On Step Sizes and Convergence Rates for SGD

Mark Oussoren
Department of Statistics
University of California, Berkeley
mark_oussoren@berkeley.edu

Abstract

SGD is the method of choice for large-scale optimization problems and the final iterate in particular is returned far more often than ensembles of iterates in practice. There is recent and beautiful literature on step-size choices and suboptimality analysis of the final iterate of SGD which motivates my discussion. In particular, I would like to explore the work of Shamir and Zhang [7] on suboptimality analysis of the final iterate as well as recent work by Harvey et al. [1] and Jain et al. [2] on tighter bounds by leveraging martingale theory and modern concentration inequalities.

1 Introduction

Here, we consider solving optimization problems of the form

$$\min_{x\in\mathcal{X}} \{f(x) = \mathbb{E}(F(x,\xi))\} \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a nonempty bounded closed convex set with diameter D, ξ is a random vector equipped with a probability measure $\mathbb{P}$ supported on $\Omega \subset \mathbb{R}^d$, and $F : \mathcal{X} \times \Omega \to \mathbb{R}$. In this setting, we assume $f(x)$ is finite and well-defined for all $x \in \mathcal{X}$ and moreover continuous and convex over $\mathcal{X}$. Under these assumptions, $f(\cdot)$ has a global minimum whose input we denote by $x^* \in \mathcal{X}$. Furthermore, we assume the existence of an oracle $\mathcal{O}$ mapping all $(x,\xi) \in (\mathcal{X},\Omega)$ to a stochastic subgradient $\hat{g}(x,\xi)$ satisfying

1. Unbiasedness: $\mathbb{E}(\hat{g}(x,\xi)) = g(x)$ for all $x \in \mathcal{X}$ and some $g(x) \in \partial f(x)$
2. Uniformly Bounded Second Moment: $\mathbb{E}\left(\|\hat{g}(x,\xi)\|^2\right) \leq G^2$ for all $x \in \mathcal{X}$.

The overarching goal of this paper is in reviewing the effectiveness of stochastic gradient descent (SGD), one of the oldest, simplest, and most popular methods, for solving (1). Loosely speaking, SGD operates by taking a small step at each iteration in the opposite direction of an unbiased estimate of the subgradient of $f(\cdot)$. Formally, SGD's update is characterized by

$$x_{t+1} \leftarrow \Pi_{\mathcal{X}}\left(x_t - \eta_t \hat{g}(x_t,\xi_t)\right)$$

where $\eta_t \in \mathbb{R}$ denotes the step-size at iteration t satisfying 1) $\eta_t \to 0$ as $t \to \infty$ and 2) $\sum_t \eta_t = \infty$, and $\xi_t \perp\!\!\!\perp \xi_u$ for $u \neq t$. Because of the randomness in the update, $\{x_t\}_{t\in\mathbb{N}}$ does not form a relaxation sequence. As a consequence, ensembles of iterates are generally returned as opposed to the final one. A natural, critical question stemming from the update above is phrased as: *how exactly do we find a sequence $\{\eta_t\}_{t\in\mathbb{N}}$ guaranteeing an ϵ-approximate solution of (1)?* This is exactly the primary focus of the paper, with emphasis on reviewing answers to the question in the case of returning the final iterate for general, convex $f(\cdot)$. Moreover, my goal is to review recent literature on SGD step-size regimes for constructing $(1-\alpha)$ probability bounds based on T queries of $\mathcal{O}$ satisfying

$$P\left(f(x_T) - f(x^*) \leq \epsilon_T(\alpha)\right) \geq 1-\alpha$$

for all $\alpha \in (0,1)$ and $\epsilon_T(\alpha) > 0$.

2 Preliminaries

2.1 Stochastic Approximation

To motivate some of the more recent results proven for SGD, I briefly review the broader class of stochastic algorithms studied extensively in probability and control theory from which most of modern day stochastic optimization is grounded.

Definition 2.1 (Stochastic Algorithm)**.** A stochastic algorithm is defined by:

$$x_{t+1} \leftarrow x_t - \gamma_{t+1}\left(h(x_t) - \Delta M_{t+1} + R_{t+1}\right)$$

where $\{\Delta M_t\}_{t\in\mathbb{N}}$ is a martingale difference sequence, $\{R_t\}_{t\in\mathbb{N}}$ is a sequence of perturbations, and $\{\gamma_t\}_{t\in\mathbb{N}}$ is a sequence of non-negative step sizes such that $\gamma_t \to 0$ as $t \to \infty$ and $\sum_t \gamma_t = \infty$.

Relevant concepts in probability supporting this definition are found in the appendix. With this definition, Robbins and Siegmund [5] proved the following theorem underpinning convergence of stochastic algorithms.

