Mirror Descent Part 2

Yanjie Ze. July 2021

Alternative forms of mirror descent

original form: $x^{t+1} = \arg\min_{x \in C} \{ \langle g^t, x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \}$

Using the Bregman divergence, we can also describe MD as:

$$D\varphi(y^{t+1}) = D\varphi(x^t) - \eta_t g^t, \text{ with } g^t \in \partial f(x^t) \quad (5.3a)$$

$$x^{t+1} \in P_{C,\varphi}(y^{t+1}) = \arg\min_{z \in C} D_\varphi(z, y^{t+1}) \quad (5.3b)$$

(projection)

This performs gradient descent in certain "dual" space.

上面的式子可以用最优情况来观察一下：

$$D_\varphi(z, y^{t+1}) = \varphi(z) - \varphi(y^{t+1}) - \langle D\varphi(y^{t+1}), z - y^{t+1} \rangle$$

- (5.3b) 的最优情况是：

$$0 \in D\varphi(x^{t+1}) - D\varphi(y^{t+1}) + N_C(x^{t+1}) \Leftarrow$$

最优时梯度为0　用上面定义替换↑　normal cone

这是因为 $\partial \varphi_C(x) = N_C(x)$ (subgradient)

$$\overset{(5.3a)}{=} \varphi(x^{t+1}) - D\varphi(x^t) + \eta_t g^t + N_C(x^{t+1})$$

- (5.1)'s optimality condition gives:

这里再一下 5.1 的式子

5.1

$$x^{t+1} = \arg\min_{x \in C} \{ \langle g^t, x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \}$$

with $g^t \in \partial f(x^t)$

$$0 \in g^t + \frac{1}{\eta_t}\{\nabla\varphi(x^{t+1}) - \nabla\varphi(x^t)\} + N_C(x^{t+1})$$

- Clearly, these two conditions are identical

这两个只差一个参数 η_t，因此这种形式与原形式是等价的

为了简化我们假设凸集 $C = R^n$

于是另一种形式是

$$x^{t+1} = \nabla\varphi^*(\nabla\varphi(x^t) - \eta g^t)$$

where $\varphi(x^*) := \sup_z\{\langle z, x\rangle - \varphi(z)\}$ is the Fenchel-conjugate of φ

When $C = R^n$, (5.3a)-(5.3b) simplifies to

$$x^{t+1} = y^{t+1} = (\nabla\varphi)^{-1}(\nabla\varphi(x^t) - \eta g^t)$$

逆运算

因此，可以得到这样的结论

$$(\nabla\varphi)^{-1} = (\nabla\varphi)^* \qquad (5.5)$$

Proof of Claim (5.5):

Suppose $y = \nabla\varphi(x)$

From the conjugate subgradient theorem, this is equivalent to:

$$\varphi(x) + \varphi^*(y) = \langle x, y\rangle$$

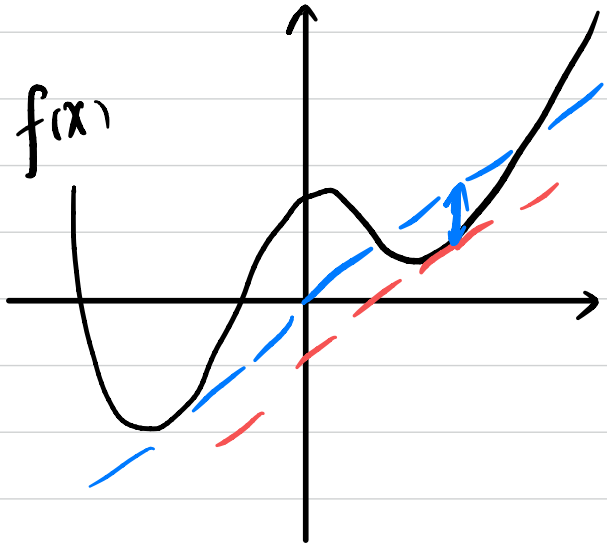
Conjugate subgradient theorem

Define (conjugate function):

For $f: R^n \to R$, $f^*: R^n \to R$ is the conjugate function of f

iff $f^*(y) = \sup_{x \in dom f} (y^T x - f(x))$

Intuition:



上确界在 $Df(x) = y$ 处取到

Theorem (Conjugate Sub-gradient Theorem)

If f is closed and convex.

