Acceleration and Stochastic Gradient Descent

Hanchao Zhang

February 14, 2022

QUIZ

Momentum Gradient Descent

$$w^{k+1} = w^k - \alpha z^{k+1}$$
$$z^{k+1} = \beta z^k + \nabla f(w^k)$$

1. Each step of momentum gradient descent should be closer to the optimal point compared to the gradient descent method with the same step size.
   - TRUE
   - FALSE
2. In the setting of $f(w) = \frac{1}{2} w^T A w - b^T w$, $A \succ 0$, only the largest eigenvalue of A controls the convergence rate of momentum gradient descent.
   - TRUE
   - FALSE

Overview

1. Stochastic Gradient Descent

2. Momentum Acceleration

Stochastic Gradient Descent

Motivation

- no access to full gradient
- to expensive to compute the full gradient

Solution

- use the nosiy (stochastic) version of the gradient
  - stochastic gradient descent
  - random coordinate descent

Quick Peek – Stochasitc Gradient Descent

Gradient Descent

$$x^{k+1} = x^k - \eta \nabla f(x^k)$$

Noisy Gradient

$$\tilde{g}(x) = \nabla f(x) + \epsilon$$

where ϵ is zero mean, and

$$E[\tilde{g}(x)] = \nabla f(x)$$

Stochasitc Optimization

Original Optimization Problem

$$\begin{aligned} &\min f(x) \\ &\text{subject to} \quad x \in \mathcal{X} \end{aligned}$$

Stochastic Optimization

$$\begin{aligned} &\min_x \; E_\xi[f(x;\xi)] \\ &\text{subject to} \quad x \in \mathcal{X} \end{aligned}$$

Example – Regression Problem

$$\begin{aligned} &\min_x \; E_\xi[(y - \xi^T x)^2] \\ &\text{subject to} \quad x \in \mathcal{X} \end{aligned}$$

Example – Regression Setting

$$
\begin{aligned}
&\min_x \; E_\xi[(y - \xi^T x)^2] \\
&= \min_x \frac{1}{n}\sum_{i=1}^{n}(y_i - \xi_i^T x)^2 \\
&= \min_x f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \\
&\text{subject to} \;\; x \in \mathcal{X}
\end{aligned}
$$

Gradient Descent

$$x_{t+1} = x_t - \eta\nabla\Big(\frac{1}{n}\sum_{i=1}^{n} f_i(x)\Big)$$

Very Expensive! Require Full pass over the data.

Example – Stochastic Optmization

1. Stochastic Gradient

$$x \Longrightarrow \square \Longrightarrow \cancel{\nabla f(x)} \Longrightarrow \tilde{g}(x) = \nabla f_I(x)$$

$$I \sim \text{uniform}(1, 2, \ldots, n)$$

Q: is $\tilde{g}$ stochastic gradient?

$$E_I[\nabla f_I(x)] = \sum_{i=1}^{n} \nabla f_i(x) \cdot \frac{1}{n} = \nabla f(x)$$

Example – Stochastic Optmization

2. Random Coordinate Descent

$$x \Longrightarrow \Box \Longrightarrow \cancel{\nabla f(x)} \Longrightarrow \tilde{g}(x) = d\nabla f_J(x) = d \cdot \begin{bmatrix} 0 \\ \vdots \\ \frac{\partial f}{\partial x_j} \\ \vdots \\ 0 \end{bmatrix}$$

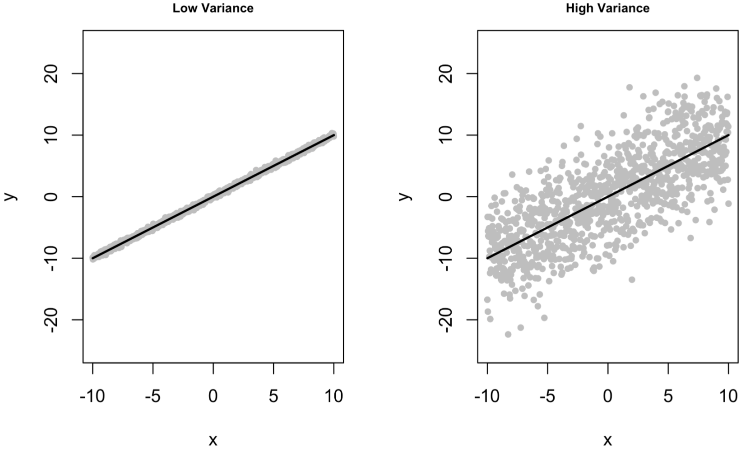
$$J \sim \text{uniform}(1, \dots, d)$$

Q: is $\tilde{g}$ stochastic gradient?

