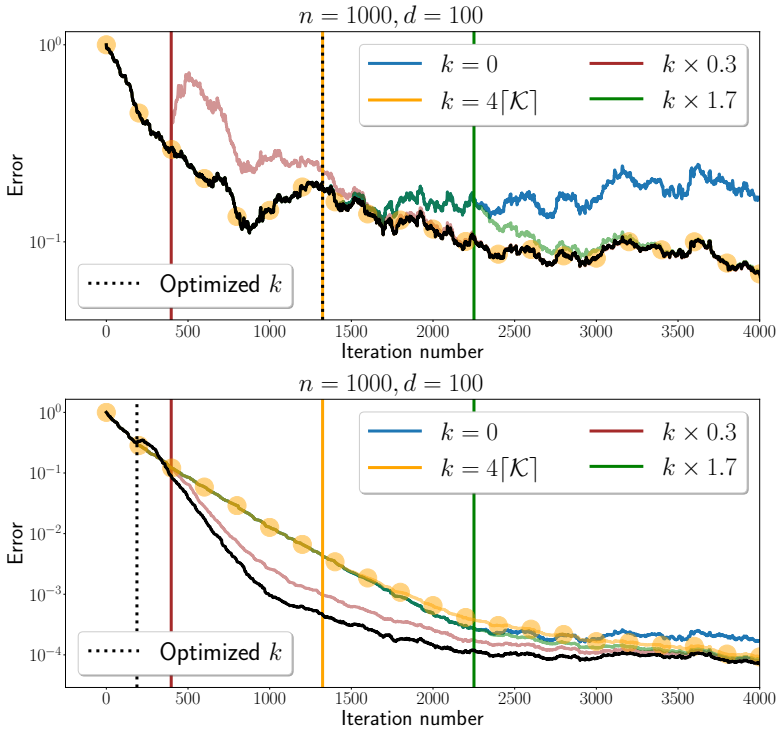

Figure 4. The first plot refers to situation when x^0 is close to x^* (for our data $\left\|r^0\right\|^2 = \left\|x^0 - x^*\right\|^2 \approx 1.0$). The second one covers the opposite case ($\left\|r^0\right\|^2 \approx 864.6$). Vertical lines denote the moments of regime switch for the curves of the same color.

According to Figure 4, when x^0 is close to x^*, the moment of regime switch does not play a significant role in minimizing the number of iteration except for a very early switch, which actually also leads to almost the same situation in the long run. The case when x^0 is far from x^* shows that preliminary one-dimensional optimization makes sense and allows to reduce the error at least during the early iterations.

J.2. More on minibatches

Figure 5 reports on the same experiment as that described in Section 6.2 (Figure 2) in the main body of the paper, but on ridge regression instead of logistic regression, and using different data sets. Our findings are similar, and corroborate the conclusions made in Section 6.2.

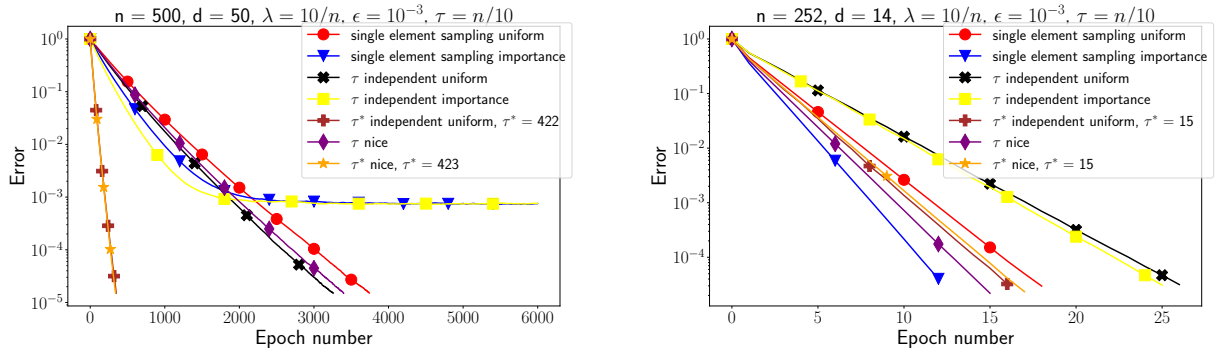

Figure 5. Performance of SGD with several minibatch strategies for ridge regression. On the left: the real data-set bodyfat from LIBSVM. On the right: synthetic data.