Theorem 2.1 (Robbins-Siegmund Theorem)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ be a sequence of sub σ-fields of $\mathcal{F}$. Let U_t, β_t, ξ_t, ζ_t be non-negative $\mathcal{F}_t$-measurable random variables such that

$$\mathbb{E}\left[U_{t+1}|\mathcal{F}_t\right] \leq (1+\beta_t)U_t + \xi_t - \zeta_t, \quad t \in \mathbb{N}.$$

Then on $\{\sum_t \beta_t < \infty, \sum_t \xi_t < \infty\}$, U_t converges a.s. to a random variable and $\sum_t \xi_t < \infty$ a.s.

This beautiful theorem appears to have an illusive appearance (if any) in modern day textbooks on optimization despite its proof ingredients remaining the basis for suboptimality analysis in literature today of the final iterate (see [1], [2], [3]). SGD can easily be seen as a stochastic algorithm and below I show the errors form a martingale difference sequence in the following lemma.

Lemma 2.2. For SGD as described above, let $\varepsilon_t = \hat{g}(x_t, \xi_t) - g(x_t)$ and $S_t = \sum_{i=1}^t \varepsilon_t$. Then, letting $\mathcal{F}_t = \sigma\left(\varepsilon_1, \ldots, \varepsilon_t\right)$ denote the σ-algebra generated by the first t errors, $\langle (S_t, \mathcal{F}_t) \rangle$ is a martingale difference sequence.

Proof. The proof is trivial, but fruitful to see once. By assumption, we have that $\mathbb{E}[S_t] < \infty$ as $\mathbb{E}[\varepsilon_l] = 0$ for all $l \in [t]$. It remains to show the sequence emulates a fair game: for all $t \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[S_t|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{i=1}^t \varepsilon_i|\mathcal{F}_{t-1}\right] && \\
&= \mathbb{E}\left[S_{t-1}|\mathcal{F}_{t-1}\right] + \mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}] && \searrow \textit{linearity of } \mathbb{E}(\cdot) \\
&= S_{t-1} + \mathbb{E}[\varepsilon_t] && \searrow \hat{g}_m \perp\!\!\!\perp \hat{g}_n \textit{ for } m \neq n \\
&= S_{t-1} && \searrow \hat{g} \textit{ is unbiased}
\end{aligned}$$

□

Thus, we can apply Robbins-Siegmund theorem to prove convergence of the SGD algorithm. Since [5] was published, martingale theory has blossomed with newer concentration inequalities (Freedman's and many other generalized exponential inequalities) giving rise to renewed efforts in understanding the algorithm's convergence and probabilities of large deviation.

2.2 Jain's Step Size

Here, I would like to introduce the step-size regime proposed by Jain et al. [2] in a concrete formulation. In the case of general convex functions, define

$$k = \inf\left\{i \in \mathbb{N} : \frac{T}{2^i} \leq i\right\}, \quad T_i = T - \left\lceil \frac{T}{2^i} \right\rceil, \quad T_{k+1} = T.$$

Then the step-size regime is given by

$$\alpha(t) = \frac{C}{\sqrt{T}} \cdot \frac{1}{2^i} \tag{2}$$

for a constant $C > 0$, $T_i \leq t \leq T_{i+1}$, and $i \in [k]$. As seen below, the step size regime decays gradually over time from T_0 at $\frac{C}{\sqrt{T}}$ to T_1 and so on towards zero. This regime contrasts with the

smoothed stochastic line search and variants where there is a warm-up period before a geometrically or polynomially decaying sequence. However, as empirically demonstrated by Zhang et al. [8], the warm-up period generally constitutes a very small number of epochs and thus, the step size schedule in [2] looks very similar to [8].

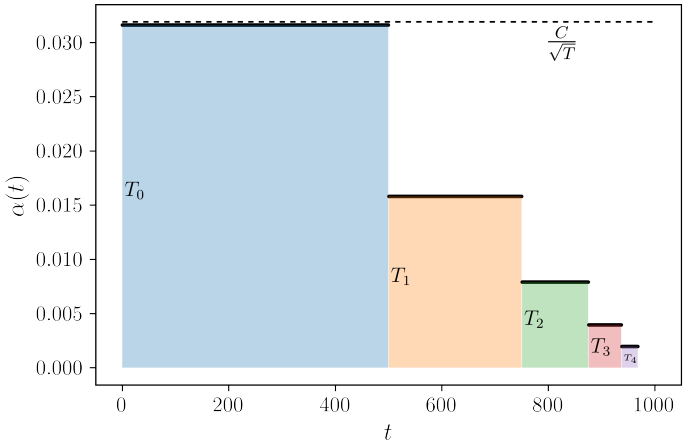


Figure 1: Jain's Step Size Schedule ($C = 1$ and $T = 1000$)

A rough intuition for this regime can be that they want to find a good point x_t and once this point is found, they desire to bound its variance from that point onward - hence the tapering of the step. However, why this regime is information-theoretically optimal or even just better than a constant $C/\sqrt{t}$ is by no means intuitive and the discussion and empirical evidence is deferred to the experiments and appendix sections.

