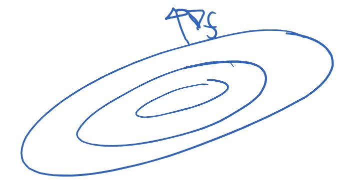
Lecture 10 - subgradient method

Thursday, February 7, 2019 13:37

<u>today</u>: subgradient method

<u>non descent methods</u>

f smooth $\Rightarrow$ smooth sublevel sets



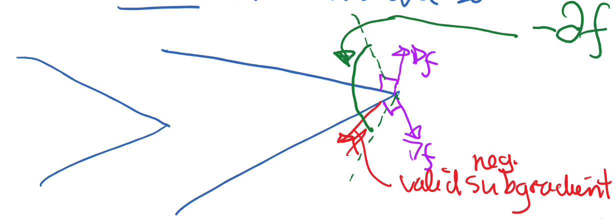
$-\nabla f$ is a <u>descent direction</u>

non-smooth f

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

(f': a subgradient)

<u>kinks</u> in the sublevel set



neg. valid subgradient but <u>going up</u>? (ascent direction)

(*) subgradient method is <u>not</u> a descent method

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

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

$x(\gamma) \triangleq x_t - \gamma f'(x_t)$ <u>and</u> for <u>any</u> $\tilde{x}$ s.t. $f(\tilde{x}) \leq f(x_t)$)

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

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

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

- $\underset{\{x_t\}}{\text{argmin}}\ f(x_t)$ [in batch setting]
- weighted average $\hat{x}_T = \sum_t \beta_t x_t$ (weights) [for stochastic setting or when too expensive to compute $f(x_t)$]

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

"Euclidean projection" of x on C

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

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 : $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ e.g. $\ell(x^{(i)}, y^{(i)}, \text{"}x\text{"}) + \frac{\lambda}{2}\|\text{"}x\text{"}\|^2$

parameter

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

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

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

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

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

4) $\mathbb{E}\|g(x,\xi)\|_2^2 \leq B^2$ (finite variance condition) ← this replaces Lipschitz gradient assumption

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

[algorithm

[sufficient condition: $\|M(x,\xi)\| \leq B \quad \forall x, \xi$]

algorithm

$x_0 \in C$ 'initialization

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

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

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

stop size

end

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

"weighted average"

where ρ_t are some convex combo. coeff.

$\sum_t \rho_t = 1 \quad \rho_t \geq 0$

14h31

convergence proof:

important inequality

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

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

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

$y = x^*$

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

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

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

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

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

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

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

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

true even if $\mu = 0$ { 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 convex setting ($\mu = 0$)

Let $\tilde{x}$ be some min pt. $\tilde{x}$ → for better rate, let $\tilde{x} = x^* = \underset{x \in X^*}{\text{argmin}} \|x - x_0\|^2$

i.e. $f(x^*) = \min_{x \in C} f(x)$

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

let $\boxed{e_t \triangleq \mathbb{E}[f(x_t) - f(x^*)]}$

let $r_t = \mathbb{E}\|x_t - x^*\|^2$

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

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

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

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

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

$$2\left(\sum_{t=0}^{T}\gamma_t\right) \min_{0\le t\le T} \varepsilon_t \le 2\sum_t \gamma_t \varepsilon_t \le \quad ''$$

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

note: $\min \varepsilon_t \to 0$ when $\dfrac{\sum_{t=0}^{T}\gamma_t^2}{\sum_t \gamma_t} \to 0$

use $\gamma_t^* = \dfrac{r_0}{B\sqrt{T+1}}$ to minimize RHS

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

b) for $\hat{x}_T = \sum_t \rho_t x_t$

by convexity

since f is convex, $f(\hat{x}_T) = f(\sum_t \rho_t x_t) \leq \sum_t \rho_t f(x_t)$