$$E_j[\tilde{g}(x)] = \sum_{j=1}^{n} d \cdot \nabla f_j(x) \cdot \frac{1}{d} = \nabla f(x)$$

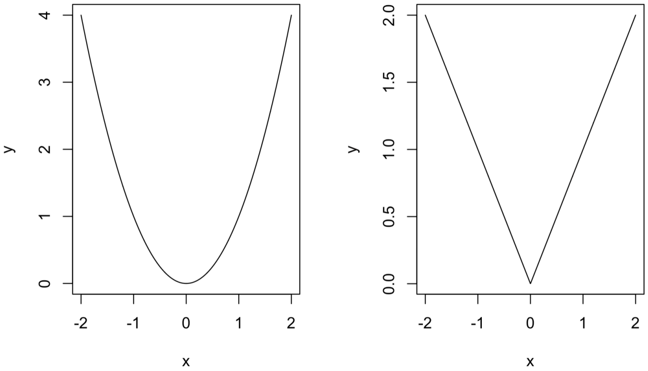
SGD – Role of Variance

Regression Setting



SGD – Why Variance Matters

Self-tuning



When Polyak-Lojasiewicz Inequality hold, $x \to x^*$, $\nabla f(x) \to 0$

SGD

Theorem: Convergence Rate

Suppose x^* exists, and $E_\xi[||\tilde{g}(x)||^2] \leq G^2 \quad \forall x$.

$$x_{t+1} = x_t - \eta \tilde{g}(x)$$

then

$$E_\xi[f(\frac{1}{T}\sum_{t=1}^{T} x_t)] - f(x^*) \leq \frac{RG^2}{\sqrt{T}}$$

$$R^2 \geq ||x_1 - x^*||_2$$

SGD

Theorem: Convergence Rate

Suppose x^* exists, and $E[||\tilde{g}(x)||^2] \leq G^2 \ \ \forall x$, and f is μ strongly convex, may not be smooth, then SGD with decreasing step size $\eta_t = \frac{2}{\mu(t+1)}$

$$x_{t+1} = x_t - \eta_t \tilde{g}(x)$$

then

$$E[f(\frac{2t}{T(T+1)}\sum_{t=1}^{T} x_t)] - f(x^*) \leq \frac{2G^2}{\mu}\frac{1}{T+1} = \mathcal{O}(\frac{1}{T})$$

Quick Summary

- convergence rate $= \mathcal{O}(\frac{1}{\sqrt{T}})$ when $E[||\tilde{g}||_2^2] \leq G^2$

Question: Can we go a little bit better, what's the key?

Quick Summary

- convergence rate $= \mathcal{O}(\frac{1}{\sqrt{T}})$ when $E[||\tilde{g}||_2^2] \leq G^2$
- convergence rate $= \mathcal{O}(\frac{1}{T})$ when we have strong convexity

Question: Can we go a little bit better, what's the key?

Quick Summary

- convergence rate $= \mathcal{O}(\frac{1}{\sqrt{T}})$ when $E[||\tilde{g}||_2^2] \leq G^2$
- convergence rate $= \mathcal{O}(\frac{1}{T})$ when we have strong convexity

Question: Can we go a little bit better, what's the key?

Quick Summary

- convergence rate $= \mathcal{O}(\frac{1}{\sqrt{T}})$ when $E[||\tilde{g}||_2^2] \leq G^2$
- convergence rate $= \mathcal{O}(\frac{1}{T})$ when we have strong convexity

Question: Can we go a little bit better, what's the key?

self-tuning property

Two Ways To Reduce Variance

- Mini-Batch
- Recentering

Convergence Rate of SGD

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

$$E(\frac{1}{T}\sum_{i=1}^{T} x_t) - f(x^*) \leq \frac{RG^2}{\sqrt{T}} \leq \frac{||x_1 - x^*||}{\sqrt{T}} \sqrt{E[||\nabla f_I(x)||_2^2]}$$

Recall that $R \leq ||x_1 - x^*||$, and $E[||\nabla f_I(x)||_2^2] \leq G^2$

Convergence Rate of SGD

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

$$E(\frac{1}{T}\sum_{i=1}^{T} x_t) - f(x^*) \leq \frac{RG^2}{\sqrt{T}} \leq \frac{||x_1 - x^*||}{\sqrt{T}} \sqrt{E[||\nabla f_I(x)||_2^2]}$$

Recall that $R \leq ||x_1 - x^*||$, and $E[||\nabla f_I(x)||_2^2] \leq G^2$

Question: Can we control $\sqrt{E[||\nabla f_I(x)||_2^2]}$?

