Introductory Lectures on Optimization

Beyond The Black-box Model (3)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

December 16, 2024

Outline

Part I

Smoothing Techniques

This section mainly refers to the lecture notes of IE598 by Niao He.

Introduction

Consider the following problem:

$$\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}),$$

where f is convex but nonsmooth, and $\mathcal{X}$ is a convex and compact set. One intuitive way to approach the above problem is to approximate the nonsmoothing function $f(\boldsymbol{x})$ by a smooth and convex function $f_u(\boldsymbol{x})$, so that we can use the standard techniques learnt so far in the course to solve the problem. Hence, we want to reduct the problem into the following:

$$\min_{\boldsymbol{x} \in \mathcal{X}} f_u(\boldsymbol{x}).$$

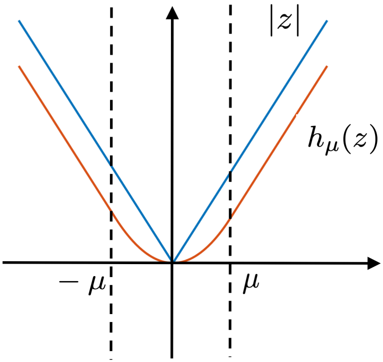
where f_u is a L_u-Lipschitz continuous, smooth and convex approximation of the function $f(\boldsymbol{x})$.

Motivation Example

Consider the simplest non-smooth and convex function, $f(\boldsymbol{x}) = |\boldsymbol{x}|$. The following function, known as the Huber function,

$$f_u(\boldsymbol{x}) = \begin{cases} \frac{\boldsymbol{x}^2}{2u}, & |\boldsymbol{x}| \leq u, \\ |\boldsymbol{x}| - \frac{u}{2}, & |\boldsymbol{x}| > u, \end{cases}$$

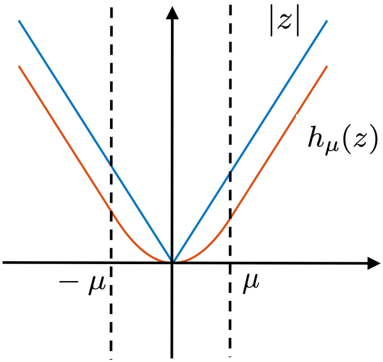
is a smooth approximation of the absolute value function. The Hubert function approximation has been widely used in machine learning to approximate non-smooth loss functions, e.g. absolute loss (robust regression), hinge loss (SVM), etc.



Motivation Example

$$f_u(\boldsymbol{x}) = \begin{cases} \frac{\boldsymbol{x}^2}{2u}, & |\boldsymbol{x}| \le u, \\ |\boldsymbol{x}| - \frac{u}{2}, & |\boldsymbol{x}| > u, \end{cases}$$

1. $f_u(\boldsymbol{x})$ is clearly continuous and differentiable everywhere.
2. $f(\boldsymbol{x}) - \frac{u}{2} \le f_u(\boldsymbol{x}) \le f(\boldsymbol{x})$.
3. If $u \to 0$, then $f_u(\boldsymbol{x}) \to f(\boldsymbol{x})$.
4. $|f''_u(\boldsymbol{x})| \le \frac{1}{u}$. This implies that $f_u(\boldsymbol{x})$ is $\frac{1}{u}$-Lipschitz continuous.



Motivation Example

Robust Regression. Suppose we have m data samples $(a_1, b_1), \ldots, (a_m, b_m)$. We intend to solve the following regression problem with absolute loss:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{i=1}^{m} |a_i^\top \boldsymbol{x} - b_i|.$$

We can approximate the absolute loss in the above optimization problem with the Huber loss and solve instead the following smooth convex optimization problem.

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \sum_{i=1}^{m} f_u(a_i^\top \boldsymbol{x} - b_i).$$

Major Techniques

1. Nesterov's Smoothing technique [Nesterov, 2005]: Nesterov's smoothing technique uses the following function to approximate $f(\boldsymbol{x})$:

$$f_u(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathrm{dom} f^*} \{\boldsymbol{x}^\top y - f^*(\boldsymbol{y}) - ud(\boldsymbol{y})\}$$

where f^* is the convex conjugate of f defined as the following:

$$f^*(\boldsymbol{y}) = \max_{\boldsymbol{x} \in \mathrm{dom}\ f} \{\boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{x})\}.$$

and $d(\boldsymbol{y})$ is some proximity function that is strongly convex and nonnegative everywhere.

