Statistical Machine Learning

Instructor: Jie Shen

Dept. of Computer Science

February 14, 2020

Gradient Descent

Consider minimization without constraints:

$$\min_{\boldsymbol{w}} \; F(\boldsymbol{w}), \;\; \boldsymbol{w} \in \mathbb{R}^d$$

Gradient Descent:

1. Initialize $\boldsymbol{w}^0$ arbitrarily, e.g. $\boldsymbol{w}^0 = \mathbf{0}$
2. For $t = 1, 2, \ldots$

$$\boldsymbol{w}^t = \boldsymbol{w}^{t-1} - \eta_t \nabla F(\boldsymbol{w}^{t-1}) \tag{1}$$

Goal:

- $\boldsymbol{w}^t \to \boldsymbol{w}^*$, where $\boldsymbol{w}^* = \arg\min F(\boldsymbol{w})$
- in few iterations (cheap computation)

Informal Analysis

$\forall t, \quad F(w^{t+1}) - F(w^t) < 0 \iff F(w^{t+1}) < F(w^t)$

Why GD "decreases" objective value (under proper conditions)?

$w^{t+1} = w^t - \eta \cdot \nabla F(w^t)$

$F(w^{t+1}) - F(w^t) = \langle \nabla F(w), \; w^{t+1} - w^t \rangle$

$= \langle \nabla F(w), \; -\eta \cdot \nabla F(w^t) \rangle$

$= -\eta \cdot \langle \nabla F(w), \nabla F(w^t) \rangle$

$\eta \cdot \nabla F(\quad) \quad \eta = 0$

$\approx \nabla F(w^t)$

$\approx 0 \qquad \eta \to 0. \qquad w \to w^t \Rightarrow \nabla F(w)$

$\approx \nabla F(w^t)$

Mean-value Thm.

$F: \mathbb{R} \to \mathbb{R}$

$\forall\ x, y \in \mathbb{R}.$

$$F(x) - F(y) = \underbrace{F'(z)}_{\mathbb{R}^d} \cdot \underbrace{(x-y)}_{\mathbb{R}^d}$$

$z \in [x, y]$

$F: \mathbb{R}^d \to \mathbb{R}$

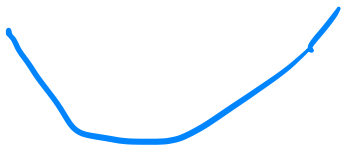
$\forall\ x, y \in \mathbb{R}^d$

$$F(x) - F(y) = \langle \nabla F(z),\ x - y \rangle$$

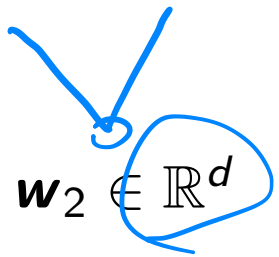
$\mathbb{R}$ $\quad$ $\mathbb{R}$ $\quad$ $\mathbb{R}^d \cdot \mathbb{R}^d$

$z = \lambda x + (1-\lambda) y, \quad \lambda \in [0, 1]$

Smoothness



$(0, +\infty)$



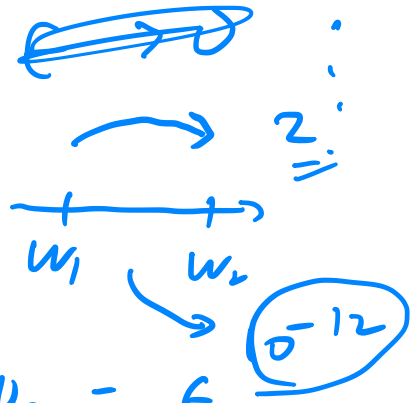
$L = 1\text{M}. \quad 10^9$

$\epsilon = \frac{1}{10^9} \quad \frac{1}{10^2}$

Smooth: $F(\boldsymbol{w})$ is smooth if for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$

$$\|\nabla F(\boldsymbol{w}_2) - \nabla F(\boldsymbol{w}_1)\|_2 \leq L \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$$

ϵ



Examples:

$F(w) = w^2$

$\nabla F(w_2) = 2w_2$

$\nabla F(w_1) = 2w_1$

$LHS = 2 \cdot \|w_2 - w_1\|$

$L = 2$. 2-smooth.