Mini-Batch SGD

$$x \Longrightarrow \square \Longrightarrow \cancel{\nabla f(x)} \Longrightarrow \tilde{g}(x) = \frac{1}{B}\sum_{j=1}^{B} \nabla f_{I_j}(x)$$

$$I_j \sim \text{uniform}(1, 2, \ldots, n)$$

Question: is $\tilde{g}(x)$ stochastic gradient?

Mini-Batch SGD

$$x \Longrightarrow \square \Longrightarrow \cancel{\nabla f(x)} \Longrightarrow \tilde{g}(x) = \frac{1}{B}\sum_{j=1}^{B} \nabla f_{I_j}(x)$$

$$I_j \sim \text{uniform}(1, 2, \ldots, n)$$

Question: is $\tilde{g}(x)$ stochastic gradient?
check:

$$E_I[\frac{1}{B}\sum_{j=1}^{B} \nabla f_{I_j}(x)] = \frac{1}{B}\sum_{j=1}^{B} \nabla f(x) = \nabla f(x)$$

Does it help?
assume variance is independent,

$$Var(\frac{1}{B}\sum_{j=1}^{B} \nabla f_{I_j}(x)) = \frac{1}{B^2}\sum_{j=1}^{B} Var(f_{I_j}(x)) = \frac{1}{B} Var(f_{I_j}(x))$$

Mini-Batch SGD

Advantages

- reduce variance
- mini-batch is parallelizable

Disadvantages

- more work per iteration
- no self-tunning when the f is smooth

Mini-Batch SGD

Advantages

- reduce variance
- mini-batch is parallelizable

Disadvantages

- more work per iteration
- no self-tunning when the f is smooth

Question: can we do better?

Mini-Batch SGD

Advantages

- reduce variance
- mini-batch is parallelizable

Disadvantages

- more work per iteration
- no self-tunning when the f is smooth

Question: can we do better?

We need self-tuning when f smooth and strongly convex

Reduce Variance by Recentering

For x and y

$$x, y \Longrightarrow \square \Longrightarrow \cancel{\nabla f(x)} \Longrightarrow \tilde{g}(x) = \nabla f_I(x) - (\nabla f_I(y) - \nabla f(y))$$

$$I \sim \text{uniform}(1, 2, \dots, n)$$

$$E_I[\tilde{g}(x)] = E_I[\nabla f_I(x) - (\nabla f_I(y) - \nabla f(y))]$$

$$= \nabla f(x) - \cancelto{0}{(\nabla f(y) - \nabla f(y))}$$

Stochastic Variance Reduced Gradient Descent Algorithm (SVRG)

Outer Loop:

On the k^{th} iteration

$x_1 = y_k$

Inner Loop:

for $t = 1, 2, \ldots T$

$$x_{t+1} = x_t - \eta\Big(\nabla f_I(x_t) - \big(\nabla f_I(y_k) - \nabla f(y_k)\big)\Big)$$

update $y_{k+1} = \frac{1}{T}\sum_{i=1}^{T} x_t$, and compute $\nabla f(y_{k+1})$

- inner loop only compute $\nabla f_I(x_t)$
- outter loop compute full gradient $\nabla f(y_{k+1})$

Variance Reduction Lemma

Variance Reduction Lemma

Let $f_1 \dots f_n$ be L-smooth, $I \sim \text{uniform}(1, \dots n)$. Then
$E_I\left[||\nabla f_I(x) - \nabla f_I(x^*)||_2^2\right] \leq 2L(f(x) - f(x^*))$

Note: $\nabla f_I(x)$ may not be small when $x \to x^*$

Proof of Variance Reduction Lemma

Let $g_i(x) = f_i(x) - [f_i(x^*) + \nabla f_i(x^*)^T(x - x^*)] \geq 0$ by convextiy