Major Techniques

2. Moreau-Yosida smoothing/regularization: Moreau-Yosida's smoothing technique uses the following function to approximate $f(\boldsymbol{x})$:

$$f_u(\boldsymbol{x}) = \min_{\boldsymbol{y} \in \text{dom } f} \{f(\boldsymbol{y}) + \frac{1}{2u} \|\boldsymbol{x} - \boldsymbol{y}\|_M^2\}$$

where $u > 0$ the approximation parameter, and the M-norm is defined as

$$\|\boldsymbol{x}\|_M^2 = \boldsymbol{x}^\top M \boldsymbol{x}.$$

This is also known as the Moreau envelope of f.

3. Ben-Tal-Teboulle smoothing based on recession function [Ben-Tal and Teboulle, 1989].
4. Randomized smoothing [Duchi et al., 2012].

Nesterov's Smoothing

We consider a more generalized problem setting as compared to the previous sections. The goal is to solve the nonsmooth convex optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \Leftrightarrow f_u(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\operatorname{dom} f^*} \{\boldsymbol{x}^\top y - f^*(\boldsymbol{y}) - ud(\boldsymbol{y})\}.$$

Assume that function f can be represented by

$$f(\boldsymbol{x}) = g(A\boldsymbol{x}+b) \triangleq \max_{\boldsymbol{y}\in\mathcal{Y}}\{\langle A\boldsymbol{x}+b,\ \boldsymbol{y}\rangle - \phi(\boldsymbol{y})\}$$

where $\phi(\boldsymbol{y})$ is a convex and continuous function and $\mathcal{Y}$ is a convex and compact set.

Remark. For many cases, we are able to construct such representation easily as compared to using the convex conjugate.

Example

Example 16

Let $f(\boldsymbol{x}) = \max_{1 \le i \le m} |a_i^\top \boldsymbol{x} - b_i|$. Computing the convex conjugate for f is a cumbersome task and f^* turns out to be very complex. But we can easily represent f as follows:

$$f(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathbb{R}^m} \left\{ (A\boldsymbol{x} - b)^\top \boldsymbol{y} \,\middle|\, \sum_i |\boldsymbol{y}_i| \le 1 \right\}$$

Remark. Let $\|\cdot\|$ be a norm in $\mathbb{R}$, The associated dual norm is defined as

$$\|\boldsymbol{z}\|_* = \sup_{\boldsymbol{x}} \{ \boldsymbol{z}^\top \boldsymbol{x} \mid \|\boldsymbol{x}\| \le 1 \}.$$

Proximity Function

Proximity Function: The function $d(\boldsymbol{y})$ should satisfy the following properties:

1. $d(\boldsymbol{y})$ is continuous and 1-strongly convex on $\mathcal{Y}$;
2. $d(\boldsymbol{y}_0) = 0$, for $\boldsymbol{y}_0 \in \text{Argmin}_{\boldsymbol{y} \in \mathcal{Y}} d(\boldsymbol{y})$;
3. $d(\boldsymbol{y}_0) \geq 0, \forall \boldsymbol{y} \in \mathcal{Y}$.

Let $b \in \mathcal{Y}$, here are some examples of valid proximity functions:

1. $d(\boldsymbol{y}) = \frac{1}{2} \|\boldsymbol{y} - b\|_2^2$;
2. $d(\boldsymbol{y}) = \frac{1}{2} \sum w_i (\boldsymbol{y}_i - b_i)^2$ with $w_i \geq 1$;
3. $d(\boldsymbol{y}) = w(\boldsymbol{y}) - w(b) - \nabla w(b)^\top (\boldsymbol{y} - b)$ with $w(\boldsymbol{x})$ being 1-strongly convex on $\mathcal{Y}$.

