"Geometrization"

Monday, February 3, 2020 7:19 PM

Part 1. Constrained Optimization & Mirror Descent

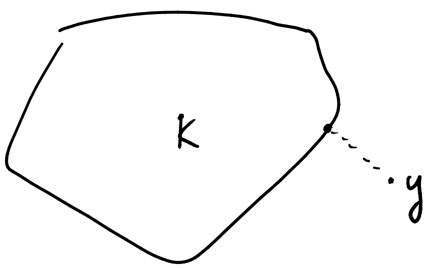
$\min_{x \in K} f(x)$

Can use CP.
But "high" polynomial time
n^2 space.

What about GD?
$-\nabla f(x)$ might point outside.
- may not be defined

$g \in \nabla f(x)$: $\forall y \quad f(y) \geq f(x) + \langle g, y-x \rangle$

Projection : $\Pi_K(y) = \arg\min_{x \in K} \|y-x\|_2$



Lemma. $\forall z \in K$

$$\|y - \Pi(y)\|^2 + \|\Pi(y) - z\|^2 \leq \|z - y\|^2.$$

Pf. Let $h(t) = \|\Pi(y) + t(z - \Pi(y)) - y\|^2$

Then $h(t) \geq \|\Pi(y) - y\|^2 = h(0)$

$\Rightarrow \quad h'(0) \geq 0$

i.e. $(\Pi(y) - y)^T (z - \Pi(y)) \geq 0$

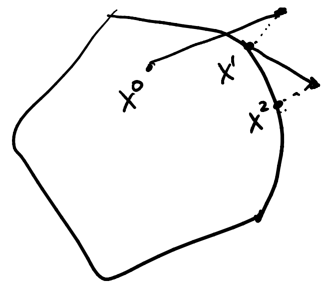
$$\therefore \|z-y\|^2 = \|z - \Pi(y) + \Pi(y) - y\|^2$$

$$= \|z - \Pi(y)\|^2 + \|\Pi(y) - y\|^2 + 2(z - \Pi(y))^T(\Pi(y) - y)$$

$$\geq \|z - \Pi(y)\|^2 + \|\Pi(y) - y\|^2.$$

Algo. Subgradient Method (aka Projected GD).

$$y^{k+1} = x^k - h\, g(x^k)$$
$$x^{k+1} = \Pi(y^{k+1}).$$



Thm. f is G-Lipschitz
$R^2 = \|x^0 - x^*\|_2^2$

Then $$f(x^T) - f(x^*) \le \frac{R}{2hT} + \frac{hG^2}{2}$$

setting $h^2 = \frac{R}{TG^2}$ $\quad \le \frac{RG}{\sqrt{T}}$

to get error ε $\quad T = \frac{R^2 G^2}{\varepsilon^2}$ suffices.

Note: for any fixed h, $f(x^T) \not\to f(x^*)$.

Pf.
$$\|x^{k+1} - x^*\|^2 = \|\Pi(y^{k+1}) - x^*\|^2$$
$$\le \|y^{k+1} - x^*\|^2 - \|\Pi(y^{k+1}) - y^{k+1}\|^2$$
$$\le \|x^k - hg^k - x^*\|^2$$
$$= \|x^k - x^*\|^2 + h^2\|g^k\|^2 - 2h\langle g^k, x^k - x^*\rangle$$

$$f(x^*) - f(x^k) \ge \langle g^k, x^* - x^k\rangle$$
$$-h\langle g^k, x^k - x^*\rangle \le -h\,(f(x^k) - f(x^*))$$

$$\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2 + h^2 G^2 - 2h\,(f(x^k) - f(x^*))$$

$\therefore\ f(x^k) - f(x^*) \le \ldots(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2) + h\,G^2$

$$\therefore \quad f(x^k) - f(x^*) \le \frac{1}{2h}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) + \frac{h}{2} G^2$$

adding up for k from 0 to k

$$\frac{1}{T}\sum_{i=0}^{T-1}\left(f(x^i) - f(x^*)\right) \le \frac{1}{2hT}\|x^0 - x^*\|^2 + \frac{h G^2}{2}$$

$$f\left(\frac{1}{T}\sum_0^{T-1} x^i\right) - f(x^*) \le \underline{\qquad\qquad} \,.$$

This leaves much to be desired.

e.g. suppose $f(x) = \|x\|_1$

then $\nabla f(x) = \text{sign}(x)$ if $\forall i \; x_i \neq 0$

$\|\nabla f(x)\| = \sqrt{n}$

e.g. if $K = \{ x : \sum_{i \in \mathbb{N}} |x_i| < \infty \}$ convex infinite-dim.

$\nabla f(x)$ unbounded.

