Making the Last Iterate of SGD Information Theoretically Optimal

Prateek Jain (MSR), Dheeraj Nagaraj (MIT) and Praneeth Netrapalli (MSR)

Abstract

Stochastic gradient descent (SGD) is one of the most widely used algorithms for large scale optimization problems. While classical theoretical analysis of SGD for convex problems studies (suffix) *averages* of iterates and obtains information theoretically optimal bounds on suboptimality, the *last point* of SGD is, by far, the most preferred choice in practice. The best known results for last point of SGD however, are suboptimal compared to information theoretic lower bounds by a $\log T$ factor, where T is the number of iterations. In fact, this additional $\log T$ factor is tight for standard step size sequences of $\Theta\left(\frac{1}{\sqrt{t}}\right)$ and $\Theta\left(\frac{1}{t}\right)$ for non-strongly convex and strongly convex settings, respectively. Similarly, even for subgradient descent (GD) when applied to non-smooth, convex functions, the best known step-size sequences still lead to $O(\log T)$-suboptimal convergence rates (on the final iterate). The main contribution of this work is to design new step size sequences that enjoy information theoretically optimal bounds on the suboptimality of *last point* of SGD as well as GD. We achieve this by designing a modification scheme, that converts one sequence of step sizes to another so that the last point of SGD/GD with modified sequence has the same suboptimality guarantees as the average of SGD/GD with original sequence. We also show that our result holds with high-probability. We validate our results through simulations which demonstrate that the new step size sequence indeed improves the final iterate significantly compared to the standard step size sequences.

Stochastic Gradient Descent

Empirical Risk Minimization: Minimize $F : \mathbb{R}^d \to \mathbb{R}$ over a closed convex set $\mathcal{W}$ where:

$$F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\,.$$

$f_i(x)$ are differentiable and convex.

SGD algorithm

- Total iterations T, step sizes $\alpha_t \geq 0$, initial point $x_1 \in \mathcal{W}$
- Sample $I_t \in \{1, 2, \ldots, n\}$
- $x_{t+1} \leftarrow \Pi_{\mathcal{W}}(x_t - \alpha_t \nabla f_{I_t}(x_t))$
- Output

$$\text{weighted average}\left((x_t)_{t=1}^{T}\right)$$

Design Choices: Weighted average, Sampling distribution for I_t, Step size sequence α_t.
We will let $I_t \sim \text{unif}(\{1, \ldots, n\})$

Existing Results

Non-Smooth, Non-Strongly Convex Case

- **Conditions:** $\mathsf{diam}(\mathcal{W}) \leq D$, $\|\nabla f_i\| \leq G$.
- **Standard Step Size:** $\gamma_t = \frac{1}{\sqrt{T}}$
- **Information Theoretic Bounds for Suboptimality:** $\Omega\left(\frac{GD}{\sqrt{T}}\right)$
- **Suboptimality with averaging:** $\hat{x} = \frac{\sum_t x_t}{T}$,

$$F(\hat{x}) - F(x^*) = O\left(\frac{GD}{\sqrt{T}}\right)$$

- **Suboptimality for Last Iterate (Sharp):**

$$F(x_T) - F(x^*) = O\left(\frac{GD\log T}{\sqrt{T}}\right)$$

Non-Smooth, Strongly Convex Case

- **Conditions:** F is μ strongly convex, $\|\nabla f_i\| \leq G$.
- **Standard Step Size:** $\gamma_t = \frac{1}{\mu t}$
- **Information Theoretic Bounds for Suboptimality:** $\Omega\left(\frac{G^2}{\mu T}\right)$
- **Suboptimality with tail averaging:** $\hat{x} = \frac{\sum_{t=T/2}^{T} x_t}{T/2}$,

$$F(\hat{x}) - F(x^*) = O\left(\frac{G^2}{\mu T}\right)$$

- **Suboptimality for Last Iterate:**

$$F(x_T) - F(x^*) = O\left(\frac{G^2\log T}{\mu\sqrt{T}}\right)$$

The rate for the last iterate is sharp for standard step sizes as shown by Harvey et. al [2018].

Our Results

Step Size Modification

Original step size: $\gamma_1, \ldots, \gamma_T$. We define the modification:

$$T_1 := T/2,\ T_2 := T/2 + T/4,\ T_3 := T/2 + T/4 + T/8 \ldots$$

Modified step size: α_t

$$\alpha_t = 2^{-i}\gamma_t \quad \text{for } T_{i-1} < t \leq T_i \tag{1}$$

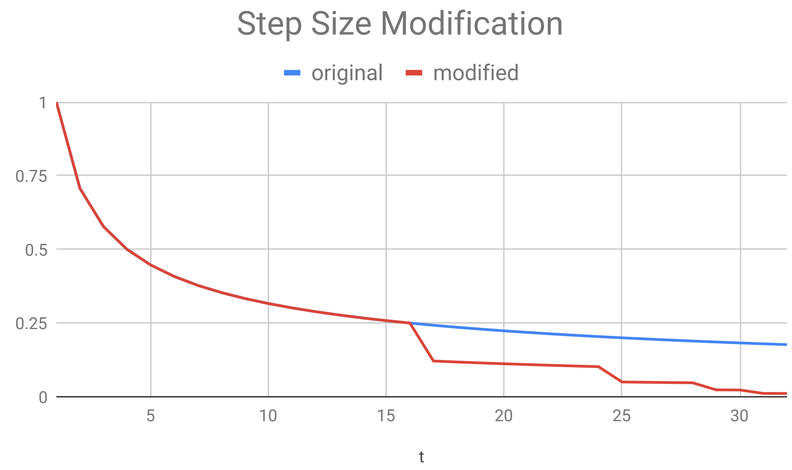

Figure 1: Step Size Modification for $T = 32$

Main Theorem

Let output of SGD with α_t step size be $x_1, \ldots, x_T$, and the output for step size γ_t be $y_1, \ldots, y_T$. Let $q()$ be any probability measure over $\{T/4, \ldots, T/2\}$. With probability atleast $1 - \delta$,

$$F(x_T) \leq O\left(\gamma_T \log\tfrac{1}{\delta}\right) + \sum_{l=T/4}^{T/2} q(l)F(y_l)$$

Non-Smooth & Non-Strongly Convex Case : Optimal Bounds

Let $\mathsf{diam}(\mathcal{W}) \leq D$, $\|\nabla f_i\| \leq G$, $\gamma_t = \frac{D}{G\sqrt{T}\sqrt{\log\frac{1}{\delta}}}$ (standard step size), step size used: α_t which is the modification of γ_t. With probability atleast $1 - \delta$:

$$F(x_T) - F(x^*) = O\left(GD\sqrt{\frac{\log\frac{1}{\delta}}{T}}\right)$$

Non-Smooth & Strongly Convex Case : Optimal Bounds

Let F be μ strongly convex, $\|\nabla f_i\| \leq G$, $\gamma_t = \frac{1}{\mu t}$ (standard step size), step size used: α_t which is the modification of γ_t. With probability atleast $1 - \delta$:

$$F(x_T) - F(x^*) = O\left(\frac{G^2\log\frac{1}{\delta}}{\mu T}\right)$$

Simulations

We consider simulated examples of Lasso regression for sparse vectors and training Support Vector Machines (SVM). The loss for averaged SGD with standard step size is denoted in green, the loss for the current iterate with standard step size is denoted in Red and the loss for current iterate with modified step size is given in Blue. The last iterate with modified step size works better since it decreases the step size when the SGD is closer to the optimum and hence allows it to approach the optimum without boucing around it.

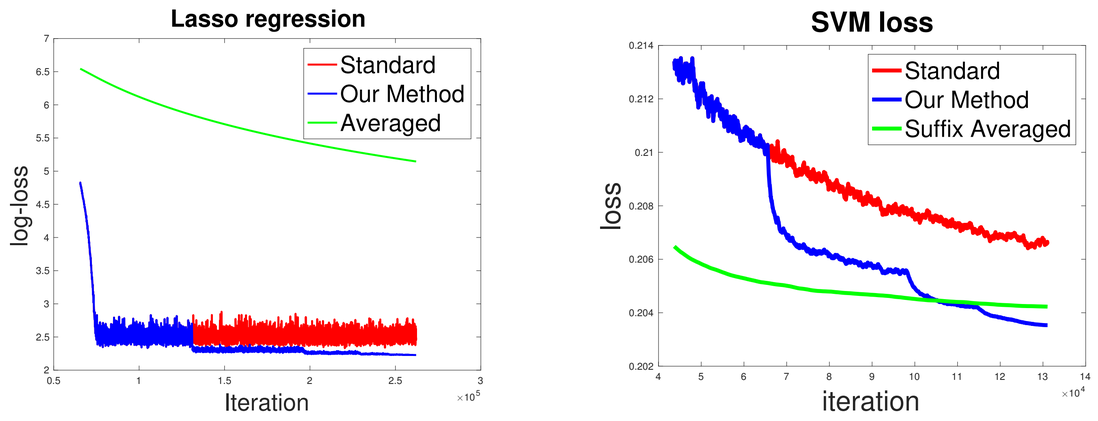

Figure 2: Simulated examples for Lasso (Non-strongly Convex) and SVM (strongly convex)

Open Question

Our results assume the knowledge of the number of iterations T to obtain optimal bounds. Are there step sizes which achieve optimal results without knowing T ?

References

1. Ohad Shamir & Tong Zhang. Stochastic gradient descent for non-smooth optimization: Conver- gence results and optimal averaging schemes. In International Conference on Machine Learning, pages 71 – 79, 2013.
2. Nicholas JA Harvey, Christopher Liaw, Yaniv Plan & Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. arXiv preprint arXiv:1812.05217, 2018.