Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent

Nicholas J. A. Harvey
Department of Computer Science
University of British Columbia
Vancouver, BC Canada
nickhar@cs.ubc.ca

Christopher Liaw
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
cvliaw@cs.ubc.ca

Sikander Randhawa
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
srand@cs.ubc.ca

Abstract

We consider stochastic gradient descent algorithms for minimizing a non-smooth, strongly-convex function. Several forms of this algorithm, including suffix averaging, are known to achieve the optimal $O(1/T)$ convergence rate *in expectation*. We consider a simple, non-uniform averaging strategy of Lacoste-Julien et al. (2011) and prove that it achieves the optimal $O(1/T)$ convergence rate *with high probability*. Our proof uses a recently developed generalization of Freedman's inequality. Finally, we compare several of these algorithms experimentally and show that this non-uniform averaging strategy outperforms many standard techniques, and with smaller variance.

1 Introduction

Stochastic gradient descent (SGD) is perhaps the single most important algorithm for minimizing strongly convex loss functions. Its popularity is a combined consequence of the simplicity of its statement and its effectiveness in both theory and practice. Gradient descent is an iterative optimization procedure, where the current solution is updated by taking a step in the opposite direction of the current gradient. In the case of using SGD for Empirical Risk Minimization, the gradient of the loss function is often too expensive to compute. So instead, we select a data point uniformly at random and compute the gradient of the loss function using only this single data point. The resulting value is not necessarily a true gradient, but it is in expectation. When it is time to output a solution, the standard textbook choices are either to report the last iterate, or the average of iterates so far.

Surprisingly, there are situations where these textbook output strategies have provably sub-optimal performance, even though the algorithm is so popular. Here, by performance we are referring to the rate at which the loss of the output convergences to the true minimum value of the loss function. Consider the setting where the loss function is strongly convex, but not smooth (for example, the regularized SVM minimization problem). In the absence of smoothness, there is no guarantee that the value of the individual iterates of SGD improve over time. In fact, Harvey et al. [2018] construct an example where the value of the iterates *increases* over time. Moreover, they show that the convergence rate of the individual iterates of even deterministic gradient descent is $\Omega(\log(T)/T)$, whereas the optimal rate is $\Theta(1/T)$ for a first-order algorithm. Rakhlin et al. [2012] show that returning the

average of all of the iterates so far is also sub-optimal by a $\log(T)$ factor (this lower bound holds in expectation).

As a result, researchers have developed several algorithms which achieve the optimal $O(1/T)$ rate in expectation, some of which are simpler than others. Because the popularity of SGD is largely due to its simplicity, a straightforward variant of the algorithm attaining the optimal rate is significantly more desirable than some other, more complex procedure. The non-uniform averaging strategy from Lacoste-Julien et al. [2012] and the suffix-averaging strategy from Rakhlin et al. [2012] are likely the simplest and closest in resemblance to textbook statements of SGD. Each method runs standard SGD until output time. In Lacoste-Julien et al. [2012], a non-uniform average over all the iterates (with iterate i given a weight proportional to i) is returned whereas in Rakhlin et al. [2012], a uniform average over the last half of the iterates (referred to as the suffix-average) is returned. There still remains another important issue to resolve though, even provided the existence of simple algorithms which obtain the optimal expected $O(1/T)$ rate.

How many random trials are needed for suffix averaging or the non-uniform averaging strategies to actually achieve the $O(1/T)$ rate? (Recall that the error is random). Usual expositions of SGD provide bounds that hold in *expectation*. This is a weak guarantee because it does not preclude the error of the algorithm from having large variance. Users of SGD want to be confident that the output of a single trial of the algorithm is extremely likely to provide the guaranteed convergence rate. In other words, they would prefer bounds that hold with high probability. Moreover, it is often impossible to run many trials of SGD and select the best one. For example, considering the Empirical Risk Minimization setting, if we are dealing with many high dimensional data points, it can be prohibitively expensive to even evaluate the loss function.

It was shown recently in Harvey et al. [2018], that suffix-averaging obtains a convergence rate of $O(\log(1/\delta)/T)$ with probability at least $1-\delta$. However, implementing suffix-averaging when the time horizon is not known ahead of time (for example, stopping SGD when the norm of the gradient is small) requires a modicum of care. Non uniform averaging could be an equally attractive alternative if its convergence rate held with high probability.

Main theoretical results. We show that running standard SGD and returning the very simple non-uniform average of the iterates from Lacoste-Julien et al. [2012] has error at most $O(\log(1/\delta)/T)$ with probability $1-\delta$. The analysis is simple and exposes a martingale which satisfies a special recursive property which was also observed in Harvey et al. [2018]. It is intriguing that this recursive property arises in multiple settings when analyzing SGD. Moreover, we show a matching lower bound of $\Omega(\log(1/\delta)/T)$ with probability at least δ. The analysis uses the simple univariate function $\frac{1}{2}x^2$. Thus, we have a tight high probability analysis of a very simple output strategy for SGD which attains the optimal rate.

Experimental results. In addition, we run detailed experiments for various return schemes of SGD for SVMs on synthetic data and real-world datasets. Our experimental results strongly suggest that the suffix averaging and non-uniform averaging schemes should be preferred over the final iterate and uniform averaging schemes.

1.1 Related Work

There are a number of other algorithms which obtain the optimal $O(1/T)$ convergence rate amongst first-order methods for minimizing a non-smooth, strongly-convex function. Hazan and Kale [2014] proved that a variant of SGD, called Epoch-GD obtains the optimal rate. Here, they partition the total time T into exponentially growing epochs. Within each epoch, they run standard SGD (with an appropriate step size) and after the end of each epoch, they reset the current point to the average of the iterates in the previous epoch.

Later, Rakhlin et al. [2012] and Lacoste-Julien et al. [2012] independently discovered simpler algorithms that also achieve the optimal $O(1/T)$ rate. In fact, both algorithms run standard SGD with the standard step size proportional to $1/t$; the only difference between the two algorithms is the return value of the algorithm. In Rakhlin et al. [2012], they show that suffix averaging, where one returns the last α fraction of the iterates (for some constant $\alpha > 0$), achieves the optimal rate. On the other hand, Lacoste-Julien et al. [2012] prove that a certain non-uniform average (see Algorithm 1)

of the iterates also achieves the optimal rate. One advantage of non-uniform averaging is that the iterates can be easily computed on the fly.

Recently, Nesterov and Shikhman [2015] have devised a modification of gradient descent for which the error of the *last iterate* converges at the optimal rate. Even more recently, Jain et al. [2019] showed that even for unmodified gradient descent the last iterate can be made to achieve the optimal rate, if the time horizon is known beforehand, and if the step-size is chosen carefully using the time horizon. Interestingly, they also show that knowing (or having a bound on) the time horizon is necessary for all the individual iterates to achieve the optimal rate.

High-probability upper bounds. All the results stated above hold only in expectation and do not rule out the possibility that the return value has high variance. Moreover, it can be expensive to compute the objective value of a point. Hence, it is desirable to have a high-probability upper bound on the return value.

To assist in this task, Harvey et al. [2018] recently developed a generalization of Freedman's Inequality. Using this, they show that if one runs SGD with the standard $1/t$ step sizes, then the last iterate and the suffix average schemes achieve error $O(\log(T)/T)$, and $O(1/T)$, respectively, with high probability. Using similar methods, Jain et al. [2019] prove that the last iterate of SGD with carefully chosen step sizes achieves an error of $O(1/T)$. (As mentioned above, this requires advance knowledge of T.) The uniform average was earlier shown by Kakade and Tewari [2008] to achieve error $O(\log(T)/T)$ with high probability. Here, we will also employ the generalized Freedman's Inequality to prove a tight high-probability upper bound on the non-uniform averaging scheme.

2 Preliminaries

Let $\mathcal{X}$ be a closed, convex subset of $\mathbb{R}^n$, $f\colon \mathcal{X} \to \mathbb{R}$ be a convex function, and $\partial f(x)$ be the subdifferential of f at x. Our goal is to solve the convex program $\min_{x\in\mathcal{X}} f(x)$. We assume that f may not be explicitly represented. Instead, the algorithm is allowed to query f via a stochastic gradient oracle, i.e., if the oracle is queried at x then it returns $\hat{g} = g - \hat{z}$ where $g \in \partial f(x)$ and $\mathrm{E}[\,\hat{z}\,] = 0$ conditioned on all past calls to the oracle. Furthermore, we assume that f is L-Lipschitz, i.e. $\|g\| \le L$ for all $x \in \mathcal{X}$ and $g \in \partial f(x)$ and that f is μ-strongly convex, i.e.

$$f(y) \;\ge\; f(x) + \langle\, g,\, y - x \,\rangle + \frac{\mu}{2}\,\|y - x\|^2 \quad \forall y, x \in \mathcal{X}, g \in \partial f(x). \tag{1}$$

Throughout this paper, $\|\cdot\|$ denotes the *Euclidean* norm in $\mathbb{R}^n$, $\Pi_{\mathcal{X}}$ denotes the projection operator onto $\mathcal{X}$ and $[T]$ denotes the set $\{1, \ldots, T\}$. For the sake of simplicity, we assume that $\|\hat{z}\| \le 1$ a.s.

