

Convexity and Stochastic Gradient Descent

Oct 26

TA: Tao Jiang

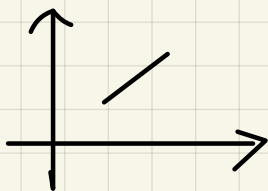
1. Convex set and convex function.

1.1 Convex set

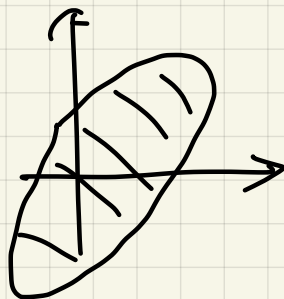
Definition: A subset $X \subseteq \mathbb{R}^n$ is convex if $\forall x, y \in X, \gamma \in [0, 1]$,

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

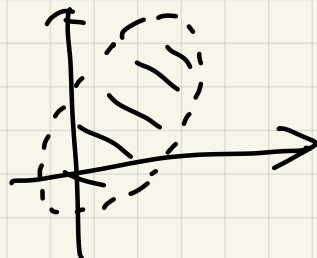
e.g. ①



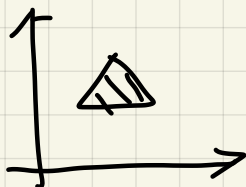
②



③



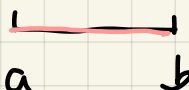
④



⑤



⑥ Any intervals in \mathbb{R}



⑦ open / closed ball

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

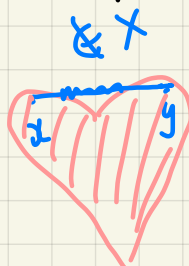
open ball

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

closed ball.

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

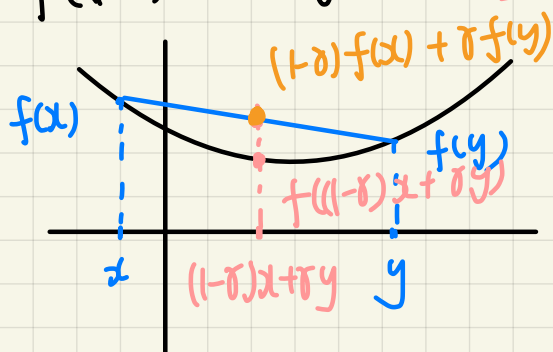
Counterexample



1.2 convex function.

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

$$f((1-\delta)x + \delta y) \leq \underbrace{(1-\delta)f(x) + \delta f(y)}$$



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

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

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

$A \geq 0$

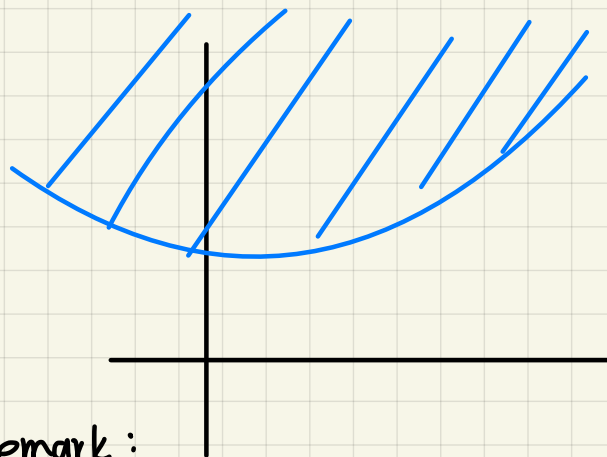
③ $f(x) = \|x\|$ norm

④ $f(x) = e^x$

⑤ $f(x) = |x|^p \quad \forall p \geq 0$

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 \uparrow 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 $f: \mathbb{R}^n \rightarrow \mathbb{R}$

at $x \in \mathbb{R}^n$ if $\mathbb{E} \tilde{g} = \nabla f(x)$.

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

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

2.3. Stochastic gradient descent.

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

for $k = 1, 2, \dots$

$$| \quad x^{(k+1)} = x^{(k)} - t_k \cdot \tilde{g}^{(k)}$$

end

t_k \leftarrow k -th step size

$\tilde{g}^{(k)}$ \leftarrow any unbiased gradient of f at $x^{(k)}$.

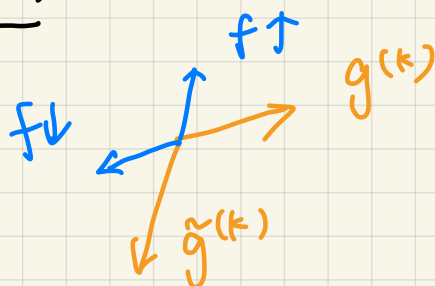
Return $x_{\text{best}} := \operatorname{argmin}_{x \in \{x_1, x_2, \dots\}} f(x)$

$$\cdot \mathbb{E}[\tilde{g}^{(k)} | x^{(k)}] = g^{(k)} = \nabla f(x^{(k)})$$

$$\cdot t_k > 0.$$

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

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



2.4. Assumptions and Notations.

Assumptions:

- f is bounded below and the optimal solution exists.

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

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

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

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

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

- 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 $f_{\text{best}}^{(k)} := \min \{f(x^{(1)}), f(x^{(2)}), \dots, f(x^{(k)})\}$.

2.5 Proof of convergence.

Result:

- convergence in expectation: $\mathbb{E} f_{\text{best}}^{(k)} \rightarrow f^*$ (proof)

- convergence in probability: for any $\varepsilon > 0$,

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

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

Proof:

key observation:

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

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

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

$$\begin{aligned} \|a - b\|_2^2 &= \|a\|_2^2 + \|b\|_2^2 - 2ab \end{aligned}$$

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

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

↙ subgradient inequality.

Rearranging yields,

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

Take expectation

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

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

$$2 \sum_{i=1}^K t_i \underbrace{\mathbb{E}(f(x^{(i)}) - f^*)}_{\geq f_{\text{best}}^{(K)} - f^*} \leq \mathbb{E}\|x^{(1)} - x^*\|_2^2 - \underbrace{\mathbb{E}\|x^{(K+1)} - x^*\|_2^2}_{\geq 0} + \sum_{i=1}^K t_i^2 G^2$$

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

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

$$\mathbb{E}(f_{\text{best}}^{(K)} - f^*) \leq \frac{\mathbb{E}\|x^{(1)} - x^*\|_2^2 - \sum_{i=1}^K t_i^2 G^2}{2 \sum_{i=1}^K t_i}$$

$C < \infty$
 \rightarrow as $k \rightarrow \infty$
 as t_i is square summable.

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

$$\Rightarrow \mathbb{E}(f_{\text{best}}^{(K)} - f^*) \rightarrow 0 \text{ as } k \rightarrow \infty \Rightarrow \text{proves convergence in expectation.}$$