So $x \leftarrow x - h \nabla f(x)$

is nonsense.

why? we are using same norm $\|\ \|$ for x and $\nabla f(x)$

But they live in different spaces.

$x \in D \qquad\qquad y = \nabla f(x) \in D^*$

$\hat{y} \leftarrow y - \nabla f(x)$

$= 0$

$\hat{x} : \nabla f(\hat{x}) = \hat{y} = 0$

that's what we want!

---

that's what we want.

Going to dual norm might be difficult.

So use an "intermediate map".

$\phi(x)$

y^{k+1}: $\nabla\phi(y^{k+1}) = \nabla\phi(x^{k+1}) - h.\nabla f(x^{k+1})$

may not be in K.

$x^{k+1} = \underset{x \in K}{\operatorname{argmin}}\ d(y^{k+1}, x)$

what distance?

<u>Bregman Divergence</u> (ϕ strictly convex)

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

E.g. 1. $\phi(x) = \|x\|^2$

$$D_\phi(y,x) = \|y\|^2 - \|x\|^2 - 2\langle x, y-x \rangle$$
$$= \|y-x\|^2.$$

2. $\phi(x) = \sum x_i \log x_i$

$$D_\phi(y,x) = \sum y_i \log y_i - \sum x_i \log x_i - \sum_i (1+\log x_i)(y_i - x_i)$$
$$= \sum_i y_i \log \frac{y_i}{x_i} - \sum_i (y_i - x_i)$$

<u>Mirror Descent</u>

$$y^{k+1}: \quad \nabla\phi(y^{k+1}) = \nabla\phi(x^k) - h g^k$$
$$g^k \in \nabla f(x^k)$$

$$x^{k+1} = \operatorname{argmin} D_\phi(x, y^{k+1})$$

$$x^{k+1} = \underset{x \in K}{\operatorname{argmin}} \; D_\phi(x, y^{k+1})$$

(call this $\Pi_\phi(y^{k+1})$).

<u>Lema</u>. $\forall z \in K$

$$D_\phi(z, \Pi_y) + D_\phi(\Pi_y, y) \le D_\phi(z, y)$$

<u>Pf</u>. $h(t) = D_\phi(\Pi_y + t(z - \Pi_y), y)$ is minized at $t = 0$

$$h'(0) = \frac{d}{dt}\Big|_{t=0} D_\phi(\Pi_y + t(z - \Pi_y), y)$$

$$= \frac{d}{dt}\Big|_{t=0} \Big(\phi(\Pi_y + t(z - \Pi_y)) - \phi(y) - \langle \nabla\phi(y), \Pi(y) - y + t(z - \Pi(y))\rangle$$

$$= \langle \nabla\phi(\Pi_y) - \nabla\phi(y), z - \Pi(y)\rangle \ge 0.$$

$$\therefore\ D_\phi(z, y) = D_\phi(z, \Pi(y)) + D_\phi(\Pi_y, y) + \langle \nabla\phi(\Pi(y)) - \nabla\phi(y), z - \Pi(y)\rangle$$

$\nabla\phi(z) - \nabla\phi(y)$ $\quad$ $\nabla\phi(z) - \nabla\phi(\Pi(y))$

$-\langle \nabla\phi(y), z - y\rangle$ $\quad$ $- \langle \nabla\phi(\Pi(y)), z - \Pi(y)\rangle$

$$\ge D_\phi(z, \Pi(y)) + D_\phi(\Pi_y, y).$$

<u>Thm</u>. f is G-Lipschitz convex function in norm $\|\cdot\|$

$\phi: D \to \mathbb{R}$ is μ-strongly convex in same norm $\|\cdot\|$.

$R^2 = \max \|\phi(x^0) - \phi(x)\|^2$.

$$R^2 = \max_{x \in D} \|\phi(x^0) - \phi(x)\|^2 .$$

$$f(x^T) - f(x^*) \leq \frac{R^2}{2hT} + \frac{hG^2}{2}$$

Pf. $$D_\phi(x^*, x^{k+1}) \leq D_\phi(x^*, y^{k+1}) - D_\phi(x^{k+1}, y^{k+1})$$

$$= D_\phi(x^*, x^k) + \underbrace{\langle \nabla\phi(x^k) - \nabla\phi(y^{k+1}), x^* - x^k \rangle} + D_\phi(x^k, y^{k+1}) - D_\phi(x^{k+1}, y^{k+1})$$