In this paper, we analyze SGD with the averaging scheme proposed by Lacoste-Julien et al. [2012]. The algorithm is given in Algorithm 1.

Algorithm 1 Stochastic, projected gradient descent for minimizing a μ-strongly convex, L-Lipschitz function with an unknown time horizon.

procedure PROJECTEDGRADIENTDESCENT($\mathcal{X} \subseteq \mathbb{R}^n,\ x_1 \in \mathcal{X}$)
  for $t \leftarrow 1, \ldots, T$ **do**
    Let $\eta_t = \frac{2}{\mu(t+1)}$
    $y_{t+1} \leftarrow x_t - \eta_t \hat{g}_t$, where $\mathrm{E}[\,\hat{g}_t \mid \hat{g}_1, \ldots, \hat{g}_{t-1}\,] \in \partial f(x_t)$
    $x_{t+1} \leftarrow \Pi_{\mathcal{X}}(y_{t+1})$
  return $\sum_{t=1}^{T} \frac{t}{T(T+1)/2} x_t$

Finally, we will use $\mathcal{F}_t$ to denote the σ-field generated by the random vectors $\hat{g}_1, \ldots, \hat{g}_t$.

Remark 2.1. *As noted in Lacoste-Julien et al. [2012], the return value of Algorithm 1 can be computed in an online manner. Indeed, we can set $z_1 = x_1$, and we can set $z_t = \rho_t x_t + (1 - \rho_t) z_{t-1}$ for $t \ge 2$, where $\rho_t = \frac{2}{t+1}$. It is a straightforward calculation to check that $z_T = \sum_{t=1}^{T} \frac{t}{T(T+1)/2} x_t$.*

2.1 Probability tools

Our main probabilistic tool is an extension of Freedman's Inequality [Freedman, 1975] developed recently by Harvey et al. [2018]. Roughly speaking, Freedman's Inequality asserts that a martingale

is bounded by the square root of its total conditional variance (TCV). As we shall see in the sequel, the martingales that arise from analyzing SGD exhibit a "chicken-and-egg" phenomenon wherein the TCV of the martingale is bounded by (a linear transformation of) the martingale itself. Here, we state a specialized form of the Generalized Freedman's Inequality which is a simple corollary from the statement given in Harvey et al. [2018].

Theorem 2.2 (Generalized Freedman, [Harvey et al., 2018, Theorem 3.3])**.** *Let $\{d_t, \mathcal{F}_t\}_{t=1}^T$ be a martingale difference sequence. Suppose that, for $t \in [T]$, v_{t-1} are non-negative $\mathcal{F}_{t-1}$-measurable random variables satisfying $\mathrm{E}\left[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2} v_{t-1}\right)$ for all $\lambda > 0$. Let $S_T = \sum_{t=1}^T d_t$ and $V_T = \sum_{t=1}^T v_{t-1}$. Suppose there exists $\alpha_1, \ldots, \alpha_T, \beta \in \mathbb{R}_{\geq 0}$ such that $V_T \leq \sum_{t=1}^T \alpha_t d_t + \beta$. Let $\alpha \geq \max_{t \in [T]} \alpha_t$. Then*

$$\Pr\left[S_T \geq x\right] \;\leq\; \exp\left(-\frac{x^2}{4\alpha \cdot x + 8\beta}\right).$$

3 Main results

Our main result is a high-probability upper bound on the final iterate of Algorithm 1. The proof is given in Section 4.

Theorem 3.1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex (with respect to $\|\cdot\|_2$) and L-Lipschitz. Assume that:*

(a) $g_t \in \partial f(x_t)$ for all t (with probability 1).

(b) $\|\hat{z}_t\| \leq 1$ (with probability 1).

Set $\eta_t = \frac{2}{\mu(t+1)}$ and $\gamma_t = \frac{t}{T(T+1)/2}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$f\left(\sum_{t=1}^T \gamma_t x_t\right) - f(x^*) \;\leq\; O\left(\frac{L \cdot \log(1/\delta) + L^2}{\mu} \cdot \frac{1}{T}\right).$$

Remark 3.2. *It is possible to strengthen the statement of Theorem 3.1 by replacing assumption (a) with the weaker assumption that $\|\hat{z}_t\|$ is subgaussian conditioned on $\mathcal{F}_{t-1}$ (for example, $\hat{z}_t \sim N(0, \frac{1}{n} I_n)$). A more detailed discussion can be found in the supplementary material.*

We also show that the bound in Theorem 3.1 is tight up to constant factors. The proof is in Section 5.

Claim 3.3. *Suppose $\sqrt{6} \leq \frac{\sqrt{2\log(1/\delta)}}{3} \leq \sqrt{T}/4$. There exists a sub-gradient oracle such that running Algorithm 1 on the function $f(x) = \frac{x^2}{2}$ with step sizes $\eta_t = \frac{1}{t+1}$ satisfies the following. With probability at least δ*

$$f\left(\sum_{t=1}^T \gamma_t x_t\right) - f(x^*) \geq \frac{\log(1/\delta)}{9 \cdot T},$$

where $\gamma_t = \frac{t}{T(T+1)/2}$.

4 Proof of high probability upper bound

The proof follows that of Lacoste-Julien et al. [2012] but we must be careful with the noise terms as our goal is obtain a high probability bound. We will need one technical lemma whose proof we relegate to the next subsection.

Lemma 4.1. *Let $Z_T = \sum_{t=1}^T t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle$. Then for any $\delta \in (0,1)$, $Z_T \leq O\left(\frac{L}{\mu} \cdot T \log(1/\delta)\right)$, with probability at least $1-\delta$.*

Proof (of Theorem 3.1). Define $\hat{z}_t = g_t - \hat{g}_t$. Since f is μ-strongly convex, we have

$$f(x_t) - f(x^*) \;\leq\; \langle\, g_t,\, x_t - x^* \,\rangle - \frac{\mu}{2} \|x_t - x^*\|_2^2$$

$$= \langle\, \hat{g}_t,\, x_t - x^* \,\rangle - \frac{\mu}{2}\,\|x_t - x^*\|_2^2 + \langle\, \hat{z}_t,\, x_t - x^* \,\rangle.$$

The first two terms can be bounded as follows.

$$\begin{aligned}
&\langle\, \hat{g}_t,\, x_t - x^* \,\rangle - \frac{\mu}{2}\,\|x_t - x^*\|_2^2 \\
&= \frac{1}{\eta_t}\langle\, x_t - y_{t+1},\, x_t - x^* \,\rangle - \frac{\mu}{2}\,\|x_t - x^*\|_2^2 \qquad \text{(by the gradient step)} \\
&= \frac{1}{2\eta_t}\Big(\,\|x_t - y_{t+1}\|_2^2 + \|x_t - x^*\|_2^2 - \|y_{t+1} - x^*\|_2^2\Big) - \frac{\mu}{2}\,\|x_t - x^*\|_2^2 \\
&\le \frac{1}{2\eta_t}\Big(\,\|x_t - y_{t+1}\|_2^2 + \|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\Big) - \frac{\mu}{2}\,\|x_t - x^*\|_2^2.
\end{aligned}$$

The last line uses a property of Euclidean projections: since x_{t+1} is the projected point $\Pi_{\mathcal{X}}(y_{t+1})$ and $x^* \in \mathcal{X}$, we have $\|x_{t+1} - x^*\|_2^2 \le \|y_{t+1} - x^*\|_2^2$.

