Making SGD Parameter-Free

Presented by Qilin Ye

March 24, 2023

§3 Moving Forward — Defining "Good Events"

(1) Key observation: when O outputs exact gradients, $g_i(\eta) \equiv \nabla f(x_i(\eta))$.

(2) This means that under exact gradient setting,

$$\sum_{i<T} \langle \Delta_i(\eta), x_i(\eta) - x^* \rangle = 0.$$

(3) Generalize above into "approximately:" for $T \in \mathbb{N}$, and $\alpha, \beta, \eta > 0$, define the "**good events**" to be

$$\mathfrak{E}(\eta) = \mathfrak{E}(\eta; T, \alpha, \beta) := \bigcap_{t \leqslant T} \left\{ \sum_{i<t} \langle \Delta_i(\eta), x_i(\eta) - x^* \rangle \geqslant -\frac{1}{4} \max(\overline{d}_t(\eta), \eta\sqrt{\beta}) \sqrt{\alpha G_t(\eta) + \beta} \right\}.$$

§3 Moving Forward — Defining "Good Events"

(1) Key observation: when O outputs exact gradients, $g_i(\eta) \equiv \nabla f(x_i(\eta))$.

(2) This means that under exact gradient setting,

$$\sum_{i<T} \langle \Delta_i(\eta), x_i(\eta) - x^* \rangle = 0.$$

(3) Generalize above into "approximately:" for $T \in \mathbb{N}$, and $\alpha, \beta, \eta > 0$, define the "**good events**" to be

$$\mathfrak{E}(\eta) = \mathfrak{E}(\eta; T, \alpha, \beta) := \bigcap_{t \leqslant T} \left\{ \sum_{i<t} \langle \Delta_i(\eta), x_i(\eta) - x^* \rangle \geqslant -\frac{1}{4} \max(\overline{d}_t(\eta), \eta\sqrt{\beta}) \sqrt{\alpha G_t(\eta) + \beta} \right\}.$$

§3 Moving Forward — Defining "Good Events"

(1) Key observation: when O outputs exact gradients, $g_i(\eta) \equiv \nabla f(x_i(\eta))$.

(2) This means that under exact gradient setting,

$$\sum_{i<T} \langle \Delta_i(\eta), x_i(\eta) - x^* \rangle = 0.$$

(3) Generalize above into "approximately:" for $T \in \mathbb{N}$, and $\alpha, \beta, \eta > 0$, define the "**good events**" to be

$$\mathfrak{E}(\eta) = \mathfrak{E}(\eta; T, \alpha, \beta) := \bigcap_{t \leqslant T} \left\{ \sum_{i<t} \langle \Delta_i(\eta), x_i(\eta) - x^* \rangle \geqslant -\frac{1}{4} \max(\bar{d}_t(\eta), \eta\sqrt{\beta}) \sqrt{\alpha G_t(\eta) + \beta} \right\}.$$

What Was That Mess?

Lemma 1 (exact gradient version)

With appropriate parameters, under exact gradient setting,

$$\eta \leqslant \varphi(\eta) \Rightarrow \overline{d}_T(\eta) \leqslant \frac{\alpha+1}{\alpha-1} \cdot d_0 \quad \text{and} \quad \overline{r}_T(\eta) \leqslant \frac{2\alpha}{\alpha-1} d_0.$$

becomes ...

Lemma 1 (stochastic version)

With appropriate parameters, under $\mathfrak{E}(\eta; T, \alpha, \beta)$, i.e., the "good event" setting, if $\eta \leqslant \varphi(\eta)$, then

$$\overline{d}_T(\eta) \leqslant \frac{3\alpha+2}{\alpha+2} d_0 \qquad \text{and} \qquad \overline{r}_T(\eta) \leqslant \frac{4\alpha}{\alpha-2} d_0.$$

What Was That Mess?

Lemma 1 (exact gradient version)

With appropriate parameters, under exact gradient setting,

$$\eta \leqslant \varphi(\eta) \Rightarrow \overline{d}_T(\eta) \leqslant \frac{\alpha+1}{\alpha-1} \cdot d_0 \quad \text{and} \quad \overline{r}_T(\eta) \leqslant \frac{2\alpha}{\alpha-1} d_0.$$

becomes ...

Lemma 1 (stochastic version)

With appropriate parameters, under $\mathfrak{E}(\eta; T, \alpha, \beta)$, i.e., the "good event" setting, if $\eta \leqslant \varphi(\eta)$, then

$$\overline{d}_T(\eta) \leqslant \frac{3\alpha+2}{\alpha+2} d_0 \qquad \text{and} \qquad \overline{r}_T(\eta) \leqslant \frac{4\alpha}{\alpha-2} d_0.$$

What Was That Mess?

