SGD: Stochastic Gradient Methods

Jingchang Liu

University of Science and Technology of China

April 6, 2017

Table of Contents

Table of Contents

Motivation

Empirical Risk minimize

- Regression: $F(w) = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i;w) - y_i\right)^2 \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(w)$
- Classification: $F(w) = \frac{1}{n}\sum_{k=1}^{n} I\left(f(x_i;w) = y_i\right) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(w)$
- Neural Networks: $E = \frac{1}{n}\sum_{k=1}^{n} E_k = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{1}{2}\sum_{j=1}^{l}\left(\hat{y}_j^k - y_j^k\right)^2\right]$

Regularization

- $F(w) = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i;w) - y_i\right)^2 + \frac{\lambda}{2}\|w\|^2 \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(w)$

Algorithm 1 Stochastic Gradient(SG) Method

Choose an initial iterate w_1.
for $k = 1, 2, \cdots$ **do**
  Generate a realization of the random variable i_k.
  Compute a stochastic vector $\nabla_{i_k} f\left(w_k\right)$
  Choose a stepsize $\alpha_k > 0$
  Set the new iterate as $w_{k+1} \leftarrow w_k - \alpha_k \nabla_{i_k} f\left(w_k\right)$
end for

Gradient Methods

Full Gradient

$$w_{k+1} \leftarrow w_k + \alpha_k \nabla F(w) = w_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} f_i(w_k)$$

Stochastic Gradient

$$w_{k+1} \leftarrow w_k + \alpha_k \nabla f_{i_k}(w)$$

Batch Gradient

$$w_{k+1} \leftarrow w_k - \frac{\alpha_k}{|S_k|} \sum_{i \in S_k} \nabla f_i(w_k)$$

Gradient Methods

Formulation

$$\min_{w} F\left(w\right) = \min_{w} \frac{1}{n} \sum_{i=1}^{n} f_i\left(w\right)$$

Gradient

$$g\left(w_k, \xi_k\right) = \begin{cases} \nabla f\left(w_k; \xi_k\right) & \text{Stochastic Gradient} \\ \frac{1}{|S_k|} \sum_{i=1}^{|S_k|} \nabla f\left(w_k; \xi_{k,i}\right) & \text{Batch Gradient} \\ \nabla F\left(w_k\right) & \text{Full Gradient} \end{cases}$$

Table of Contents

Assumption

Lipschitz-continuous gradient

The objective function $f : \Re^d \rightarrow \Re$ is continuously differentiable and the gradient function of f, namely, $\nabla f : \Re^d \rightarrow \Re^d$, is Lipschitz continuous with Lipschitz constant $L > 0$,i.e,

$$\left\| \nabla f\left(w\right) - \nabla f\left(\bar{w}\right) \right\|_2 \leq L \left\| w - \bar{w} \right\|_2 \quad \text{for all } \{w, \bar{w}\} \subset \Re^d \quad (1)$$

Strongly convex

f is strongly convex with parameter $c > 0$ if

$$g\left(x\right) = f\left(x\right) - \frac{c}{2} x^T x \quad \text{is convex}$$

Equal Properties

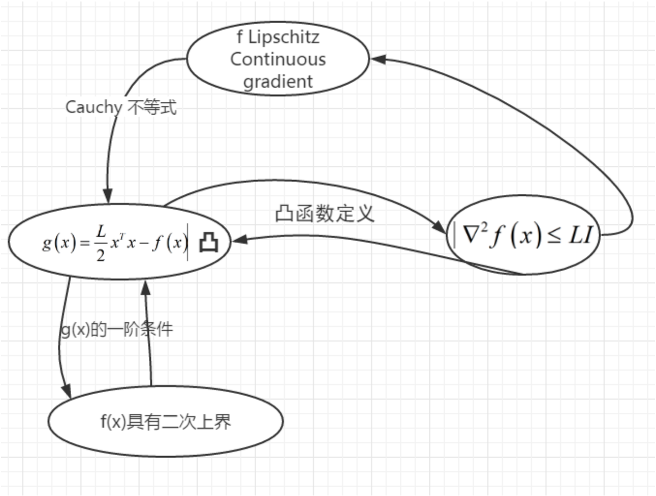

Figure: Equal Properties introduced from Lipschitz

Tips 1

1. $$\|\nabla f(x) - \nabla f(x)\|_2 \leq L\|x - y\|_2$$

2. $$\begin{aligned}(\nabla f(x) - \nabla f(y))^T(x - y) &\leq \|\nabla f(x) - \nabla f(x)\|_2 \|x - y\|_2 \\ &\leq L\|x - y\|_2^2\end{aligned}$$

3. $$\text{def.} \quad g(x) = \frac{L}{2}x^T x - f(x) \qquad \text{Then} \quad \nabla g(x) = Lx - \nabla f(x)$$

Montonicity of gradient

A differentiable function g is convex if and only if $dom\, g$ is convex and

$$(\nabla g(x) - \nabla g(y))^T(x - y) \geq 0 \quad \text{for all } x, y \in \mathrm{dom}\, g$$