3 Expected Suboptimality Bounds

Here I review the work of [1], [2], and [7]. A summary of existing work can be seen in the below table where suboptimality rates are given in expectation (several lower bound results are included for completeness).

Work	Step Size	Convexity	Assumptions	Rate
Nemirovski et al. [4]	$1/\sqrt{t}$	Convex	Smooth	$\Theta(1/T)$
Nemirovski et al. [4]	$1/\lambda t$	Strongly-Convex	Smooth	$\Theta(1/\sqrt{T})$
Shamir and Zhang [7]	$1/\sqrt{t}$	Convex		$\Theta(/\sqrt{T})$
Shamir and Zhang [7]	$1/\lambda t$	Strongly-Convex		$\Theta(\log(T)/T)$
Harvey et al. [1]	$1/\sqrt{t}$	Convex	GD, not SGD; $n = T$	$\Omega(\log(T)/\sqrt{T})$
Harvey et al. [1]	$1/\lambda t$	Strongly-Convex	GD, not SGD; $n = T$	$\Omega(\log(T)/T)$
Jain et al. [2]	2	Convex	T known	$\Theta(1/\sqrt{T})$
Jain et al. [2]	2	Strongly-Convex	T known	$\Theta(1/T)$
Liu and Lu [3]	$1/\sqrt{t}$	Convex	GD, not SGD; $n \leq T$	$\Omega(\log(n)/\sqrt{T})$
Liu and Lu [3]	$1/\lambda t$	Strongly-Convex	GD, not SGD; $n \leq T$	$\Omega(\log(n)/T)$

Table 1: Suboptimality of the Gradient Method's Final Iterate

While I will not discuss lower bounds in this review, Liu and Lu [3]'s results for the final iterate of GD are especially interesting in that they round off the claim made by Harvey et al. [1] that the lower bound scales logarithmically with the dimension of the domain and proved it for $n \leq T$ by harnessing results on hitting times.

3.1 Basic Decreasing Step Size

Shamir and Zhang [7] were the first (to the best of my knowledge) to develop suboptimality bounds for SGD's final iterate (however Zhang actually proved the result in 2004). Below, I work out the main result of their paper in a bit more detail.

Theorem 3.1 (Theorem 2 in [7])**.** For SGD with constant step size $\eta_t = 1/\sqrt{t}$, then for $T > 1$

$$\mathbb{E}\left[f(x_T) - f(x^*)\right] \leq \left(D^2 + 2G^2\right)\frac{2 + \ln(T)}{2\sqrt{T}}$$

Proof. For all $x \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x\|^2\right] &= \mathbb{E}\left[\|\textstyle\prod_{\mathcal{X}}(x_t - \eta_t \hat{g}(x_t, \xi_t)) - x\|^2\right] \\
&\leq \mathbb{E}\left[\|x_t - \eta_t \hat{g}(x_t, \xi_t) - x\|^2\right] && \text{as } \textstyle\prod_{\mathcal{X}}(\cdot) \text{ is non-expanding} \\
&\leq \mathbb{E}\left[\|x_t - x\|^2 - \langle \hat{g}(x_t, \xi_t), x_t - x\rangle + \eta_t^2 \|\hat{g}(x_t, \xi_t)\|^2\right] && \text{triangle inequality} \\
&\leq \mathbb{E}\left[\|x_t - x\|^2\right] - 2\eta_t \mathbb{E}\left[\langle g(x_t), x_t - x\rangle\right] + \eta_t^2 G^2 && \text{by subgradient assumptions} \\
&\Leftrightarrow \mathbb{E}\left[\langle g(x_t), x_t - x\rangle\right] \leq \tfrac{1}{2\eta_t}\left(\mathbb{E}\left[\|x_t - x\|^2\right] - \mathbb{E}\left[\|x_{t+1} - x\|^2\right]\right) + \tfrac{\eta_t G^2}{2}.
\end{aligned}$$

From here, we let $k \in [T-1]$ and sum over the last $k+1$ of these inequalities to obtain

$$\sum_{t=T-k}^{T} \mathbb{E}\left[\langle g(x_t), x_t - x\rangle\right] \leq \sum_{t=T-k}^{T} \frac{\mathbb{E}\left[\|x_t - x\|^2\right] - \mathbb{E}\left[\|x_{t+1} - x\|^2\right]}{2\eta_t} + \frac{\eta_t G^2}{2}$$

$$= \frac{1}{2\eta_{T-k}}\mathbb{E}\left[\|x_{T-k} - x\|^2\right] + \frac{1}{2}\sum_{t=T-k+1}^{T} \mathbb{E}\left[\|x_t - x\|^2\right]\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=T-k}^{T}\eta_t.$$

