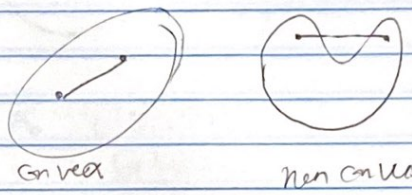
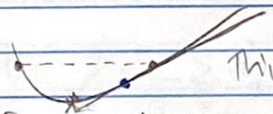
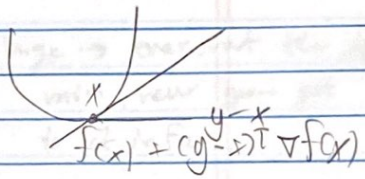
Gradient Descent & SGD



K is convex
$\Leftrightarrow \forall x, y.\ \lambda x + (1-\lambda) y \in K$





A function is convex $\Leftrightarrow \forall x, y.$
$\lambda f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

- For diff f, $f(y) \geq f(x) + \nabla f(x)^T (y - x)$
- For thrice diff function

$$(Hf)_{ij} = (\nabla^2 f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

f is convex $\Leftrightarrow$ Hf is positive semidefinite for all x

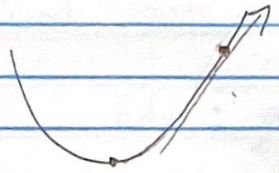
$\forall i, \lambda_i(A) \geq 0.$

Def A function $f: \mathbb{R}^n \to \mathbb{R}$ is λ-Lipschitz if
$\forall x, y \in \mathbb{R}^n,\ |f(x) - f(y)| \leq \lambda \|x - y\|_2$

Minimize a convex func: $f: \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(\vec{x}) = 0$$

$f(x) = \frac{1}{2} x^T A x - bx.$
$\nabla f(x) \overset{?}{=} Ax - b = 0.$
$f(x) = -\frac{x^T A x}{x^T x}$



$x_0 - \eta_1 \nabla f(x_0)$

how ~~for~~ big?

step size η: too small → too long to reach local min

too large → overshoot the ~~last~~ min, never gonna get to it, in fact.

$\vec{x}_0$ = starting point

for $t = 1 \dots T$:

$\vec{x}_t = \vec{x}_{t-1} - \eta_t \nabla f(\vec{x}_{t-1})$

return $\hat{x} = \frac{1}{T} \sum_t \vec{x}_t$

Thm: Let x^* be the minimizer of $f: \mathbb{R}^n \to \mathbb{R}$

If f is convex & diff and satisfies

$\forall x \in \mathbb{R}^n, \|\nabla f(x)\|_2 \le G$, then

* ——— setting $T = \frac{1}{\varepsilon^2} G^2 \|x_0 - x^*\|_2^2$ and $\eta_t = \eta = \frac{\|x_0 - x^*\|_2}{G\sqrt{T}}$

gives $f(\hat{x}) \le f(x^*) + \varepsilon$

Pf

Lemma A:

$$\sum_t f(\vec{x}_t) \le \left(\sum_t f(x^*)\right) + \frac{1}{2} G^2 T \eta + \frac{1}{2\eta} \|x_0 - x^*\|_2^2$$

Proof of Thm:

$$f(\hat{x}) = f\left(\frac{1}{T}\sum_t \vec{x}_t\right) \overset{\text{by convexity of } f}{\le} \frac{1}{T}\left(\sum_t f(\vec{x}_t)\right)$$

$$\overset{\text{by def}}{=}$$

$$\overset{\text{Lemma A}}{\le} \frac{1}{T}\left(T f(x^*)\right) + \frac{1}{2} G^2 \eta + \frac{1}{2\eta T} \|\vec{x}_0 - x^*\|_2^2$$

$$\le f(x^*) + \frac{G}{\sqrt{T}} \|\vec{x}_0 - x^*\|_2$$

$$\le f(x^*) + \frac{G \cdot \|\vec{x}_0 - x^*\|_2}{\frac{1}{\varepsilon} G \|x_0 - x^*\|_2} \le f(x^*) + \varepsilon$$

* plug in * η_t

* plug in T

If we don't know about $f(x^*)$, $f(\hat{x}) = f(x^*) + O\left(\frac{1}{\sqrt{T}}\right)$