$F(w) = |w|$.

$w_1 = -\epsilon$. $w_2 = \epsilon$

$\nabla F_1 = -1$, $\nabla F_2 = 1$

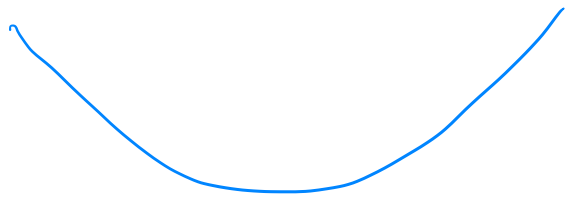
$LHS = 2 \quad RHS = L \cdot 2\epsilon$

can we find L.

$2 < 2L\epsilon$?

$\Leftrightarrow \quad L \cdot \epsilon > 1$

$F(w) = w^4$



$\nabla = 4 \cdot w^3$.

$LHS = 4 \cdot |w_1^3 - w_2^3|$

$RHS = L \cdot |w_1 - w_2|$

Can find $L > 0$. s.t. $\forall w_1, w_2 \in \mathbb{R}$

$4 \cdot |w_1^3 - w_2^3| \leq L \cdot |w_1 - w_2|$

$\Leftrightarrow$ $L \geq 4 \cdot |w_1^2 + w_1 w_2 + w_2^2|$

$\forall w_1, w_2 \in \mathbb{R}$.

non-smooth.

domain $= [-1, 1]$ ✓

When GD fails to find global optimum

- when it terminates

- when it gets stuck at a non-optimal point

The Big Picture

Convex Optimization

Convex set: $\mathcal{C} \subset \mathbb{R}^d$ is said to be a convex set if for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and any $0 \leq \lambda \leq 1$, $\lambda \boldsymbol{u} + (1 - \lambda)\boldsymbol{v} \in \mathcal{C}$

- illustration, examples

Convex Function

Convex function: $F(\boldsymbol{w})$ is said to be a convex function if the set $\mathcal{E} = \{(\boldsymbol{w}, y) : \ y \geq F(\boldsymbol{w})\}$ is convex

- $\mathcal{E}$ is called the epigraph of $F(\boldsymbol{w})$
- illustration

Characterization of Convex Functions

Theorem 1

Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. The following are equivalent:

1. *F is convex;*
2. *$F(\boldsymbol{w}_2) \geq F(\boldsymbol{w}_1) + \langle \nabla F(\boldsymbol{w}_1), \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle$;*
3. *$\nabla^2 F(\boldsymbol{w})$ is positive semi-definite.*

Typically use 3 to check the convexity.

Let f and h be convex functions.

- $a \cdot f + b \cdot h$ is convex when $a \geq 0$ and $b \geq 0$
- $f(h)$ may NOT be convex

Convex Program

$$\min_{\boldsymbol{w}} \; F(\boldsymbol{w}), \quad \boldsymbol{w} \in \mathcal{C}.$$

Convex Program: both $F(\boldsymbol{w})$ and $\mathcal{C}$ are convex

- optimality: local optimum $\Longleftrightarrow$ global optimum
- works well
- easy to solve

Convergence Analysis

$\epsilon \in (0,1)$. $t = ?$ $F(w^t) - F(w^*) \leq \epsilon$

Theorem 2

Suppose $F(\boldsymbol{w})$ is convex and L-smooth. Pick $0 < \eta \leq 1/L$. Then for all $t \geq 1$,

$$F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*) \leq \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2}{2\eta} \cdot \frac{1}{t}$$

$t = 1/\epsilon$

In particular, picking $\eta = 1/L$ gives

$$F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*) \leq \frac{2L\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2}{2t}.$$

