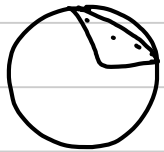
LECTURE - 6

Note Title 27-Jun-19

MIRROR DESCENT

$$\min_{x \in \{x_i \geq 0,\ \sum_i x_i = 1\}} \underbrace{\frac{1}{2}\|Ax - b\|^2}_{\hookrightarrow = f(x)}$$



PGD: $x_{t+1} = P_{\mathcal{X}}\left(x_t - \eta \nabla f(x_t)\right)$ | $O\left(\frac{\|A^TA\|\sqrt{2}}{t}\right)$

$L = \|A^TA\|$; $\|x_0 - x^*\| \leq Diam(\mathcal{X}) \leq \sqrt{2}$

↑ Smoothness

each Column of A : $\|A_i\| \leq 1$

$\|A^T A\| = \|A\|^2$ Could be $\approx \underline{\underline{d}}$.

$[\,|\,|\,|\,|\,|\,]$ ↓ unit norm columns.

When $\mathcal{X}$ and $f(\cdot)$ have good properties w.r.t. $\|\cdot\|$ which is not Euclidean, Can we do better?

x $\quad\quad$ ∇f

$\downarrow$ $\quad\quad$ $\downarrow$

$\|\cdot\|$ $\quad\quad$ $\|\cdot\|_*$

e.g., for probabilities $\|\cdot\|_1 \longmapsto \|\cdot\|_\infty$

<u>Defn. of dual norm</u>: Given a norm $\|\cdot\|$ on x, the dual norm on the space of linear operators $\mathcal{L}(x)$ is defined as: $\|l\|_* = \sup_{\substack{\|x\| \le 1 \\ x \in x}} l(x) \quad \forall\, l \in \mathcal{L}(x)$.

$\|x\| \triangleq \|x\|_1 = \sum_i |x_i|$ $\quad x \in \mathbb{R}^d$ $\xrightarrow{\text{Dual norm}}$ $\|y\|_* = \|y\|_\infty = \max_i |y_i|$ $\quad y \in \mathbb{R}^d$

$\|y\|_* = \sup_{\|x\|_1 \le 1} \sum_i x_i y_i$

$= \|y\|_\infty$.

$x_{i^*} = \text{sign}(y_{i^*})$ on $i^* = \arg\max_i |y_i|$

$= 0$ o.w.

$\|\cdot\|$ on x $\quad$, $\quad$ $\|\cdot\|_*$ on $\nabla f(x)$

$\|\cdot\|_1$ $\quad$ $\|\cdot\|_\infty$

$f(y) \ge f(x) + \langle \underbrace{\nabla f(x)}_{\nabla f(x)}, \underbrace{y - x}_{\in x} \rangle$

f-Convex & $\|\nabla f(x)\|_* \le G$ → Lipschitz

$\|\cdot\|$

GD: $x_{t+1} = \arg\min_x \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|_2^2$

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

① Not clear how to solve this step $\left(\sum_i |x_i - x_{t,i}|\right)^2$

② Not clear what the potential function is.

Defn. **Bregman divergence**: Given a **Convex** function $\phi : \mathcal{X} \to \mathbb{R}$,

$$D_\phi(x; y) \triangleq \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle \geq 0$$

$\because \phi$ is convex.

Mirror descent:

$$x_{t+1} = \arg\min_{x} \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} D_\phi(x; x_t)$$

Requirement: ϕ is 1-strongly convex w.r.t. $\|\cdot\|$ on $\mathcal{X}$.

$$\phi(y) \geq \phi(x) + \langle \nabla\phi(x), y - x \rangle + \frac{1}{2}\|x - y\|^2.$$

ASSUMPTIONS

① f is convex

② $\|\nabla f(x)\|_* \leq G$

③ $\phi(\cdot)$ is 1-Strongly Convex in $\|\cdot\|$ on X.

$D_\phi(x^*; x_t)$

$x \in X$ and $l \in \mathcal{L}(x) = \nabla f(x)$.

Hölder's ineq.: $\langle l, x \rangle \leq \|l\|_* \cdot \|x\|$

$$x_{t+1} = \arg\min_x \underbrace{f(x_t)}_{\text{Const.}} + \underbrace{\langle \nabla f(x_t), x - x_t \rangle}_{\text{Linear in } x} + \underbrace{\frac{1}{2\eta} D_\phi(x; x_t)}_{\text{Convex in } x}.$$

$$\nabla f(x_t) + \frac{1}{2\eta} \nabla_x \underbrace{D_\phi(x; x_t)}_{\phi(x) - \phi(x_t) - \langle \nabla\phi(x_t), x - x_t \rangle} = 0.$$

$$\nabla f(x_t) + \frac{1}{2\eta} \left[\nabla\phi(x_{t+1}) - 0 - \nabla\phi(x_t) \right] = 0.$$

$$\nabla_{t+1}: \quad \underbrace{\nabla\phi(x_{t+1})}_{= x_{t+1}} = \underbrace{\nabla\phi(x_t)}_{x_t} - 2\eta \nabla f(x_t). \quad \Big] \text{ MD Step.}$$