Then for a pair of vectors x, y, the followings are equivalent:

1. $x^T y = f(x) + f^*(y)$
2. $y \in \partial f(x)$
3. $x \in \partial f^*(y)$

We continue the proof:

$$(\nabla \varphi)^{-1} = (\nabla \varphi)^* \qquad (5.5)$$

Proof of Claim (5.5):

Suppose $y = \nabla \varphi(x)$

From the conjugate subgradient theorem, this is equivalent to:

$$\varphi(x) + \varphi^*(y) = \langle x, y \rangle$$

Since $\varphi^{**} = \varphi$, we further have

$$\varphi^*(y) + \varphi^{**}(x) = \langle x, y \rangle$$

which combined with the conjugate subgradient theorem, yields

$$x = \nabla \varphi^*(y)$$

Since $y = \nabla \varphi(x)$, we plug into:

$$x = \nabla \varphi^*(\nabla \varphi(x))$$

Hence,

$$(\nabla \varphi^*)^{-1}(x) = \nabla \varphi(x)$$

Thus, $\nabla \varphi^* = \nabla \varphi^{-1}$

Convergence analysis

Convex and Lipschitz problems

$$\begin{aligned} &\underset{x}{\text{minimize}} && f(x) \\ &\text{subject to} && x \in C \end{aligned}$$

给出一个定理：

Theorem 5.3

Suppose f is convex and Lipschitz continuous (in the sense that $\|g\|_* \le L_f$ for any subgradient g of f) on C. Suppose φ is ρ-strongly convex w.r.t. $\|\cdot\|$. Then

$$f^{\text{best},t} - f^{\text{opt}} \le \frac{\sup_{x \in C} D_\varphi(x, x^0) + \frac{L_f^2}{2\rho} \sum_{k=0}^{t} \eta_k^2}{\sum_{k=0}^{t} \eta_k}$$

$\|z\|_* = \sup\{z^T x \mid \|x\| \le 1\}$

我们可以带入一个值看看

If $\eta_t = \frac{\sqrt{2\rho R}}{L_f} \frac{1}{\sqrt{t}}$ with $R := \sup_{x \in C} D_\varphi(x, x^0)$, then

$$f^{\text{best},t} - f^{\text{opt}} \le \frac{R + \frac{L_f^2}{2\rho} \sum_{k=0}^{t} \frac{2\rho R}{L_f^2 t}}{\sum_{k=0}^{t} \frac{\sqrt{2\rho R}}{L_f} \frac{1}{\sqrt{t}}}$$

$$= \frac{R + \sum_{k=0}^{t} \frac{R}{t}}{\sum_{k=0}^{t} \frac{1}{\sqrt{t}}} \cdot \frac{L_f}{\sqrt{2\rho R}} = O\left(\frac{L_f \sqrt{R}}{\sqrt{\rho}} \frac{\log t}{\sqrt{t}}\right)$$

Example: optimization over probability simplex

Suppose $C = \Delta$ is the probability simplex, and pick $x^0 = n^{-1}1$

(1) set $\varphi(x) = \frac{1}{2}\|x\|_2^2$, which is $\underset{\rho=1}{1}$-strong convex w.r.t $\|\cdot\|_2$

$$D_\varphi(x, x^0) = \varphi(x) - \varphi(x_0) - \langle \nabla\varphi(x_0), x - x_0 \rangle$$

$$\sup_{x\in\Delta} D_\varphi(x, x^0) = \sup_{x\in\Delta} \frac{1}{2}\|x - n^{-1}1\|_2^2$$

