Note Title 27-Jun-19

LECTURE-5

Faster algorithms (better than the black-box bounds) for structured non-smooth problems

$$f(x) = \underbrace{g(x)}_{\text{Convex \& smooth}} + \underbrace{h(x)}_{\text{Convex \& perhaps non-smooth}}$$

Access to : $\arg\min_{x} \langle w, x \rangle + h(x) + \frac{1}{2\eta}\|x - y\|^2$

$\forall\, w, \eta, y.$

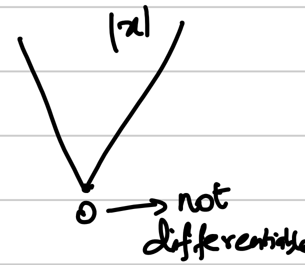
$$f(x) = g(x) + h(x).$$

$\boxed{\text{GD with Prox Algorithm}}$: $x_{t+1} = \arg\min_x \; g(x_t) + \langle \nabla g(x_t), x - x_t \rangle$
$\qquad + h(x) + \frac{1}{2\eta} \|x - x_t\|^2.$

GD : $x_{t+1} = \arg\min_x \; g(x_t) + \langle \nabla g(x_t), x - x_t \rangle$
$\qquad + h(x_t) + \langle \nabla h(x_t), x - x_t \rangle$
$\qquad + \frac{1}{2\eta} \|x - x_t\|^2.$

LASSO: $f(x) = \underbrace{\frac{1}{2}\|Ax-b\|^2}_{g(x)} + \underbrace{\lambda \|x\|_1}_{h(x)}$

(Compressed Sensing)



$g(x)$: Linear regression

$h(x)$: encourage sparsity of x.

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

$$= \arg\min_x \langle \nabla g(x_t), x \rangle + \underbrace{h(x)}_{\lambda \|x\|_1} + \frac{1}{2\eta}\|x - x_t\|^2$$

$$= \arg\min_{x} \sum_{i=1}^{d} x_i \nabla g(x_t)_i + |x_i| + \frac{1}{2\eta}\left(x_i - (x_t)_i\right)^2$$

$$(x_{t+1})_i = \arg\min_{x_i} x_i \nabla g(x_t)_i + |x_i| + \frac{1}{2\eta}\left(x_i - (x_t)_i\right)^2$$

$$= \arg\min_{x_i} \frac{1}{2\eta}\left(\underbrace{x_i - (x_t)_i} + \eta \nabla g(x_t)_i\right)^2 + |x_i|.$$

Exercise:

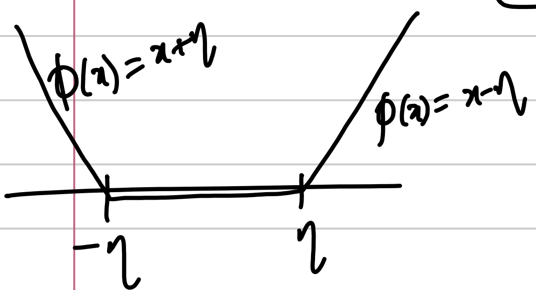
$$
\begin{aligned}
(x_{t+1})_i &= (x_t)_i - \eta \nabla g(x_t)_i - \eta && \text{if } \boxed{(x_t)_i - \eta \nabla g(x_t)_i} \geq \eta \\
&= \qquad \text{"} \qquad + \eta && \text{if } \qquad \text{"} \qquad \leq -\eta \\
&= 0 && \text{Otherwise.}
\end{aligned}
$$

$$= \phi\left(x_t - \eta \nabla g(x_t)\right),$$ where ϕ is applied to each coordinate.

$$x_{t+1} = \underbrace{x_t - \eta \nabla g(x_t)}_{\triangleq\, \hat{x}_{t+1}} + \eta\, \mathbb{1}_{\{\text{Coordinates} < -\eta\}}$$

$$- \eta\, \mathbb{1}_{\{\qquad > \eta\}}$$

$$- \hat{x}_{t+1}\, \mathbb{1}_{\{-\eta \le \text{Coord.} \le \eta\}}$$



$$\hat{f}_\eta(x_t; x) \triangleq \underbrace{g(x_t)} + \underbrace{\langle \nabla g(x_t), x - x_t \rangle} + \underbrace{h(x)} + \underbrace{\frac{1}{2\eta} \|x - x_t\|^2}.$$

