Convexity and Stochastic Gradient Descent

Oct 26 TA: Tao Jiang

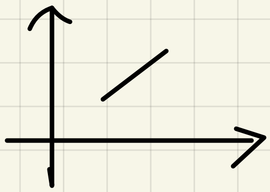
1. Convex set and convex function.

1.1 Convex set

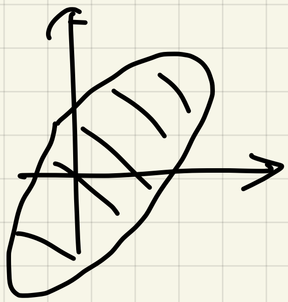
Definition: A subset $X \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in X$, $\gamma \in [0,1]$,

we have $(1-\gamma)x + \gamma y \in X$

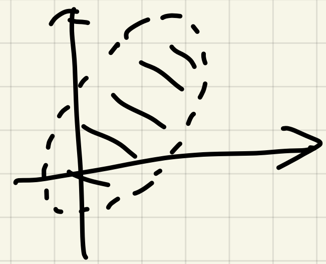
e.g. ①



②



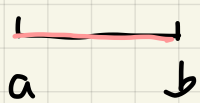
③



④

⑤

⑥ Any intervals in $\mathbb{R}$



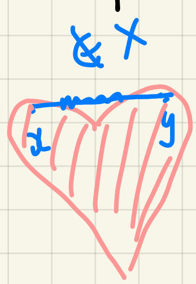
⑦ open / closed ball

$B(x_0, r) = \{x \in \mathbb{R}^n : \|x - x_0\| < r\}$ open ball

$\bar{B}(x_0, r) = \{x \in \mathbb{R}^n : \|x - x_0\| \leq r\}$ closed ball.

⑧ Polyhedron $\{x \in \mathbb{R}^n : Ax \geq b\}$.

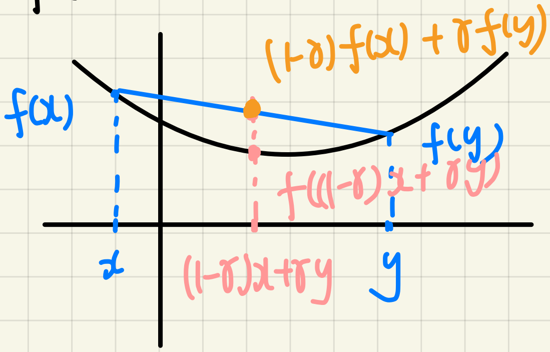
Counterexample



1.2 convex function.

Definition : $f: X \to \mathbb{R}$ is convex if $\forall x, y \in X$, $\forall \gamma \in [0,1]$,

$$f((1-\gamma)x + \gamma y) \leq (1-\gamma)f(x) + \gamma f(y)$$



e.g. ① Linear function $f(x) = a^T x + b$

② convex quadratic function $f(x) = ax^2 + bx + c$, $a \geq 0$

$$\frac{1}{2} x^T A x + b^T x + c$$

$A \succeq 0$

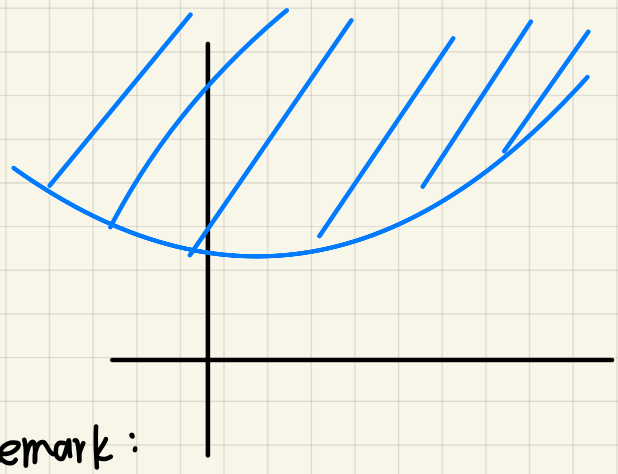
③ $f(x) = \|x\|$ norm

④ $f(x) = e^x$

⑤ $f(x) = |x|^p$ $\forall p \geq 0$

⑥ $f(x) = -\ln x$

⑦ distance function



Remark :

$\text{epi}(f) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f(x)\}$.

is convex

$\Leftrightarrow f$ is convex

2. Stochastic gradient method (SGD)

2.1 Motivation

$$\min_x \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

e.g. · least square : $f_i(x) = (a_i^T x - b_i)^2$

· logistic regression : $f_i(x) = -\log(1+\exp(-b_i + a_i^T x))$

→ solving a problem with more data (higher m) is easier

→ complexity ↑ with m.