It is convenient to scale by t in order to later obtain a telescoping sum. Using the definition of the gradient step, i.e. $x_t - y_{t+1} = \eta_t \hat{g}_t$, we have

$$\begin{aligned}
&t \cdot \Big(f(x_t) - f(x^*) - \langle\, \hat{z}_t,\, x_t - x^* \,\rangle\Big) \\
&\le \frac{t\,\|\eta_t \hat{g}_t\|_2^2}{2\eta_t} + t\Big(\frac{1}{2\eta_t} - \frac{\mu}{2}\Big)\,\|x_t - x^*\|_2^2 - \frac{t}{2\eta_t}\,\|x_{t+1} - x^*\|_2^2 \\
&= \frac{t\,\|\hat{g}_t\|_2^2}{\mu(t+1)} + \Big(\frac{\mu t(t+1)}{4} - \frac{2\mu t}{4}\Big)\,\|x_t - x^*\|_2^2 - \frac{t(t+1)\mu}{4}\,\|x_{t+1} - x^*\|_2^2 \\
&\le \frac{(L+1)^2}{\mu} + \frac{\mu}{4} \cdot \Big(t(t-1)\,\|x_t - x^*\|_2^2 - t(t+1)\,\|x_{t+1} - x^*\|_2^2\Big).
\end{aligned}$$

Now, summing over t, the right-hand side telescopes and we obtain

$$\sum_{t=1}^{T} t \cdot \big(f(x_t) - f(x^*)\big) \;\le\; \sum_{t=1}^{T} t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle + \frac{T \cdot (L+1)^2}{\mu}$$

Dividing by $T(T+1)/2$ and applying Jensen's inequality, we obtain

$$\begin{aligned}
f\Big(\sum_{t=1}^{T} \gamma_t x_t\Big) - f(x^*) \;&\le\; \sum_{t=1}^{T} \gamma_t \cdot \big(f(x_t) - f(x^*)\big) \\
&\le\; \frac{2}{T(T+1)} \underbrace{\sum_{t=1}^{T} t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle}_{=:\, Z_T} + \frac{2 \cdot (L+1)^2}{\mu(T+1)}.
\end{aligned}$$

Finally, we can use Lemma 4.1 to obtain a high probability bound on Z_T, completing the proof of the theorem. □

4.1 Bounding Z_T

Observe that Z_T is a sum of a martingale difference sequence. Define $d_t = t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle$, $v_{t-1} := t^2\,\|x_t - x^*\|$, and $V_T = \sum_{t=1}^{T} v_{t-1}$. Note that v_{t-1} is $\mathcal{F}_{t-1}$-measurable. The next claim shows that v_{t-1} and d_t satisfy the assumptions of Generalized Freedman's inequality (Theorem 2.2).

Claim 4.2. *For all t and $\lambda > 0$, we have* $\mathrm{E}\,[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}] \le \exp\Big(\frac{\lambda^2}{2} v_{t-1}\Big)$.

Proof. First, we can apply Cauchy-Schwarz to get that $|t\langle\, \hat{z}_t,\, x_t - x^* \,\rangle| \le t \cdot \|\hat{z}_t\| \cdot \|x_t - x^*\| \le t \cdot \|x_t - x^*\|$ because $\|\hat{z}_t\| \le 1$ a.s. Next, applying Hoeffding's Lemma ([Massart, 2007, Lemma 2.6]), we have $\mathrm{E}\,[\exp(\lambda t\langle\, \hat{z}_t,\, x_t - x^* \,\rangle) \mid \mathcal{F}_{t-1}] \le \exp\Big(\frac{\lambda^2}{2} t^2\,\|x_t - x^*\|^2\Big)$. □

To bound Z_T, we will show that we can bound its TCV by a linear combination of the increments. This will allow us to use the Generalized Freedman Inequality (Theorem 2.2).

Lemma 4.3. *There exists non-negative constants $\alpha_1, \ldots, \alpha_T$ such that $\max_{i\in[T]}\{\alpha_i\} = O\left(\frac{T}{\mu}\right)$ and $\beta = O\left(\frac{L^2}{\mu^2}T^2\right)$ such that $V_T \le \sum_{t=1}^T \alpha_t d_t + \beta$.*

Proof (of Lemma 4.1). By Claim 4.2, we have $\mathrm{E}\left[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}\right] \le \exp\left(\frac{\lambda^2}{2}v_{t-1}\right)$ for all $\lambda > 0$. By Lemma 4.3, we have $V_T \le \sum_{t=1}^T \alpha_t d_t + \beta$. Plugging $\alpha = O\left(\frac{T}{\mu}\right)$, $\beta = O\left(\frac{L^2}{\mu^2}T^2\right)$, and $x = O\left(\frac{L}{\mu}\cdot T\log(1/\delta)\right)$ into Theorem 2.2 proves the lemma. □

It remains to prove Lemma 4.3. To do so, we will need the following two lemmata, which are adapted from Rakhlin et al. [2012] to use the step sizes $\eta_t = \frac{2}{\mu(t+1)}$. For completeness, we provide a proof in the supplementary material.

Lemma 4.4 ([Rakhlin et al., 2012, Lemma 5])**.** *With probability 1, and for all t, $\|x_t - x^*\| \le \frac{2L}{\mu}$.*

Lemma 4.5 ([Rakhlin et al., 2012, Lemma 6])**.** *For all $t \ge 3$, there exists non-negative numbers $a_1(t), \ldots, a_t(t)$ with $a_i(t) = \Theta(i^3/t^4)$ and $b_1(t), \ldots, b_t(t)$ with $b_i(t) = \Theta(i^2/t^4)$, such that with probability 1*

$$\|x_{t+1} - x^*\|^2 \le \frac{4}{\mu}\sum_{i=3}^t a_i(t)\langle\, \hat{z}_i,\, x_i - x^*\,\rangle + \frac{4}{\mu^2}\sum_{i=3}^t b_i(t)\,\|\hat{g}_t\|^2 .$$

Remark 4.6. *Lemma 4.4 and Lemma 4.5 are true regardless of the assumption we place on $\hat{z}_t$.*

Proof (of Lemma 4.3). Recall $\|\hat{g}_i\| \le L + 1$ because f is L-Lipschitz and $\|\hat{z}_i\| \le 1$ almost surely. By Lemma 4.4 and Lemma 4.5, we have:

$$\begin{aligned}
&V_T \\
&= \sum_{t=1}^T t^2\cdot\|x_t - x^*\|^2 \\
&\le \frac{56L^2}{\mu^2} + \sum_{t=4}^T t^2\left(\frac{4}{\mu}\sum_{i=3}^{t-1} a_i(t-1)\langle\,\hat{z}_i,\,x_i - x^*\,\rangle + \frac{4}{\mu^2}\sum_{i=3}^{t-1} b_i(t-1)\,\|\hat{g}_i\|^2\right) \\
&\le \frac{56L^2}{\mu^2} + \sum_{t=4}^T t^2\left(\frac{4}{\mu}\sum_{i=3}^{t-1} a_i(t-1)\langle\,\hat{z}_i,\,x_i - x^*\,\rangle + \frac{4(L+1)^2}{\mu^2}\sum_{i=3}^{t-1} b_i(t-1)\right) \\
&= \frac{4}{\mu}\sum_{t=4}^T t^2\left(\sum_{i=3}^{t-1} a_i(t-1)\langle\,\hat{z}_i,\,x_i - x^*\,\rangle\right) + \frac{4(L+1)^2}{\mu^2}\sum_{t=4}^T t^2\left(\sum_{i=3}^{t-1} b_i(t-1)\right) + \frac{56L^2}{\mu^2} \\
&= \sum_{i=3}^{T-1}\underbrace{\frac{4}{\mu}\left(\sum_{t=i+1}^T t^2\cdot\frac{a_i(t-1)}{i}\right)}_{:=\alpha_i}\cdot i\langle\,\hat{z}_i,\,x_i - x^*\,\rangle + \underbrace{\frac{4(L+1)^2}{\mu^2}\sum_{t=4}^T t^2\left(\sum_{i=3}^{t-1} b_i(t-1)\right) + \frac{56L^2}{\mu^2}}_{:=\beta}
\end{aligned}$$

Define $\alpha_1, \alpha_2, \alpha_T = 0$. We have $V_T \le \sum_{i=1}^T \alpha_i\cdot i\cdot\langle\hat{z}_i,\,x_i - x^*\rangle + \beta$. It remains to show $\max\{\alpha_i\} = O\left(\frac{T}{\mu}\right)$ and $\beta = O\left(\frac{L^2}{\mu^2}T^2\right)$. To bound $\max\{\alpha_i\}$, observe that for $i \in \{3, \ldots, T-1\}$,

$$\sum_{t=i+1}^T t^2\cdot\frac{a_i(t-1)}{i} = \sum_{t=i+1}^T t^2 O\left(\frac{i^2}{t^4}\right) = \sum_{t=i+1}^T t^2 O\left(\frac{1}{t^2}\right) = O\left(T - i\right).$$

To bound β, observe

$$\sum_{t=4}^T t^2\left(\sum_{i=3}^{t-1} b_i(t-1)\right) = \sum_{t=4}^T t^2\sum_{i=3}^{t-1} O\left(\frac{i^2}{t^4}\right) = \sum_{t=4}^T t^2\sum_{i=3}^{t-1} O\left(\frac{1}{t^2}\right) = \sum_{t=4}^T O(t) = O(T^2).$$

□

5 Description of high probability lower bound

Setup of the lower bound. Consider the one dimensional, 1-strongly convex function $f(x) = \frac{1}{2}x^2$ with feasible region $\mathcal{X} = [-6, 6]$. Suppose, that at any point x_t, the gradient oracle returns a value of the form $x_t - \hat{z}_t$, where $\mathrm{E}[\,\hat{z}_t\,] = 0$. Clearly, this is a valid subgradient oracle. Suppose we run Algorithm 1 with a *slightly modified step size of* $\eta_t = \frac{1}{t+1}$ starting from initial point $x_1 = 0$.

