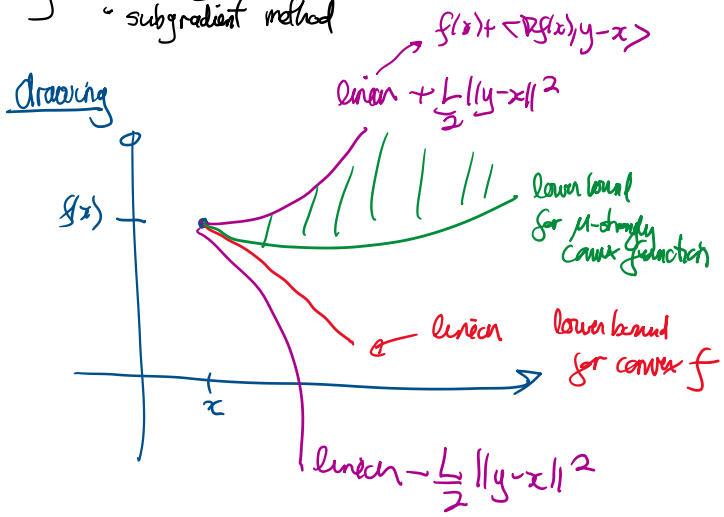
Lecture 10 - subgradient method

Thursday, February 15, 2024 9:35 AM

today: • basic gradient methods
• subgradient method

drawing



when f is twice differentiable (Hessian)

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

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

f is μ-strongly convex $\Leftrightarrow$ $f(\cdot) - \frac{\mu}{2}\|\cdot\|^2$ is convex

gradient descent:

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

a) when f is convex & L-smooth

$f(x_t) - \min_x f(x) \leq O\left(\frac{L r_0^2}{t}\right)$ where $r_0 \geq dist(x_0, X^*)$, $\triangleq \min_{x \in X^*} \|x_0 - x\|_2$

$\min_x f(x) \triangleq f^*$ $\quad$ $X^* = \operatorname{argmin}_x f(x)$

↳ [see Nesterov book for proof]

"sublinear"

note: no guarantee on $dist(x_t, X^*)$

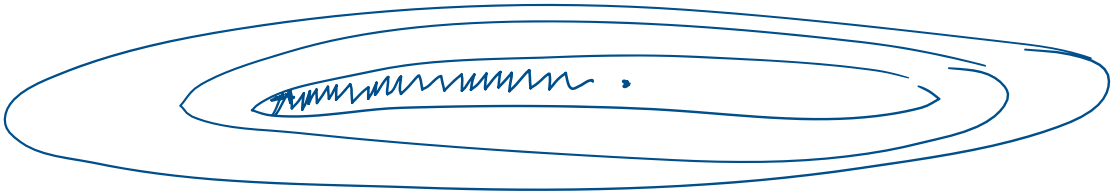
(for general L-smooth convex fcts for $t \leq dim(x_t)$)

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

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



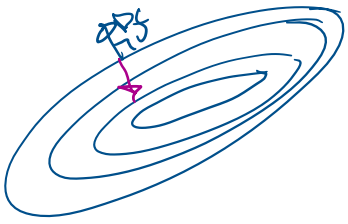
Newton's method $x_{t+1} = x_t - \gamma_t [H(x_t)^{-1}] \nabla f(x_t)$

$Hd = \nabla f(x_t)$

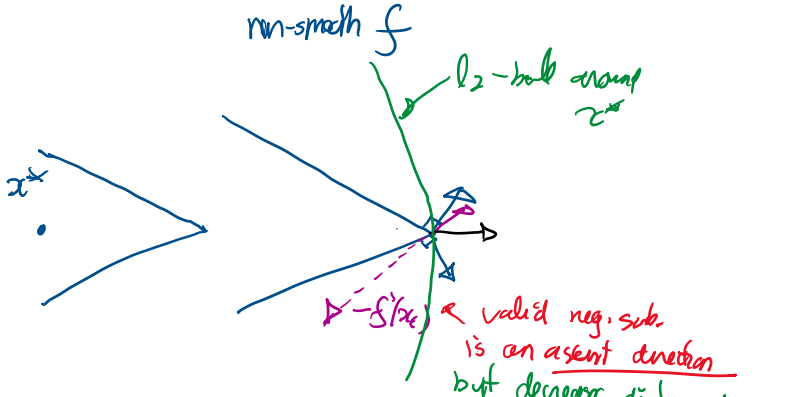
$\min_d \| Hd - \nabla f(x_t) \|_2^2$

subgradient method

non-descent methods



$-\nabla f$ is a descent direction



$-f'(x_t)$ is not nec. a descent direction

$-f'(x_t)$ a subgradient

(*) subgradient method is not nec. a descent method (on function values)

