Lecture 10 - scribbles - convex opt

Friday, February 23, 2018
14:32

today: continue stochastic subgradient method (SGD)

[sidenote: $g(w) = \mathbb{E}_{(x,y)\sim p} f(x,y,w)$]
$\rightsquigarrow L(w)$ using calibration function

last time: we showed that $\min_{0\le t\le T} \underbrace{\left(\mathbb{E} f(x_t) - f(x^*)\right)}_{\triangleq \varepsilon_t} \le \frac{BR_0}{\sqrt{T+1}}$ when $\gamma_t = \frac{R_0}{B\sqrt{T+1}}$ (constant)

for stochastic subgradient method

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

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

strongly convex case ($\mu > 0$)

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

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

$$\varepsilon_t \le \frac{1}{2}\left(\gamma_t^{-1} - \mu\right) r_t^2 - \frac{\gamma_t^{-1}}{2} r_{t+1}^2 + \frac{\gamma_t B^2}{2} \quad \Rightarrow \quad \gamma_t^{-1} = \frac{\mu(t+2)}{2}$$

multiply inequalities by $(t+1)$ (to get a telescoping sum)

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

$$\frac{(t+1)}{(t+2)} \le 1$$

$$(t+1)\,\varepsilon_t \le \frac{\mu}{4}(t+1)t\,r_t - \frac{\mu}{4}(t+1)(t+2)\,r_{t+1} + \frac{B^2}{\mu}$$

$$\Rightarrow \sum_{t=0}^{T}(t+1)\varepsilon_t \le \frac{\mu}{4}\sum_{t=0}^{T}\underbrace{\left[\underbrace{(t+1)t\,r_t}_{\triangleq u_t} - \underbrace{(t+1)(t+2)\,r_{t+1}}_{u_{t+1}}\right]}_{u_0 - u_{T+1}} + (T+1)\frac{B^2}{\mu}$$

$$u_0 - u_{T+1} = 0 - (T+1)(T+2)\,r_{T+1}$$

Let $p_t \triangleq \frac{(t+1)}{S_T}$

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

$$(\ddagger)\quad \boxed{\sum_{t=0}^{T} p_t\varepsilon_t + \underbrace{\frac{\mu}{4}\frac{(T+1)(T+2)}{S_T}}_{=1} r_{T+1} \le \frac{(T+1)}{\mu}\frac{B^2}{S_T}}$$

Let $\hat{x}_T \triangleq \sum_{t=0}^{T} p_t x_t$ (weighted average)

by convexity, $f(\hat{x}_T) = f\left(\sum_{t=0}^{T} p_t x_t\right) \le \sum_{t=0}^{T} p_t f(x_t)$

$$\Rightarrow \mathbb{E}f(\hat{x}_T) - f(x^*) \le \sum_{t=0}^{T} p_t \underbrace{\left[\mathbb{E}f(x_t) - f(x^*)\right]}_{\varepsilon_t} \overset{(\ddagger)}{\le} \frac{(T+1)}{\mu S_T}B^2$$

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

$\Rightarrow \sum_{t=0}^{T} p_t = 1$

$$\frac{(t+1)}{S_T} = \frac{2}{T+2}$$

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

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

Strategy:

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

$$\sum_t 2p_t\gamma_t\varepsilon_t \le \sum_t p_t\left[\quad\quad\right]$$

note: $p_{t,T} = \frac{(t+1)}{S_T} = \frac{2(t+1)}{(T+1)(T+2)}$

Landscape of global convergence rates

f is convex — $f(x_t) - f(x^*) \le$ ____ — $r_0 \ge \mathrm{dist}(x_0, X^*)$

$\mathbb{E} f(\hat{x}_t) - f(x^*) \le$ ____ (stochastic setting)

assumptions	rate deterministic (batch)		stochastic setting	finite sum special case $\frac{1}{n}\sum_{i=1}^n f_i(x)$	
1) non-smooth $\|\partial f\| \le B$	$O\left(\frac{B r_0}{\sqrt{t}}\right)$	subgradient method	$O\left(\frac{B r_0}{\sqrt{t}}\right)$		
2) smooth L-Lipschitz ∇f	$O\left(\frac{L r_0^2}{t}\right)$	gradient method	$O\left(\frac{\square}{\sqrt{t}}\right)$ SGD	$O\left(\frac{\sqrt{n} L}{t}\right)$	SAG/A
	$O\left(\frac{L r_0^2}{t^2}\right)$ (lower bound)	Nesterov method "optimal method"			
f is μ-strongly convex 3) non-smooth $\|\partial f\| \le B$	$O\left(\frac{B^2}{\mu t}\right)$	subgradient method	$O\left(\frac{B^2}{\mu t}\right)$		
smooth L-Lipschitz	$O(\exp(-\frac{\mu}{L} t))$	gradient method	$O\left(\frac{\square}{\mu t}\right)$	$O(\exp(-\square t))$	SAG/A
	$O(\exp(-\sqrt{\frac{\mu}{L}}\, t))$	Nesterov method			

(*) note: projecting gives the same rates

more generally, proximal gradient method as well

$\min_x f(x) + h(x)$ "composite smooth optimization"

smooth non-smooth

constrained opt: $h(x) = \delta_C(x) \triangleq \begin{cases} +\infty & \text{if } x \notin C \\ 0 & \text{o.w.} \end{cases}$

proximal gradient method: prox step

related to step-size ...

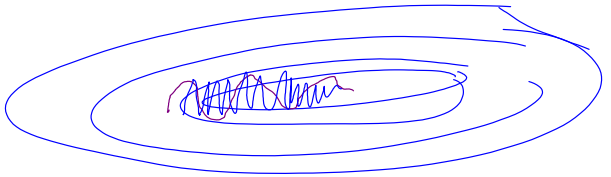
proximal gradient method:

$$x_{t+1} = \arg\min_x \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{L}{2}\|x - x_t\|^2 + h(x)$$

related to step-size...

projected gradient method = proximal gradient method when $h(x) = \delta_C(x)$

example: $h(x) = \|x\|_1$



"linear coupling" paper