The Stochastic Gradient Method and Variance Reduction

Presented on August 4, 2021

Outline

The Stochastic Gradient Descent Algorithm

Variance Reduction Methods

Recap - The Gradient Descent Algorithm

- The workhorse:

$$\min_{\theta \in \mathbb{R}^d} \left\{ f^N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) \right\}$$

- GD iteration:

$$\theta^{k+1} = \theta^k - \gamma^k \cdot \nabla f^N(\theta).$$

- Computation cost per iteration: N gradient evaluations!
- It would be nice if we can compute less per iteration...

The SGD Algorithm

Empirical risk minimization

$$\min_{\theta} f^N(\theta) = \frac{1}{N}\sum_{i=1}^{N} f_i(\theta^k)$$

- Pick an index i^k: $\theta^{k+1} = \theta^k - \gamma^k \cdot \nabla f_{i^k}(\theta^k)$
- Reduced workload per iteration
- Convergence?
- SGD vs. GD?

Convergence Analysis - Convex Function

Deterministic GD:

Theorem

Let f be convex with bounded gradient, then the sequence $(x^k)_{k\in\mathbb{N}}$ generated by GD with step size $\gamma = \frac{\|x^0 - x^\star\|}{\sqrt{T+1}B}$ satisfies

$$f(\bar{\theta}^T) - f(\theta^\star) \leq \frac{\|x^0 - x^\star\| B}{2\sqrt{T+1}},$$

where $\bar{\theta}^T = \sum_{k=0}^{T} \theta^k / (T+1)$

Some preliminaries

Conditional expectation

- $\mathbb{E}(X|Y)$ is a random variable: "best guess" of X knowing Y
- Law of total expectation: $\mathbb{E}X = \mathbb{E}(\mathbb{E}(X|Y))$

Filtration

- $\mathcal{F}^k = \sigma(i^0, \ldots, i^k)$
- If i^0 up to i^{k-1} are given, then θ^k is determined: $\theta^k \in \mathcal{F}^{k-1}$
- Perfect information: $\mathbb{E}(\theta^k|\mathcal{F}^{k-1}) = \theta^k$
- Partial information: $\mathbb{E}(\theta^k \cdot Y|\mathcal{F}^{k-1}) = \theta^k \mathbb{E}(Y|\mathcal{F}^{k-1})$

A First Proof of SGD

Let's mimic the proof in the deterministic case

$$\begin{aligned}\|\theta^{k+1} - \theta^\star\|^2 &= \|\theta^k - \gamma \nabla f_{i^k}(\theta^k) - \theta^\star\|^2 \\ &= \|\theta^k - \theta^\star\|^2 - 2\gamma \nabla f_{i^k}(\theta^k)^\top (\theta^k - \theta^\star) + \gamma^2 \|\nabla f_{i^k}(\theta^k)\|^2\end{aligned}$$

Observation: i^k is independent of θ^k

- $\mathbb{E}[\nabla f_{i^k}(\theta^k) \mid \mathcal{F}^{k-1}] = \nabla f^N(\theta^k)$
- $\mathbb{E}[\nabla f_{i^k}(\theta^k)^\top (\theta^k - \theta^\star) \mid \mathcal{F}^{k-1}] \geq f^N(\theta^k) - f^N(\theta^\star)$

Use law of total expectation:

$$\begin{aligned}\mathbb{E}\|\theta^{k+1} - \theta^\star\|^2 &= \mathbb{E}[\mathbb{E}\|\theta^{k+1} - \theta^\star\|^2 | \mathcal{F}^{k-1}] \\ &\leq \mathbb{E}\|\theta^k - \theta^\star\|^2 - 2\gamma \mathbb{E}[f^N(\theta^k) - f^N(\theta^\star)] + \gamma^2 B^2\end{aligned}$$

We get "on average" the same inequality in the deterministic case.

Stochastic GD:

Theorem

Let f^N be convex with bounded gradient, then the sequence $(x^k)_{k\in\mathbb{N}}$ generated by SGD with step size $\gamma = \frac{\|x^0 - x^\star\|}{\sqrt{T+1}B}$ satisfies

$$\mathbb{E}\big[f(\bar{\theta}^T) - f(\theta^\star)\big] \leq \frac{\|x^0 - x^\star\|B}{2\sqrt{T+1}},$$

where $\bar{\theta}^T = \sum_{k=0}^{T} \theta^k/(T+1)$.

Convergence Analysis - Strongly Convex Smooth Function

- Deterministic setting

Theorem

Let f be μ-strongly convex and L-smooth, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by GD with step size $\gamma = 1/L$ satisfies

$$f(\theta^{k+1}) - f^\star \leq \left(1 - \frac{\mu}{L}\right)(f(\theta^k) - f^\star)$$

Can we hope for the same result?