Remark 5.1. *Note that we are using a step size of $\frac{1}{t+1}$ instead of the step size $\frac{2}{t+1}$ used in the statement of Algorithm 1. It is possible to modify the analysis to use the step size as stated in Algorithm 1, however the analysis is much cleaner using $\frac{1}{t+1}$ and still captures the main ideas.*

Claim 5.2. *Suppose $x_1 = 0$ and assume $|\hat{z}_t| \leq 6$. Then, $x_t = \frac{1}{t}\sum_{i=1}^{t-1} \hat{z}_i$ for all $2 \leq t \leq T$.*

Definition of gradient oracle. Let $\hat{z}_t = 0$ if $t \leq \frac{T}{2}$ or $T > \frac{3T}{4}$ and otherwise for $T/2 + 1 \leq t \leq \frac{3T}{4}$, define $\hat{z}_t = \frac{T+1}{T-t} X_t$ where X_t is uniform in $\{+1, -1\}$. Note that this gradient oracle satisfies the conditions of Claim 5.2. That is $|\hat{z}_t| \leq 6$ for all t, as long as $T \geq 2$.

By definition of $\hat{z}_t$ and Claim 5.2, one can check that $\sum_{i=1}^{T} \gamma_t x_t$ is an average of Bernoulli random variables. Applying a reverse Chernoff bound from Klein and Young [2015] completes the proof. The complete details can be found in the supplementary materials.

6 Experimental results

The four return strategies discussed in this paper have fairly similar theoretical guarantees. The aim of this section is to compare the strategies on real data sets, focusing on two aspects of their performance: the expectation and the concentration of the objective value. The results are shown in Figure 1. Additional experimental results can be found in the supplementary material (Section D).

The results of the experiments reveal a clear message. The final iterate and the uniform average return strategies perform noticeably worse than the suffix average and non-uniform average, both in terms of expectation and concentration. This is consistent with the fact that their theoretical guarantees are also worse. The performance of the suffix average and the non-uniform average are nearly indistinguishable, with the suffix average having a slight advantage in expectation.

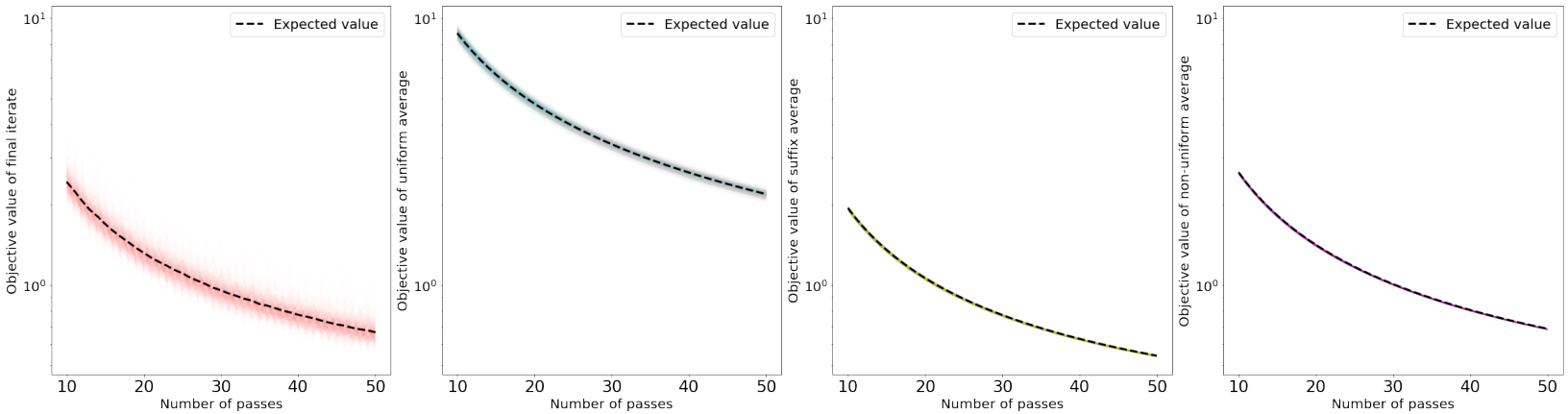
Methodology. We consider the regularized SVM optimization problem

$$f(w) \;\coloneqq\; \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{m} \max\left\{0, 1 - y_i w^{\mathsf{T}} x_i\right\},$$

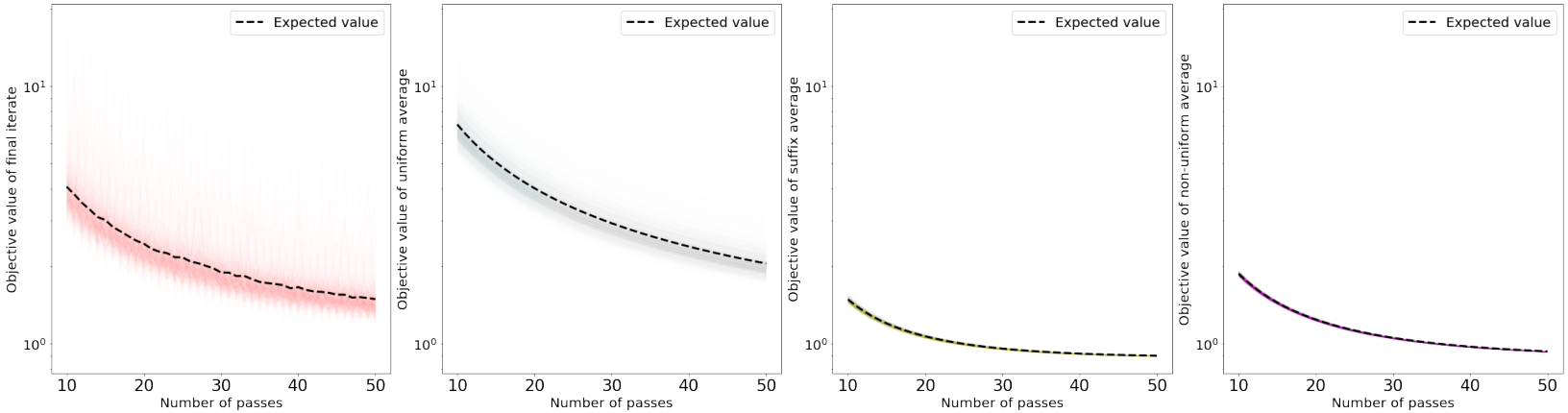
where m is the number of data points and we use n to denote the dimension of each data point. We run SGD with step size $\eta_t = \frac{2}{t+1}$ and with regularization parameter $\lambda = 1/m$. This particular step size is required for Theorem 3.1, and the analyses for the other averaging schemes can also accommodate this choice of step size. Furthermore, we found that the relative performance of the different averaging schemes is not particularly sensitive to the choice in step size. We plot the value of f for each return strategy every m iterations (which we refer to as an 'effective pass'). Since the output of SGD is random, there is a distribution over the outputs which we would like to capture. We run 1000 trials of SGD. The colored curves are exactly these 1000 trials, which are plotted with low opacity. At any point in time, the darkness of the plot at a specific objective value indicates the number of trials that achieved that value at that time. The dotted dark lines represent the average amongst the trials.

Figure 1 suggests that practitioners should consider using the suffix average or non-uniform average in lieu of the final iterate or uniform average. It is possible to implement suffix averaging and non-uniform averaging with minimal effort, and the performance boost is significant. Implementing non-uniform averaging (even when the time horizon is not fixed ahead of time) only requires a single additional line of code.

Data sets. We performed our experiments on a set of freely available binary classification data sets. The experiments from this section use the *cina0* ($n = 16033$ and $d = 132$) and the *protein* ($m = 145751$ and $n = 74$) data sets. We ran the same experiments on the *rcv1* ($m = 20242$ and $n = 47236$), *covtype* ($m = 581,012$ and $n = 54$) and *quantum* ($m = 50000$ and $n = 78$) data sets. The results for these data sets can be found in Section D. Sparse features were scaled to $[0, 1]$ whereas dense features were scaled to have zero mean and unit variance. Data sets *quantum* and *protein* can be found at the KDD cup 2004 website, *cina0* can be found at the Causality Workbench website and *covtype* and *rcv1* can be found at the LIBSVM website.



(a) cina0



(b) protein

Figure 1: Number of effective passes vs. objective value. The first row plots the results using the *cina0* data set, whereas the second plots results using the *protein* data set. From left to right, we plot the objective value over time of the final iterate, uniform average, suffix average and non-uniform average for 1000 trials of SGD.

References

David A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. *CoRR*, abs/1812.05217, 2018. URL `http://arxiv.org/abs/1812.05217`.

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

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. *arXiv preprint arXiv:1904.12443*, 2019.

Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pages 801–808, 2008.

Philip Klein and Neal E Young. On the number of iterations for Dantzig–Wolfe optimization and packing-covering approximation algorithms. *SIAM Journal on Computing*, 44(4):1154–1172, 2015.