J.3. Stepsize as a function of the minibatch size

In our last experiment we calculate the stepsize γ as a function of the minibatch size τ for τ-nice sampling using equation (38). Figure 6 depicts three plots, for three synthetic data sets of sizes $(n, d) \in \{(50, 5), (100, 10), (500, 50)\}$. We consider regularized ridge regression problems with $\lambda = 1/n$. Note that the stepsize is an increasing function of τ.

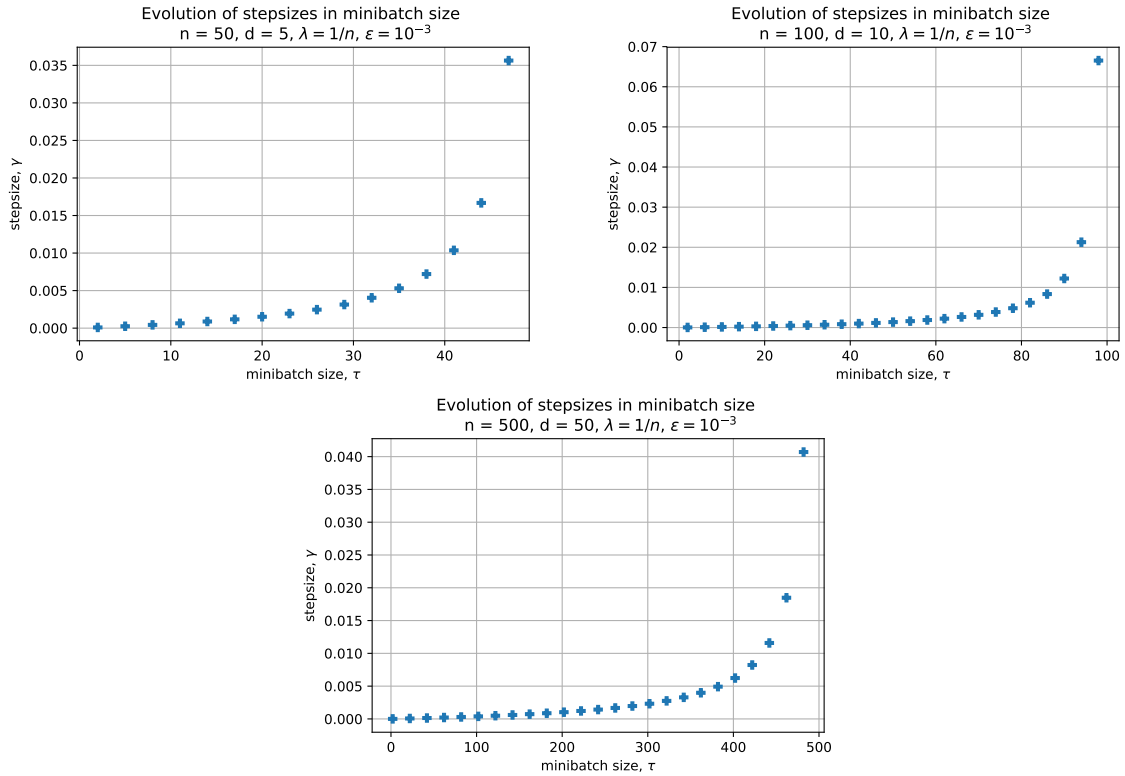

Figure 6. Evolution of stepsize with minibatch size τ for τ nice sampling.