Input: $\epsilon \in (0,1)$

$F(w)$

Output: w^t. $F(w^t) - F(w^*) \le \epsilon$

Alg: $\eta = 1/L$

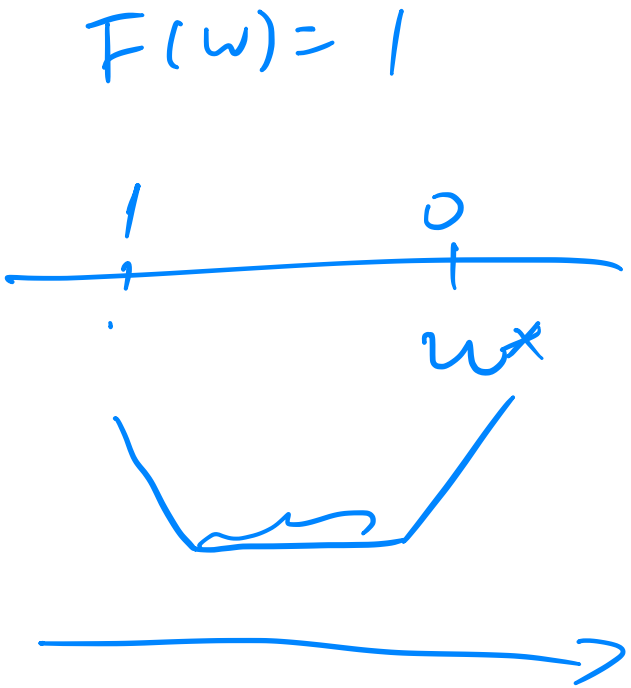
$T = 1/\epsilon$

$t = 1, \cdots T$

$w^{t+1} \leftarrow w^t - \eta \cdot \nabla F(w^t)$

Implications

1. $F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*)$ v.s. $\|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2$



Implications

2. Iteration complexity

$t = 1/\epsilon$

$t : \quad F(w^t) - F(w^*) \leq \epsilon$

t

Implications

$$\sup_{w_1, w_2} \frac{\| \nabla^2 F(w) \cdot (w_2 - w_1) \|}{\| w_2 - w_1 \|} = \sup_{v} \frac{\| \nabla^2 F(w) \cdot v \|}{\| v \|}$$

3. Estimate L

L = max-eigen value of Hessian

$\lambda \in \mathbb{R}$, $v \in \mathbb{R}^d$

(λ, v), M.

if $M \cdot v = \lambda \cdot v$

$\| M \cdot v \| = |\lambda| \cdot \| v \|$

$$\frac{\| \nabla F(w_2) - \nabla F(w_1) \|}{\| w_2 - w_1 \|} \leq L$$

$$\Leftrightarrow L = \sup_{w_1, w_2} \frac{\| \nabla F(w_2) - \nabla F(w_1) \|}{\| w_2 - w_1 \|}$$

$$\frac{\| M \cdot v \|}{\| v \|} = |\lambda|$$

$$\frac{\| \nabla F(w) \cdot v \|}{\| v \|} = \lambda$$

$F: \mathbb{R} \to \mathbb{R}$ ✓

$F: \mathbb{R}^d \to \mathbb{R}$ ✓

$g: \mathbb{R}^d \to \mathbb{R}^d$

w_1, w_2

matrix-ve. multi.

$$g(w_2) - g(w_1) = \nabla g(w) \cdot (w_2 - w_1)$$

$\mathbb{R}^d \quad \mathbb{R}^d \quad \mathbb{R}^{d \times d} \quad \mathbb{R}^d$

$g \triangleq \nabla F$

$$\nabla F(w_2) - \nabla F(w_1) = \nabla^2 F(w) \cdot (w_2 - w_1)$$

Given $F(w)$

$\nabla^2 F(w)$ by yourself.

↓

python / Matlab

Faster Rate of Convergence

w^* is unique. $F(w) = w^4$ $\times$

not unique. $w_2 \neq w^*$. $\frac{\|\nabla - \|}{\|w - \cdot\|} = 4(w_1^2 + w_1 w_2 + w_2^2)$

Strongly Convex: for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$

$\nabla^2 F \geq \lambda \cdot I$ $\|\nabla F(\boldsymbol{w}_2) - \nabla F(\boldsymbol{w}_1)\|_2 \geq \alpha \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$ $\alpha \in (1, \infty)$

$\geq 4 \cdot 3$

- functions satisfying SC
- not satisfying

$0 \geq \alpha \cdot \|w_2 - w^*\| = 12$

$\Rightarrow w_2 = w^*$ 7:40

$\nabla^2 \geq 0$

$F(w) = w^2$ $\checkmark \lambda \cdot I$ $\alpha =$ min-eigenvalue of Hessian.

2-SC. convex

$F(w) = g(w) + \lambda \cdot \|w\|^2$ λ-SC.