Simon Lacoste-Julien, Mark W. Schmidt, and Francis R. Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012. URL `http://arxiv.org/abs/1212.2002`.

Pascal Massart. *Concentration inequalities and model selection*. Springer, 2007.

Yu. Nesterov and V. Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, Jun 2015.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of ICML*, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

A Proof of Lemma 4.4 and Lemma 4.5

Both of the proofs in this section are slight modifications of the proofs found in Rakhlin et al. [2012].

Proof (of Lemma 4.4). Due to strong convexity and the fact that $f(x_t) - f(x^*) \geq 0$, we have

$$L \left\| x_t - x^* \right\| \geq \left\| g_t \right\| \left\| x_t - x^* \right\| \geq \left\langle\, g_t,\, x_t - x^* \,\right\rangle \geq \frac{\mu}{2} \left\| x_t - x^* \right\|^2 ,$$

where we used L-Lipschitzness of f to bound $\|g_t\|$ by L.

□

Proof (of Lemma 4.5). The definition of strong convexity yields

$$\left\langle\, g_t,\, x_t - x^* \,\right\rangle \geq f(x_t) - f(x^*) + \frac{\mu}{2} \left\| x_t - x^* \right\|^2 .$$

Strong convexity and the fact that $0 \in \partial f(x^*)$ implies

$$f(x_t) - f(x^*) \geq \frac{\mu}{2} \left\| x_t - x^* \right\|^2 .$$

Next, recall that for any $x \in \mathcal{X}$, and for any z, we have $\left\| \Pi_{\mathcal{X}}(z) - x \right\| \leq \left\| z - x \right\|$. Lastly, recall $\eta_t = \frac{2}{\mu(t+1)}$. Using these, we have

$$\begin{aligned}
\left\| x_{t+1} - x^* \right\|^2 &= \left\| \Pi_{\mathcal{X}}(x_t - \eta_t \hat{g}_t) - x^* \right\|^2 \\
&\leq \left\| x_t - \eta_t \hat{g}_t - x^* \right\|^2 \\
&= \left\| x_t - x^* \right\|^2 - 2\eta_t \left\langle\, \hat{g}_t,\, x_t - x^* \,\right\rangle + \eta_t^2 \left\| \hat{g}_t \right\|^2 \\
&= \left\| x_t - x^* \right\|^2 - 2\eta_t \left\langle\, g_t,\, x_t - x^* \,\right\rangle + 2\eta_t \left\langle\, \hat{z}_t,\, x_t - x^* \,\right\rangle + \eta_t^2 \left\| \hat{g}_t \right\|^2 \\
&\leq \left\| x_t - x^* \right\|^2 - 2\eta_t \left(f(x_t) - f(x^*) \right) - \eta_t \mu \left\| x_t - x^* \right\|^2 \qquad (2) \\
&\quad + 2\eta_t \left\langle\, \hat{z}_t,\, x_t - x^* \,\right\rangle + \eta_t^2 \left\| \hat{g}_t \right\|^2 \\
&\leq (1 - 2\eta_t \mu) \left\| x_t - x^* \right\|^2 + 2\eta_t \left\langle\, \hat{z}_t,\, x_t - x^* \,\right\rangle + \eta_t^2 \left\| \hat{g}_t \right\|^2 \\
&= \left(1 - \frac{4}{t+1} \right) \left\| x_t - x^* \right\|^2 + \frac{4}{\mu(t+1)} \left\langle\, \hat{z}_t,\, x_t - x^* \,\right\rangle + \frac{4}{\mu^2 (t+1)^2} \left\| \hat{g}_t \right\|^2 . \qquad (3)
\end{aligned}$$

Repeatedly applying Eq. (3) until $t = 4$, yields the following

$$\begin{aligned}
\left\| x_{t+1} - x^* \right\|^2 \leq\ & \frac{4}{\mu} \sum_{i=4}^{t} \left[\frac{1}{i+1} \prod_{j=i+1}^{t} \left(1 - \frac{4}{j+1} \right) \right] \cdot \left\langle\, \hat{z}_i,\, x_i - x^* \,\right\rangle \\
&+ \frac{4}{\mu^2} \sum_{i=4}^{t} \left[\frac{1}{(i+1)^2} \prod_{j=i+1}^{t} \left(1 - \frac{4}{j+1} \right) \right] \cdot \left\| \hat{g}_t \right\| . \qquad (4)
\end{aligned}$$

Observing that

$$\prod_{j=i+1}^{t} \left(1 - \frac{4}{j+1} \right) = \prod_{j=i+1}^{t} \frac{j-3}{j+1} = \frac{(i-2) \cdot (i-1) \cdot i \cdot (i+1)}{(t-2) \cdot (t-1) \cdot t \cdot (t+1)},$$

proves the lemma by taking $a_i(t) = \frac{1}{i+1} \cdot \frac{(i-2) \cdot (i-1) \cdot i \cdot (i+1)}{(t-2) \cdot (t-1) \cdot t \cdot (t+1)}$ and $b_i(t) = \frac{1}{(i+1)^2} \cdot \frac{(i-2) \cdot (i-1) \cdot i \cdot (i+1)}{(t-2) \cdot (t-1) \cdot t \cdot (t+1)}$

□

B Proof of high probability lower bound

In this section we show that the error of SGD when returning $\sum_{t=1}^{T} \frac{t}{T(T+1)/2} x_t$ is $\Omega(\log(1/\delta)/T)$ with probability at least δ. We begin by stating a useful lemma.

Lemma B.1 ([Klein and Young, 2015, Lemma 4]). *Let $X_1, \ldots, X_n$ be independent random variables taking value $\{-1, +1\}$ uniformly at random and $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Suppose $\sqrt{6} \leq c \leq \sqrt{n}/2$, then*

$$\Pr\left[X \geq \frac{c}{\sqrt{n}} \right] \geq \exp(-9c^2/2).$$

Proof (of Claim 3.3). Since the gradient oracle satisfies the assumption in Claim 5.2 (which we prove below), we may apply Claim 5.2 to obtain:

$$\begin{aligned}
\sum_{t=1}^{T} \gamma_t x_t &= \sum_{t=2}^{T} \gamma_t \left[\frac{1}{t} \sum_{i=1}^{t-1} \hat{z}_i \right] && \text{(by Claim 5.2)} \\
&= \frac{2}{T(T+1)} \sum_{t=2}^{T} \sum_{i=1}^{t-1} \hat{z}_i && \text{(definition of } \gamma_t) \\
&= \frac{2}{T(T+1)} \sum_{i=1}^{T-1} \hat{z}_i \cdot (T-i) && \text{(swap order of summation)} \\
&= \frac{2}{T(T+1)} \sum_{i=T/2+1}^{3T/4} \hat{z}_i \cdot (T-i) && (\hat{z}_i = 0 \text{ for all other } i) \\
&= \frac{2}{4} \left(\frac{1}{T/4} \sum_{i=T/2+1}^{3T/4} X_i \right) && \text{(definition of } \hat{z}_i).
\end{aligned}$$

Now, we may apply Lemma B.1 with $c = \frac{\sqrt{2\log(1/\delta)}}{3}$, and $n = T/4$ to obtain:

$$f\left(\sum_{t=1}^{T} \gamma_t x_t \right) = \frac{1}{2} \left(\sum_{t=1}^{T} \gamma_t x_t \right)^2 \geq \frac{1}{2} \left(\frac{1}{2} \frac{\sqrt{2\log(1/\delta)}}{3\sqrt{T/4}} \right)^2 = \frac{\log(1/\delta)}{9 \cdot T},$$

with probability at least δ. □

The proof of Claim 3.3 required the use of Claim 5.2. We now provide a proof of this claim

Proof (of Claim 5.2). We prove the claim via induction. For the base case consider $x_2 = \Pi_{\mathcal{X}}(x_1 - \eta_1 \hat{g}_1)$. Recall that $\hat{g}_1 = g_1 - \hat{z}_1$ where g_1 is the gradient of f at x_1. Since $x_1 = 0$, we have $g_1 = 0$ and $x_2 = \Pi_{\mathcal{X}}(\eta_1 \hat{z}_1) = \frac{1}{2}\hat{z}_1$ because $|\hat{z}_t| \leq 1$ for all t and $\eta_t = \frac{1}{t+1}$.

Next, assume that $x_t = \frac{1}{t}\sum_{i=1}^{t-1} \hat{z}_i$. Then, $x_{t+1} = \Pi_{\mathcal{X}}(y_t)$ where $y_t = x_t - \eta_t \hat{g}_t$ where $\hat{g}_t = \nabla f(x_t) - \hat{z}_t$. Hence, we have

$$y_t = \frac{1}{t}\sum_{i=1}^{t-1} \hat{z}_t - \eta_t \left(\frac{1}{t}\sum_{i=1}^{t-1} \hat{z}_i - \hat{z}_t \right) = \frac{1}{t+1}\sum_{i=1}^{t} \hat{z}_t.$$

Clearly, $y_t \in \mathcal{X}$, and therefore $x_{t+1} = y_t = \frac{1}{t+1}\sum_{i=1}^{t} \hat{z}_t$ as desired.