Proposition 2 (exact gradient version)

Let $\eta_0 = \texttt{RootFindingBisection}(\eta_{\text{low}}, \eta_{\text{high}}; T, \alpha, \beta)$, where $\alpha > 1, \beta > 0, T \in \mathbb{N}$, and each $\eta > 0$. Asssume $\eta_{\text{high}} > \varphi(\eta_{\text{high}})$. Let $\overline{x} = T^{-1} \sum_{i<T} x_i(\eta_0)$ be the average iterate. Under exact gradient setting:

(1) if $\eta_{\text{low}} \leqslant \varphi(\eta_{\text{low}})$ then for some $\eta' \in [\eta_0, 2\eta_0]$,

$$\|\overline{x} - x_0\| \leqslant \frac{2\alpha}{\alpha - 1} d_0 \qquad \text{and} \qquad f(\overline{x}) - f(x^*) \leqslant \frac{2\alpha}{\alpha - 1} \cdot \frac{d_0\sqrt{\alpha G_T(\eta') + \beta}}{T};$$

(2) if $\eta_{\text{low}} > \varphi(\eta_{\text{low}})$, then $\eta_0 = \eta_{\text{low}}$, and

$$\|\overline{x} - x_0\| \leqslant \eta_0\sqrt{\alpha G_T(\eta_0) + \beta} \quad \text{and} \quad f(\overline{x}) - f(x^*) \leqslant \frac{d_0\sqrt{\alpha G_T(\eta_0) + \beta} + \eta_0 G_T(\eta_0)}{T}.$$

What Was That Mess?

Proposition 2 (exact gradient version)

Let $\eta_0 = \texttt{RootFindingBisection}(\eta_{\text{low}}, \eta_{\text{high}}; T, \alpha, \beta)$, where $\alpha > 1, \beta > 0, T \in \mathbb{N}$, and each $\eta > 0$. Asssume $\eta_{\text{high}} > \varphi(\eta_{\text{high}})$. Let $\overline{x} = T^{-1} \sum_{i<T} x_i(\eta_0)$ be the average iterate. Under exact gradient setting:

(1) if $\eta_{\text{low}} \leqslant \varphi(\eta_{\text{low}})$ then for some $\eta' \in [\eta_0, 2\eta_0]$,

$$\|\overline{x} - x_0\| \leqslant \frac{2\alpha}{\alpha - 1} d_0 \quad \text{and} \quad f(\overline{x}) - f(x^*) \leqslant \frac{2\alpha}{\alpha - 1} \cdot \frac{d_0\sqrt{\alpha G_T(\eta') + \beta}}{T};$$

(2) if $\eta_{\text{low}} > \varphi(\eta_{\text{low}})$, then $\eta_0 = \eta_{\text{low}}$, and

$$\|\overline{x} - x_0\| \leqslant \eta_0\sqrt{\alpha G_T(\eta_0) + \beta} \quad \text{and} \quad f(\overline{x}) - f(x^*) \leqslant \frac{d_0\sqrt{\alpha G_T(\eta_0) + \beta} + \eta_0 G_T(\eta_0)}{T}.$$

What Was That Mess?

Proposition 2 (stochastic version)

Let $\eta_0 = \texttt{RootFindingBisection}(\eta_{\text{low}}, \eta_{\text{high}}; T, \alpha, \beta)$, where $\alpha > 2, \beta > 0, T \in \mathbb{N}$, and $\eta_{\text{high}} = 2^{2^k}\eta_{\text{low}}$ for some k. Assume $\eta_{\text{high}} > \varphi(\eta_{\text{high}})$. Let $\overline{x} = T^{-1}\sum_{i<T} x_i(\eta_0)$ be the average iterate. Assume the "good events" $\bigcap_{j=0}^{2^k} \mathfrak{E}(2^j\eta_{\text{low}}; T, \alpha, \beta)$ all hold.

(1) If $\eta_{\text{low}} \leqslant \varphi(\eta_{\text{low}})$, then for some $\eta' \in [\eta_0, 2\eta_0]$,

$$\|\overline{x} - x_0\| \leqslant \frac{4\alpha}{\alpha - 2} d_0 \quad \text{and} \quad f(\overline{x}) - f(x^*) \leqslant \frac{9\alpha - 2}{2(\alpha - 2)} \cdot \frac{d_0\sqrt{\alpha G_T(\eta') + \beta}}{T};$$

(2) If $\eta_{\text{low}} > \varphi(\eta_{\text{low}})$ and in addition $\mathfrak{E}(\eta_{\text{low}}; T, \alpha, \beta)$ holds, then $\eta_0 = \eta_{\text{low}}$, and

$$\|\overline{x} - x_0\| \leqslant \eta_{\text{low}}\sqrt{\alpha G_T(\eta_{\text{low}}) + \beta} \quad \text{and} \quad f(\overline{x}) - f(x^*) \leqslant \frac{5}{4}\frac{d_0\sqrt{\alpha G_T(\eta_{\text{low}}) + \beta} + \eta_{\text{low}}(\alpha G_T(\eta_{\text{low}} + \beta))}{T}.$$

What Was That Mess?

Lemma 2 (exact gradient version)

With appropriate parameters, under exact gradient setting, if the following holds, then $\eta > \varphi(\eta)$:

$$\eta > \eta_{\max} := \frac{2\alpha}{\alpha - 1} \cdot \frac{d_0}{\sqrt{\alpha\|g_0\|^2 + \beta}}.$$

Consequently, when our algorithm terminates, $k \leqslant 2 \log \log^{+}(\eta_{\max}/\eta_{\epsilon})$.

becomes...