Convergence of SGD - Strongly Convex

Descent Lemma

$$f^N(\theta^{k+1}) \leq f^N(\theta^k) + \nabla f^N(\theta^k)^\top(\theta^{k+1} - \theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2.$$

Conditioning on the past $\mathcal{F}^{k-1}$

$$\begin{aligned}
&\mathbb{E}[f^N(\theta^{k+1})|\mathcal{F}^{k-1}] \\
\leq& f^N(\theta^k) - \gamma\mathbb{E}[\nabla f^N(\theta^k)^\top \nabla f_{i^k}(\theta^k)|\mathcal{F}^{k-1}] + \frac{\gamma^2 L}{2}\mathbb{E}[\|\nabla f_{i^k}(\theta^k)\|^2|\mathcal{F}^{k-1}] \\
=& f^N(\theta^k) - \gamma \cdot \|\nabla f^N(\theta^k)\|^2 + \frac{\gamma^2 L}{2}\mathbb{E}[\|\nabla f_{i^k}(\theta^k)\|^2|\mathcal{F}^{k-1}].
\end{aligned}$$

Quite unfortunately...

$$\mathbb{E}[\|\nabla f_{i^k}(\theta^k)\|^2|\mathcal{F}^{k-1}] = \frac{1}{m}\sum_{i=1}^{m} \|\nabla f_i(\theta^k)\|^2 \geq \|\nabla f^N(\theta^k)\|^2.$$

Convergence of SGD - Strongly Convex

Assumption: bounded variance

$$\mathbb{E}[\|\nabla f_{i^k}(\theta^k)\|^2 | \mathcal{F}^{k-1}] - \|\nabla f^N(\theta^k)\|^2 \leq \sigma^2.$$

Plug in and use the gradient dominance property

$$\begin{aligned}
\mathbb{E}[f^N(\theta^{k+1}) | \mathcal{F}^{k-1}] &\leq f^N(\theta^k) - \gamma \cdot \|\nabla f^N(\theta^k)\|^2 + \frac{\gamma^2 L}{2}\left(\|\nabla f^N(\theta^k)\|^2 + \sigma^2\right) \\
&= f^N(\theta^k) - \gamma\left(1 - \frac{\gamma L}{2}\right) \cdot \|\nabla f^N(\theta^k)\|^2 + \frac{\gamma^2 L}{2}\sigma^2 \\
&\leq f^N(\theta^k) - \gamma\left(1 - \frac{\gamma L}{2}\right) \cdot 2\mu\left(f^N(\theta^k) - f^\star\right) + \frac{\gamma^2 L}{2}\sigma^2
\end{aligned}$$

Convergence analysis - strongly convex smooth function

- Stochastic setting

Theorem

Let f be μ-strongly convex and L-smooth, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by SGD with step size γ satisfies

$$\mathbb{E}[f(\theta^{k+1})] - f^\star \leq \left(1 - 2\mu\gamma\left(1 - \frac{\gamma L}{2}\right)\right)(\mathbb{E}[f(\theta^k)] - f^\star) + \frac{\gamma^2 L}{2}\sigma^2$$

- Optimization error does not go to zero!
- Send γ to zero to reduce the bad term $\longrightarrow$ kills the rate

Convergence analysis - strongly convex smooth function

Theorem

Let f^N be μ-strongly convex and L-smooth, and let γ^k be chosen such that

$$\gamma^k = \frac{\beta}{c+k} \quad \textit{for some} \quad \beta > \frac{1}{\mu},\ c > 0 \quad \textit{such that } \gamma^0 \leq \frac{1}{L}$$

then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by SGD satisfies

$$\mathbb{E}[f^N(\theta^k)] - f^\star \leq \frac{1}{c+k} \max\left\{ \frac{\beta^2\sigma^2 L}{2(\beta\mu - 1)}, (c+1)(f^N(\theta^0) - f^\star) \right\}$$

- Constant γ: linear rate to $\mathcal{N}(\theta^\star)$
- Diminishing γ^k: sublinear rate to $\theta^\star$

We want "GD convergence rate" + "SGD workload per iteration"

Outline

The Stochastic Gradient Descent Algorithm

Variance Reduction Methods

What is "wrong" with SGD

The key inequality

$$\mathbb{E}[f^N(\theta^{k+1})|\mathcal{F}^{k-1}] \leq f^N(\theta^k) - \gamma \left(1 - \frac{\gamma L}{2}\right) \cdot 2\mu \left(f^N(\theta^k) - f^\star\right) + \frac{\gamma^2 L}{2}\sigma^2$$

Decrease γ to kill the last term: sublinear rate

Constant learning rate γ

SGD iteration: $\theta^{k+1} = \theta^k - \gamma \cdot \nabla f_{i^k}(\theta^k)$

Sanity check: assume $\theta^k \to \theta^*$ (not granted), then $\gamma \cdot \nabla f_{i^k}(\theta^k) \to 0$

θ^* cannot be stationary: $\nabla f_i(\theta^*) \neq 0$ for any i.