Nesterov's smoothing

Consider the following smooth approximation of f

$$f_u(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathcal{Y}}\{\langle A\boldsymbol{x}+b,\ \boldsymbol{y}\rangle - \phi(\boldsymbol{y}) - ud(\boldsymbol{y})\}.$$

Propsition 17

1. $f_u(\boldsymbol{x})$ is continuously differentiable.
2. $\nabla f_u(\boldsymbol{x}) = A^\top y(\boldsymbol{x})$, where $y(\boldsymbol{x}) = \operatorname{argmax}_{\boldsymbol{z}\in\mathcal{Y}}\{\langle A\boldsymbol{x}+b,\ \boldsymbol{z}\rangle - \phi(\boldsymbol{z}) - ud(\boldsymbol{z})\}$.
3. $f_u(\boldsymbol{x})$ is M-Lipschitz smooth, where $M = \frac{\|A\|_2^2}{u}$ $(\|A\|_2 = \max_{\boldsymbol{x}} \|A\boldsymbol{x}\|_2^2 \mid \|\boldsymbol{x}\|_2 < 1)$.

See Theorem 1 of [Nesterov, 2005] for proofs.

Nesterov's smoothing

Theorem 18 (Approximation Accuracy)

For any $u > 0$, let $D_{\mathcal{Y}}^2 = \max_{\boldsymbol{y}\in\mathcal{Y}} d(\boldsymbol{y})$, we have

$$f(\boldsymbol{x}) - uD_{\mathcal{Y}}^2 \leq f_u(\boldsymbol{x}) \leq f(\boldsymbol{x}).$$

Proof. The result can be derived directly from

$$f_u(\boldsymbol{x}) \leq f_0(\boldsymbol{x}) = f(\boldsymbol{x}),$$

and

$$f(\boldsymbol{x}) - uD_{\mathcal{Y}}^2 \leq f_u(\boldsymbol{x})$$

can be easily obtained. $\square$

Nesterov's smoothing

Remark.

$$
\begin{aligned}
f(\boldsymbol{x}) - uD_{\mathcal{Y}}^2 &= \langle Ax + b,\ \boldsymbol{y}^* \rangle - \phi(\boldsymbol{y}^*) - uD_{\mathcal{Y}}^2 \\
&\le \langle A\boldsymbol{x} + b,\ \boldsymbol{y}^* \rangle - \phi(\boldsymbol{y}^*) - ud(\boldsymbol{y}^*) \\
&\le f_u(\boldsymbol{x}),
\end{aligned}
$$

where

$$
\begin{aligned}
f(\boldsymbol{x}) &= \max_{\boldsymbol{y} \in \mathcal{Y}} \{ \langle A\boldsymbol{x} + b,\ \boldsymbol{y} \rangle - \phi(\boldsymbol{y}) \} \\
&= \langle A\boldsymbol{x} + b,\ \boldsymbol{y}^* \rangle - \phi(\boldsymbol{y}^*).
\end{aligned}
$$

Nesterov's smoothing

Analysis of Nesterov's smoothing:

1. Let $f_* = \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ and $f_{u,*} = \min_{\boldsymbol{x} \in \mathcal{X}} f_u(\boldsymbol{x})$, we have

$$f_{u,*} \leq f_*.$$

Moreover, for any $\boldsymbol{x}_t$ generated by an algorithm,

$$\begin{aligned} & f(\boldsymbol{x}_t) - f_* \leq f(\boldsymbol{x}_t) - f_{u,*}, \\ \Leftrightarrow\ & f(\boldsymbol{x}_t) - f_* \leq \underbrace{f(\boldsymbol{x}_t) - f_u(\boldsymbol{x}_t)}_{\text{approximation error}} + \underbrace{f_u(\boldsymbol{x}_t) - f_{u,*}}_{\text{optimization error}}. \end{aligned}$$

Nesterov's smoothing

Analysis of Nesterov's smoothing:

2. If we apply projected gradient descent to solve the smooth problem, we have

$$f(\boldsymbol{x}_t) - f^* \leq O\left(\frac{\|A\|_2^2 D_{\mathcal{X}}^2}{ut} + uD_{\mathcal{Y}}^2\right).$$

Therefore, if we want the error to be less tan a threshold ϵ, we need to set $u = O\left(\frac{\epsilon}{D_{\mathcal{Y}}^2}\right)$ and the total number of iterations is at most $T_\epsilon = O\left(\frac{\|A\|_2^2 D_{\mathcal{X}}^2}{\epsilon u}\right) = O\left(\frac{\|A\|_2^2 D_{\mathcal{X}}^2 D_{\mathcal{Y}}^2}{\epsilon^2}\right)$.