Recall $\nabla\phi(y^{k+1}) = \nabla\phi(x^k) - h g^k$

$$= D_\phi(x^*, x^k) + h \langle g^k, x^* - x^k \rangle + D_\phi(x^k, y^{k+1}) - D_\phi(x^{k+1}, y^{k+1})$$

$$f(x^*) - f(x^k) \geq \langle g^k, x^* - x^k \rangle$$

$$D_\phi(x^*, x^{k+1}) - D_\phi(x^*, x^k) \leq -h(f(x^k) - f(x^*)) + \underbrace{D_\phi(x^k, y^{k+1}) - D_\phi(x^{k+1}, y^{k+1})}$$

$$= \phi(x^k) - \phi(x^{k+1}) - \langle \nabla\phi(y^{k+1}), x^k - x^{k+1} \rangle$$

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

$$= h \langle g^k, x^k - x^{k+1} \rangle - \frac{\rho}{2} \|x^k - x^{k+1}\|^2$$

$$\leq h \cdot G \cdot \|x^k - x^{k+1}\| - \frac{\rho}{2} \|x^k - x^{k+1}\|^2$$

$$\leq \frac{h^2 G^2}{2\rho} .$$

$\therefore$

$$f(x^k) - f(x^*) \leq \frac{1}{h} (D_\phi(x^*, x^k) - D_\phi(x^*, x^{k+1})) + \frac{hG^2}{2\rho}$$

adding up

$$\leq \frac{1}{\ } (D_\phi(x^*, x^0) - D_\phi(x^*, x^{T+1})) + hG^2$$

cleaning up

$$\frac{1}{T}\sum_{i=0}^{T-1}\left(f(x^i)-f(x^*)\right) \le \frac{1}{hT}\left(D_\phi(x^*,x^0)-D_\phi(x^*,x^{T+1})\right)+\frac{hG^2}{2\rho}$$

$$\le \frac{D_\phi(x^*,x^0)}{hT}+\frac{hG^2}{2\rho}$$

<u>Example</u>. $D=\{x: \sum x_i=1,\ x_i\geq 0\}$

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

$$D_\phi(y,x)=\sum y_i\log\frac{y_i}{x_i}$$

$$x^{k+1}=\operatorname{argmin} h\langle g^k,x\rangle+D_\phi(x,x^k)$$

$$=\operatorname*{argmin}_{\sum x_i=1,\ x_i\geq 0} h\langle g^k,x\rangle+\sum_i x_i\log\frac{x_i}{x^k}.$$

$$\frac{\partial}{\partial x_i}=0 \qquad h\cdot g_i^k+\log\frac{x_i^{k+1}}{x_i^k}+1=\lambda$$

$$\boxed{x_i^{k+1}=x_i^k e^{-hg_i}\cdot C.}$$

Start with $x^0=\frac{1}{n}\cdot\mathbb{1}$.

then $D_\phi(x,x^0)\leq\log n.=R^2$

Then $D_\phi(x, x^*) = \cdots$

ϕ is 1-strongly convex w.r.t. $\|\cdot\|_1$: $\quad \nabla^2 \phi(z)_i = \dfrac{1}{z_i}$

$$\phi(y) = \phi(x) + \langle \nabla\phi(x), y-x \rangle + \frac{1}{2}(y-x)^T \nabla^2\phi(z)(y-x)$$

$\sum z_i = 1$ So, $\displaystyle \frac{1}{2}\sum \frac{(y_i - x_i)^2}{z_i} \geq \frac{1}{2}\sum |y_i - x_i|$

$\Rightarrow \rho = 1.$

$$\therefore f\left(\frac{1}{T}\sum_{k=0}^{T-1} x^k\right) - f(x^*) \leq RG\sqrt{\frac{2}{\rho T}} \leq \sqrt{\frac{2\log n}{T}}.$$

for f 1-Lipschitz.

(Much better than $\sqrt{\frac{n}{T}}$!)

$$x^{k+1} = \operatorname*{argmin}_{x \in D} D_\phi(x, y^{k+1})$$

$$= \operatorname*{argmin}_{x \in D} \phi(x) - \phi(y^{k+1}) - \langle \nabla\phi(y^{k+1}), x - y^{k+1} \rangle$$

$$= \operatorname*{argmin}_{x \in D} \phi(x) - \langle \nabla\phi(y^{k+1}), x \rangle$$

$$= \operatorname{argmin} \phi(x) - \langle \nabla\phi(x^k) - h g^k, x \rangle$$

$$= \underset{x \in D}{\operatorname{argmin}} \ \phi(x) - \langle \nabla \phi(x^k) - h g^k, x \rangle$$

$$= \underset{x \in D}{\operatorname{argmin}} \ h \langle g^k, x \rangle + \phi(x) - \langle \nabla \phi(x^k), x \rangle$$

$$= \underset{x \in D}{\operatorname{argmin}} \ h \cdot \langle g^k, x \rangle + D_\phi(x, x^k)$$

in analogy with

$$x^{k+1} = \underset{x}{\operatorname{argmin}} \ h \langle g^k, x \rangle + \|x - x^k\|^2$$