□

C Subgaussian noise extension

The main result in this section is a strengthening of Theorem 3.1 by weakening the bounded noise assumption on the stochastic gradient oracle. First, we require a definition.

Definition C.1. *A random variable X is said to be κ-subgaussian if $\mathrm{E}\left[\exp\left(X^2/\kappa^2\right)\right] \leq 2$. In additional, we say that X is κ-subgaussian conditioned on $\mathcal{F}$ if $\mathrm{E}\left[\exp\left(X^2/\kappa^2\right)\right] \leq 2$. Note that κ^2 in this setting may itself be a random variable.*

Remark C.2. *Note that the class of subgaussian random variables contains bounded random variables. Furthermore, this class also contains Gaussian random variables (which, of course, are not bounded). Therefore, the following theorem is indeed a strengthening of Theorem 3.1, which only dealt with stochastic gradient oracles that used almost surely bounded noise.*

Theorem C.3. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Suppose that $f : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex (with respect to $\|\cdot\|_2$) and L-Lipschitz. Assume that:*

(a) $g_t \in \partial f(x_t)$ for all t (with probability 1).

(b) $\|\hat{z}_t\|$ is κ-subgaussian conditioned on $\mathcal{F}_{t-1}$ for some $\kappa \in \mathbb{R}$.

Set $\eta_t = \frac{2}{\mu(t+1)}$. Let $\gamma_t = \frac{t}{T(T+1)/2}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have,

$$f\left(\sum_{t=1}^{T} \gamma_t x_t\right) - f(x^*) \;\leq\; O\left(\frac{(L+\kappa)^2}{\mu} \cdot \frac{\log(1/\delta)}{T}\right).$$

Proof (of Theorem C.3). We may follow the proof of Theorem 3.1 from Section 4 and remove any bound used on $\|\hat{g}_t\|^2$ to obtain

$$f\left(\sum_{t=1}^{T} \gamma_t x_t\right) - f(x^*) \;\leq\; \frac{2}{T(T+1)} \underbrace{\sum_{t=1}^{T} t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle}_{:=Z_T} + \frac{2}{\mu T(T+1)} \sum_{t=1}^{T} \|\hat{g}_t\|^2 . \qquad (5)$$

Theorem C.3 follows trivially from the following two lemmata:

Lemma C.4. *For any $\delta \in (0,1)$, $\sum_{t=1}^{T} \|\hat{g}_t\|^2 = O\left((L+\kappa)^2 T \cdot \log(1/\delta)\right)$ with probability at least $1-\delta$.*

Lemma C.5. *Let $Z_T = \sum_{t=1}^{T} t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle$. Then, for any $\delta \in (0,1)$, we have $Z_T = O\left(\frac{(L+\kappa)^2}{\mu} T \cdot \log(1/\delta)\right)$ with probability at least $1-\delta$.*

□

C.1 Proof of Lemma C.4

We begin with a fact about subgaussian random variables.

Claim C.6. *Let X be a random variable. Define $\|X\|_{\psi_2}$ as $\inf\left\{\, t > 0 \,:\, \mathrm{E}\left[\exp\left(X^2/t^2\right)\right] \leq 2 \,\right\}$. Then, $\|\cdot\|_{\psi_2}$ is a norm.*

Observe that X is κ-subgaussian if and only if $\|X\|_{\psi_2} \leq \kappa$. As a consequence of Claim C.6, we have the following claim.

Claim C.7. *There exists $\xi = O(L+\kappa)$, such that $\|\hat{g}_t\|$ is ξ-subgaussian conditioned on $\mathcal{F}_{t-1}$.*

Proof. Using the triangle inequality, we have

$$\|\hat{g}_t\| \;=\; \|g_t - \hat{z}_t\| \;\leq\; \|g\|_t + \|\hat{z}_t\| .$$

Therefore,

$$\|\|\hat{g}_t\| \mid \mathcal{F}_{t-1}\|_{\psi_2} \;\leq\; \|\|g_t\| \mid \mathcal{F}_{t-1}\|_{\psi_2} + \|\|\hat{z}_t\| \mid \mathcal{F}_{t-1}\|_{\psi_2} \;\leq\; \|\|g\|_t \mid \mathcal{F}_{t-1}\|_{\psi_2} + k,$$

because we assumed $\|\hat{z}_t\|$ is conditionally κ-subgaussian. Also, note that $\|g_t\|$ is conditionally $(L/\ln 2)$-subgaussian because f is L-Lipschitz and so $\|g_t\| \leq L$. □

Now, we proceed to prove Lemma C.4 using an MGF bound:

Claim C.8. *There exists $\xi = O(L+\kappa)$ such that $\mathrm{E}\left[\exp\left(\sum_{i=1}^{T} \|\hat{g}_i\|^2 / \left(T \cdot \xi^2\right)\right)\right] \leq 2$.*

Using Claim C.8 we can prove Lemma C.4:

Proof (of Lemma C.4). Using the exponentiated Markov inequality we have for any $\lambda > 0$:

$$\Pr\left[\sum_{t=1}^{T} \|\hat{g}_t\|^2 \geq x\right] \;\leq\; \frac{\mathrm{E}\left[\exp\left(\lambda \sum_{t=1}^{T} \|\hat{g}_t\|^2\right)\right]}{\exp(\lambda x)}.$$

Plugging in $\lambda = O\left(\frac{1}{T\cdot\xi^2}\right)$ and $x = O\left(T \cdot \xi^2 \log(1/\delta)\right)$ completes the proof. □

It remains to prove Claim C.8.

Proof (of Claim C.8). We will show that for every $1 \leq t \leq T$,

$$\mathrm{E}\left[\exp\left(\sum_{i=1}^{t} \|\hat{g}_i\|^2 / (T\cdot\xi^2)\right)\right] \;\leq\; 2^{1/T}\,\mathrm{E}\left[\exp\left(\sum_{i=1}^{t-1} \|\hat{g}_i\|^2 / (T\cdot\xi^2)\right)\right]. \tag{6}$$

Indeed,

$$\begin{aligned}
&\mathrm{E}\left[\exp\left(\sum_{i=1}^{t} \|\hat{g}_i\|^2 / (T\cdot\xi^2)\right)\right] \\
&= \mathrm{E}\left[\exp\left(\sum_{i=1}^{t-1} \|\hat{g}_i\|^2 / (T\cdot\xi^2)\right)\mathrm{E}\left[\exp\left(\|\hat{g}_t\|^2 / (T\cdot\xi^2)\right) \mid \mathcal{F}_{t-1}\right]\right].
\end{aligned}$$

Furthermore,

$$\mathrm{E}\left[\exp\left(\|\hat{g}_t\|^2 / (T\cdot\xi^2)\right) \mid \mathcal{F}_{t-1}\right] \leq 2^{1/T}$$

for all $1 \leq t \leq T$ using Claim C.7 and Jensen's inequality. Therefore, Eq. (6) is true for all $1 \leq t \leq T$.

Hence,

$$\mathrm{E}\left[\exp\left(\sum_{i=1}^{t} \|\hat{g}_i\|^2 / (T\cdot\xi^2)\right)\right] \;\leq\; \left(2^{1/T}\right)^T \;=\; 2,$$

as desired. □

C.2 Proof of Lemma C.5

We may follow the proof of Lemma 4.1 from Subsection 4.1. Define $d_t = t \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle$, $\tilde{v}_{t-1} := 2\kappa^2 \cdot t^2 \|x_t - x^*\|^2$, and $\tilde{V}_T = \sum_{t=1}^{T} \tilde{v}_{t-1}$. Note that $\tilde{v}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable.

Claim C.9. *For all t and $\lambda > 0$, we have*

$$\mathrm{E}\left[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\tilde{v}_{t-1}\right).$$

The proof of this requires a lemma from Vershynin [2018].