Nesterov's smoothing

Analysis of Nesterov's smoothing:

3. If we apply accelerated gradient descent to solve the smooth problem, the we have

$$f(\boldsymbol{x}_t) - f^* \leq O\left(\frac{\|A\|_2^2 D_{\mathcal{X}}^2}{ut^2} + uD_{\mathcal{Y}}^2\right).$$

Therefore, if we want the error to be less tan a threshold ϵ, we need to set $u = O\left(\frac{\epsilon}{D_{\mathcal{Y}}^2}\right)$ and the total number of iterations is at most $T_\epsilon = O(\frac{\|A\|_2 D_{\mathcal{X}}}{\sqrt{\epsilon u}}) = O(\frac{\|A\|_2 D_{\mathcal{X}} D_{\mathcal{Y}}}{\epsilon})$.

In the later case the overall complexity $O(1/\epsilon)$ is substantially better than the $O(1/\epsilon^2)$ complexity when we dirctly apply subgradient descent to solve the original nonsmooth convex problem.

Examples

Consider objective $f(\boldsymbol{x}) = |\boldsymbol{x}|$. Note that f admits the following two different representation:

$$f(\boldsymbol{x}) = \sup_{|\boldsymbol{y}|\leq 1} \boldsymbol{y}\boldsymbol{x}$$

or

$$f(\boldsymbol{x}) = \sup_{\substack{\boldsymbol{y}_1,\boldsymbol{y}_2\geq 0\\ \boldsymbol{y}_1+\boldsymbol{y}_2=1}} (\boldsymbol{y}_1 - \boldsymbol{y}_2)x$$

Hence, $\mathcal{Y} = \{\boldsymbol{y} : |\boldsymbol{y}| \leq 1\}$ or $\mathcal{Y} = \{\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) : \boldsymbol{y}_1, \boldsymbol{y}_2 \geq 0, \boldsymbol{y}_1 + \boldsymbol{y}_2 = 1\}$; and function $\phi(\boldsymbol{y}) \triangleq 0$.

Examples

Example 19 ($d(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^2$)

$d(\cdot)$ is 1-strongly convex on $\mathcal{Y} = \{\boldsymbol{y} : |\boldsymbol{y}| \le 1\}$, and $d(\boldsymbol{y}) \ge 0$.

$$f_u(\boldsymbol{x}) = \sup_{|\boldsymbol{y}|\le 1} \left\{ \boldsymbol{y}\boldsymbol{x} - \frac{u}{2}\boldsymbol{y}^2 \right\} = \begin{cases} \frac{\boldsymbol{x}^2}{2u}, & |\boldsymbol{x}| \le u, \\ |\boldsymbol{x}| - \frac{u}{2}, & |\boldsymbol{x}| > u, \end{cases}$$

which is the well-known Huber function.

Examples

Remark.

$$\operatorname*{argmax}_{\boldsymbol{y}\in Y}\left\{-\frac{u}{2}\left(\boldsymbol{y}-\frac{\boldsymbol{x}}{u}\right)^2+\frac{\boldsymbol{x}^2}{2u}\right\}.$$

We have to discuss the constraint $|\boldsymbol{y}| \le 1$:

(1) when $-1 \le \frac{\boldsymbol{x}}{u} \le 1$, we have $\boldsymbol{y}_* = \frac{\boldsymbol{x}}{u}$ and $f_u(\boldsymbol{x}) = \frac{\boldsymbol{x}^2}{2u}$.

(2) when $\frac{\boldsymbol{x}}{u} \ge 1 > 0$, we have $\boldsymbol{y}_* = 1$ and $f_u(\boldsymbol{x}) = \boldsymbol{x} - \frac{u}{2}$.