Theoretical Guarantee

Theorem 3

Suppose $F(\boldsymbol{w})$ is α-strongly convex and L-smooth. Let $\{\boldsymbol{w}^t\}_{t\geq 1}$ be the iterates generated by GD where $0 < \eta \leq 2/(\alpha + L)$. Then for all $t \geq 1$,

$$\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2 \leq \sqrt{1 - \frac{2\eta\alpha L}{\alpha + L}} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2 .$$

In particular, picking $\eta = 2/(\alpha + L)$ gives

$$\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2 \leq \left(1 - \frac{2}{c+1}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2$$

where $c \overset{\text{def}}{=} L/\alpha$ is the condition number.

- converges linearly / geometric rate of convergence
- typically the best one can hope

Implications

For all $t \geq 1$,

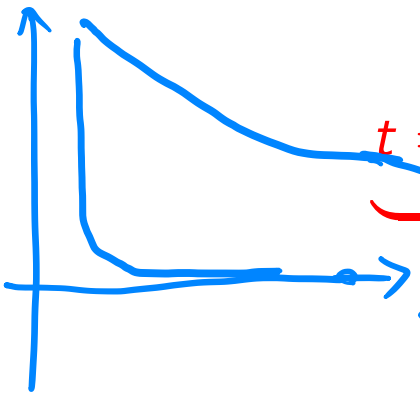
$$\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2 \leq \left(1 - \frac{2}{c+1}\right)^t \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2$$

$$\leq e^{-\frac{2t}{c+1}} \left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2 \quad (\text{by } 1 + x \leq e^x)$$

For any pre-defined error $0 < \epsilon < 1$,

$$t = c \log \frac{\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2}{\epsilon} \implies \left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2 \leq \epsilon$$

iteration complexity



Implications

3. Estimate α

∇^2.

$L \geq \alpha$

$\eta \leq \frac{2}{\alpha + \eta 2}$.

$\eta = \frac{2}{L+L} \leq \frac{2}{\alpha+L}$

$= \frac{1}{L}$.

Overall Computational Complexity

Below #Iter hides the dependence on $\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2$, $c = L/\alpha$

Condition	Guarantee	#Iter
α-SC, L-smooth	$\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2 \leq \epsilon$	$c \log(1/\epsilon)$
L-smooth	$F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*) \leq \epsilon$	L/ϵ

- illustration

$\{(x_i, y_i)\} \subseteq \mathbb{R}^d \times \mathbb{R}$

$F(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w \cdot x_i)^2 + \lambda \cdot \|w\|^2$

- GD solves linear regression efficiently

$d^2(n+d)$ v.s. $nd \cdot c \log(1/\epsilon)$

- note on c

$d^3 \rightarrow d$.

Improve Gradient Descent

Program

$$\min_{\boldsymbol{w}} \; F(\boldsymbol{w}), \quad \text{s.t. } \boldsymbol{w} \in \mathbb{R}^d.$$

- $F(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$ for $y \in \mathbb{R}$
- $F(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \max\{1 - y_i \boldsymbol{x}_i \cdot \boldsymbol{w}, 0\} + 0.5\lambda \|\boldsymbol{w}\|_2^2$ for $y \in \{+1, -1\}$

GD:

- $O(nd)$ to evaluate $\nabla F(\boldsymbol{w})$
- always converges to opt

If computational cost is major concern,

can we boost the efficiency?

Explore the problem structure

Investigate Problem Structure

Suppose

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

new assumption

- linear regression

$$F(\boldsymbol{w}) = \frac{1}{n}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right\|_2^2, \quad f_i(\boldsymbol{w}) = (y_i - \boldsymbol{x}_i \cdot \boldsymbol{w})^2$$

- SVM

$$F(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \max\{1 - y_i \boldsymbol{x}_i \cdot \boldsymbol{w}, 0\} + 0.5\lambda \left\|\boldsymbol{w}\right\|_2^2$$

$$f_i(\boldsymbol{w}) = \max\{1 - y_i \boldsymbol{x}_i \cdot \boldsymbol{w}, 0\} + 0.5\lambda \left\|\boldsymbol{w}\right\|_2^2$$

