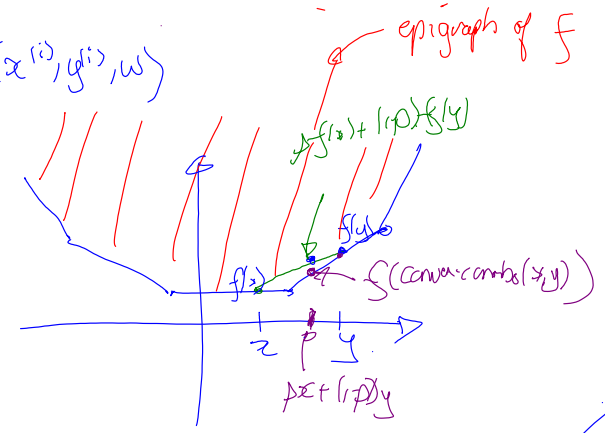
Lecture 9 - scribbles - convex analysis

Tuesday, February 20, 2018
14:31

today: convex optimization

convex surrogate loss

motivation: $\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}, y^{(i)}, w)$



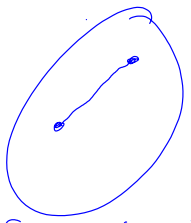
convex analysis recap:

$f: \mathbb{R}^d \to \mathbb{R}$

f is convex $\iff$

$f(px+(1-p)y) \le pf(x)+(1-p)f(y) \quad \forall x,y \quad p\in[0,1]$

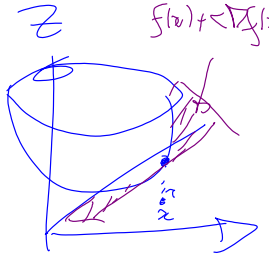
convex combination between x & y → $y+p(x-y)$



epigraph $f = \{(x,y): y \ge f(x)\}$

if f is differentiable at x, and convex

$\Rightarrow f(y) \ge f(x) + \langle \nabla f(x), y-x \rangle \quad \forall y$

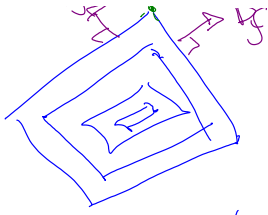


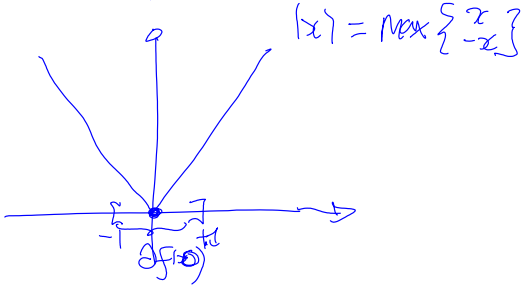
(f convex)

subdifferential of f at x

subgradient: v of f at x. $v \in \partial f(x)$

$\Leftrightarrow y \in \mathrm{dom}(f)$, $f(y) \geq f(x) + \langle v, y-x \rangle$



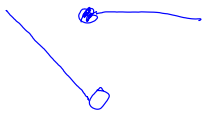


Some standard assumptions:

f is μ-strongly convex $\Leftrightarrow f(y) \geq f(x) + \langle \partial f(x), y-x \rangle + \frac{\mu}{2}\|x-y\|^2$

μ: strong convexity constant

$\forall x, y \in \mathrm{dom}(f)$ $\quad \langle v, y-x \rangle$ for any $v \in \partial f(x)$



$\mathrm{dom}(f) \triangleq \{x : f(x) < +\infty\}$

$f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$

$f^{-1}(]a, +\infty[)$ is open for all a

f is L-smooth i.e. f has a L-Lipschitz gradient for all x

$\Leftrightarrow \|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$

$\|w\|_* \triangleq \sup_{\|v\| \leq 1} \langle w, v \rangle$

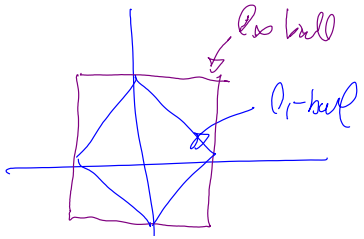
$(\|\cdot\|_p)^* = \|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$

$p = 2 \Rightarrow q = 2$

$p = 1 \Rightarrow q = \infty$

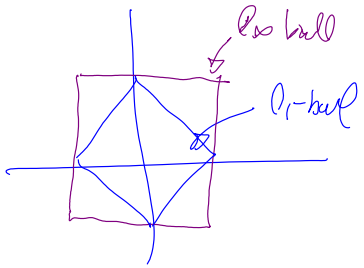
generalized C-S.