If h is convex and L-smooth, $h(x - \frac{1}{L}\nabla h(x)) \leq h(x) - \frac{1}{2L}||\nabla h(x)||_2^2$ and applies this to g

$$
\begin{aligned}
0 \leq g_i(x - \frac{1}{L}\nabla g_i(x)) &\leq g_i(x) - \frac{1}{2L}||\nabla g_i(x)||_2^2 \\
&\downarrow \\
-g_i(x) &\leq -\frac{1}{2L}||\nabla g_i(x)||_2^2 \\
&\downarrow \\
||\nabla g_i(x)||_2^2 &\leq 2Lg_i(x)
\end{aligned}
$$

Proof of Variance Reduction Lemma – Continuous

$$g_i(x) = f_i(x) - [f_i(x^*) + \nabla f_i(x^*)^T(x - x^*)] \geq 0$$

$$||\nabla g_i(x)||_2^2 = ||\nabla f_i(x) - \nabla f_i(x^*)||_2^2 \leq 2L\Big(f_i(x) - f_i(x^*) + \nabla f_i(x^*)^T(x - x^*)\Big)$$

$$E_I\Big[||\nabla f_I(x) - \nabla f_I(x^*)||\Big] \leq 2LE\Big[f_I(x) - f_I(x^*) + \nabla f_I(x^*)^T(x - x^*)\Big]$$

$$\leq 2L\Big(f(x) - f(x^*) + \cancelto{0}{\nabla f(x^*)^T}(x - x^*)\Big)$$

The recentered gradient $E_I\Big[||\nabla f_I(x) - \nabla f_I(x^*)||\Big] \rightarrow 0$, when $x \rightarrow x^*$

Stochastic Variance Reduction Griadent Desecent

SVRG Theorem

Let $f = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, f_i is L-smooth, and f is μ strongly convex. SVRG algorithm with step size $\eta = \frac{1}{10 \cdot L}$, and inner loop size $T = 10 \cdot (\frac{L}{\mu})$.
Then after $s+1$ iterations of the outer loop

$$E[f(y^{s+1})] - f(x^*) \leq 0.9^s (f(y^1) - f(x^*))$$

Key Feature:

- linear convergence
- L/μ does not appear in the convergence rate

Proof of SVRG Algorithm

good enough to prove $E[f(y^{s+1})] - f(x^*) \leq 0.9(f(y^s) - f(x^*))$

Recall: $y^{s+1} = \frac{1}{T}\sum x_t$, where x_t is provided in s^{th} inner loop

$$\left|\left|x_{t+1} - x^*\right|\right|_2^2 = \left|\left|x_t - \eta\Big(\nabla f_{I_t}(x_t) - (\nabla f_{I_t}(y) - \nabla f(y))\Big) - x^*\right|\right|_2^2$$

$$= ||x_t - x^*||_2^2 - 2\eta \underbrace{\Big(\nabla f_{I_t}(x_t) - \nabla f_{I_t}(y) + \nabla f(y)\Big)^T}_{V_t}(x_t - x^*) + \eta^2||V_t|||_2^2$$

$$= \underbrace{||x_t - x^*||_2^2}_{a} - \underbrace{2\eta V_t^T(x_t - x^*)}_{b} + \underbrace{\overbrace{\eta^2||V_t||_2^2}^{\text{variance term}}}_{c}$$

$a \to 0$ and $b \to 0$ when $x_t \to x^*$, we want the term c goes to 0 as well

Proof of SVRG Algorithm

Let's just look at the term c

$$
\begin{aligned}
E\left[||V_t||_2^2\right] &= E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(y) + \nabla f(y)||_2^2\right] \\
&= E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(x^*) + \nabla f_{i_t}(x^*) - \nabla f_{i_t}(y) + \nabla f(y)||_2^2\right] \\
&\leq 2E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(x^*)||_2^2\right] + \underbrace{2E\left[||\nabla f_{I_t}(x^*) - \nabla f_{I_t}(y) + \nabla f(y)||_2^2\right]}_{=0} \\
&\leq 2E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(x^*)||_2^2\right] + 2E\left[||\nabla f_{I_t}(y) - f_{I_t}(x^*)||_2^2\right] \\
&\leq 4L(f(x_t) - f(x^*) + f(y) - f(x^*)) \quad \text{by variance reduction lemma twice}
\end{aligned}
$$

