Lecture 10 - subgradient method

Thursday, February 6, 2020 13:30

today: • fundamental descent lemma
• subgradient method

fundamental descent lemma:

when ∇f is L-Lipschitz (lemma holds even if f is not convex)

$$f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|^2 \quad \forall x, y$$

$$* \quad f(\underbrace{x - \gamma \nabla f(x)}_{y_\gamma}) \le f(x) - \gamma \langle \nabla f(x), \nabla f(x) \rangle + \frac{\gamma^2 L}{2}\|\nabla f(x)\|^2$$

$$= f(x) - \left[\gamma\left(1 - \frac{\gamma L}{2}\right)\right]\|\nabla f(x)\|^2$$

$$> 0 \iff \boxed{0 < \gamma < \frac{2}{L}}$$

→ minimize RHS with respect to γ

gives $\boxed{\gamma^* = \frac{1}{L}}$

$$f(y_{\gamma^*}) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

---

descent lemma:

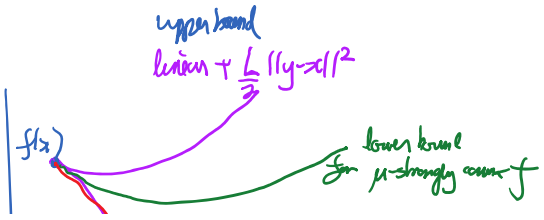
think of 2nd order Taylor expansion

$$f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}\int_{\gamma=0}^{1} \langle y-x, \underbrace{H(x+\gamma(y-x))}_{\text{Hessian of } f}\, y-x \rangle d\gamma$$

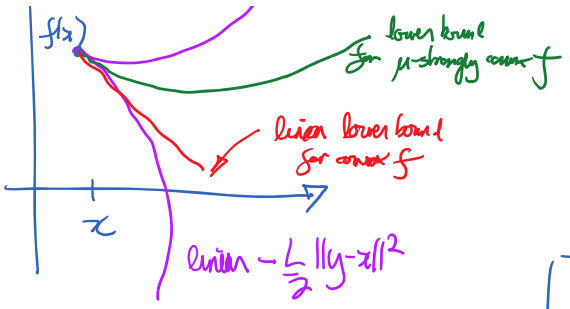
? integral form of remainder

top e-value of H in absolute value $\le L$

$$v^T H v \le \lambda_{\max}(H)\|v\|^2 \le L\|y-x\|^2$$



when f is twice differentiable



when f is twice differentiable (Hessian)

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

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

f is μ-strongly convex $\Leftrightarrow$ $f - \frac{\mu}{2}\|\cdot\|^2$ 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 $r_0 \geq dist(x_0, X^*)$, $X^* \triangleq \operatorname{argmin}_{x \in X} \|x_0 - x'\|_2$

$\min_x f(x) = f^*$, $X^* = \operatorname{argmin} f(x)$

"sublinear rate"

note: no guarantee on $dist(x_t, X^*)$ (for general L-smooth convex f's for $t \leq dim(x_t)$) → Nesterov lower bound

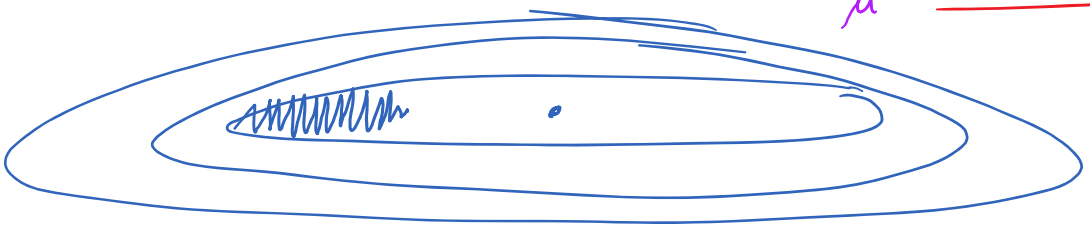
[see Nesterov book for proof]

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

"linear rate"

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

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



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

14h27

Subgradient method

non-descent methods

f smooth $\Rightarrow$ smooth subdevel sets

non-smooth f

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

$-\nabla f$ is a descent direction

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

g a subgradient

Kinks in the sublevel set

valid neg. subg. but is an ascent direction ??

(*) subgradient method is not a descent method

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

thus gets 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 $\hat{x}_T$

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

p_t convex weights

* 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

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

and thus $\|P_C(x) - y\|_2 \leq \|x - y\|_2 \quad \forall y \in C$

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

some subgradient of f at x

[example:

a) f is differentiable in x & "well behaved"

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

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

b) ERM example: $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ eg $f_i(x) = \ell(x^{(i)}, y^{(i)}; \text{"}x\text{"})$ (parameter) $+ \frac{\lambda}{2} \|\text{"}x\text{"}\|^2$

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

$\hookrightarrow \{1, \dots, n\}$

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

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

this replaces the Lipschitz gradient assumption

(i.e. f is Lipschitz instead)

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

Stochastic subgradient method

algorithm

$x_0 \in C$ 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 $\hat{x}_T \triangleq \sum_{t=0}^{T} P_{T,t} x_t$ ("weighted average")

where $P_{T,t}$ are some convex comb. coefs

$\sum_t P_{T,t} = 1 \quad P_{T,t} \ge 0$

Convergence proof:

important inequality

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

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

→ use this on $\langle -f'(x_t), x_t - x^* \rangle$

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

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

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

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

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

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

$$\overset{\text{using } (\dagger)}{\le} \|x_t-\hat{x}\|^2 + \gamma_t^2 B^2 - 2\gamma_t \left[f(x_t)-f(\hat{x}) + \frac{\mu}{2}\|x_t-\hat{x}\|^2\right]$$

$\mathbb{E}[\mathbb{E}[\,\cdot \mid x_t]] = \mathbb{E}[\,\cdot\,]$ (also true for $\mu=0$)

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