Proof Lemma A

Define: $\phi_t = \frac{1}{2\eta}\|\vec{x}_t - x^*\|_2^2$

you are free from x_t

- $\phi_0 = \frac{1}{2\eta}\|\vec{x}_0 - x^*\|_2^2$

Claim: $f(\vec{x}_t) + \phi_{t+1} - \phi_t \le f(x^*) + \frac{1}{2}G^2\eta$

Proof

$$f(\vec{x}_t) + \phi_{t+1} - \phi_t = f(\vec{x}_t) + \frac{1}{2\eta}\left(\|\vec{x}_{t+1} - x^*\|_2^2 - \|\vec{x}_t - x^*\|_2^2\right)$$

$$\|(\vec{x}_{t+1} - \vec{x}_t) + \vec{x}_t - x^*\|_2^2$$

Δx : $\vec{x}_{t+1} - \vec{x}_t$, z : $\vec{x}_t - x^*$

$$= (\Delta x + z)^T(\Delta x + z) - \|z\|_2^2$$

$$= \|\Delta x\|_2^2 + 2\Delta x^T z + \|z\|_2^2 - \|z\|_2^2$$

$$= f(\vec{x}_t) + \frac{1}{2\eta}\left[\|\Delta x\|_2^2 + 2\Delta x^T(\vec{x}_t - x^*)\right]$$

$\Delta x = \vec{x}_{t+1} - \vec{x}_t = -\eta \nabla f(\vec{x}_t)$

plug Δx in: $\le f(\vec{x}_t) + \frac{1}{2\eta}\left[\|\eta \nabla f(\vec{x}_t)\|_2^2 - 2\eta \nabla f(\vec{x}_t)^T(\vec{x}_t - x^*)\right]$ ($\|\nabla f(\vec{x}_t)\|_2^2 \le G^2$)

$$= \frac{1}{2}\eta G^2 + f(\vec{x}_t) - \nabla f(\vec{x}_t)^T(\vec{x}_t - x^*)$$

$$= \frac{1}{2}\eta G^2 + f(\vec{x}_t) + \nabla f(\vec{x}_t)^T(x^* - \vec{x}_t)$$

linear approximation

$$\le \frac{1}{2}\eta G^2 + f(x^*) \qquad \le f(x^*)$$

$\blacksquare$

Prof Lemma A

$$\sum_t f(\vec{x}_t) = \left(\sum_t f(x^*)\right) + \frac{1}{2}G^2\eta + \frac{1}{2\eta}\|\vec{x}_0 - x^*\|_2^2$$

$$= \left(\sum_t f(x^*) + \phi_T - \phi_{T-1} + \phi_{T-1} \cdots \phi_0\right) -$$

$$\leq \left(\sum_t f(x^*)\right) + \frac{1}{2}G\eta T$$

$$\Rightarrow \sum_t f(\vec{x}_t) \leq \left(\sum_t f(x^*)\right) + \frac{1}{2}G^2\eta T + \phi_0 - \phi_T$$

$$\leq \left(\sum_t f(x^*)\right) + \frac{1}{2}G^2\eta T + \frac{1}{2\eta}\|x_0 - x^*\|_2^2$$

$\phi_0 = 0$

from cluster.

SGD

minimize $f(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\vec{\theta})$

$(\vec{x}_i, y_i)$

$$f(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i^T\vec{\theta} - y_i)^2$$

$$\nabla f = \frac{1}{n}\sum_i^n \nabla f_i(\vec{\theta}) \qquad O(n \cdot d) \text{ work per iteration}$$

$$\mathbb{E}_i[f_i(\vec{\theta})] = \frac{1}{n}\sum_{i=1}^{n} f_i(\vec{\theta}) = f(\vec{\theta})$$

$$\mathbb{E}_i[\nabla f_i(\vec{\theta})] = \nabla f(\vec{\theta})$$

$\mathbb{E}[\nabla f_i(\vec{x}_{t-1})]$

For $t = T \cdots T-1$

① pick i at random.

② $\vec{x}_t = \vec{x}_{t-1} - \gamma_t \nabla f_i(\vec{x}_{t-1})$

per iter, now, $O(d)$ work $\leftarrow$ SGD $\rightarrow$ space reduction -

(overall ?)

Thm SGD. Sum

- $\|\nabla f_i(x)\|_2 \le G$, then

$$\mathbb{E}[f(\hat{x})] \le f(x^*) + \varepsilon$$

~~Theorem~~

~~Minimize~~ Expectation means nothing.

- Minibatch SGD

pick $B \subseteq [N]$, $|B| = b$

$$\vec{x}_t = \vec{x}_{t-1} - \frac{\eta_t}{b} \sum_{i \in B} \nabla f_i(\vec{x}_{t-1})$$

→ reducing a variance.

$$\mathbb{E}\left[\frac{1}{b}\sum \nabla f_i(\vec{x}_{t-1})\right] = \nabla f(\vec{x}_{t-1}), \quad \text{Var} : \frac{1}{b}$$

$O(\sqrt{b/T} + b/T)$ convergence.

if not batch $O\left(\frac{1}{\sqrt{T}}\right)$