$\longrightarrow$ ①

By optimality of x_{t+1},

$$f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{1}{2\eta} D_\phi(x_{t+1}; x_t) \leq f(x_t) + \langle \nabla f(x_t), x_t - x_t \rangle + \frac{1}{2\eta} D_\phi(x_t; x_t)$$

$$= f(x_t)$$

$$\frac{1}{4\eta} \|x_t - x_{t+1}\|^2 \overset{1\text{-Str. Conv. of } \phi}{\leq} \frac{1}{2\eta} D_\phi(x_{t+1}; x_t) \leq - \langle \nabla f(x_t), x_{t+1} - x_t \rangle$$

$$\overset{\text{Hölder's}}{\leq} \|\nabla f(x_t)\|_* \cdot \|x_{t+1} - x_t\|$$

$$\|x_t - x_{t+1}\| \leq 4\eta \|\nabla f(x_t)\|_* \leq 4\eta G \longrightarrow (2)$$

$$D_\phi(x; x_{t+1}) = \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_{t+1}),\ x - x_{t+1}\rangle$$

$$\overset{①}{=} \phi(x) - \phi(x_{t+1}) - \langle \nabla\phi(x_t) - 2\eta\nabla f(x_t),\ x - x_{t+1}\rangle$$

$$\overset{\text{Conv. of } \phi}{\leq} \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t),\ x_{t+1} - x_t\rangle$$

$$- \langle \nabla\phi(x_t) - 2\eta\nabla f(x_t),\ x - x_{t+1}\rangle$$

$$= \phi(x) - \phi(x_t) - \langle \nabla\phi(x_t),\ x - x_t\rangle$$

$$- 2\eta\langle \nabla f(x_t),\ x_{t+1} - x\rangle$$

$$= D_\phi(x; x_t) - 2\eta\langle \nabla f(x_t),\ x_t - x\rangle$$

$$-2\eta \langle \nabla f(x_t), x_{t+1} - x_t \rangle .$$

$$\overset{\text{Hölder's}}{\leq} D_\phi(x; x_t) - 2\eta \langle \nabla f(x_t), x_t - x \rangle + 2\eta \| \nabla f(x_t) \|_* \cdot \| x_{t+1} - x_t \|$$

(Convexity: $\langle \nabla f(x_t), x_t - x \rangle \geq f(x_t) - f(x)$; ② : $\| x_{t+1} - x_t \| \leq 4\eta G$)

$$\leq D_\phi(x; x_t) - 2\eta [f(x_t) - f(x)] + 2\eta \underbrace{\| \nabla f(x_t) \|_*}_{\leq G} \cdot 4\eta G .$$

$$\leq D_\phi(x; x_t) - 2\eta [f(x_t) - f(x)] + 8\eta^2 G^2 .$$

$$D_\phi(x; x_{t+1}) \le D_\phi(x; x_t) - 2\eta[f(x_t) - f(x)] + 8\eta^2 G^2.$$

Telescoping and averaging gives us

$$\frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x)] \le \frac{D_\phi(x; x_0) - D_\phi(x; x_T)}{2\eta T} + 4\eta G^2.$$

$$\le \frac{D_\phi(x; x_0)}{2\eta T} + 4\eta G^2.$$

$$\eta = \frac{1}{G}\sqrt{\frac{D_\phi(x; x_0)}{8T}}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[f(x_t)-f(x)\right] \leq C\cdot\frac{G\sqrt{D_\phi(x;x_0)}}{\sqrt{T}}.$$

$$\leq C\cdot\frac{G\,\|x-x_0\|}{\sqrt{T}}$$

① If $\|\cdot\| = \|\cdot\|_2$ (↳ Euclidean), then choose $\phi = \frac{1}{2}\|x\|_2^2$ → 1-str. cvx. w.r.t. $\|\cdot\|_2$

$$D_\phi(x;x_t) = \frac{1}{2}\|x-x_t\|^2.$$

In this case a) Mirror descent = GD

b) Obtained bounds are the same.

