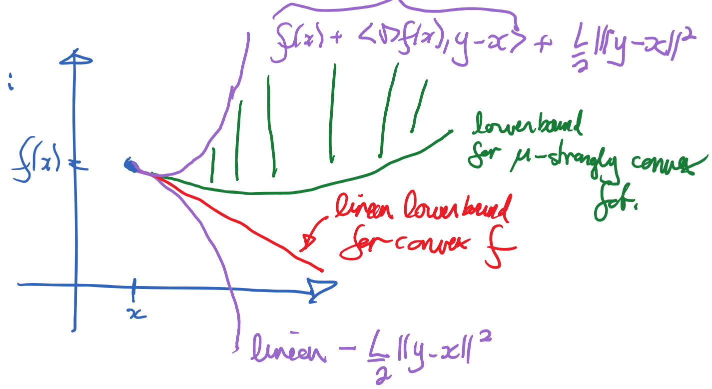
Lecture 10 - subgradient method

Thursday, February 16, 2023 1:35 PM

today: • basic gradient methods
• subgradient method

drawing:



Hessian

if f is twice differentiable

$L = \sup_x (\lambda_{max}(H(x)))$

$\mu = \inf_x (\lambda_{min}(H(x)))$

gradient descent:

$$x_{t+1} = x_t - \gamma \nabla f(x_t) \qquad \gamma = \frac{1}{L}$$

a) when f is convex & L-smooth

$$f(x_t) - \underbrace{\min_x f(x)}_{\triangleq f^*} \leq O\left(\frac{L r_0^2}{t}\right)$$

[see Nesterov book for proof]

where $r_0 \geq \mathrm{dist}(x_0, X^*)$ with $\mathrm{dist}(x_0, X^*) \triangleq \underset{x' \in X^*}{\mathrm{argmin}} \|x_0 - x'\|_2$ and $X^* = \underset{x}{\mathrm{argmin}}\, f(x)$

"sublinear"

note: no guarantee on $\mathrm{dist}(x_t, X^*)$
(for general L-smooth convex fcts
for $t \leq \dim(x_t)$)
↳ Nesterov lower bound

$x^T H x$

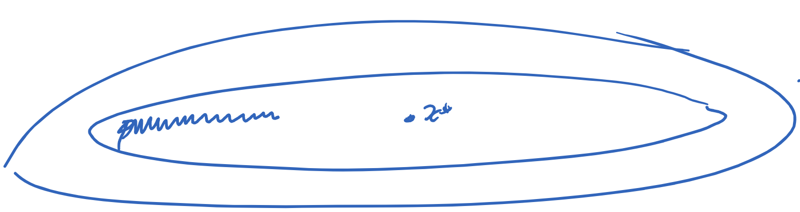
b) if f is μ-strongly convex & L-smooth

$$f(x_t) - f(x^*) \leq O\left(\exp\left(-\frac{\mu}{L} t\right)\right)$$

"linear rate"

$\frac{L}{\mu} \triangleq \underline{\text{condition \# of } f} = \kappa$

≥ 1



$\exp\left(-\frac{t}{\kappa}\right)$

Newton's method

$$x_{t+1} = x_t - \gamma_t [H(x_t)]^{-1} \nabla f(x_t)$$

[linear rate sketch: $f(x_\gamma) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$

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

μ-strongly convex fct. $\Rightarrow$ $f(x) - f^* \le \frac{1}{2\mu}\|\nabla f(x)\|^2$

$\Downarrow$

$-\|\nabla f\|^2 \le -2\mu(f(x)-f^*)$

$$\underbrace{f(x_{t+1}) - f^*}_{\triangleq h_{t+1}} \le f(x_t) - f^* - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

$$\le f(x_t) - f^* - \frac{\mu}{L}\underbrace{(f(x_t) - f^*)}_{\triangleq h_t}$$

$$h_{t+1} \le \left(1 - \frac{\mu}{L}\right) h_t$$

$$\le \left(1 - \frac{\mu}{L}\right)^t h_0 \qquad (1-x)^t \le \exp(-xt) \quad \forall x$$

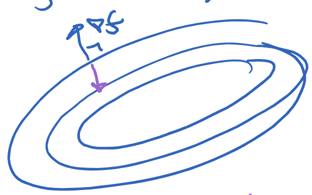
$$\le \exp\left(-\frac{\mu}{L} t\right) h_0 \quad //$$