As $f(\cdot)$ is convex, and $\eta_t = 1/\sqrt{t}$, we further have

$$\sum_{t=T-k}^{T} \mathbb{E}\left[f(x_t) - f(x)\right] \leq \sum_{t=T-k}^{T} \mathbb{E}\left[\langle g(x_t), x_t - x\rangle\right]$$

$$\leq \frac{\sqrt{T-k}}{2}\mathbb{E}\left[\|x_{T-k} - x\|^2\right] + \frac{1}{2}\sum_{t=T-k+1}^{T} \mathbb{E}\left[\|x_t - x\|^2\right]\left(\sqrt{t} - \sqrt{t-1}\right) + \frac{G^2}{2}\sum_{t=T-k+1}^{T}\frac{1}{\sqrt{t}}.$$

Here comes the crucial trick that deviates from standard SGD suboptimality proofs that let $x = x^*$. Instead, we take $x = x_{T-k}$ and use the fact the diameter of $\mathcal{X}$ is bounded by D to obtain

$$\sum_{t=T-k}^{T} \mathbb{E}\left[f(x_t) - f(x_{T-k})\right] \leq \frac{1}{2}\sum_{t=T-k+1}^{T} \mathbb{E}\left[\|x_t - x_{T-k}\|^2\right]\left(\sqrt{t} - \sqrt{t-1}\right) + \frac{G^2}{2}\sum_{t=T-k}^{T}\frac{1}{\sqrt{t}}$$

$$\leq \frac{D^2}{2}(\sqrt{T} - \sqrt{T-k}) + \frac{G^2}{2}\sum_{T-k}^{T}\frac{1}{\sqrt{t}} \leq \frac{D^2 + 2G^2}{2}(\sqrt{T} - \sqrt{T-k-1}).$$

As $\left(\sqrt{T} - \sqrt{T-(k+1)}\right)\left(\sqrt{T} + \sqrt{T-(k+1)}\right) = k+1$, we divide both sides of the above inequality by $k+1$ to obtain

$$\frac{1}{k+1}\sum_{t=T-k}^{T} \mathbb{E}\left[f(x_t) - f(x_{T-k})\right] \leq \frac{D^2 + 2G^2}{2\left(\sqrt{T} + \sqrt{T-(k+1)}\right)}.$$

Let $S_k = \frac{1}{k+1}\sum_{t=T-k}^{T}\mathbb{E}[f(x_t)]$, then we expand the RHS and rearrange the sum as

$$-\mathbb{E}[f(x_{T-k})] \leq -S_k + \frac{D^2 + 2G^2}{2\left(\sqrt{T} + \sqrt{T-(k+1)}\right)} \leq -S_k + \frac{D^2 + 2G^2}{2\sqrt{T}}.$$

Adding $(k+1)S_k$ to both sides, the above simplifies to

$$(k+1)S_k - \mathbb{E}[f(x_{T-k})] = kS_{k-1} \leq kS_k + \frac{D^2 + 2G^2}{2\sqrt{T}}$$

$$\iff S_{k-1} \leq S_k + \frac{D^2 + 2G^2}{2\sqrt{T}}.$$

Recursively applying this inequality and subtracting both sides by $\mathbb{E}[f(x^*)]$ yields

$$\mathbb{E}\left[f(x_T) - f(x^*)\right] = S_0 - \mathbb{E}[f(x^*)] \leq S_{T-1} - \mathbb{E}[f(x^*)] + \frac{D^2 + 2G^2}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[f(x_t) - f(x^*)\right] + \frac{D^2 + 2G^2}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k}.$$

Now we are left with the task of bounding the sum on the RHS. Reusing our general convexity argument above where $x = x^*$ and $k = T - 1$ gives us that

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[f(x_t) - f(x^*)\right] \leq \frac{1}{2T}\left(D^2 + 2G^2\right)\left(\sqrt{T} - \sqrt{T - (T-1) - 1}\right) \leq \frac{D^2 + 2G^2}{2\sqrt{T}}. \quad (3)$$

Finally, the last ingredient to obtain the result is a crude bound for $\sum_{k=1}^{T-1} \frac{1}{k}$ which can be obtained easily as

$$\sum_{k=1}^{T-1} \frac{1}{k} \leq \sum_{k=1}^{T} \frac{1}{k} = \sum_{k=1}^{T} \frac{1}{k} \int_{k-1}^{k} dx = 1 + \sum_{k=2}^{T} \frac{1}{k} \int_{k-1}^{k} dx \leq 1 + \sum_{k=2}^{T} \int_{k-1}^{k} \frac{1}{x} dx$$

$$= 1 + \sum_{k=2}^{T} \ln(k) - \ln(k-1) = 1 + \ln(T).$$

Plugging these two bounds into our suboptimality inequality rounds off the last touches on the proof:

$$\mathbb{E}\left[f(x_T) - f(x^*)\right] \leq \frac{D^2 + 2G^2}{2\sqrt{T}} + \frac{D^2 + 2G^2}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k} \leq \frac{D^2 + 2G^2}{2\sqrt{T}} \left(2 + \ln(T)\right)$$

□

3.2 Jain's Step Size

In this section, I review the convergence of the final iterate in the case of Jain's step size defined here. The constants in their analysis are actually slightly off at the end - the factor in front of G^2 should be 12 not 11 and 2 not 4 for D^2, and I correct this below in theorem 3.3.