Proof of SVRG Algorithm

Let's just look at the term c

$$
\begin{aligned}
E\left[||V_t||_2^2\right] &= E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(y) + \nabla f(y)||_2^2\right] \\
&= E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(x^*) + \nabla f_{i_t}(x^*) - \nabla f_{i_t}(y) + \nabla f(y)||_2^2\right] \\
&\leq 2E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(x^*)||_2^2\right] + \underbrace{2E\left[||\nabla f_{I_t}(x^*) - \nabla f_{I_t}(y) + \nabla f(y)||_2^2\right]}_{=0} \\
&\leq 2E\left[||\nabla f_{I_t}(x_t) - \nabla f_{I_t}(x^*)||_2^2\right] + 2E\left[||\nabla f_{I_t}(y) - f_{I_t}(x^*)||_2^2\right] \\
&\leq 4L(f(x_t) - f(x^*) + f(y) - f(x^*)) \quad \text{by variance reduction lemma twice}
\end{aligned}
$$

ps $E[(Y - E(Y))^2] \leq E[Y^2]$

Proof of SVRG Algorithm

term b

$$E\left[2\eta V_t^T(x_t - x^*)\right] = 2\eta E[V_t]^T(x_t - x^*)$$
$$= 2\eta \nabla f(x_t)^T(x_t - x^*)$$
$$\geq 2\eta\big(f(x_t) - f(x^*)\big) \quad \text{by convexity}$$

Then, we have

$$E\left[||x_{t+1} - x^*||_2^2\right] = E[a + b + c]$$
$$\leq ||x_t - x^*||_2^2 - 2\eta\big(f(x_t) - f(x^*)\big) + 4\eta^2 L(f(x_t) - f(x^*) + f(y) - f(x^*))$$
$$= ||x_t - x^*||_2^2 - 2\eta(1 - 2\eta L)(f(x_t) - f(x^*)) + 4\eta^2 L(f(y) - f(x^*))$$
$$\downarrow \text{iterating}$$
$$= ||x_1 - x^*||_2^2 - 2\eta(1 - 2\eta L) \cdot E[\sum_{k=1}^{t}(f(x_k) - f(x^*))] + 4\eta^2 L \cdot t(f(y) - f(x^*)$$

Proof of SVRG Algorithm

Recall:

- $x_1 = y^k$
- $||y - x^*||_2^2 \leq \frac{2}{\mu}(f(y) - f(x^*))$

$$E\left[||x_{t+1} - x^*||_2^2\right] \leq ||x_1 - x^*||_2^2 - 2\eta(1-2\eta)L \cdot E[\sum_{k=1}^{t}(f(x_k) - f(x^*))] + 4\eta^2 L \cdot t(f(y) - f(x^*$$

$$2\eta(1-2\eta)L \cdot E[f(\frac{1}{T}\sum x_t) - f(x^*)] \leq (\frac{2}{\mu} + \eta^2 4LT)\frac{1}{T}E(f(y) - f(x^*))$$

$$E[f(y^{s+1})] - f(x^*) \leq 0.9\big(E[f(y^s)] - f(x^*)\big)$$

Accelerate Gradient Descent

Key Idea: use gradient computed at previous step to accelerate the algorithm

Methods:

- Momentum
- Nesterov

Momentum Acceleration

Momentum Method:

$$x_{k+1} = x_k - \eta z_k$$
$$z_k = \nabla f(x_k) + \beta z_{k-1}$$

Nesterov Method:

$$x_{t+1} = x_t + d_t - \eta \nabla f(x_t + d_t)$$
$$d_t = \gamma_t (x_t - x_{t-1})$$

Theorem

Let f to be L-smooth and μ-convex.

$$f(z_t) - f(x^*) \leq \frac{L+\mu}{2} ||x_1 - x^*||_2^2 \cdot \exp\{\frac{(t+1)}{\sqrt{k}}\}$$

Summary

- gradient descent: $\mathcal{O}(n\frac{L}{\mu}\log(\frac{1}{\epsilon}))$ iterations for ϵ-accuracy
- momentum acceleration: $\mathcal{O}(n\sqrt{\frac{L}{\mu}}\log(\frac{1}{\epsilon}))$ iterations for ϵ-accuracy
- stochastic gradient descent: $\mathcal{O}(\frac{1}{\mu\epsilon})$ for ϵ-accuracy
- SVRG: $\mathcal{O}((n+\frac{L}{\mu})\log(\frac{1}{\varepsilon}))$ for ϵ-accuracy

The End