Making SGD Parameter-Free

Presented by Qilin Ye

March 22, 2023

... the same old SGD:

$$x_{t+1} := x_t - \eta \nabla F(x_t)$$

where F is convex & differentiable.

Non-differentiable? Use unbiased **subgradients**: $x_{t+1} := x_t - \eta g_t$.[1]

[1] A subgradient of f satisfies $f(z) \geqslant f(x) + g^T(z - x)$ for all z.

... the same old SGD:

$$x_{t+1} := x_t - \eta \nabla F(x_t)$$

where F is convex & differentiable.

Non-differentiable? Use unbiased **subgradients**: $x_{t+1} := x_t - \eta g_t$.[1]

[1] A subgradient of f satisfies $f(z) \geqslant f(x) + g^T(z - x)$ for all z.

Apparently, choosing the correct learning rate is not a trivial job.

(1) Too large? Possible oscillation. Too small? Slow!

(2) Distance between starting point and optimum matters.

(3) The rate of convergence is affected by scaling.

(4) ...

§0 Stolchastic Gradient Descent... Problems?

Apparently, choosing the correct learning rate is not a trivial job.

(1) Too large? Possible oscillation. Too small? Slow!

(2) Distance between starting point and optimum matters.

(3) The rate of convergence is affected by scaling.

(4) ...

Apparently, choosing the correct learning rate is not a trivial job.

(1) Too large? Possible oscillation. Too small? Slow!

(2) Distance between starting point and optimum matters.

(3) The rate of convergence is affected by scaling.

(4) ...

§0 Parameter-Free Optimizations and Regrets

We aim to design **parameter-free** algorithms that "automatically" tune the learning rate.

And we aim to obtain "good" **regret** guarantees.

We aim to design **parameter-free** algorithms that “automatically” tune the learning rate.

And we aim to obtain “good” **regret** guarantees.

§0 Notations and Problem Setup

(1) Let $\mathcal{X} \subset \mathbb{R}^d$ be convex closed and let $f : \mathcal{X} \to \mathbb{R}$ be convex.

(2) Let x^* a minimum of f, assuming existence.

(3) Let $\mathcal{O}$ be an oracle that is a subgradient of f in expectation: $\mathbb{E}[\mathcal{O}(x) \mid x] \in \partial f(x)$.

(4) Denote the iterates by $x_0, x_1(\eta), x_2(\eta), \ldots$ and (sub)gradients $g_0, g_1(\eta), g_2(\eta), \ldots$. Define $\overline{x}(\eta) := T^{-1} \sum_{i<T} x_i(\eta)$.

(5) Distance to optimum and running maximum distance:

$$d_t(\eta) := \|x_t(\eta) - x^*\| \qquad \overline{d}_t(\eta) := \max_{i \leqslant t} d_i(\eta).$$

(6) Distance to x_0 and running max distance: $r_t(\eta), \overline{r}_t(\eta)$.

(7) Oracle error & running squared norms of oracles:

$$\nabla_i := g_i - \nabla f(x_i(\eta)) \qquad G_t(\eta) := \sum_{i<t} \|g_i(\eta)\|^2.$$

§0 Notations and Problem Setup

(1) Let $\mathcal{X} \subset \mathbb{R}^d$ be convex closed and let $f : \mathcal{X} \to \mathbb{R}$ be convex.

(2) Let x^* a minimum of f, assuming existence.

(3) Let $\mathcal{O}$ be an oracle that is a subgradient of f in expectation: $\mathbb{E}[\mathcal{O}(x) \mid x] \in \partial f(x)$.

(4) Denote the iterates by $x_0, x_1(\eta), x_2(\eta), \ldots$ and (sub)gradients $g_0, g_1(\eta), g_2(\eta), \ldots$. Define $\overline{x}(\eta) := T^{-1} \sum_{i<T} x_i(\eta)$.

(5) Distance to optimum and running maximum distance:

$$d_t(\eta) := \|x_t(\eta) - x^*\| \qquad \overline{d}_t(\eta) := \max_{i \leqslant t} d_i(\eta).$$

(6) Distance to x_0 and running max distance: $r_t(\eta), \overline{r}_t(\eta)$.