② $\|\ \| = \|\cdot\|_1$ and $\mathcal{X} = \{ x : x_i \geq 0 ;\ \Sigma x_i = 1 \}$.

$$\phi(x) = \sum_i x_i \log x_i$$

$$D_\phi(x;y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y \rangle$$

$$= \sum_i x_i \log x_i - \sum_i y_i \log y_i - \sum_i (1+\log y_i)(x_i - y_i)$$

$$= \sum_i x_i \log \frac{x_i}{y_i} + \underbrace{\sum_i (x_i - y_i)}_{=0}$$

$$= KL(x \| y).$$

ASIDE

$x_i \geq 0$: $\phi(x) = \sum_i x_i \log x_i - \sum_i x_i$ $\quad [\sum_i x_i \neq 1]$.

$\sum_i x_i = 1, \quad x_i \geq 0. \longrightarrow \phi(x) = \sum_i x_i \log x_i$

$D_\phi(x; y) = KL(x \| y).$

① ϕ [or eq. D_ϕ] is 1-Strongly Convex

Pinsker's inequality: $KL(x \| y) \geq 2 \|x-y\|_1^2$

② $x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \underbrace{f(x_t)}_{\text{Const.}} + \langle \nabla f(x_t), x - \underbrace{x_t}_{\text{Const.}} \rangle + \frac{1}{2\eta} D_\phi(x; x_t)$

$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \langle \nabla f(x_t), x \rangle + \frac{1}{2\eta} \sum_i x_i \log \frac{x_i}{(x_t)_i}$

$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_i x_i \left[\nabla f(x_t)_i + \underbrace{\frac{1}{2\eta}} \log \frac{x_i}{(x_t)_i} \right]$.

$$= \operatorname*{argmin}_{x \in \mathcal{X}} \sum_i x_i \log \frac{x_i}{(x_t)_i \exp\left(-2\eta (\nabla f(x_t))_i\right)}$$

$$z_i \triangleq \frac{(x_t)_i \exp\left(-2\eta (\nabla f(x_t))_i\right)}{\sum_{j=1}^{d} (x_t)_j \exp\left(-2\eta (\nabla f(x_t))_j\right)}$$

$$= \operatorname*{argmin}_{x \in \mathcal{X}} \sum_i x_i \log \frac{x_i}{z_i} - \underbrace{\sum_i x_i}_{=1} \underbrace{\log\left(\sum_{j=1}^{d} (x_t)_j \exp\left(-2\eta (\nabla f(x_t))_j\right)\right)}_{\text{Constant}}$$

Constant

$$= \underset{x \in X}{\operatorname{argmin}} \quad KL(x \| z) = z.$$

Ind. discovered

$$x_{t+1} = z = \frac{x_t \odot \exp(-2\eta \nabla f(x_t))}{\| x_t \odot \exp(-2\eta \nabla f(x_t)) \|_1}.$$

MULTIPLICATIVE WEIGHTS UPDATE

EXPONENTIATED GRADIENT ALGORITHM

$$\min_{x:\, x_i \geq 0,\ \sum x_i = 1} \frac{1}{2}\cdot\|Ax-b\|^2.$$

$\|\cdot\|_1$ on X $\longrightarrow$ $\|\cdot\|_\infty$ on $\nabla f(X)$.

$$\phi(x) = \sum_i x_i \log x_i$$

$$\text{Av. Suboptimality} \leq C \cdot \frac{\|\nabla f(x)\|_\infty \sqrt{D_\phi(x; x_0)}}{\sqrt{T}}$$

$$\nabla f(x) = A^T(Ax-b) \longrightarrow \|A^T(Ax-b)\|_\infty \leq \|A^TAx\|_\infty + \|A^Tb\|_\infty$$

$$\leq \max_i \|(A^TA)_i\|_\infty + \|A^Tb\|_\infty$$

Hölder's inequality

i^{th} row of A^TA

Non-Euclidean		Euclidean
$\|A^TA\|_\infty$	$\leq$	$\|A^TA\|_2$ Could be d times larger.

$\downarrow$ largest element in A^TA

$$G \coloneqq \|A^TA\|_\infty + \|A^Tb\|_\infty \leq \|A^TA\|_2 + \|A^Tb\|_2.$$

$D_\phi(x; x_0)$: $x_0 = \frac{1}{d} \cdot \mathbb{1}$

$$= \sum_i x_i \log \frac{x_i}{(x_0)_i} = \underbrace{\sum_i x_i \log x_i}_{\leq 0} + \underbrace{\sum_i x_i \log d}_{= \log d}$$

$$\text{Av. Suboptimality} \leq C \cdot \frac{(\|A^TA\|_\infty + \|A^Tb\|_\infty) \cdot \sqrt{\log d}}{\sqrt{T}} \cdot \sqrt{\log d} \cdot$$

Could be $\Omega(d)$

$$C \cdot \frac{(\|A^TA\|_2 + \|A^Tb\|_2) \cdot 1}{\sqrt{T}} \cdot$$