- any sample-wise loss

Stochastic Gradient Descent

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

1. Initialize $\boldsymbol{w}^0$, say $\boldsymbol{w}^0 = \mathbf{0}$
2. For $t = 1, 2, \ldots$

   Uniformly draw i_t from $\{1, 2, \ldots, n\}$, and update

$$\boldsymbol{w}^t = \boldsymbol{w}^{t-1} - \eta_t \nabla f_{i_t}(\boldsymbol{w}^{t-1}) \tag{2}$$

$\nabla f_3(w)$

- example, intuition

$E_{i_t}[w^t] = E[w^{t-1}] - \eta_t E_{i_t}[\nabla f_{i_t}(\cdot)]$

- time cost per iteration is $O(d)$ (GD needs nd)
- total time = cost/iter · #iter

$\nabla f_1(w) \cdot \frac{1}{n} + \nabla f_2(w) \cdot \frac{1}{n} + \nabla f_3(w) \cdot \frac{1}{n}$

$+ \cdots + \nabla f_n(w) \cdot \frac{1}{n} = \nabla F(w)$

necessary condition:

$$w^{t+1} = w^t - \eta_t \cdot \nabla f_{i_t}(w^t)$$

$$\downarrow \qquad \downarrow$$

$$w^* \qquad w^*$$

$$\Rightarrow \eta_t \cdot \nabla f_{i_t}(w^*) = 0$$

$$\Rightarrow \eta_t = 0 \quad \text{as} \quad t \to \infty$$

$$\nabla f_{i_t}(w^*) \neq 0$$

$$F(w) = \underbrace{\frac{1}{2} \cdot w^2}_{f_1(w)} + \underbrace{\frac{1}{2} \cdot (w-1)^2}_{f_2(w)}$$

$w^* = 1/2$

$$\nabla f_1(w^*) = \frac{1}{2} \qquad \nabla f_2(w^*) = -\frac{1}{2}$$

Convergence Rate for SGD

① $\nexists M > 0$. $\forall |F(w_2) - F(w_1)| \le M \cdot \|w_2 - w_1\|$

② $\exists M > 0$. $\|\nabla F(w)\| \le M$

α-SC, Lipschitz

$$\eta_t \le \frac{1}{\alpha t}, \quad \mathbb{E}\left[\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2\right] \le \frac{\log t}{t}$$

$1/\epsilon$

SGD total cost = $d \cdot \frac{1}{\epsilon} \cdot c.$

convex, Lipschitz

GD: $= nd \cdot \log\frac{1}{\epsilon}$

$n^2 d$

$$\eta_t \le \frac{1}{\sqrt{t}}, \quad \mathbb{E}\left[F(\boldsymbol{w}^t) - F(\boldsymbol{w}^*)\right] \le \frac{\log t}{\sqrt{t}}$$

GD: $c \cdot nd \cdot \log\frac{1}{\epsilon} \Rightarrow (c+nd)\log\frac{1}{\epsilon}$

$c = O(n)$ $\quad O(n)$

$1/\sqrt{t}$

We can modify SGD for faster rate (a rich literature). $t = 1/\epsilon^2$

Rie Johnson & SVRG

Tong Zhang '14 stochastic variance reduced

GD v.s. SGD

Oct 19.

TA

Oct. 26

6:30 - 9:00 pm

Table 1: Overall computational cost to obtain ϵ opt. error

Condition	GD	SGD
SC	$n \log(1/\epsilon)$	$1/\epsilon$
Convex	n/ϵ	$(1/\epsilon)^2$

Week 1 - Oct 12

Open book.

Oct 21

SGD wins if

- large-scale data

GD wins if

- small data set
- need high accuracy (i.e. ϵ is small)

No calculator

In Practice...

SGD is used in

and more...

Other Practical Concerns

- Storage
- real-time decision-making

Online Learning

Initialize the model.

for $t = 1, 2, \ldots$

- receive $\boldsymbol{x}_t$
- make prediction $\hat{\boldsymbol{y}}_t$
- receive $\boldsymbol{y}_t$
- evaluate loss $\ell(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t)$
- update model

Compare to SGD

Mid-term

March 12, 18:30 - 21:00 EST

- linear algebra, calculus
- advanced probability
- linear regression
- gradient descent
- stochastic gradient
- online learning
- statistical machine learning (Mar. 5)