Goal : find algorithms that work better given more data.

Idea : throw away data (partial data)

2.2. Stochastic gradient.

Definition : random $\tilde{g} \in \mathbb{R}^n$ is a stochastic gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ if $\mathbb{E}\tilde{g} = \nabla f(x)$.

i.e. for all z, $f(z) \geq f(x) + (\mathbb{E}\tilde{g})^T (z - x)$ subgradient inequality

$$\tilde{g} = \underset{\substack{\uparrow \\ g = \nabla f(x)}}{g} + v \leftarrow \text{error}, \quad \mathbb{E}v = 0.$$

2.3. Stochastic gradient descent.

Initialize $x_1 \in \mathbb{R}^n$

for $k = 1, 2, \ldots$

$\quad x^{(k+1)} = x^{(k)} - t_k \cdot \tilde{g}^{(k)}$ ← any unbiased gradient of f at $x^{(k)}$.

k-th step size (t_k); k-th iterate ($x^{(k)}$)

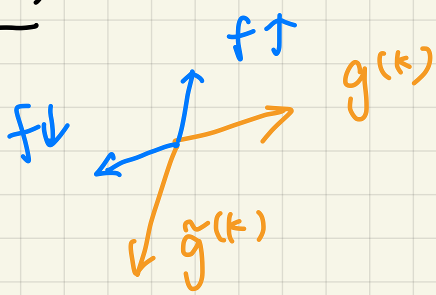
end

Return $x_{best} := \underset{x \in [x_1, x_2 \cdots]}{\arg\min} f(x)$

- $\mathbb{E}[\tilde{g}^{(k)} \mid x^{(k)}] = g^{(k)} = \nabla f(x^{(k)})$
- $t_k > 0$.

Remark: Stochastic gradient method may not a descent method. To see this,

$$\underline{(\tilde{g}^{(k)})^T} g^{(k)} < 0 \quad \text{is possible.}$$



2.4. Assumptions and Notations.

Assumptions:

- f is bounded below and the optimal solution exists.

$$\underset{\text{optimal value}}{f^*} := \min_x f(x) > -\infty, \quad f(\underset{\text{optimal solution.}}{x^*}) = f^*$$

- Stochastic gradient $\tilde{g}$ has bounded second moment:

$$\underbrace{\mathbb{E}\, \|\tilde{g}^{(k)}\|_2^2}_{\text{second moment}} \le G \quad \text{for all } k.$$

- Initial iterate is not too far from the optimal solution.

$\mathbb{E}\|x^{(1)} - x^*\|_2^2 \le R^2$.

- step sizes t_k are square-summable but not summable.

$$t_k \geq 0, \quad \sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty.$$

e.g. $t_k = \frac{1}{k}$

- f is convex and smooth

Notations:

- Best optimal value $f_{best}^{(k)} := \min\{f(x^{(1)}), f(x^{(2)}), \cdots, f(x^{(k)})\}$.

2.5 Proof of convergence.

Result:

- convergence in expectation: $\mathbb{E} f_{best}^{(k)} \to f^*$ (proof)
- convergence in probability: for any $\varepsilon > 0$,

$$\lim_{k \to \infty} \operatorname{Prob}(f_{best}^{(k)} \geq f^* + \varepsilon) = 0$$

- almost sure convergence: $\lim_{k \to \infty} f_{best}^{(k)} = f^*$.

Proof:

key observation:

$$\mathbb{E}(\|x^{(k+1)} - x^*\|_2^2 \mid x^{(k)})$$

$$= \mathbb{E}(\|x^{(k)} - t_k \tilde{g}^{(k)} - x^*\|_2^2 \mid x^{(k)})$$

$\|a - b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 - 2ab$

$$= \|x^{(k)} - x^*\|_2^2 - 2t_k (x^{(k)} - x^*) \underbrace{\mathbb{E}(\tilde{g}^{(k)} \mid x^{(k)})}_{\nabla f(x^{(k)})} + t_k^2 \underbrace{\mathbb{E}(\|\tilde{g}^{(k)}\|_2^2 \mid x^{(k)})}_{\text{by assumption} \le G^2}$$

$$\le \|x^{(k)}-x^*\|_2^2 - 2t_k (x^{(k)}-x^*) \nabla f(x^{(k)}) + t_k^2 G^2$$

subgradient inequality.