(7) Oracle error & running squared norms of oracles:

$$\nabla_i := g_i - \nabla f(x_i(\eta)) \qquad G_t(\eta) := \sum_{i<t} \|g_i(\eta)\|^2.$$

§0 Notations and Problem Setup

(1) Let $\mathcal{X} \subset \mathbb{R}^d$ be convex closed and let $f : \mathcal{X} \to \mathbb{R}$ be convex.

(2) Let x^* a minimum of f, assuming existence.

(3) Let $\mathcal{O}$ be an oracle that is a subgradient of f in expectation: $\mathbb{E}[\mathcal{O}(x) \mid x] \in \partial f(x)$.

(4) Denote the iterates by $x_0, x_1(\eta), x_2(\eta), \ldots$ and (sub)gradients $g_0, g_1(\eta), g_2(\eta), \ldots$. Define $\overline{x}(\eta) := T^{-1} \sum_{i<T} x_i(\eta)$.

(5) Distance to optimum and running maximum distance:

$$d_t(\eta) := \|x_t(\eta) - x^*\| \qquad \overline{d}_t(\eta) := \max_{i \leqslant t} d_i(\eta).$$

(6) Distance to x_0 and running max distance: $r_t(\eta), \overline{r}_t(\eta)$.

(7) Oracle error & running squared norms of oracles:

$$\nabla_i := g_i - \nabla f(x_i(\eta)) \qquad G_t(\eta) := \sum_{i<t} \|g_i(\eta)\|^2.$$

§0 Notations and Problem Setup

(1) Let $\mathcal{X} \subset \mathbb{R}^d$ be convex closed and let $f : \mathcal{X} \to \mathbb{R}$ be convex.

(2) Let x^* a minimum of f, assuming existence.

(3) Let $\mathcal{O}$ be an oracle that is a subgradient of f in expectation: $\mathbb{E}[\mathcal{O}(x) \mid x] \in \partial f(x)$.

(4) Denote the iterates by $x_0, x_1(\eta), x_2(\eta), \ldots$ and (sub)gradients $g_0, g_1(\eta), g_2(\eta), \ldots$. Define $\overline{x}(\eta) := T^{-1} \sum_{i<T} x_i(\eta)$.

(5) Distance to optimum and running maximum distance:

$$d_t(\eta) := \|x_t(\eta) - x^*\| \qquad \overline{d}_t(\eta) := \max_{i \leqslant t} d_i(\eta).$$

(6) Distance to x_0 and running max distance: $r_t(\eta), \overline{r}_t(\eta)$.

(7) Oracle error & running squared norms of oracles:

$$\nabla_i := g_i - \nabla f(x_i(\eta)) \qquad G_t(\eta) := \sum_{i<t} \|g_i(\eta)\|^2.$$

§0 Notations and Problem Setup

(1) Let $\mathcal{X} \subset \mathbb{R}^d$ be convex closed and let $f : \mathcal{X} \to \mathbb{R}$ be convex.

(2) Let x^* a minimum of f, assuming existence.

(3) Let $\mathcal{O}$ be an oracle that is a subgradient of f in expectation: $\mathbb{E}[\mathcal{O}(x) \mid x] \in \partial f(x)$.

(4) Denote the iterates by $x_0, x_1(\eta), x_2(\eta), \ldots$ and (sub)gradients $g_0, g_1(\eta), g_2(\eta), \ldots$. Define $\overline{x}(\eta) := T^{-1} \sum_{i<T} x_i(\eta)$.

(5) Distance to optimum and running maximum distance:

$$d_t(\eta) := \|x_t(\eta) - x^*\| \qquad \overline{d}_t(\eta) := \max_{i \leqslant t} d_i(\eta).$$

(6) Distance to x_0 and running max distance: $r_t(\eta), \overline{r}_t(\eta)$.