(3) when $\frac{\boldsymbol{x}}{u} \le -1 < 0$, we have $\boldsymbol{y}_* = -1$ and $f_u(\boldsymbol{x}) = -\boldsymbol{x} - \frac{u}{2}$.

Examples

Example 20 ($d(\boldsymbol{y}) = 1 - \sqrt{1 - \boldsymbol{y}^2}$)

$d(\cdot)$ is 1-strongly convex on $\mathcal{Y} = \{\boldsymbol{y} : |\boldsymbol{y}| \leq 1\}$, and $d(\boldsymbol{y}) \geq 0$.

$$\begin{aligned} f_u(\boldsymbol{x}) &= \sup_{|\boldsymbol{y}| \leq 1} \left\{ \boldsymbol{y}\boldsymbol{x} - u\left(1 - \sqrt{1 - \boldsymbol{y}^2}\right)\right\} \\ &= \sqrt{\boldsymbol{x}^2 + u^2} - u. \end{aligned}$$

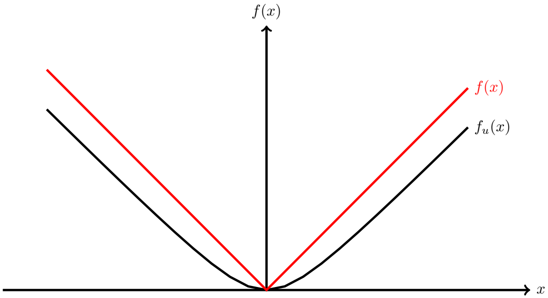
Examples

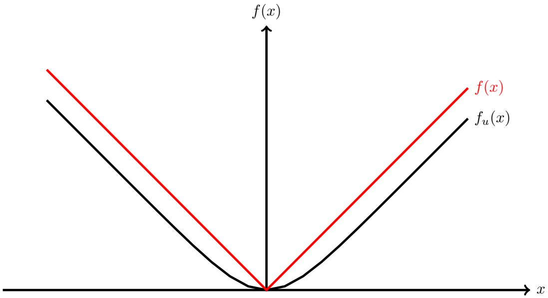
Example 21 ($d(\boldsymbol{y}) = \boldsymbol{y}_1 \log \boldsymbol{y}_1 + \boldsymbol{y}_2 \log \boldsymbol{y}_2 + \log 2$)

$d(\cdot)$ is 1-strongly convex on $\mathcal{Y} = \{\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) : \boldsymbol{y}_1, \boldsymbol{y}_2 \geq 0, \boldsymbol{y}_1 + \boldsymbol{y}_2 = 1\}$, and $d(\boldsymbol{y}) \geq 0$.

$$
\begin{aligned}
f_u(\boldsymbol{x}) &= \sup_{\substack{\boldsymbol{y}_1, \boldsymbol{y}_2 \geq 0 \\ \boldsymbol{y}_1 + \boldsymbol{y}_2 = 1}} \left\{ (\boldsymbol{y}_1 - \boldsymbol{y}_2)\boldsymbol{x} - u\left(\boldsymbol{y}_1 \log \boldsymbol{y}_1 + \boldsymbol{y}_2 \log \boldsymbol{y}_2 + \log 2\right) \right\} \\
&= u \log \left(\frac{e^{-\frac{\boldsymbol{x}}{u}} + e^{\frac{\boldsymbol{x}}{u}}}{2} \right).
\end{aligned}
$$

Example





Moreau-Yosida Regularization

Consider function

$$f(\boldsymbol{x}) = \max_{\boldsymbol{y}}\{\boldsymbol{y}^\top \boldsymbol{x} - f^*(\boldsymbol{y})\}.$$

We can show that

$$\begin{aligned} f_u(\boldsymbol{x}) &= \max_{\boldsymbol{y}} \left\{\boldsymbol{y}^\top \boldsymbol{x} - f^*(\boldsymbol{y}) - \frac{u}{2}\|\boldsymbol{y}\|_2^2\right\} = \left(f^* + \frac{u}{2}\|\cdot\|_2^2\right)^* (\boldsymbol{x}) \\ &= \inf_{\boldsymbol{y}} \left\{ f(\boldsymbol{y}) + \frac{1}{2u}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right\} \end{aligned}$$

Let f and g be two proper, convex and semi-continuous functions, then

1. $(f+g)^*(\boldsymbol{x}) = \inf_{\boldsymbol{y}} \{f^*(\boldsymbol{y}) + g^*(\boldsymbol{x} - \boldsymbol{y})\}$.
2. $(\alpha f)^*(\boldsymbol{x}) = \alpha f^*(\frac{\boldsymbol{x}}{\alpha})$ for $\alpha > 0$.

