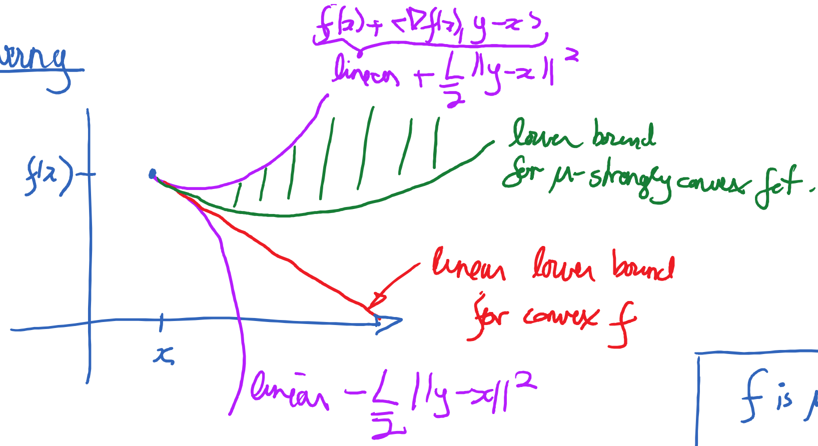
Lecture 10 - subgradient method

Tuesday, February 16, 2021 14:33

today:
- basic gradient methods
- Subgradient method

drawing



when f is twice differentiable

$L = \sup_x \left(\lambda_{max}(H(x))\right)$ — Hessian

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

$$\boxed{f \text{ is } \mu\text{-strongly convex} \iff f - \frac{\mu\|\cdot\|^2}{2} \text{ is convex}}$$

gradient descent:

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

a) when is f is convex & L-smooth

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

"sublinear"

↳ [see Nesterov book for proof]

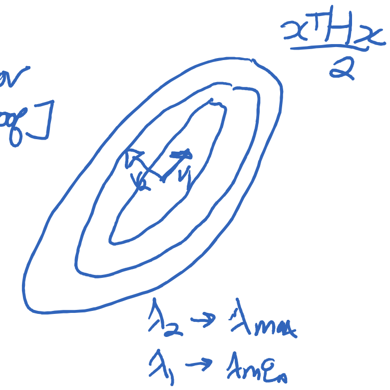
where $r_0 \geq \text{dist}(x_0, X^*)$

$\triangleq \underset{x' \in X^*}{\text{argmin}} \|x_0 - x'\|_2$

$X^* = \underset{x}{\text{argmin}} f(x)$

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

↳ Nesterov lower bound

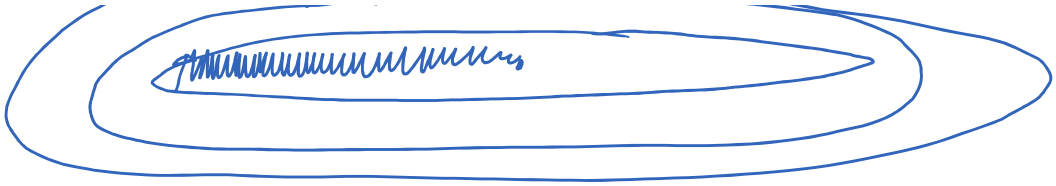


b) if f is μ-strongly convex & L-smooth

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

"linear rate"

$\frac{L}{\mu} \triangleq$ condition # of f

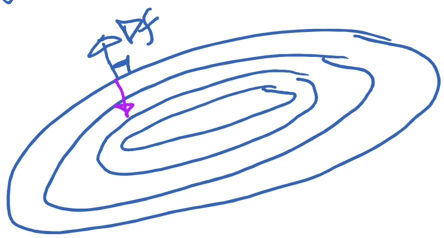


<u>Newton's method</u> $\quad x_{t+1} = x_t - \gamma_t \left[H(x_t)\right]^{-1} \nabla f(x_t)$