$\langle w, v \rangle \leq \|w\|_* \|v\|$



$p=2 \Rightarrow q=2$

$p=1 \Rightarrow q=\infty$



fundamental descent lemma:

when ∇f is L-Lipschitz (f though is not nec. convex)

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

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

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

minimize RHS with respect to γ gives $\boxed{\gamma^* = \frac{1}{L}}$

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

[think of 2nd order Taylor expansion

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

Hessian $\quad g(x,y) \to$ some pt. between x & y

top eigenvalue of $H \le L$ in absolute value

$$\lambda_{\min}(H)\|v\|^2 \le v^T H v \le \lambda_{\max}(H)\|v\|^2$$

or

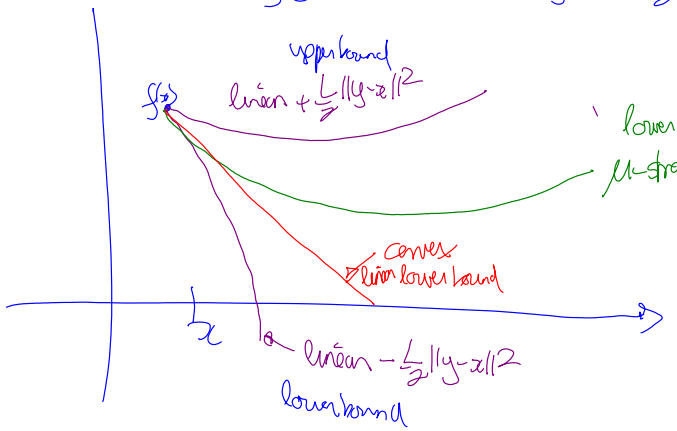
"pre-way" is to use fundamental thm. of calculus

$$f(y) = f(x) + \int_0^1 \frac{d}{d\gamma} f(x + \gamma(y-x))\, d\gamma$$

]

Lec 4

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



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

f twice differentiable

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

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

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) $f(x_t) - \min_x f(x) \le O\left(\frac{L r_0^2}{t}\right)$ when f is convex & L-smooth

where $r_0 \ge dist(x_0, X^*)$ $\quad X^* = \arg\min_x f(x)$

rate:

* no guarantee on $dist(x_t, X^*)$ (in general)
  * Nesterov lower bounds

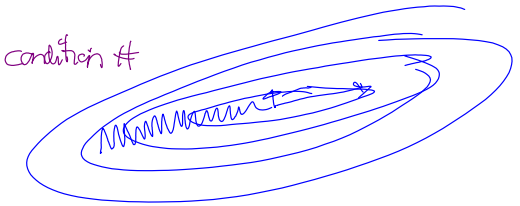
"sublinear"

[see Nesterov book for $O(\frac{1}{\varepsilon})$ rate]

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

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

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

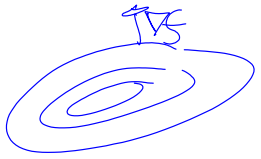


Newton's method $x_{t+1} = x_t - \gamma H(x_t)^{-1} \nabla f(x_t)$

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

$\gamma_t \nabla f(x_t)$

f smooth $\Rightarrow$ smooth sublevel sets

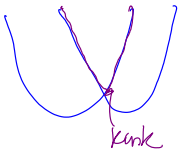


$-\nabla f$ is a descent direction

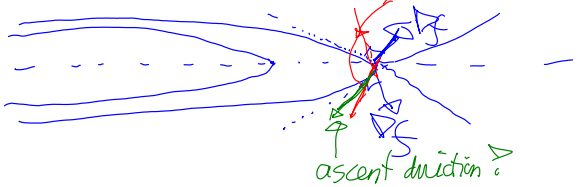
non-smooth f

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

$f'(x_t)$ subgradient



Kinks in the sublevel set $-\partial f(x)$



⭐ subgradient method is not a descent method

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

(i.e. $\|x_t - \gamma f'(x_t) - \tilde{x}\|^2 \leq \|x_t - \tilde{x}\|^2$ for γ small enough)

$x(\gamma) = x_t - \gamma f'(x_t)$

16h43

thus get closer to x^*

[in non-smooth optimization, $f(x_t)$ can go up and down]

$\Rightarrow$ combine multiple points x_t to get $\hat{x}_T$

$\operatorname{argmin}_{\{x_t\}} f(x_t)$ (in batch setting)

weights

or weighted average $\hat{x}_T = \sum_t p_t^{\text{weights}} x_t$

for stochastic setting or loss expensive to compute $f(x_t)$

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

"Euclidean projection"