Moreau-Yosida Regularization

Interpretation of Proximal Point Algorithm: Apply gradient method to minimize Moreau envelop

$$\min \left\{ f_u(\boldsymbol{x}) = \inf_{\boldsymbol{y}} \left(f(\boldsymbol{y}) + \frac{1}{2u} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right) \right\}.$$

This is an exact smooth reformulation of problem of minimizing $f(\boldsymbol{x})$:

1. solution $\boldsymbol{x}$ is minimizer of f.
2. f_u is differentiable with Lipschitz continuous gradient ($L = 1/t$).

Gradient Update: with fixed $t_k = 1/L = u$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - u \nabla f_u(\boldsymbol{x}_k) = \mathrm{prox}_{uf}(\boldsymbol{x}_k).$$

Remark: $\nabla f_u(\boldsymbol{x}_k) = \frac{1}{u}(\boldsymbol{x}_k - \mathrm{prox}_{uf}(\boldsymbol{x}_k))$.

Moreau-Yosida Regularization

Remark.
Since

$$f_u(\boldsymbol{x}) = \max_{\boldsymbol{y}} \left\{ \boldsymbol{y}^\top \boldsymbol{x} - f^*(\boldsymbol{y}) - \frac{u}{2} \|\boldsymbol{y}\|_2^2 \right\},$$

we have

$$\nabla f_u(\boldsymbol{x}) = A \top \boldsymbol{y}(\boldsymbol{x}), \text{ where } A = I \text{ and } \boldsymbol{y} \text{ attained the above } \max_{\boldsymbol{y}}$$

$$\boldsymbol{x} - u\boldsymbol{y} \in \partial f^*(\boldsymbol{y}) \Leftrightarrow \boldsymbol{y} \in \partial f(\boldsymbol{x} - u\boldsymbol{y}) \Leftrightarrow \boldsymbol{y} = \frac{1}{u}(\boldsymbol{x} - \operatorname{prox}_{uf}(\boldsymbol{x})).$$

The last equality follows from the following properey.

$$\boldsymbol{z} = \operatorname{prox}_h(p) \Leftrightarrow p - \boldsymbol{z} \in \partial h(\boldsymbol{z}).$$

$$\text{That is } \boldsymbol{x} - u\boldsymbol{y} = \operatorname{prox}_{uf}(\boldsymbol{x}) \Leftrightarrow \boldsymbol{x} - (\boldsymbol{x} - u\boldsymbol{y}) \in u\partial f(\boldsymbol{x} - u\boldsymbol{y}).$$

Part II
Mirror Descent

Introduction

Generally, each iteration of gradient descent, Newton method, subgradient descent can be regarded as a local optimization, and the objective functions are respectively :

$$x_{k+1} = \operatorname{argmin}\left\{ f(x_k) + \langle \nabla f(x_k),\ x - x_k \rangle + \frac{1}{2h_k}\|x - x_k\|_2^2 \right\},$$

$$x_{k+1} = \operatorname{argmin}\left\{ f(x_k) + \langle \nabla f(x_k),\ x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k),\ (x - x_k) \rangle \right\},$$

$$x_{k+1} = \operatorname{argmin}\left\{ f(x_k) + \langle g,\ x - x_k \rangle + \frac{1}{2h_k}\|x - x_k\|_2^2 \right\}.$$

To generalize the method beyond Euclidean distance, it is straightforward to use Bregman divergence as a measure of displacement.