14h19

subgradient method

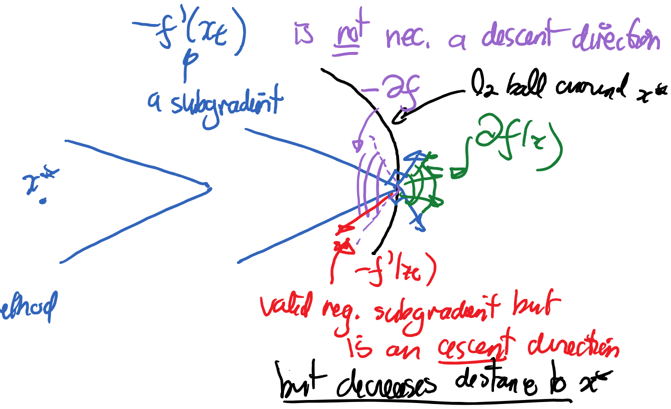
non-descent methods

f smooth $\Rightarrow$ smooth sublevel sets



$-\nabla f$ is a descent direction on f

non-smooth f



(*) subgradient method is not nec. a descent method

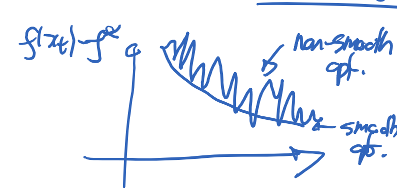
but $-f'(x_t)$ is a descent direction on $\|x(\gamma) - \tilde{x}\|_2^2$ for any $\tilde{x}$ in sublevel set at x

$\uparrow$

$x_t - \gamma f'(x_t)$

$x(\gamma)$ gets closer to any $\tilde{x}$ s.t. $f(\tilde{x}) \le f(x_t)$ for small enough γ

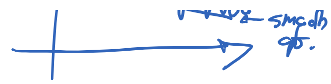
thus get closer to any x^*



* in non-smooth optimization, $f(x_t)$ can go up & down

to stabilize $\Rightarrow$ combine multiple pt. x_t to get $\hat{x}_T$

* in non-smooth optimization, $f(x_t)$ can go up & down



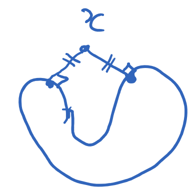
to stabilize $\Rightarrow$ combine multiple pt. x_t to get $\hat{x}_T$

- $\underset{\{x_t\}_{t=1}^{T}}{\operatorname{argmin}} f(x_t)$ [in batch setting]
- weighted average $\hat{x}_T = \sum_t \underbrace{\rho_t^{(T)}}_{\text{convex weights}} x_t$

(*) projection operator on a closed convex set C

$$P_C(x) \triangleq \underset{y \in C}{\operatorname{argmin}} \|x-y\|_2^2$$

"Euclidean projection" of x on C



$P_C(\cdot)$ is non-expansive i.e. $\|P_C(x) - P_C(y)\|_2 \leq \|x-y\|_2 \quad \forall x \,\&\, y$

- if $y \in C$, then $P_C(y) = y$

$$\min_{x \in C} f(x) \qquad P_C(x^*) = x^*$$

$$\|P_C(x) - \underbrace{x^*}_{P_C(x^*)}\|_2 \leq \|x - x^*\|_2$$

$\Rightarrow$ thus projection on C just makes iterates closer to x^*

stochastic subgradient method

setup: want to solve $\min_{x \in C} f(x)$ where $f(x) \triangleq \mathbb{E}_\xi[h(x,\xi)]$

assumptions: 1) f & C are convex

[2) projection on C is "cheap"]

3) we have a stochastic oracle which gives $g(x,\xi)$ for random ξ

s.t. $\boxed{\mathbb{E}_\xi[g(x,\xi) \mid x] = \underbrace{f'(x)}_{\text{some subgradient of } f \text{ at } x}}$