$$= \sup_{x\in\Delta} \frac{1}{2}\left(\|x\|_2^2 - \frac{1}{n}\right)$$

$\because \|x\|_2^2 = x_1^2 + x_2^2 + \cdots x_n^2 \le 1$

$$\therefore \sup_{x\in\Delta} D_\varphi(x, x^0) \le \frac{1}{2}\left(1 - \frac{1}{n}\right) \le \frac{1}{2}$$

$$\therefore R \le \frac{1}{2}$$

If any subgradient g obeys $\|g\|_2 \le L_{f,2}$

再根据上面代入算出的结论，以及 Theorem 5.3，可得

$$f^{best} - f^{opt} \le O\left(L_{f,2}\frac{\log t}{\sqrt{t}}\right)$$

(2) 再设 $\phi(x) = -\sum_{i=1}^{n} x_i \log x_i$, which is 1- strongly convex w.r.t. $\|\cdot\|_1$, $\rho = 1$

然后 (因为负熵的 D_ϕ 就是KL散度)

$$\sup_{x \in \Delta} D_\phi(x, x^0) = \sup_{x \in \Delta} KL(x \| x^0)$$

$$= \sup_{x \in \Delta} \sum_{i=1}^{n} x_i \log x_i - \sum_{i=1}^{n} x_i \log \frac{1}{n}$$

$$\sup_{x \in \Delta} -\sum_{i=1}^{n} x_i \log \frac{1}{n} = \log n \cdot \frac{1}{n} \cdot n = \log n$$

$$\sup_{x \in \Delta} \sum_{i=1}^{n} x_i \log x_i \leq 0 \quad \text{因为 } \log x_i \leq 0$$

因此:

$$\sup_{x \in \Delta} D_\phi(x, x^0) \leq \log n$$

FW:

If any subgradient g obeys $\|g\|_\infty \leq L_{f,\infty}$

$$f^{best,t} - f^{opt} \leq O\left(L_{f,\infty} \sqrt{\log n} \frac{\log t}{\sqrt{t}} \right)$$