$x^* = \arg\min_x f(x)$

$$x_{t+1} = \arg\min_x \hat{f}_\eta(x_t; x).$$

g is L-Smooth

$\eta \le \frac{1}{L}.$

$$\begin{aligned}
f(x^*) &= g(x^*) + h(x^*) \\
&\overset{\text{Conv. of } g(\cdot)}{\geq} g(x_t) + \langle \nabla g(x_t), x^* - x_t \rangle + h(x^*) \\
&\qquad + \frac{1}{2\eta}\|x^* - x_t\|^2 - \frac{1}{2\eta}\|x^* - x_t\|^2 \\
&= \hat{f}_\eta(x_t; x^*) - \frac{1}{2\eta}\|x^* - x_t\|^2 \\
&\geq \hat{f}_\eta(x_t; x_{t+1}) + \frac{1}{2\eta}\|x^* - x_{t+1}\|^2 - \frac{1}{2\eta}\|x^* - x_t\|^2
\end{aligned}$$

$\left[x_{t+1} = \arg\min \hat{f}_\eta \ \& \ \hat{f}_\eta - \frac{1}{\eta} \text{ str. cvx.} \right]$

$$\hat{f}_\eta(x_t; x) = \underbrace{g(x_t) + \langle \nabla g(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|^2}_{} + h(x)$$

L-smoothness of $g(\cdot)$ & $\eta \leq \frac{1}{L}$

$$\geq g(x) + h(x) = f(x). \quad \forall x.$$

x_t x

smoothness Lemma: $g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$

$$f(x^*) \geq f(x_{t+1}) + \frac{1}{2\eta}\|x^* - x_{t+1}\|^2 - \frac{1}{2\eta}\|x^* - x_t\|^2.$$

Rearranging,

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\eta\,[f(x_{t+1}) - f(x^*)].$$

[Telescope] $$\leq \|x_0 - x^*\|^2 - 2\eta \sum_{s=1}^{t+1} [f(x_s) - f(x^*)].$$

$$\frac{1}{t+1} \sum_{s=1}^{t+1} [f(x_s) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2 - \|x_{t+1} - x^*\|^2}{2\eta\,(t+1)}.$$

$\eta = \frac{1}{L}$ $$\leq \frac{L\|x_0 - x^*\|^2}{2(t+1)}.$$

When the non-smooth component is simple, we can get $O(\frac{1}{T})$ convergence rate. [ISTA]

Iterative Shrinkage Thresholding Algorithm

We can infact improve this to $O(\frac{1}{T^2})$ using Nesterov's AGD. [FISTA]

STOCHASTIC ALGORITHMS

1)
$$f(x) = \mathbb{E}_{(a,b)}\left[(a^T x - b)^2\right]$$

$$\nabla f(x) = \mathbb{E}_{(a,b)}\left[(a^T x - b)a\right].$$

Oracle : (a_i, b_i)

$$\hat{\nabla} f(x) \triangleq (a_i^T x - b_i)a_i \Longrightarrow \mathbb{E}[\hat{\nabla} f(x)] = \underline{\nabla f(x)}.$$

② Empirical Risk Minimization (ERM)

$$f(x) = \frac{1}{2n} \sum_{i=1}^{n} (a_i^T x - b_i)^2$$

If $a_i \in \mathbb{R}^d$

$O(nd)$ time $\longleftarrow \nabla f(x) = \frac{1}{n} \sum_{i=1}^{n} (a_i^T x - b_i) a_i$

$O(d)$ time $\longleftarrow \hat{\nabla} f(x) \triangleq (a_i^T x - b_i) a_i$ where $i \sim \mathrm{Unif}[1, \ldots, n]$.

Setting: $x \longrightarrow \boxed{\text{oracle}} \longrightarrow \hat{\nabla} f(x)$

$\hat{\nabla} f(x)$ independent of everything else $\Big($ $\mathbb{E}[\hat{\nabla} f(x)] = \nabla f(x)$; $\mathbb{E}\left[\| \hat{\nabla} f(x) - \nabla f(x) \|^2\right] \leq \sigma^2$.