Theorem 3.2 (Theorem 3 in [2])**.** Let $\{y_t\}_{t=1}^T$ be the iterates of SGD with a decreasing step size with $\beta = 1$ decay and $\{x_t\}_{t=1}^T$ be the iterates of SGD with step size defined in 2. Then for all $T \geq 4$,

$$\mathbb{E}\left[f(x_T)\right] - \inf_{\lceil \frac{T}{4} \rceil \leq t \leq T_1} \mathbb{E}\left[f(y_t)\right] \leq \frac{10G^2}{\sqrt{T}}$$

Theorem 3.3 (Theorem 1 in [2])**.** For SGD with the step size defined in 2 and $C = 1$, then for $T \geq 4$

$$\mathbb{E}\left[f(x_T) - f(x^*)\right] \leq \frac{2D^2 + 12G^2}{\sqrt{T}}$$

Proof. We first apply Theorem 3.2, to obtain

$$\begin{aligned}
\mathbb{E}\left[f(x_T) - f(x^*)\right] &\leq \inf_{\lceil \frac{T}{4} \rceil \leq t \leq T_1} \mathbb{E}\left[f(y_t)\right] - f(x^*) + \frac{10G^2}{\sqrt{T}} && \Big) \textit{ mean} \geq \textit{infimum} \\
&\leq \frac{10G^2}{\sqrt{T}} + \frac{1}{T_1 - \lceil \frac{T}{4} \rceil + 1} \sum_{t=\lceil \frac{T}{4} \rceil}^{T_1} \mathbb{E}\left[f(y_t) - f(x^*)\right] && \Big) T_1 \leq 2(T_1 - \lceil \tfrac{T}{4} \rceil + 1) \\
&\leq \frac{10G^2}{\sqrt{T}} + \frac{2}{T_1} \sum_{t=\lceil \frac{T}{4} \rceil}^{T_1} \mathbb{E}\left[f(y_t) - f(x^*)\right] && \Big) y_t \geq x^* \\
&\leq \frac{10G^2}{\sqrt{T}} + \frac{2}{T_1} \sum_{t=1}^{T_1} \mathbb{E}\left[f(y_t) - f(x^*)\right] && \Big) \textit{ standard: see 3} \\
&\leq \frac{10G^2}{\sqrt{T}} + \frac{1}{\sqrt{T_1}} \left[D^2 + G^2\right] && \Big) \lceil \tfrac{T}{4} \rceil \leq T_1 \leq \lceil \tfrac{T}{2} \rceil \\
&\leq \frac{10G^2}{\sqrt{T}} + \frac{2}{\sqrt{T}} \left[D^2 + G^2\right] \\
&= \frac{2D^2 + 12G^2}{\sqrt{T}}
\end{aligned}$$

□

Notice that the removal of the $\log(T)$ term in the numerator ensures that these bounds are information-theoretically optimal. I constrained the step size here for the case where $\gamma_t = 1/\sqrt{T}$, but the paper generalizes this result to any decreasing step size with at most polynomial decay. The only minor discrepancy with Jain's claim that I mentioned is that in the standard analysis with constant step size $1/\sqrt{T}$, we have $\mathbb{E}\left[f(x_T) - f(x^*)\right] \leq \frac{D^2+G^2}{2\sqrt{T}}$. High probability bounds are much more difficult to develop, and I explicate key motifs for this type of analysis in the appendix for sake of space.

4 Experiments

In this section, I review convergence of SGD for several step size regimes and optimization problems. Jain et al. [2] experimented with SGD under LASSO regression; however, I believe this is a meaningless (and most likely cherry-picked) experiment as Schmidt et al. [6] have analyzed many more suitable methods for this type of problem. Instead, I will consider a stochastic utility problem as done by Nemirovski et al. [4], alongside a support vector machine (SVM) problem as done by Jain et al. [2] and much of literature. I consider the following step sizes in each case (replacing T with $\sqrt{T}$ in the case of general, convex f)

- Constant step size: $\eta_t = \frac{1}{T}$ returning the final iterate x_T
- Constant decay: $\eta_t = \frac{1}{t}$ returning the final iterate x_T
- Jain's step size: defined 2 returning the final iterate x_T
- Constant step size: $\eta_t = \frac{1}{T}$ returning the average of iterates $T^{-1} \sum_{t=1}^{T} x_t$.