Solution: correct the gradient to kill $\sigma \implies$ VR methods

Basic idea: replace $\nabla f_{i^k}(\theta^k)$ by g^k such that $g^k \to \nabla f^N(\theta^k)$

Stochastic average gradient

Let us rewrite the gradient as

$$\nabla f^N(\theta^k) = \frac{1}{N}\sum_{i=1}^{N} f_i(\theta^k) = \frac{1}{N}\Big(\nabla f_{i^k}(\theta^k) + \underbrace{\sum_{j\neq i^k} \nabla f_j(\theta^k)}_{\text{not available}}\Big)$$

Replace $\nabla f_j(\theta^k)$ by its latest evaluation $\nabla f_j(\theta^{k-d_j})$.

Implementation:

- Maintain a gradient table v_i storing the latest evaluation of $\nabla f_i(\theta)$
- At iteration k, update table

$$v_i^k = \begin{cases} \nabla f_i(\theta^k), & \text{if } i = i^k \\ v_i^{k-1}, & \text{otherwise.} \end{cases}$$

- Summation can be done cheaply by recycling previous computations

SAG - Convergence Rate

Theorem

Let f^N be μ-strongly convex and **each** *f_i $L_{\max}$-smooth, then the sequence $(\theta^k)_{k\in\mathbb{N}}$ generated by SGD with step size $\gamma = 1/(16L_{\max})$ satisfies*

$$\mathbb{E}\left[f^N(\theta^k)\right] - f^\star \le \left(1 - \min\left\{\frac{\mu}{L_{\max}}, \frac{1}{8m}\right\}\right)^k \times \left(\frac{3}{2}(f^N(\theta^0) - f^\star) + \frac{4L_{\max}}{m}\|\theta^0 - \theta^\star\|^2\right)$$

- Linear rate $O\big((m + L_{\max}/\mu)\log 1/\varepsilon\big)$
- Compare to full gradient in terms of gradient evaluations $O\big(m \cdot (L/\mu)\log 1/\varepsilon\big)$
- Which is better? $(L_{\max} \le mL)$ [prove it]
- Proof is hard – the gradient surrogate $g^k = \frac{1}{N}\sum_{i=1}^N v_i^k$ is biased.

Control Variates

Basic idea: Suppose we want to estimate $\mu = \mathbb{E}X$, and we have some random variable $Y \approx X$ with known mean $\zeta = \mathbb{E}Y$.

Given (X_i, Y_i), let $\widetilde{X}_i = X_i - Y_i + \zeta$, then

Unbiased	$\mathbb{E}(\widetilde{X}_i) = \mathbb{E}X_i = \mu$
Reduced variance	$V(\widetilde{X}_i) \leq \mathbb{E}\|X_i - Y_i\|^2 \approx 0.$

Apply this idea to the gradient estimator

$$\nabla f^N(\theta) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\theta) = \frac{1}{m}\sum_{i=1}^{m} (\nabla f_i(\theta) - v_i + v_i) = \frac{1}{m}\sum_{i=1}^{m} (\nabla f_i(\theta) - v_i + \bar{v})$$

Let $g^k = \underbrace{\nabla f_{i^k}(\theta^k)}_{X_i} - \underbrace{(v_{i^k}^k - \bar{v}^k)}_{Y_i}$.

- g^k is unbiased
- choose v_i such that $v_i^k \to \nabla f_i(\theta^k)$ for variance reduction

SAGA

- Maintain a gradient table v_i storing the latest evaluation of $\nabla f_i(\theta)$
- At iteration k, update table

$$v_i^k = \begin{cases} \nabla f_i(\theta^k), & \text{if } i = i^k \\ v_i^{k-1}, & \text{otherwise.} \end{cases}$$

- SAGA gradient estimator

$$g^k = \nabla f_{i^k}(\theta^k) - v_{i^k}^k + \frac{1}{m}\sum_{i=1}^m v_i^k$$

Recall SAG gradient estimator takes form

$$g^k = \frac{1}{m}\nabla f_{i^k}(\theta^k) - \frac{1}{m} v_{i^k}^k + \frac{1}{m}\sum_{i=1}^m v_i^k$$

SAGA is very similar to SAG

- $\gamma = O(1/L_{\max})$, linear rate $O\big((m + L_{\max}/\mu)\log 1/\varepsilon\big)$
- g^k is unbiased simplifies the proof

SVRG

Drawback of SAG and SAGA: table maintainance cost $O(md)$

How to reduce memory requirement *without sacrificing the rate*?

The idea of SVRG: align the reference points of the v_i's.

Every t iterations, do

- store $\bar{\theta} = \theta^k$
- compute full gradient $\bar{v} = \nabla f^N(\theta^k)$

SVRG gradient estimator

$$g^k = \nabla f_{i^k}(\theta^k) - \nabla f_{i^k}(\bar{\theta}) + \bar{v}$$

Convergence: If $t \sim U\{1, \ldots, m\}$, γ depends on $\mu, L_{\max}, t$, linear rate $O\big((m + L_{\max}/\mu) \log 1/\varepsilon\big)$

- Memory requirement $O(d)$
- Full gradient computation once in a while