Lemma 2 (stochastic version)

With appropriate parameters, if "good event" $\mathfrak{E}(\eta; T, \alpha, \beta)$ holds, then if the following implies $\eta > \varphi(\eta)$:

$$\eta > \eta_{\max} := \frac{4\alpha}{\alpha - 2} \cdot \frac{d_0}{\sqrt{\alpha\|g_0\|^2 + \beta}}.$$

Consequently, if $\bigcap_{k=2,4,8,\ldots} \mathfrak{E}(2^{2^k}\eta_{\epsilon}; T_k, \alpha^{(k)}, \beta^{(k)})$ holds, when our algorithm terminates, $k \leqslant 2 \log \log^{+}(\eta_{\max}/\eta_{\epsilon})$.

What Was That Mess?

Lemma 2 (exact gradient version)

With appropriate parameters, under exact gradient setting, if the following holds, then $\eta > \varphi(\eta)$:

$$\eta > \eta_{\max} := \frac{2\alpha}{\alpha - 1} \cdot \frac{d_0}{\sqrt{\alpha \|g_0\|^2 + \beta}}.$$

Consequently, when our algorithm terminates, $k \leqslant 2 \log \log^+(\eta_{\max}/\eta_\epsilon)$.

becomes...

Lemma 2 (stochastic version)

With appropriate parameters, if "good event" $\mathfrak{E}(\eta; T, \alpha, \beta)$ holds, then if the following implies $\eta > \varphi(\eta)$:

$$\eta > \eta_{\max} := \frac{4\alpha}{\alpha - 2} \cdot \frac{d_0}{\sqrt{\alpha \|g_0\|^2 + \beta}}.$$

Consequently, if $\bigcap_{k=2,4,8,\ldots} \mathfrak{E}(2^{2^k} \eta_\epsilon; T_k, \alpha^{(k)}, \beta^{(k)})$ holds, when our algorithm terminates, $k \leqslant 2 \log \log^+(\eta_{\max}/\eta_\epsilon)$.

For the remainder of the analysis, just like Fact 1, we assume the gradient oracle is uniformly bounded by $L > 0$.

Lemma 3: "good events" are likely

Let $T \in \mathbb{N}, \eta > 0, \delta \in (0, 1)$ be given. Define $C = \log(60\delta^{-1} \log^2(6T))$.
If $\alpha \geqslant 1024C$ and $\beta \geqslant 1024C^2L^2$ then $\mathbb{P}(\mathfrak{E}(\eta; T, \alpha, \beta)) \geqslant 1 - \delta$.

For the remainder of the analysis, just like Fact 1, we assume the gradient oracle is uniformly bounded by $L > 0$.

Lemma 3: "good events" are likely

Let $T \in \mathbb{N}, \eta > 0, \delta \in (0,1)$ be given. Define $C = \log(60\delta^{-1}\log^2(6T))$.
If $\alpha \geqslant 1024C$ and $\beta \geqslant 1024C^2L^2$ then $\mathbb{P}(\mathfrak{E}(\eta; T, \alpha, \beta)) \geqslant 1 - \delta$.

Proposition 3

Let budget B, initial step size $\eta_\epsilon > 0$, and failure probability $\delta \in (0,1)$ be given. Let $\alpha^{(k)} = 1024 C_k$ and $\beta^{(k)} = 1024 C_k^2 L^2$, where $C_k = 2k + \log(60\delta^{-1}\log^2(6B))$. Then, $\mathbb{P}(\bigcap_{k=2,4,8,\dots} \bigcap_{j=0,1,\dots,2^k} \mathfrak{E}(2^j n_\epsilon; B, \alpha^{(k)}, \beta^{(k)})) \geqslant 1-\delta$.

Proof. Notice that $C_k = \log(60\log^2(6B)/(2^{-2k}\delta))$ so by the previous lemma, with $T = B$, $\alpha = \alpha^{(k)}, \beta = \beta^{(k)}$, and failure probability $2^{-2k}\delta$, for any η,

$$1 - \mathbb{P}(\mathfrak{E}(\eta; B, \alpha^{(k)}, \beta^{(k)})) \leqslant 2^{-2k}\delta.$$

By union bound

$$1 - \mathbb{P}\left(\bigcap_{j=0}^{2^k} \mathfrak{E}(2^j \eta_\epsilon; B, \alpha^{(k)}, \beta^{(k)})\right) \leqslant (2^k+1)2^{-2k}\delta \leqslant 2^{-(k-1)}\delta$$

and finally

$$1 - \mathbb{P}\left(\bigcap_{k=2,4,8,\dots} \bigcap_{j=0}^{2^k} \mathfrak{E}(2^j \eta_\epsilon; B, \alpha^{(k)}, \beta^{(k)})\right) \leqslant \sum_{k\geqslant 1} 2^{-k}\delta = \delta.$$