比较上面两个不同 ϱ 的结果：

Euclidean：$O(L_{f,2}\frac{\log t}{\sqrt{t}})$

KL：$O(L_{f,\infty}\sqrt{\log n}\cdot\frac{\log t}{\sqrt{t}})$

因为 $\|g\|_\infty = \max_i |g_i|$

$\|g\|_2 = \sqrt{\sum_i g_i^2}$

所以 $\|g\|_\infty \le \|g\|_2 \le \sqrt{n}\,\|g\|_\infty$

所以 $\frac{1}{\sqrt{n}} \le \frac{L_{f,\infty}}{L_{f,2}} \le 1$

所以 KL yields much better performance

收敛更快

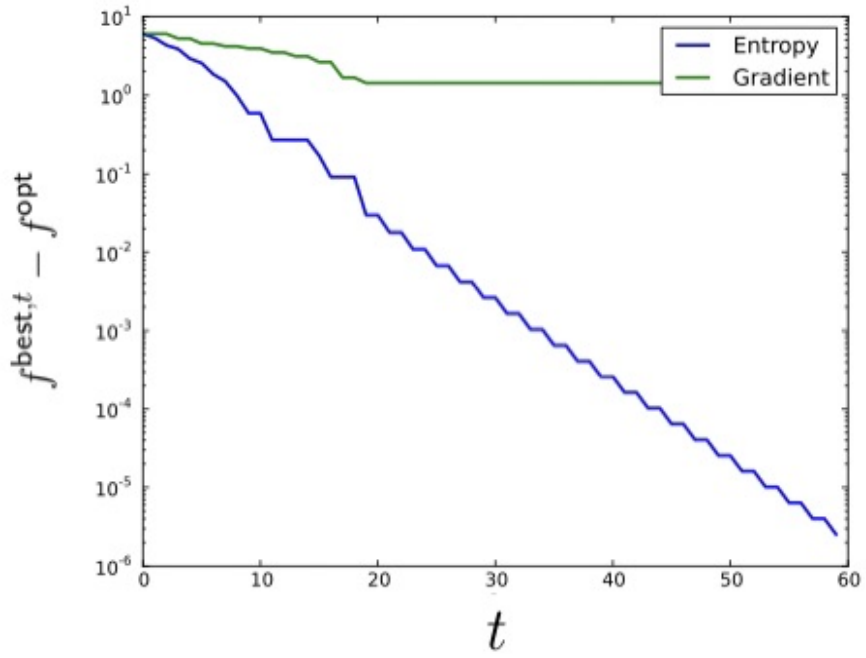
Numerical example: robust regression

$$\underset{x}{\text{minimize}} \quad f(x) = \sum_{i=1}^{m} |a_i^T x - b_i|$$

$$\text{subject to} \quad x \in \Delta = \{x \in \mathbb{R}^n_+ \mid 1^T x = 1\}$$

with $a_i \sim N(0, I_{n \times n})$, $b_i = \frac{a_{i,1} + a_{i,2}}{2} + N(0, 10^{-2})$

$m = 20$, $n = 3000$



Fundamental inequality for mirror descent

基本不等式

Lemma 5.4

$$\eta_t \left(f(\boldsymbol{x}^t) - f^{\mathsf{opt}} \right) \leq D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}) + \frac{\eta_t^2 L_f^2}{2\rho}$$

- $D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1})$ motivates us to form a telescopic sum later

Proof of Lemma 5.4:

$$f(x^t) - f(x^*) \leq \langle g^t, x^t - x^* \rangle \quad \text{(subgradient 的性质)}$$

$$= \frac{1}{\eta_t} \langle \nabla\varphi(x^t) - \nabla\varphi(y^{t+1}), x^t - x^* \rangle \quad \text{(MD 的更新法则)}$$

$$= \frac{1}{\eta_t} \{ D_\varphi(x^*, x^t) + D_\varphi(x^t, y^{t+1}) - D_\varphi(x^*, y^{t+1}) \}$$

(three point lemma)

For any three points x, y, z,

$$D_\varphi(x, z) = D_\varphi(x, y) + D_\varphi(y, z) - \langle \nabla\varphi(z) - \nabla\varphi(y), x - y \rangle$$

$$\leq \frac{1}{\eta_t} \{ D_\varphi(x^*, x^t) + D_\varphi(x^t, y^{t+1}) - D_\varphi(x^*, x^{t+1}) - D_\varphi(x^{t+1}, y^{t+1}) \}$$

(Pythagorean)

If $x_{C,\varphi} = P_{C,\varphi}(x)$, then

$$D_\varphi(z, x) \geq D_\varphi(z, x_{C,\varphi}) + D_\varphi(x_{C,\varphi}, x), \quad \forall z \in C$$

$$= \frac{1}{\eta_t}\{D_\varphi(x^*, x^t) - D_\varphi(x^*, x^{t+1})\}$$

$$+ \frac{1}{\eta_t}\{D_\varphi(x^t, y^{t+1}) - D_\varphi(x^{t+1}, y^{t+1})\}$$

所以根据我们要证的，我们 bound 和二项系数即证：

$$D_\varphi(x^t, y^{t+1}) - D_\varphi(x^{t+1}, y^{t+1}) \le \frac{(\eta_t L_f)^2}{2\rho}$$