4.1 SVM

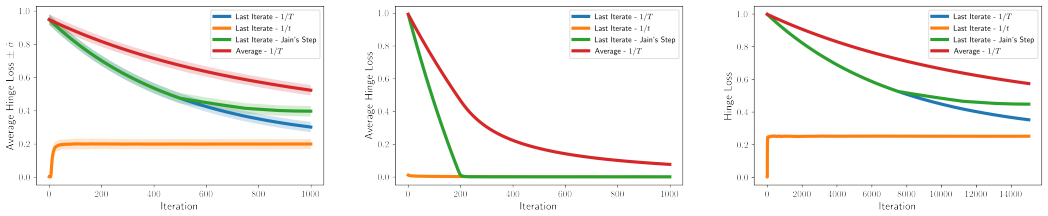
In this section, I consider applying SGD to the soft-margin SVM problem with 500 samples under the hinge-loss where $x_i \sim \mathcal{N}(0, \sigma^2 I_{30})$, $y_i = \text{sign}(x_i[1] + z_i)$, and $z_i \sim \mathcal{N}(0, 2)$:

$$\min_{w \in \mathbb{R}^{30}, b \in \mathbb{R}} \left\{ w^T w + C \sum_{i=1}^{m} \max\{0, (1 - \epsilon_i) - y_i(w^T x_i - b)\} \right\}.$$

Then we compute the subderivatives of the objective to arrive at the gradient method update for each sample (x_i, y_i):

$$\begin{bmatrix} w_{t+1} \\ b_{t+1} \end{bmatrix} \leftarrow \begin{cases} \begin{bmatrix} w_t - \eta_t(2w_t - Cy_i x_i) \\ b_t - \eta_t C y_i \end{bmatrix} & 1 - \epsilon_i > y_i(w_t^T x_i - b_t) \\ \begin{bmatrix} w_t(1 - 2\eta_t) \\ b_t \end{bmatrix} & \text{Otherwise.} \end{cases}$$

From this, I mirror Jain's setup by considering $C = \frac{1}{n}$ and $b = 0$ (as the noise is centered). However, I deviate here from his experiment by considering intervals of the parameters T, ϵ, and σ^2. The justification for this deviation and robustness check is clear: 1) we never exactly know T, 2) we also usually do not implement hard-margin SVM solutions in practice, and 3) it would be interesting to analyze the robustness of the step sizes with varying variances. As displayed in the first figure, it



(a) (b) (c)

Figure 2: SGD SVM Robustness Checks. (a) Varying $\epsilon \in [10^{-7}, 0.1]$; (b) Varying $\sigma^2 \in [1, 1000]$ (here $\sigma^2 = 10$); (c) Varying $T \in [1000, 100, 000]$ (here $T = 10,000$).

is evident that Jain's step-size is very dependent on knowledge of T and his experiments were not robust with respect to the parameters. An interesting point that I do not have great intuition for is that Jain's step and the average step are much more robust in hinge loss for X generated with large variance. In the last figure (c), it is apparent that over time, the constant $1/T$ tends towards the $1/t$ regime; however, the computational power required for this problem makes this a very unlikely favorite for practitioners (Jain considers a very, very large T that is extremely expensive). More figures stemming from this experiment can be found in the notebook here.

4.2 Stochastic Utility

The second problem I will investigate is in the optimization of a stochastic utility model similar to an experiment done by Nemirovski et al. [4] for analyzing mirror stochastic approximation and variants (I replace i.i.d normals which allow for trivial computation with independent exponentials with varying rates):

$$\min_{x \in \mathcal{X}} \left\{ f(x) = \mathbb{E}\left[\phi\left(\sum_{i=1}^{n} \xi_i x_i \right) \right] \right\}$$

where $\mathcal{X} = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^{n} x_i = 1\}$, $\xi_i \sim \text{Exp}(i)$ are independent, and $\phi(\cdot)$ is a piece-wise linear convex function given by $\phi(t) = \max\{s_1 t, \ldots, s_m t\}$ for a known constant sequence $\{s_k\}_{k=1}^m$ with $s_k \geq 0$. This setup allows the problem to be reduced easily as

$$\sum_{i=1}^{n} \xi_i x_i \sim \text{Exp}\left(\sum_{i=1}^{n} i x_i \right), \quad \phi\left(\sum_{i=1}^{n} \xi_i x_i \right) = \max_{k \in [m]} \left\{ s_k \sum_{i=1}^{n} i x_i \right\}$$

where we can further deduce, denoting $\mathcal{P}[m]$ as the powerset of the set of the first m natural numbers excluding the empty set, that

$$\mathbb{E}\left[\max_{k \in [m]} \left\{ s_k \sum_{i=1}^{n} \xi_i x_i \right\} \right] = \int_0^\infty 1 - \prod_{k=1}^{m} \left(1 - \exp\{-y s_k \sum_{i=1}^{n} i x_i\} \right) dy$$

$$= \sum_{\mathcal{K} \in \mathcal{P}[m]} (-1)^{|\mathcal{K}|-1} \int_0^\infty \exp\{-y \sum_{k \in \mathcal{K}} s_k \sum_{i=1}^{n} i x_i\} dy = \sum_{\mathcal{K} \in \mathcal{P}[m]} (-1)^{|\mathcal{K}|-1} \left(\sum_{k \in \mathcal{K}} s_k \sum_{i=1}^{n} i x_i \right)^{-1}.$$