$$\le \|x^{(k)}-x^*\|_2^2 - 2t_k \mathbb{E}(f(x^{(k)})-f^*) + t_k^2 G^2$$

Rearranging yields,

$$2t_k \mathbb{E}(f(x^{(k)})-f^*) \le \|x^{(k)}-x^*\|_2^2 - \mathbb{E}(\|x^{(k+1)}-x^*\|_2^2 \mid x^{(k)}) + t_k^2 G^2$$

Take expectation

$$2t_k \mathbb{E}(f(x^{(k)})-f^*)) \le \mathbb{E}\|x^{(k)}-x^*\|_2^2 - \mathbb{E}\|x^{(k+1)}-x^*\|_2^2 + t_k^2 G^2 \qquad (*)$$

sum $(*)$ over $k=1,\dots,K$,

$$2\sum_{i=1}^{k} t_i \mathbb{E}(\underbrace{f(x^{(i)})-f^*}_{\ge f_{best}^{(k)}-f^*}) \le \mathbb{E}\|x^{(1)}-x^*\|_2^2 - \cancel{\underbrace{\mathbb{E}\|x^{k}-x^*\|_2^2}} + \sum_{i=1}^{k} t_i^2 G^2$$

$$\text{LHS} \ge 2\sum_{i=1}^{k} t_i \mathbb{E}(f_{best}^{(k)}-f^*)$$

Divide both sides by $2\sum_{i=1}^{k} t_i$ to obtain

$$\mathbb{E}(f_{best}^{(k)}-f^*) \le \frac{\underbrace{\mathbb{E}\|x^{(1)}-x^*\|_2^2}_{\le R^2} - \sum_{i=1}^{k} t_i^2 G^2}{\underbrace{2\sum_{i=1}^{k} t_i}}$$

$c < \infty$ as $k\to\infty$ as t_i is square summable.

$\to \infty$ as $k \to \infty$ because t_i's are not summable

$\to 0$ as $k\to\infty$

$\Rightarrow \mathbb{E}(f_{best}^{(k)}-f^*) = 0 \Rightarrow$ proves convergence in expectation. ∎

$$X_j := |v_j - \mu|$$

$$\mathbb{P}(\underbrace{\max_j X_j}_{Y} > \varepsilon) \quad \leftarrow \mathbb{P}(Y > \varepsilon)$$

$$= 1 - \mathbb{P}(Y \le \varepsilon)$$

$$= 1 - \mathbb{P}(\max_j X_j \le \varepsilon)$$

$$= 1 - \underbrace{\mathbb{P}(X_1 \le \varepsilon, X_2 \le \varepsilon, \dots, X_p \le \varepsilon)}_{\prod_{j=1}^{p} \mathbb{P}(X_j \le \varepsilon) \;=\; (\mathbb{P}(X_1 \le \varepsilon))^p}$$

$$X_1 = |v_1 - \mu| \qquad n v_1 \sim \mathrm{Bin}(n, \mu)$$

$$\mathbb{P}(X_1 \le \varepsilon) = \mathbb{P}(|v_1 - \mu| \le \varepsilon) \qquad \mathbb{P}(n v_1 = k) = (\quad)$$

$$= \mathbb{P}(|n v_1 - n\mu| \le n\varepsilon)$$

$$= \mathbb{P}(-n\varepsilon \le n v_1 - n\mu \le n\varepsilon)$$

$$= \mathbb{P}(n\mu - n\varepsilon \le n v_1 \le n\mu + \varepsilon)$$

nV_j = # of Biden voters at $j \sim \text{Bin}(n, \mu)$

$\mathbb{P}(\text{\# of Biden voters} = k) = \binom{n}{k} \mu^k (1-\mu)^{n-k}$

① Define $X_j := |V_j - \mu|$

$$P(\max_j X_j > \varepsilon)$$

$$= 1 - \underbrace{P(\max_j X_j \le \varepsilon)}$$

$$= P(X_1 \le \varepsilon, \ldots, X_p \le \varepsilon)$$

$$= \prod_{j=1}^{p} P(X_j \le \varepsilon)$$

$$= \left(P(X_1 \le \varepsilon)\right)^p$$

$$= P(|V_j - \mu| \le \varepsilon)$$

$$= P(-\varepsilon \le V_j - \mu \le \varepsilon)$$

say $\mu = 0.5$
$\varepsilon = 0.1$
$n = 2$

$$= P(\mu - \varepsilon \le V_j \le \mu + \varepsilon)$$

$$= P(\underbrace{n(\mu - \varepsilon)}_{2 \cdot 0.4} \le nV_j \le \overbrace{n(\mu+\varepsilon)}^{2 \cdot 0.6}) = ?$$

$nV_j \sim \text{Binomial}(n, \mu)$