$\|\nabla f(x)\| \le G$

f : Convex & G-Lipschitz
Random noise has bounded variance

SGD [Robbins & Monro]

$$x_{t+1} = x_t - \eta \hat{\nabla} f(x_t).$$

$$\mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] = \mathbb{E}\left[\|x_t - x^* - \eta \hat{\nabla} f(x_t)\|^2\right]$$

$$= \mathbb{E}\left[\|x_t - x^*\|^2 - 2\eta \langle \hat{\nabla} f(x_t), x_t - x^* \rangle + \eta^2 \|\hat{\nabla} f(x_t)\|^2\right]$$

$$= \mathbb{E}\left[\|x_t - x^*\|^2\right] - 2\eta \, \underline{\mathbb{E}\left[\langle \hat{\nabla} f(x_t), x_t - x^* \rangle\right]} + \eta^2 \, \underline{\mathbb{E}\left[\|\hat{\nabla} f(x_t)\|^2\right]}$$

1) $\mathbb{E}\left[\left\langle \hat{\nabla} f(x_t), x_t - x^* \right\rangle \Big| x_t\right] = \left\langle \mathbb{E}\left[\hat{\nabla} f(x_t) \Big| x_t\right] x_t - x^* \right\rangle$

$= \left\langle \nabla f(x_t), x_t - x^* \right\rangle$

[Convexity] $\geqslant f(x_t) - f(x^*)$.

2) $\mathbb{E}\left[\left\langle \hat{\nabla} f(x_t), \hat{\nabla} f(x_t) \right\rangle \Big| x_t\right] = \mathbb{E}\left[\left\langle \nabla f(x_t) + n_t, \nabla f(x_t) + n_t \right\rangle\right]$

$n_t \triangleq \hat{\nabla} f(x_t) - \nabla f(x_t)$

$= \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \mathbb{E}\left[\|n_t\|^2\right]$

$\leq G^2 + \sigma^2$

Since $f(\cdot)$ is G-Lipschitz

Since noise variance $\leq \sigma^2$

So, $\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \mathbb{E}[\|x_t - x^*\|^2]$

$$- 2\eta\,[\mathbb{E}[f(x_t)] - f(x^*)]$$

$$+ \eta^2 (G^2 + \sigma^2).$$

Using the above inequality iteratively

$$\leq \mathbb{E}[\|x_0 - x^*\|^2] - 2\eta \sum_{s=0}^{t} [\mathbb{E}[f(x_s)] - f(x^*)]$$

$$+ \eta^2 (t+1)(G^2 + \sigma^2).$$

$$\frac{1}{t+1} \sum_{s=0}^{t} \mathbb{E}[f(x_s)] - f(x^*) \leq \frac{1}{2\eta(t+1)} [\|x_0 - x^*\|^2 - \mathbb{E}[\|x_{t+1} - x^*\|^2]]$$

$$+ \frac{\eta}{2}\left(G^2+\sigma^2\right).$$

Choose $\eta = \sqrt{\dfrac{\|x_0 - x^*\|^2}{(t+1)(G^2+\sigma^2)}}$

function parameter

stochastic oracle

Av. expected suboptimality $\leq \dfrac{\sqrt{(G^2+\sigma^2)} \cdot \|x_0 - x^*\|}{\sqrt{t+1}}$.

For nonsmooth: this is the best possible rate $\left(\frac{1}{\sqrt{t}}\right)$

For smooth: Cannot improve on $\frac{1}{\sqrt{t}}$; but can obtain $\dfrac{L\|x_0 - x^*\|^2}{t^2} + \dfrac{\sigma\|x_0 - x^*\|}{\sqrt{t}}$.

Exercise: SGD for L-Smooth $f(\cdot)$ gets rate of

$$O\left(\frac{L\|x_0 - x^*\|^2}{t} + \frac{\sigma\|x_0 - x^*\|}{\sqrt{t}}\right).$$

Exercise: If f is G-Lipschitz and μ-Strongly convex, then SGD [with diff. η] gets rate of

$$O\left(\frac{G^2+\sigma^2}{\mu t}\right).$$