[examples:

a) if h is diff. in x & well behaved

$$g(x,\xi) \triangleq \nabla_x h(x,\xi)$$ "Leibniz rule"

then $\mathbb{E}_\xi[\nabla_x h(x,\xi)] = \nabla_x [\mathbb{E}_\xi[h(x,\xi)]] = \nabla f(x)$

b) <u>ERM example</u> (finite sum) $\quad f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ $\quad$ e.g. $f_i(w) = \ell(x^{(i)}, y^{(i)}; w) + \frac{\lambda \|w\|^2}{2}$

$$h(x,\xi) \triangleq f_\xi(x)$$

$$\xi \in \{1, \dots, n\}$$

at step t, sample $i_t \overset{\text{unif}}{\sim} \{1,\dots,n\}$

use $g_t \triangleq g(x_t, i_t) \triangleq f'_{i_t}(x_t)$

here $\mathbb{E}_\xi[f'_{i_t} \mid x] = \frac{1}{n}\sum_{i=1}^{n} f'_i(x_t) = f'(x_t)$

4) $\mathbb{E}_\xi \|g(x,\xi)\|_2^2 \leq B^2$ $\quad$ (finite variance condition)

this replaces the Lipschitz gradient ass. for non-smooth / stochastic opt.

[sufficient condition $\|h'(x,\xi)\| \leq B \quad \forall x \,\&\, \xi$]

<u>algorithm</u> — (projected stochastic subgradient method)

$x_0 \in C$ initialization

for $t = 0, \dots, T-1$

let g_t be $g(x_t, \xi_t)$ [from oracle]

let $x_{t+1} = P_C[x_t - \gamma_t g_t]$

step-size

end

output $\hat{x}_T \triangleq \sum_{t=0}^{T} \rho_{t,T}\, x_t$ $\quad$ where $\rho_{t,T}$ are some <u>convex combo</u> coef.

"weighted avg." $\quad \sum_t \rho_{t,T} = 1 \quad \rho_{t,T} \geq 0$

$\rightarrow O\left(\frac{1}{\sqrt{t}}\right)$ rate

<u>convergence proof</u>:

important inequality (by μ-strong convexity)

$$f(y) \geq f(x) + \langle f'(x), y-x \rangle + \frac{\mu}{2}\|y-x\|^2 \quad (\mu \geq 0)$$

$$\boxed{-\langle f'(x), y-x \rangle \leq -\left(f(x) - f(y) + \frac{\mu}{2}\|y-x\|^2\right) \quad \forall x \,\&\, y} \quad (+)$$

$x_{t+1} = P_C(x_t - \gamma_t g_t)$ by def.

$$\|x_{t+1} - \tilde{x}\|_2^2 \overset{\text{by } P_C}{\leq} \|x_t - \gamma_t g_t - \tilde{x}\|_2^2$$

any feasible pt $\tilde{x} \in C$

$$= \|x_t - \tilde{x}\|^2 + \gamma_t^2 \|g_t\|^2 - 2\gamma_t \langle g_t, x_t - \tilde{x} \rangle \quad (\text{valid } \forall \tilde{x} \in C)$$

$$\mathbb{E}\left[\|x_{t+1} - \tilde{x}\| \mid x_t\right] \leq \|x_t - \tilde{x}\|^2 + \gamma_t^2 \underbrace{\mathbb{E}\left[\|g_t\|^2 \mid x_t\right]}_{\leq B^2} - 2\gamma_t \langle \underbrace{\mathbb{E}[g_t \mid x_t]}_{f'(x_t)}, x_t - \tilde{x} \rangle$$

$\boxed{\text{using } (+)}$

$$\leq \|x_t - \tilde{x}\|^2 + \gamma_t^2 B^2 - 2\gamma_t \left[f(x_t) - f(\tilde{x}) + \frac{\mu}{2}\|x_t - \tilde{x}\|^2\right]$$

$$\mathbb{E}[\mathbb{E}[\ \mid x_t]] = \mathbb{E}[\ ]$$

$$\mathbb{E}\left[\|x_{t+1} - \tilde{x}\|^2\right] \leq (1 - \mu\gamma_t)\, \mathbb{E}\left[\|x_t - \tilde{x}\|^2\right] - 2\gamma_t \left[\mathbb{E} f(x_t) - f(\tilde{x})\right] + \gamma_t^2 B^2$$

true even if $\mu = 0$

let $r_t \triangleq \mathbb{E}\|x_t - \tilde{x}\|^2$; for γ_t small enough

r_t <u>decreases</u> with t for any $\tilde{x} \in C$ s.t. $f(\tilde{x}) \leq \mathbb{E} f(x_t)$

i.e. we have $r_{t+1} \leq r_t$