Lemma C.10 ([Vershynin, 2018, Proposition 2.5.2])**.** *Suppose X is a mean-zero random variable such that $\mathrm{E}\left[\exp\left(X/\kappa^2\right)\right] \leq 2$. Then, $\mathrm{E}\left[\exp(\lambda X)\right] \leq \exp\left(\lambda^2\kappa^2\right)$ for all $\lambda > 0$.*

Proof (of Claim C.9). Because $\|\hat{z}_t\|$ is κ-subgaussian conditioned on $\mathcal{F}_{t-1}$, we have by Cauchy-Schwarz

$$\mathrm{E}\left[\exp\left(\frac{t^2 \cdot \langle\, \hat{z}_t,\, x_t - x^* \,\rangle^2}{\kappa^2 \cdot t^2 \|x_t - x^*\|^2}\right) \mid \mathcal{F}_{t-1}\right] \;\leq\; \mathrm{E}\left[\exp\left(\|\hat{z}_t\|^2/\kappa^2\right) \mid \mathcal{F}_{t-1}\right] \;\leq\; 2.$$

Therefore, by Lemma C.10 we have

$$\mathrm{E}\left[\exp(\lambda) \mid \mathcal{F}_{t-1}\right] \;\leq\; \exp\left(\frac{\lambda^2}{2}(2\kappa^2 \cdot t^2 \|x_t - x^*\|^2\right),$$

as desired. □

To bound Z_T, we will proceed similarly as in Subsection 4.1. We will bound the TCV of Z_T by a linear combination of the increments. The only difference is, we will show that this bound holds with high probability, instead of with probability one. This will allow us to use a form of the Generalized Freedman Inequality (Theorem C.12) which the case where we can bound the total conditional variance by a linear transformation of the increments of the martingale with high probability.

Lemma C.11. *There exists non-negative constants* $\alpha_1, \ldots, \alpha_T = O\left((L+\kappa)^2 \frac{T}{\mu}\right)$ *and* $\beta = O\left(\frac{(L+\kappa)^4}{\mu^2} T^2\right)$ *such that for every* $\delta \in (0,1)$, $\tilde{V}_T \leq \sum_{t=1}^T \alpha_t d_t + \beta \log(1/\delta)$ *with probability at least* $1-\delta$.

Given Lemma C.11, we are ready to prove Lemma C.5. But first, we require a slightly more general version of the Generalized Freedman Inequality where the bound on the TCV by a linear transformation of the increments of the martingale holds only with arbitrarily high probability, rather than with probability 1.

Theorem C.12 (Generalized Freedman, [Harvey et al., 2018, Theorem 3.3])**.** *Let* $\{d_t, \mathcal{F}_t\}_{t=1}^T$ *be a martingale difference sequence. Suppose that, for* $t \in [T]$, v_{t-1} *are non-negative* $\mathcal{F}_{t-1}$*-measurable random variables satisfying* $\mathrm{E}\left[\exp(\lambda d_t) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2} v_{t-1}\right)$ *for all* $\lambda > 0$. *Let* $S_T = \sum_{t=1}^T d_t$ *and* $V_T = \sum_{t=1}^T v_{t-1}$. *Suppose there exists* $\alpha_1, \ldots, \alpha_T, \beta \in \mathbb{R}_{\geq 0}$ *such that for every* $\delta \in (0,1)$, $V_T \leq \sum_{t=1}^T \alpha_t d_t + \beta \log(1/\delta)$. *Let* $\alpha \geq \max_{t \in [T]} \alpha_t$. *Then*

$$\Pr\left[S_T \geq x\right] \;\leq\; \exp\left(-\frac{x^2}{4\alpha \cdot x + 8\beta}\right) + \delta.$$

Proof (of Lemma C.5). By Claim C.9 we have $\mathrm{E}\left[\exp\left(\lambda d_t\right) \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\tilde{v}_{t-1}\right)$. By Lemma C.11 we have that for every $\delta \in (0,1)$ $\tilde{V}_T \leq \sum_{t=1}^T \alpha_t d_t + \beta \log(1/\delta)$, with probability at least $1-\delta$. Plugging $\alpha = O\left((L+\kappa)^2 \frac{T}{\mu}\right)$, $\beta = O\left(\frac{(L+\kappa)^4}{\mu^2} T^2 \log(1/\delta)\right)$ and $x = O\left(\frac{(L+\kappa)^2}{\mu} \cdot T \log(1/\delta)\right)$ into Theorem C.12, proves Lemma C.5. □

It remains to prove Lemma C.11.

Proof (of Lemma C.11). Observe that $\tilde{V}_T = 2\kappa^2 V_T = \sum_{t=1}^T t^2 \cdot \|x_t - x^*\|^2$ where V_T was defined in Subsection 4.1. We focus our attention on bounding V_T, and then scale up accordingly at the end.

We may follow the proof of Lemma 4.3 with a key modification: Do not bound $\|\hat{g}_i\|$ by $L+1$ as this is no longer valid, because we no longer are using the bounded noise assumption.

This yields:

$$V_T \tag{7}$$

$$\leq \sum_{i=3}^{T-1} \underbrace{\frac{4}{\mu}\left(\sum_{t=i+1}^{T} t^2 \frac{a_i(t-1)}{i}\right)}_{:=\alpha_i} \cdot i \cdot \langle\, \hat{z}_i,\, x_i - x^* \,\rangle + \frac{4}{\mu^2} \underbrace{\sum_{t=4}^{T} t^2 \left(\sum_{i=3}^{t-1} b_i(t-1) \left\|\hat{g}_i\right\|^2\right)}_{:=G_T} + \frac{56L^2}{\mu^2}. \tag{8}$$

Define $\alpha_1, \alpha_2, \alpha_T = 0$. We already showed in the proof of Lemma 4.3 that $\alpha_i = O\left(\frac{T}{\mu}\right)$. Therefore, it remains to bound G_T by $O\left((L+\kappa)^2 T^2 \cdot \log(1/\delta)\right)$, with probability at least $1-\delta$. We rewrite G_T as

$$G_T \;=\; \sum_{i=3}^{T-1} \underbrace{\left(\sum_{t=i+1}^{T} t^2 b_i(t-1)\right)}_{:=s_i} \cdot \left\|\hat{g}_i\right\|^2 \;=\; \sum_{i=3}^{T-1} s_i \left\|\hat{g}_i\right\|^2.$$

We use the following MGF bound on G_T, which we prove below.

Claim C.13. $\mathrm{E}\left[\exp\left(\lambda G_T\right)\right] \leq \exp\left(\lambda O\left(\xi^2\right)\sum_{i=3}^{T-1} s_i\right)$ *for all* $\lambda = O\left(\frac{1}{\xi^2 \max\{s_i\}}\right)$.

Therefore, via an exponentiated Markov inequality and Claim C.13, we have

$$\Pr\left[G_T \geq x\right] \;\leq\; \frac{\mathrm{E}\left[\exp\left(\lambda G_T\right)\right]}{\exp\left(\lambda x\right)} \;\leq\; \exp\left(\lambda O\left(\xi^2\right)\sum_{i=3}^{T-1} s_i - \lambda x\right).$$

Setting $\lambda = O\left(\frac{1}{\xi^2\sum_{i=3}^{T-1} s_i}\right)$ and $x = O\left(\xi^2\sum_{i=3}^{T-1} s_i \cdot \log(1/\delta)\right)$ shows $G_T \leq O\left(\xi^2\sum_{i=3}^{T-1} s_i \cdot \log(1/\delta)\right)$ with probability at least $1-\delta$. Observe that $s_i = O(T)$:

$$s_i \;=\; \sum_{t=i+1}^{T} t^2 b_i(t-1) \;=\; \sum_{t=i+1}^{T} t^2 O\left(\frac{i^2}{t^4}\right) \;=\; \sum_{t=i+1}^{T} O\left(1\right) \;=\; O\left(T\right).$$

Therefore $x = O\left(\xi^2 T^2 \log(1/\delta)\right) = O\left((L+\kappa)^2 T^2 \log(1/\delta)\right)$, with probability at least $1-\delta$. That is, with probability at least $1-\delta$

$$G_T \leq O\left(\left(L+\kappa\right)^2 T^2 \log(1/\delta)\right).$$

Plugging this back in to Eq. (7), we obtain

$$V_T \;\leq\; \sum_{i=1}^{T} \alpha_i d_i + \beta \log(1/\delta)$$

where $\alpha_i = O(\frac{T}{\mu})$ and $\beta = O\left(\frac{(L+\kappa)^2}{\mu^2} T^2 \cdot \log(1/\delta)\right)$. Multiplying both sides by $2\kappa^2$ yields the desired bound on $\tilde{V}_T$.

□

Now it remains to prove Claim C.13.