This translates to an interesting problem that is dependent on our sequence $\{s_k\}_{k=1}^m$. For simplicity, consider $s_k = k^{1/10}$, then our objective is given by

$$\min_{x \in \mathcal{X}} f(x) = \min_{x \in \mathcal{X}} \left\{ \left(\sum_{i=1}^{n} i x_i \right)^{-1} \sum_{\mathcal{K} \in \mathcal{P}[m]} (-1)^{|\mathcal{K}|-1} \left(\sum_{k \in \mathcal{K}} k^{1/10} \right)^{-1} \right\}$$

which is minimized with $x^* = e_n$ (the nth basis vector) as

$$\sum_{\mathcal{K} \in \mathcal{P}[m]} (-1)^{|\mathcal{K}|-1} \left(\sum_{k \in \mathcal{K}} k^{1/10} \right)^{-1} > 0.$$

I would like to employ stochastic gradient descent here and analyze suboptimality, so I consider obtaining estimates of the gradient by means of Monte Carlo where the partial with respect to x_i is given by

$$\frac{\partial f(x)}{\partial x_i} = \int_0^\infty 1 - \frac{\partial}{\partial x_i} \prod_{k=1}^m \left(1 - \exp\{-y s_k \sum_{j \neq i} j x_j\} e^{-y s_k i x_i} \right) dy$$

$$= \int_0^\infty 1 - \sum_{k=1}^m \left[y s_k i \exp\{-y s_k \sum_{i=1}^n i x_i\} \times \prod_{j \neq i} \left(1 - \exp\{-y s_j \sum_{i=1}^n i x_i\} \right) \right] dy.$$

To perform Monte Carlo on this integral I, we must first transform the integral bounds for sampling (by partitioning the integral into two regions, $[0, 1)$ and $[1, \infty)$, then performing a change of variables on the one defined on $[1, \infty)$), we obtain:

$$I = \int_0^1 1 - \sum_{k=1}^m \left[y s_k i \exp\{-y s_k \sum_{i=1}^n i x_i\} \times \prod_{j \neq k} \left(1 - \exp\{-y s_j \sum_{i=1}^n i x_i\} \right) \right] dy$$

$$+ \int_0^1 \frac{1}{z^2} \left(1 - \sum_{k=1}^m \left[z^{-1} s_k i \exp\{-z^{-1} s_k \sum_{i=1}^n i x_i\} \times \prod_{j \neq k} \left(1 - \exp\{-z^{-1} s_j \sum_{i=1}^n i x_i\} \right) \right] \right) dz.$$

From here, with $m = 100$, $n = 50$, and $l = 100$ for Monte Carlo, we obtain the following results in 3. Again, akin to the first experiment, the theoretical optimality of Jain's step-size contrasts with what the experiment tells us. This notebook with documented code can be found in this repository.

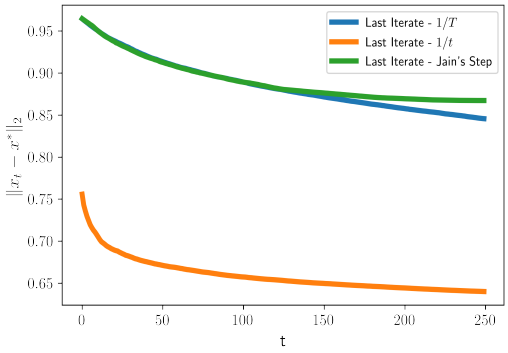


Figure 3: Last Iterate Suboptimality

5 Conclusion

We have reviewed step size regimes, suboptimality bounds and experiments of the final iterate of SGD. Empirically, we have demonstrated concern that the non-standard step size proposed by Jain et al. [2] is not robust to deviations of T nor the parameters or convexity constraints. The empirical results rather allude to the robustness of standard step size sequences even for the last iterate. In particular, knowledge of T a priori appears to be an impossible assumption in practice to satisfy - the removal of the $\log(\cdot)$ term as discussed in Liu and Lu [3] and Shamir and Zhang [7] is not possible without it. These lower bound results rather forebode possibly tight limits of what is plausible with SGD irrespective of the step size regime, but much more research and review is still impending.

References

[1] N. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. *PMLR*, 99(2):1579–1613, 2019.

[2] P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021.