(*) can also show with $\gamma_t = \frac{A}{\sqrt{t+1}}$, $\min_{0 \leq t \leq T} \varepsilon_t \leq O\left(\frac{\log(T+1)}{\sqrt{T+1}}\right)$

and if set C is bounded, can show $O\left(\frac{\text{diam}(C)}{\sqrt{T+1}}\right)$ rate

strongly convex case ($\mu > 0$)

$$r_{t+1} \leq (1-\mu\gamma_t) r_t - 2\gamma_t \varepsilon_t + \gamma_t^2 B^2$$

divide by this

use $\boxed{\gamma_t = \frac{2}{\mu(t+2)}}$

$$\varepsilon_t \leq \frac{1}{2}(\gamma_t^{-1} - \mu) r_t - \frac{\gamma_t^{-1}}{2} r_{t+1} + \frac{\gamma_t}{2} B^2$$

$\gamma_t^{-1} = \frac{\mu(t+2)}{2}$

multiply ineq. by $(t+1)$

$$(t+1)\varepsilon_t \leq \frac{1}{2}(t+1)\left[\frac{t\mu + \cancel{2\mu} - \cancel{2\mu}}{2}\right] r_t - \frac{\mu}{4}(t+1)(t+2) r_{t+1} + \underbrace{(t+1)\frac{B^2}{\cancel{2}}\frac{\cancel{2}}{\mu(t+2)}}_{\leq \frac{B^2}{\mu}}$$

$$\leq \frac{\mu}{4}\left[\underbrace{(t+1)t\, r_t}_{\triangleq u_t} - \underbrace{(t+1)(t+2) r_{t+1}}_{u_{t+1}}\right] + \frac{B^2}{\mu}$$

telescoping sum!

$$\Rightarrow \sum_{t=0}^{T} (t+1)\varepsilon_t \leq \frac{\mu}{4}\left[\underset{0}{u_0} - u_{T+1}\right] + (T+1)\frac{B^2}{\mu}$$

let $\boxed{\rho_t \triangleq \frac{(t+1)}{S_T}}$

$S_T \triangleq \sum_{t=0}^{T}(t+1) = \frac{(T+1)(T+2)}{2}$

$$S_T \sum_{t=0}^{T} \rho_t \varepsilon_t \leq \frac{\mu}{4}\left[0 - (T+1)(T+2) r_{T+1}\right] + \frac{T+1}{\mu} B^2$$

$$S_T \sum_{t=0} \rho_t \varepsilon_t \leq \frac{\mu}{4}\left[0 - (T+1)(T+2)\, r_{T+1}\right] + \frac{T+1}{\mu} B$$

$$(\ddagger)\quad \boxed{\sum_{t=0}^{T} \rho_t \varepsilon_t + \frac{\mu}{4}\frac{(T+1)(T+2)}{S_T}\, r_{T+1} \leqslant \frac{(T+1)}{\mu}\frac{B^2}{S_T}}$$

let $\hat{x}_T \triangleq \sum_{t=0}^{T} \rho_t x_t$ (weighted avg.)

by convexity, $f(\hat{x}_T) = f\left(\sum_t \rho_t x_t\right) \leqslant \sum_t \rho_t f(x_t)$

$$\frac{T+1}{S_T} = \frac{2}{T+2}$$

$$\Rightarrow \quad \mathbb{E}\left[f(\hat{x}_T) - f(x^*)\right] \leqslant \sum_t \rho_t \underbrace{\left[\mathbb{E} f(x_t) - f(x^*)\right]}_{\varepsilon_t} \overset{(\ddagger)}{\leqslant} \frac{T+1}{\mu S_T} B^2$$

thus $\boxed{\mathbb{E} f(\hat{x}_T) - f(x^*) \leq \frac{2B^2}{\mu(T+2)}}$ vs. $O\left(\frac{1}{\sqrt{T}}\right)$ rate when $\mu = 0$

also $\underbrace{\mathbb{E}\|x_{T+1} - x^*\|^2}_{r_{T+1}} \leq \frac{4B^2}{\mu^2(T+2)}$ //