Proof (of Claim C.13). We will show that for every t and for all $\lambda \leq 1/\max\{s_i\}$,

$$\mathrm{E}\left[\exp\left(\lambda \sum_{i=3}^{t} s_i \left\|\hat{g}_i\right\|^2\right)\right] \leq \exp\left(\lambda O\left(\xi^2\right) s_t\right) \mathrm{E}\left[\exp\left(\lambda \sum_{i=3}^{t-1} s_i\right)\right].$$

The claim then follows by recursively applying the above inequality. Note that s_i is $\mathcal{F}_{i-1}$ measurable. So, we have

$$\mathrm{E}\left[\exp\left(\lambda \sum_{i=3}^{t} s_i \left\|\hat{g}_i\right\|^2\right)\right] \;=\; \mathrm{E}\left[\exp\left(\lambda \sum_{i=3}^{t-1} s_i \left\|\hat{g}_i\right\|^2\right) \mathrm{E}\left[\exp\left(\lambda s_t \left\|\hat{g}_t\right\|^2\right) \;\middle|\; \mathcal{F}_{t-1}\right]\right].$$

Note that because $\left\|\left\|\hat{g}_t\right\| \mid \mathcal{F}_{t-1}\right\|_{\psi_2} \leq \xi$, this implies

$$\mathrm{E}\left[\exp\left(\left\|\hat{g}_t\right\|^2 / \xi^2\right) \;\middle|\; \mathcal{F}_{t-1}\right] \leq 2.$$

Therefore, if $\lambda = O\left(\frac{1}{\xi^2}\right)$. Then by Jensens inequality, raising both sides of the above inequality to the power of $\lambda\xi^2$ yields

$$\mathrm{E}\left[\exp\left(\lambda \left\|\hat{g}_t\right\|^2\right) \;\middle|\; \mathcal{F}_{t-1}\right] \leq \exp\left(\lambda O\left(\xi^2\right)\right).$$

Hence, if $\lambda = O\left(\frac{1}{\xi^2 \max\{s_i\}}\right)$, then for every t we have

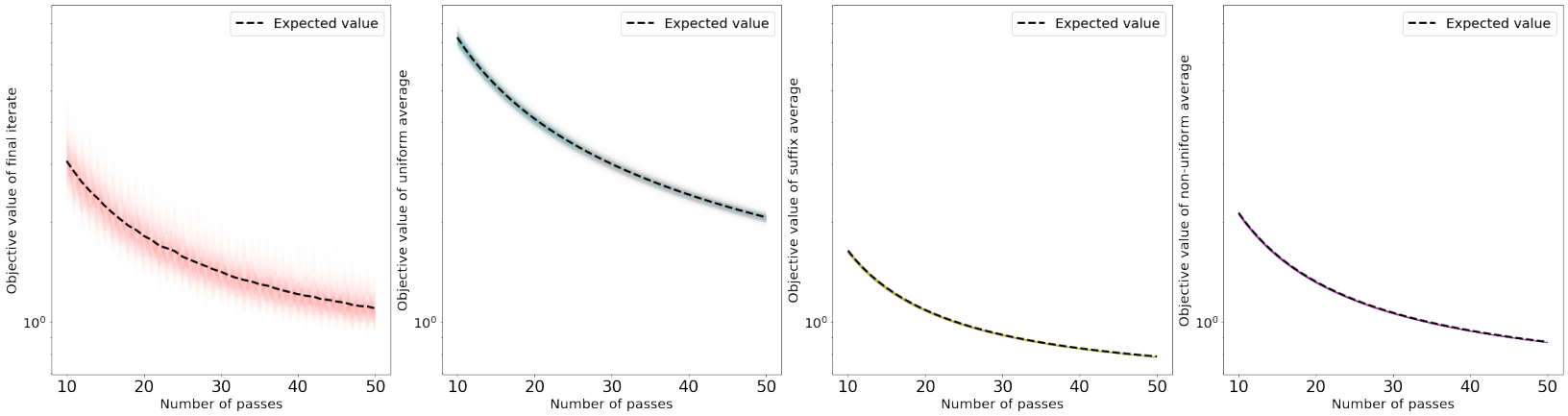
$$\mathrm{E}\left[\exp\left(\lambda s_t \left\|\hat{g}_t\right\|^2\right) \;\middle|\; \mathcal{F}_{t-1}\right] \leq \exp\left(\lambda O\left(\xi^2\right) s_t\right),$$

which completes the proof. □

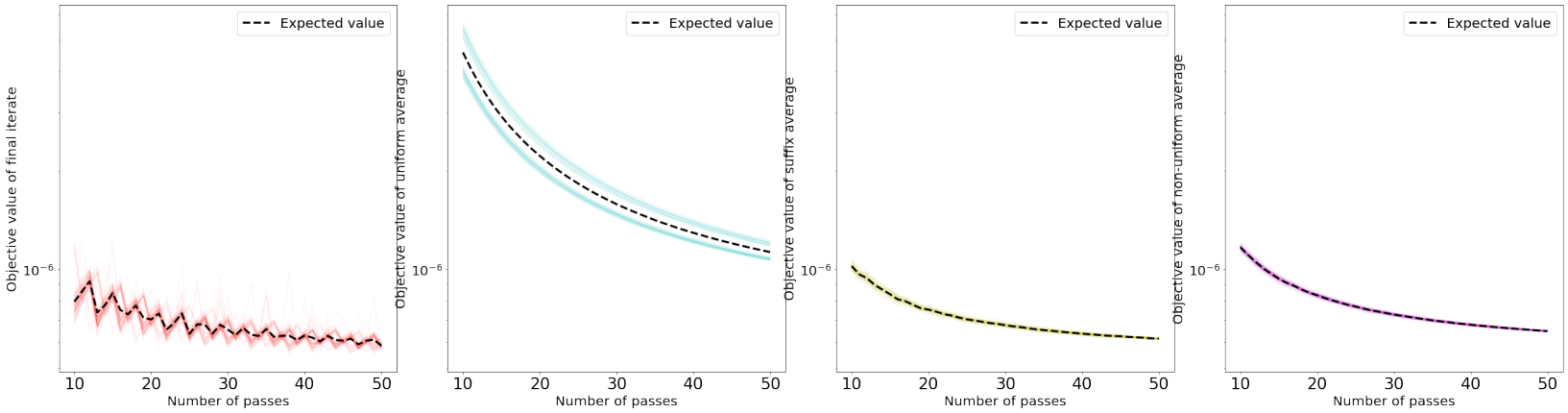
D Additional experiments

In each experiment we run SGD for the regularized SVM optimization problem described in Section 6. We use regularization parameter $\lambda = 1/n$ and step size $\eta_t = \frac{2}{\mu(t+1)}$. For each return strategy, we run many trials of SGD and plot the objective value over time for every trial. At any point in time, the darkness of the plot at a specific objective value indicates the number of trials that achieved that value at that time.

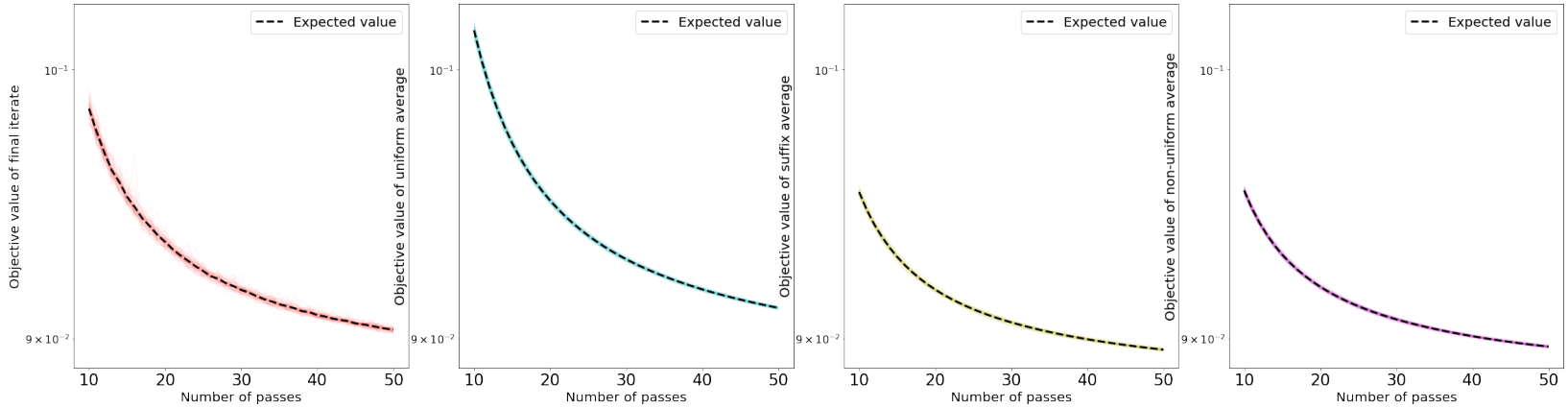
We use the freely available data sets *quantum* ($m = 50000$ and $n = 78$), *covtype* ($m = 581012$ and $n = 54$) and *rcv1* ($m = 20242$ and $n = 47236$). We run 1000 trials of SGD on the *quantum* data set, 80 trials of SGD on the *covtype* data set and 70 trials of SGD on the *rcv1* data set. The *quantum* data set can be found at the KDD cup 2004 website and *covtype* and *rcv1* can be found at the LIBSVM website.



(a) quantum



(b) covtype (80 trials)



(c) rcv1 (70 trials)

Figure 2: Number of effective passes vs. objective value. Figure 2a plots the results for the *quantum* dataset; Figure 2b plots the results for *covtype* dataset; Figure 2c plots the results for the *rcv1* dataset. From left to right, we plot the objective value over time of the final iterate, uniform average, suffix average and non-uniform average for 1000 trials of SGD.