but $-f'(x_t)$ is a descent direction

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

$\left[\triangleq x_t - \gamma f'(x_t) \right]$

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

thus get closer to any x^*

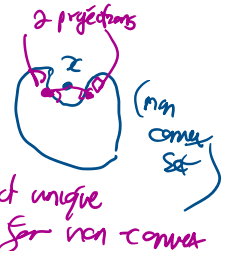
* in non-smooth optimization, $f(x_t)$ can go up & down

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

- $\operatorname{argmin}_{\{x_t\}_{t=1}^T} f(x_t)$ [in batch setting]
- weighted average $\hat{x}_T = \sum_t p_t^{(T)} x_t$ [for stochastic setting or when $f(x_t)$ is too expensive]

(*) projection operator on a closed convex set C $\quad P_C(x) \triangleq \operatorname{argmin}_{y \in C} \|x - y\|_2^2$

"Euclidean projection" of x on C



"Euclidean projection" of x on C

(not unique for non-convex set)

$P_C(\cdot)$ 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 $P_C(y) = y$

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

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

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

$\min_x f(x) \longrightarrow \min_{x \in C} f(x)$

$P_C(x)$

10h28

stochastic subgradient method

some random objective function

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

computational assumptions:

2) projection on C is "cheap"

3) we have a stochastic oracle which gives $g(x, \xi)$ for random ξ s.t. $\mathbb{E}_\xi [g(x, \xi) \mid x] = f'(x)$

some subgradient of f at x

examples:

a) f is differentiable 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) ERM (finite sum) $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ e.g. $f_i(x) = \ell(x^{(i)}, y^{(i)}; x)$ ("x" = parameter)

$= \mathbb{E}_I [f_I(x)]$ $I \overset{\text{unif}}{\sim} \{1, \dots, n\}$ $+ \frac{\lambda}{2} \|x\|^2$

$h(x, \xi) \triangleq f_\xi(x) \quad \xi \in \{1, \dots, n\}$

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

$h(x,\xi) = f_\xi(x) \quad \xi \in \{1,\dots,n\}$

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

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

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

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

↑ this replaces the Lipschitz gradient assump. for non-smooth & non-stochastic opt.

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

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

(↑ γ_t: step-size)

end

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

where $\rho_{t,T}$ are some ~~convex~~ combo. coeff.

$\sum_t \rho_{t,T} = 1 \quad \rho_{t,T} \geq 0$

(we'll see) → $O(\frac{1}{\sqrt{t}})$ rate

convergence proof:

important inequality:

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

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

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

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

(↑ $\tilde{x}$: any feasible pt. $\tilde{x} \in C$)

$$= \|x_t - \tilde{x}\|_2^2 + \gamma_t^2 \|g_t\|^2 - 2\gamma_t \langle g_t, x_t - \tilde{x} \rangle \quad [\text{valid } \forall \tilde{x} \in C]$$

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

$$= \|x_t - \tilde{x}\|_2^2 + \gamma_t^2 \|g_t\|^2 - 2\gamma_t \langle g_t, x_t - \tilde{x} \rangle \quad [\text{valid } \forall \tilde{x} \in C]$$

$$\mathbb{E}[\|x_{t+1} - \tilde{x}\|^2 | x_t] \leq \|x_t - \tilde{x}\|_2^2 + \gamma_t^2 \underbrace{\mathbb{E}[\|g_t\|^2 | x_t]}_{\leq B^2} - 2\gamma_t \langle \underbrace{\mathbb{E}[g_t | x_t]}_{f'(x_t)}, x_t - \tilde{x} \rangle \quad (\dagger)$$

using $(\dagger)$ $\leq \|x_t - \tilde{x}\|_2^2 + \gamma_t^2 B^2 - 2\gamma_t [f(x_t) - f(\tilde{x}) + \frac{\mu}{2} \|x_t - \tilde{x}\|^2]$

$\mathbb{E}[\mathbb{E}[\,|x_t]] = \mathbb{E}[\,]$

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

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

we have $\mathbb{E}\|x_t - \tilde{x}\|^2 \triangleq r_t$ decreases for any $\tilde{x} \in C$ s.t. $f(\tilde{x}) \leq \mathbb{E} f(x_t)$

ie. we have $r_{t+1} < r_t$