(7) Oracle error & running squared norms of oracles:

$$\nabla_i := g_i - \nabla f(x_i(\eta)) \qquad G_t(\eta) := \sum_{i<t} \|g_i(\eta)\|^2.$$

Fact 1

If all $\|g_i\|$'s are uniformly bounded by $L > 0$, then setting η to be the fixed point of

$$\eta \mapsto \frac{\|x_0 - x^*\|}{(\sum_{i<T} \|g_i(\eta)\|^2)^{1/2}} = \frac{d_0}{\sqrt{G_T(\eta)}}$$

satisfies the optimal error bound for the average iterate after T iterations:

$$f(\overline{x}) - f(x^*) \leqslant \frac{d_0\sqrt{G_T(\eta)}}{T} = O(d_0 L T^{-1/2}).$$

Fact 2: SoTA w/out Knowing $d_0 = \|x_0 - x^*\|$ a priori

... gains an additional logarithmic factor:

$$O\left(d_0\sqrt{\log(1 + Td_0^2\epsilon^{-2})/T} + \epsilon/T\right).$$

§0 What Did This Paper Do?

(1) For any prescribed $\epsilon > 0$ and $\delta \in (0, 1)$, this paper provides a $1 - \delta$ probability optimality gap with an additional log factor:

$$O\Big((d_0 T^{-1/2} + \epsilon T^{-1}) \cdot \log^2(\delta^{-1} \log(d_0 T \epsilon^{-1})\Big)$$

(2) Strong localization guarantee: the average iterate (as well as other intermediate outputs) $\overline{x}$ satisfies $\|\overline{x} - x^*\| = O(\|x_0 - x^*\|)$.

(3) Good adaptivity to gradient norms and other scenarios.

§0 What Did This Paper Do?

(1) For any prescribed $\epsilon > 0$ and $\delta \in (0, 1)$, this paper provides a $1 - \delta$ probability optimality gap with an additional log factor:

$$O\left((d_0 T^{-1/2} + \epsilon T^{-1}) \cdot \log^2(\delta^{-1} \log(d_0 T \epsilon^{-1})\right)$$

(2) Strong localization guarantee: the average iterate (as well as other intermediate outputs) $\overline{x}$ satisfies $\|\overline{x} - x^*\| = O(\|x_0 - x^*\|)$.

(3) Good adaptivity to gradient norms and other scenarios.

(1) For any prescribed $\epsilon > 0$ and $\delta \in (0, 1)$, this paper provides a $1 - \delta$ probability optimality gap with an additional log factor:

$$O\Big((d_0 T^{-1/2} + \epsilon T^{-1}) \cdot \log^2(\delta^{-1} \log(d_0 T \epsilon^{-1})\Big)$$

(2) Strong localization guarantee: the average iterate (as well as other intermediate outputs) $\overline{x}$ satisfies $\|\overline{x} - x^*\| = O(\|x_0 - x^*\|)$.

(3) Good adaptivity to gradient norms and other scenarios.

This page is intentionally left blank.

§1 High-Level Idea: Using Proxy for d_0

In SGD, the output iterates $x_t(\eta)$ should ideally converge to x^*

$$\Rightarrow \frac{r_t(\eta)}{\sqrt{G_T(\eta)}} \text{ converges to } \frac{d_0}{\sqrt{G_T(\eta)}}.$$

Instead of computing the uncomputable fixed point, we resort to approximating the fixed point of

$$\eta \mapsto \frac{\bar{r}_T(\eta)}{\sqrt{\alpha G_T(\eta) + \beta}}. \tag{FP1}$$

(Why $\bar{r}_T$ instead of r_T?)

Propostion 1

Assuming we magically found the η satisfying (FP1), and with probability 1 our oracle $\mathcal{O}(x) = \nabla f(x)$ (i.e. *true* gradient):

Proposition 1

If $\alpha > 1, \beta = 0$, then the average iterate $\bar{x} := T^{-1} \sum_{i<T} x_i(\eta)$ satisfies

$$\|\bar{x} - x^*\| \leqslant \frac{2\alpha}{\alpha - 1}\|x_0 - x^*\| = \frac{2\alpha}{\alpha - 1} d_0$$

and

$$f(\bar{x}) - f(x^*) \leqslant \frac{\alpha^{3/2}}{\alpha - 1} \cdot \frac{d_0\sqrt{G_T(\eta)}}{T} \sim \frac{d_0\sqrt{G_T(\eta)}}{T}.$$

(*This is the SoTA regret bound!*)