$P_C(\cdot)$ is a contraction 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 a random $g(x,\xi)$

s.t. $\mathbb{E}_\xi [g(x,\xi) \mid x] = f'(x)$

↑ any subgradient of f at x

[if f is differentiable in x and "well behaved"]

[if f is differentiable in x and "well behaved"]

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

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

ERM example: $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ $\rightsquigarrow$ $\ell(x^{(i)}, y^{(i)}, w) + \frac{\lambda\|w\|^2}{2}$ (parameter w) $\quad h(x,\xi) = f_\xi(x)$, $\xi \in \{1,\dots,n\}$

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

use $g_t \triangleq g(x_t,\xi_t) \triangleq \nabla f_{i_t}(x_t)$

here $\mathbb{E}_{i_t}[\nabla f_{i_t}(x) \mid x] = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \nabla f(x)$

4) $\mathbb{E}\|g(x,\xi)\|^2 \le B^2$ (finite variance condition) [if $\|h'(x,\xi)\| \le B$ is sufficient]

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: $\hat{x}_T \triangleq \sum_{t=0}^{T} \lambda_{T,t} x_t$ ("weighted average")

where λ_t are some convex combo coefficients $\sum_{t=0}^{T} \lambda_t = 1$, $\lambda_t \ge 0$

convergence proof:

important inequality

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

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

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

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

$\|x_{t+1} - \tilde{x}\|^2 \leq \|x_t - \gamma_t g_t - \tilde{x}\|^2$

($\tilde{x}$: any feasible point, i.e. $\tilde{x} \in C$; Π_C: projection onto C)

$= \|x_t - \tilde{x}\|^2 + \gamma_t^2 \|g_t\|^2 - 2\gamma_t \langle g_t, x_t - \tilde{x} \rangle$

$$\mathbb{E}\left[\|x_{t+1}-\tilde{x}\|^2 \mid x_t\right] \leq \|x_t - \tilde{x}\|^2 + \gamma_t^2 \mathbb{E}\left[\|g_t\|^2 \mid x_t\right] - 2\gamma_t \langle \mathbb{E}[g_t \mid x_t], x_t - \tilde{x} \rangle$$

($\mathbb{E}[g_t \mid x_t] = f'(x_t)$)

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

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

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

γ_t small enough, we have $\mathbb{E}\|x_t - \tilde{x}\|^2$ decreases for any $\tilde{x} \in C$ s.t. $f(\tilde{x}) \leq \mathbb{E}[f(x_t)]$

* non-strongly setting: $\mu = 0$

Let $\tilde{x}$ be some min point x^*, i.e. $f(x^*) = \min_{x \in C} f(x)$

Let $r_t \triangleq \mathbb{E}\|x_t - x^*\|^2$

Let $\boxed{\varepsilon_t \triangleq \mathbb{E}[f(x_t) - f(x^*)]}$ expected suboptimality error

$$r_{t+1} \le r_t - 2\gamma_t \varepsilon_t + \gamma_t^2 B^2$$

$$\Rightarrow \quad 2\gamma_t [\varepsilon_t] \le r_t - r_{t+1} + \gamma_t^2 B^2 \quad \forall t$$

* sum inequalities from $t=0$ to $t=T$

$$\Rightarrow \underbrace{2\sum_{t=0}^{T} \gamma_t \varepsilon_t}_{\downarrow} \le \underbrace{r_0 - r_{T+1}}_{\text{telescoping sum } \sum_{t=0}^{T}(r_t - r_{t+1})} + \left(\sum_{t=0}^{T} \gamma_t^2\right) B^2$$

$$2\left(\sum_{t=0}^{T} \gamma_t\right) \min_{0 \le t \le T} \varepsilon_t \le \downarrow$$

$$\Rightarrow \boxed{\min_{0 \le t \le T} \varepsilon_t \le \frac{r_0 + \left(\sum_{t=0}^{T} \gamma_t^2\right) B^2}{2\sum_{t=0}^{T} \gamma_t}}$$

Use $\gamma_t = \frac{r_0}{B\sqrt{T+1}}$ to minimize RHS

$$\Rightarrow \boxed{\min_{0 \le t \le T} \varepsilon_t \le \frac{B r_0}{\sqrt{T+1}}}$$

* since f is convex, $f\left(\sum_t p_t x_t\right) \le \sum_t p_t f(x_t)$