Equal Properties

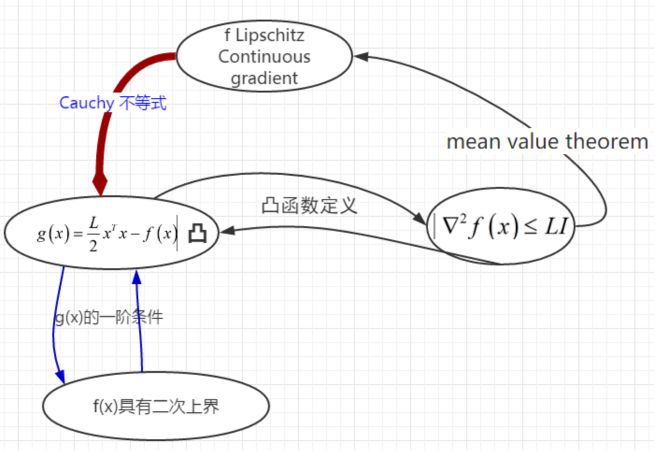

Figure: Equal Properties introduced from Lipschitz

Tips 2

First-order Condition of Convex Function g

$$g(y) \geq g(x) + \nabla g(x)^T (x - y) \quad \text{for all } x, y \in \mathbf{dom}\, g$$

Quadratic upper bound on f

$$g(x) = \frac{L}{2} x^T x - f(x) \qquad \text{Then} \quad \nabla g(x) = Lx - \nabla f(x)$$

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$$

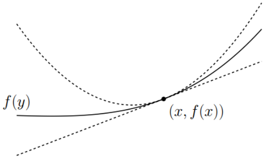

Figure: Quadratic upper bound

Equal Properties

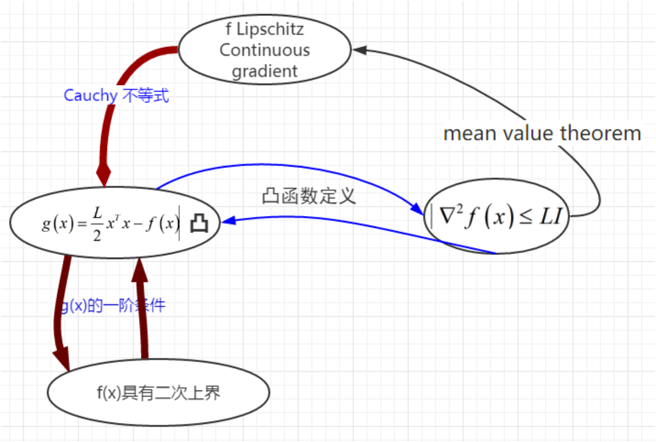

Figure: Equal Properties introduced from Lipschitz

Tip3

Second-order Condition of Convex Function g

$$\nabla^2 g(x) \succeq 0$$

Second-order Gradient of f

$$\nabla g(x) = Lx - \nabla f(x)$$
$$\nabla^2 g(x) = L - \nabla^2 f(x)$$
$$\nabla^2 f(x) \preceq L$$

Equal Properties

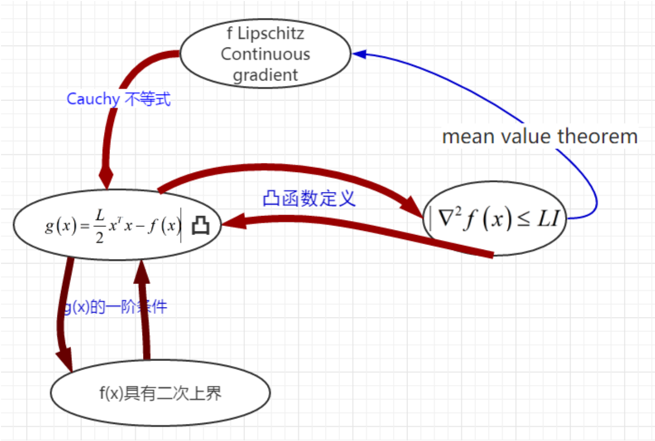

Figure: Equal Properties introduced from Lipschitz

Tips 4

Mean value theorem of $f^{'}$

A function f is continuous on the closed interval $[a, b]$, and differentiable on the open interval (a, b), then there exsists a point c in (a, b) such that:

$$f^{'}(c) = \frac{f(b) - f(a)}{b - a}$$

Mean value theorem of $\nabla^2 f$

$$\frac{\nabla f(x) - f(x)}{x - y} = \nabla^2 f(\zeta) \leq L$$

Equal Properties

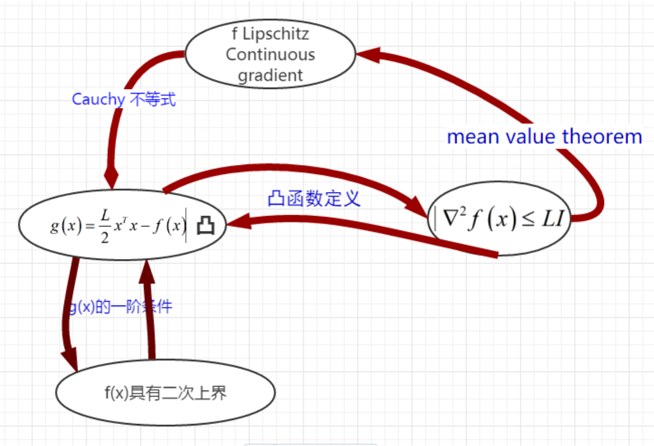

Figure: Equal Properties introduced from Lipschitz

Table of Contents

Theorem

Th

Under Lipschitz-continuous Assumption, the iterations of SG satisfy the following inequality for all $k \in N$:

$$\begin{aligned} E_{\xi_k}\left[F\left(w_{k+1}\right)\right] - F\left(w_k\right) \leq & -\alpha_k \nabla F\left(w_k\right)^T E_{\xi_k}\left[g\left(w_k, \xi_k\right)\right] \\ & + \frac{1}{2}\alpha_k^2 L E_{\xi_k}\left[\left\|g\left(w_k, \xi_k\right)\right\|_2^2\right] \end{aligned}$$

Proof

By Lipschitz-continuous Assumption, the iterates generated by SG satisfy

$$\begin{aligned} F\left(w_{k+1}\right)-F\left(w_k\right) &\le \nabla F\left(w_k\right)^T\left(w_{k+1}-w_k\right)+\frac{1}{2}L\left\|w_{k+1}-w_k\right\|_2^2 \\ &\le -\alpha_k\nabla F\left(w_k\right)^T g\left(w_k,\xi_k\right)+\frac{1}{2}\alpha_k^2 L\left\|g\left(w_k,\xi_k\right)\right\|_2^2 \end{aligned}$$

Noting that w_{k+1} but not w_k depends on ξ_k. Take expectation with respect to ξ_k, we'll get

$$\begin{aligned} E_{\xi_k}\left[F\left(w_{k+1}\right)\right]-F\left(w_k\right) \le -\alpha_k\nabla F\left(w_k\right)^T E_{\xi_k}\left[g\left(w_k,\xi_k\right)\right] \\ +\frac{1}{2}\alpha_k^2 L E_{\xi_k}\left[\left\|g\left(w_k,\xi_k\right)\right\|_2^2\right] \end{aligned}$$

Variance Reduce

Trade-Offs of Mini-Batch

$$g(w_k, \xi_k) = \frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(w_k)$$

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] = \frac{\mathbb{V}_{\xi_k}[\nabla_{i_k} f(w_k)]}{|S_k|}$$

Popular Variance Reduce Methods

SAG, SAGA, SVRG

Convergence Rate of Full Gradient

Th

Assume F is convex and L-Lipschitz continuous gradient. With step size $\alpha = \frac{1}{L}$. Then

$$F\left(x_{k+1}\right) - F^* \leq \frac{2L\left\|x_1 - x^*\right\|_2^2}{k}$$

Th

Assume F is c-strongly convex and L-Lipschitz continuous gradient. With step size $\alpha = \frac{2}{M+L}$. Then

$$F\left(x_{k+1}\right) - F^* \leq \frac{L}{2}\exp\left(-\frac{4k}{Q+1}\right)\left\|x_1 - x^*\right\|_2^2$$

where $Q = \frac{L}{c}$

Convergence Rate of SG, Fixed Stepsize

Th

Assume F is c-strongly convex and L-Lipschitz continuous gradient. Under some mild assumption. Suppose that the SG method is run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}$$

Then, the expected optimality gap satisfies the following inequality for all $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}\left[F\left(w_k\right) - F_*\right] &\leq \frac{\alpha \bar{L} M}{2c\mu} + \left(1 - \bar{\alpha} c \mu\right)^{k-1}\left(F\left(w_1\right) - F_* - \frac{\bar{\alpha} LM}{2c\mu}\right) \\ &\xrightarrow{k\to\infty} \frac{\alpha \bar{L} M}{2c\mu}\end{aligned}$$

Convergence Rate of SG, Diminishing Stepsize

Assume F is c-strongly convex and L-Lipschitz continuous gradient. Under some mild assumption. Suppose that the SG method is run with a stepsize sequence such that, for all $k \in \mathbb{N}$,

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{for some } \beta > \frac{1}{c\mu} \text{ and } \gamma > 0 \quad \text{such that } \alpha_1 \le \frac{\mu}{LM_G}$$

Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}\left[F\left(w_k\right) - F_*\right] \le \frac{\upsilon}{\gamma + k}$$

where

$$\upsilon := \max\left\{\frac{\beta^2 LM}{2\left(\beta c\mu - 1\right)}, \left(\gamma + 1\right)\left(F\left(w_1\right) - F_*\right)\right\}$$

Q&A