<u>subgradient method</u>

<u>non-descent methods</u>

f smooth $\Rightarrow$ smooth sublevel sets

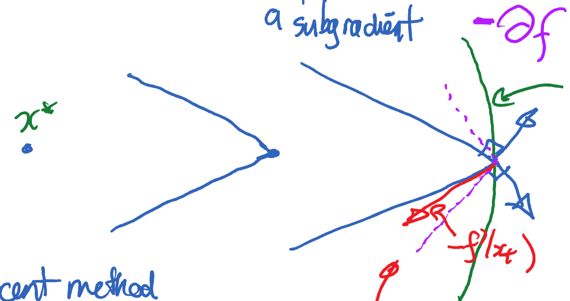


$-\nabla f$ is a <u>descent direction</u>

non-smooth f

$-f'(x_t)$ is <u>not</u> nec. a descent direction

a subgradient



valid neg. subgr. but is an <u>ascent direction</u>

but decrease distance to x^*

⊛ subgradient method is <u>not</u> nec. a descent method

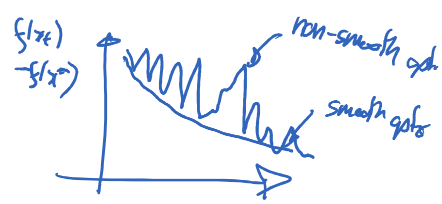
<u>but</u> $-f'(x_t)$ is a <u>descent direction</u>

on $\|x(\gamma) - \tilde{x}\|_2^2$ for any $\tilde{x}$ in sublevel set at x

$x_t - \gamma f'(x_t)$

$x(\gamma)$ gets closer to any $\tilde{x}$ s.t. $f(\tilde{x}) \leq f(x_t)$ for <u>small enough</u> γ

thus get <u>closer to any x^*</u>



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

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

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

convex weights

when too expensive to compute $f(x_t)$

(*) projection operator on a <u>closed</u> <u>convex set</u> C

$$P_C(x) \triangleq \arg\min_{y \in C} \|x-y\|_2^2$$

"Euclidean projection" of x on C

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

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

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

$\|P_C(y)-x^*\|_2 \leq \|y-x^*\|_2$

$\Rightarrow$ thus projection on C just make iterates <u>closer</u> to x^*

15h33

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. $\boxed{\mathbb{E}_\xi[g(x,\xi) \mid x] = f'(x)}$

some subgradient of f at x

[examples:

a) f is diff. in x & "well behaved"

$g(x,\xi) \triangleq \nabla_x h(x,\xi)$

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

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

(finite sum)

$$h(x,\xi) \triangleq f_\xi(x) \qquad + \frac{\lambda}{2}\|"x"\|^2$$

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

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

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) $\mathbb{E}\|g(x,\xi)\|_2^2 \leq B^2$ (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 sub. method)

$x_0 \in C$ initialization

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

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

$$\hat{x}_T \triangleq \sum_{t=0}^{T} \rho_{t,T}\, x_t$$

where $\rho_{t,T}$ are some <u>convex</u> combo coef.

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

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

<u>convergence proof</u>:

important inequality

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

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

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

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

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

any feasible pt. $\hat{x} \in C$

$$= \|x_t - \hat{x}\|^2 + \gamma_t^2\|g_t\|^2 - 2\gamma_t \langle g_t, x_t - \hat{x} \rangle \quad [\text{valid } \forall \hat{x} \in C]$$

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

$\circledast$ $(\dagger)$

using $(\dagger)$

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

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

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

true even if $\mu = 0$; for γ_t small enough

we have $\underset{\triangleq\, r(t)}{\mathbb{E}\|x_t - \hat{x}\|^2}$ $\underset{\text{with } t}{\underline{\text{decreases}}}$ for any $\hat{x} \in C$ s.t. $f(\hat{x}) \leq \mathbb{E}f(x_t)$

ie we have $r(t+1) \leq r(t)$