证明：

$$\begin{aligned}
&D_\varphi(x^t, y^{t+1}) - D_\varphi(x^{t+1}, y^{t+1}) \\
&= \varphi(x^t) - \varphi(x^{t+1}) - \langle \nabla\varphi(y^{t+1}), x^t - x^{t+1}\rangle \quad \text{(定义)} \\
&\le \langle \nabla\varphi(x^t), x^t - x^{t+1}\rangle - \frac{\rho}{2}\|x^t - x^{t+1}\|^2 - \langle \nabla\varphi(y^{t+1}), x^t - x^{t+1}\rangle \quad \Leftarrow \rho\text{-strong convexity} \\
&\quad \text{(strong convexity of } \varphi) \\
&= \langle \nabla\varphi(x^t) - \nabla\varphi(y^{t+1}), x^t - x^{t+1}\rangle - \frac{\rho}{2}\|x^t - x^{t+1}\|_2^2 \\
&= \eta_t\langle g^t, x^t - x^{t+1}\rangle - \frac{\rho}{2}\|x^t - x^{t+1}\|^2 \quad \text{(MD update rule)} \\
&\le \eta_t L_f\|x^t - x^{t+1}\| - \frac{\rho}{2}\|x^t - x^{t+1}\|^2 \quad \text{(Cauchy-Schwarz)} \\
&\le \frac{(\eta_t L_f)^2}{2\rho} \quad \text{(optimize quadratic function in } \|x^t - x^{t+1}\|)
\end{aligned}$$

$\varphi(x^{t+1}) \ge \varphi(x^t) + \langle \nabla\varphi(x^t), x^{t+1} - x^t\rangle + \frac{\rho}{2}\|x^{t+1} - x^t\|^2$

$\langle g^t, x^t - x^{t+1}\rangle \le \|g^t\|\,\|x^t - x^{t+1}\|$

二次函数最大值在 $-\frac{b}{2a} = \frac{\eta_t L_f}{\rho}$ 处

MD update rule

$$\nabla\varphi(y^{t+1}) = \nabla\varphi(x^t) - \eta_t g^t, \text{ with } g^t \in \partial f(x^t) \quad (5.3a)$$

$$x^{t+1} \in P^{\varphi}_{C}(y^{t+1}) = \underset{z\in C}{\arg\min}\, D_\varphi(z, y^{t+1}) \quad (5.3b)$$

projection

Theorem 5.3

Suppose f is convex and Lipschitz continuous (in the sense that $\|\boldsymbol{g}\|_* \le L_f$ for any subgradient $\boldsymbol{g}$ of f) on $\mathcal{C}$. Suppose φ is ρ-strongly convex w.r.t. $\|\cdot\|$. Then

$$f^{\text{best},t} - f^{\text{opt}} \le \frac{\sup_{\boldsymbol{x}\in\mathcal{C}} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0) + \frac{L_f^2}{2\rho}\sum_{k=0}^t \eta_k^2}{\sum_{k=0}^t \eta_k}$$

有了 Lemma 5.4后，可以证明刚才的 Theorem 5.3

From Lemma 5.4, one has

(Lemma 5.4)

$$\eta_k \left(f(\boldsymbol{x}^k) - f^{\text{opt}}\right) \le D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^k) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{k+1}) + \frac{\eta_k^2 L_f^2}{2\rho}$$

Taking this inequality for $k = 0, \cdots, t$ and summing them up give

对从k=0到t求和

$$\begin{aligned}\sum_{k=0}^t \eta_k \left(f(\boldsymbol{x}^k) - f^{\text{opt}}\right) &\le D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^0) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}) + \frac{L_f^2 \sum_{k=0}^t \eta_k^2}{2\rho} \\ &\le \sup_{\boldsymbol{x}\in\mathcal{C}} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0) + \frac{L_f^2 \sum_{k=0}^t \eta_k^2}{2\rho}\end{aligned}$$

3-point Lemma: $D_\varphi(x^*, x^0) \le D(x^*, x^{t+1}) + D(x^{t+1}, x^0)$

This together with $f^{\text{best},t} - f^{\text{opt}} \le \frac{\sum_{k=0}^t \eta_k \left(f(\boldsymbol{x}^k) - f^{\text{opt}}\right)}{\sum_{k=0}^t \eta_k}$ concludes the proof

$$\le \frac{\sup_{x\in\mathcal{C}} D_\varphi(x, x^0) + \frac{L_f^2 \sum_{k=0}^t \eta_k^2}{2\rho}}{\sum_{k=0}^t \eta_k}$$