[3] D. Liu and Z. Lu. The convergence rate of sgd's final iterate: Analysis on dimension dependence. 2021. doi: 10.48550/arXiv.2106.14588.

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

[5] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. *Optimizing Methods in Statistics*, pages 233–257, 1971.

[6] M. Schmidt, G. Fung, and R. Rosaless. Optimization methods for l1-regularization. 2008. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.230.7982&rep=rep1&type=pdf`.

[7] O. Shamir and T. Zhang. *Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79. PMLR, 2013.

[8] P. Zhang, H. Lang, Q. Liu, and L. Xiao. Statistical adaptive stochastic optimization. 2020. URL `https://openreview.net/forum?id=B1gkpR4FDB`.

6 Appendix

6.1 Martingales & Concentration Inequalities

Here I introduce some beautiful concepts from probability theory that are key ingredients for developing tight probability bounds of stochastic approximation algorithms. Concentration properties for sums of random variables was studied thoroughly in classical probability theory, but it wasn't until the 1970s when the appearance of martingale methods sparked renewed interest and more powerful tools for handling more general functions of independent random variables.

Definition 6.1. Let $X_n \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{F}_n \subset \mathcal{F}$ be σ-fields. We call $\langle \mathcal{F}_n \rangle$ a filtration if $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ and $\langle X_n \rangle$ adapted to $\langle \mathcal{F}_n \rangle$ if $X_n \in \mathcal{F}_n$. We call $\langle (X_n, \mathcal{F}_n) \rangle$ a (martingale, submartingale, supermartingale) if $\langle X_n \rangle$ is adapted to $\langle \mathcal{F}_n \rangle$ and

$$X_n \ (=, \leq, \geq) \ \mathbb{E}\left[X_{n+1} | \mathcal{F}_n\right] \quad \forall n \in \mathbb{N} \ \ a.s.$$

and $\mathbb{E}[X_n] < \infty$. We call $\langle (X_n, \mathcal{F}_n) \rangle$ a (martingale, submartingale, supermartingale) difference sequence if

$$\mathbb{E}\left[X_{n+1} | \mathcal{F}_n\right] \ (=, \geq, \leq) \ 0 \ \ \forall n \in \mathbb{N} \ \ a.s.$$

Theorem 6.1. (Freedman's Inequality) Consider a sequence of real-valued supermartingale differences $\langle (Y_i, \mathcal{G}_i) \rangle_{i=1}^n$ defined on a probability space $(\Omega, \mathcal{G}, \mathbb{P})$ with $Y_0 = 0$ and $Y_i \leq \epsilon$ for some $\epsilon > 0$. Let

$$S_k = \sum_{i=1}^{k} Y_i, \quad k \in [n].$$

Then S_k is a supermartingale and for all $s, v > 0$, there is some $k \in [n]$ such that

$$P\left(S_k \geq s \bigcap \sum_{i=1}^{k} \mathbb{E}[Y_i^2 | \mathcal{G}_{i-1}] \leq v^2 \right) \leq \exp\left\{ -\frac{s^2}{2(v^2 + s\epsilon)} \right\}$$

Theorem 6.2. (Chernoff's Inequality) For a random variable X, let $\psi_X(\lambda)$ denote the logarithm of the moment generating function (sometimes referred to as the cumulant generating function) and the Cramér transform of X (or the Fenchel-Legendre dual of ψ_X over the positive reals) by

$$\psi_X^*(t) = \sup_{\lambda \geq 0} \lambda t - \psi_X(\lambda).$$

Then for all $t \geq 0$,

$$P(X \geq t) \leq \exp\{-\psi_X^*(t)\}.$$

Theorem 6.3. (Hoeffding's Inequality) Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Let

$$S = \sum_{i=1}^{n} X_i - \mathbb{E}[X_i].$$

Then for every $t > 0$,

$$P(S \geq t) \leq \exp\left\{ -\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right\}.$$

Theorem 6.4. (Bernstein's Inequality) Let $X_1, \ldots . X_n$ be independent real-valued random variables. Assume that there exist $v, c > 0$ such that $\sum_{i=1}^{n} \mathbb{E}[X_i^2] \leq v$ and

$$\sum_{i=1}^{n} \mathbb{E}\left[\max\{X_i, 0\}^q\right] \leq \frac{q!}{2} v c^{q-2}$$

for all $q \geq 3$. Let $S = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])$, then for $\lambda \in (0, c^{-1})$ and $t > 0$,

$$\psi_S(\lambda) \leq \frac{v\lambda^2}{2(1 - c\lambda)}$$

and

$$\psi_S^*(t) \geq \frac{v}{c^2} h_1\left(\frac{ct}{v}\right)$$

where $h_1(x) = 1 + x - \sqrt{1 + 2x}$ for $u > 0$. In particular, for all $t > 0$,

$$P\left(S \geq \sqrt{2vt} + ct\right) \leq e^{-t}.$$