Bregman Divergence

Definition 22 (Bregman Divergence)

Let $\psi : \Omega \to \mathbb{R}$ be a function that is : a) strictly convex, b) continuously differentiable, c) defined on a closed convex set Ω. Then the Bregman divergence is defined as

$$\triangle_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y),\ x - y\rangle, \forall x, y \in \Omega.$$

That is, the difference between the value of ψ at x and the first order Taylor expansion of ψ around y evaluated at point x.

Examples of Bregman Divergence

1. **Euclidean distance.** Let $\psi(x) = \frac{1}{2}\|x\|_2^2$. Then
$$\triangle_\psi(x, y) = \frac{1}{2}\|x - y\|_2^2 .$$

2. **Kullback-Leibler divergence.** For $\Omega = \{x \in \mathbb{R}_+^n : \sum_i x_i = 1\}$, and $\psi(x) = \sum_i x_i \log x_i$. Then
$$\triangle_\psi(x, y) = \sum_i x_i \log \frac{x_i}{y_i}$$
for $x, y \in \Omega$. This is called relative entropy, or Kullback-Leibler divergence, commonly used between probability distributions x and y.

Examples of Bregman Divergence

1. **Based on ℓ_p norm.** Let $p \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. $\psi(x) = \frac{1}{2}\|x\|_q^2$. Then

$$\triangle_\psi(x, y) = \frac{1}{2}\|x\|_q^2 - \frac{1}{2}\|y\|_q^2 - \langle x, \nabla \frac{1}{2}\|y\|_q^2 \rangle.$$

Note $\frac{1}{2}\|y\|_q^2$ is not necessarily continuously differentiable, which makes this case not precisely consistent with our definition.

Remark. The subgradient is a linear oracle on the dual sphere. (see Frank-Wolfe section)

Properties of Bregman Divergence

1. Strict convexity in the first argument x. Trivial by the strict convexity of ψ.
2. Nonnegativity: $\triangle_\psi(x,y) \geq 0$ for all x and y. $\triangle_\psi(x,y) = 0$ if and only if $x = y$.
3. Asymmetry: in general, $\triangle_\psi(x,y) \neq \triangle_\psi(y,x)$.
4. Linearity in ψ. For any $\alpha > 0$, $\triangle_{\psi+\alpha\varphi}(x,y) = \triangle_\psi(x,y) + \alpha \triangle_\varphi(x,y)$.
5. Gradient in x: $\frac{\partial}{\partial x} \triangle_\psi(x,y) = \nabla\psi(x) - \nabla\psi(y)$.
6. Generalized triangle inequality:

$$\begin{aligned} \triangle_\psi(x,y) + \triangle_\psi(y,z) &= \psi(x) - \psi(y) - \langle \nabla\psi(y),\ x - y\rangle \\ &\quad + \psi(y) - \psi(z) - \langle \nabla\psi(z),\ y - z\rangle \\ &= \triangle_\psi(x,z) + \langle x - y,\ \nabla\psi(z) - \nabla\psi(y)\rangle \end{aligned}$$

Properties of Bregman Divergence

7. Duality. Suppose ψ is strongly convex. Then

$$(\nabla\psi^*)(\nabla\psi(x)) = x, \qquad \triangle_\psi(x, y) = \triangle_{\psi^*}(\nabla\psi(y), \nabla\psi(x)).$$

Proof. (for the first equality only) Recall

$$\psi^*(y) = \sup_{z\in Q} \left\{\langle z,\ y\rangle - \psi(z)\right\}.$$

Here, sup must be attainable because ψ is strongly convex and Q is closed.

Remark. $(\nabla\psi)^{-1} = \nabla\psi^*$.

Properties of Bregman Divergence

7. Duality.

Proof. (Continued.) x is a maximier if and only if $y = \nabla\psi(x)$. So

$$\psi^*(y) = -\psi(x) + \langle x,\ y\rangle \Leftrightarrow y = \nabla\psi(x).$$

Since $\psi = \psi^{**}$, so $\psi^*(y) + \psi^{**}(x) = \langle x,\ y\rangle$, which means y is the maximizer in

$$\psi^{**}(x) = \sup_z \left\{\langle x,\ z\rangle - \psi^*(z)\right\}.$$

This means $x = \nabla\psi^*(y)$. $\square$

Properties of Bregman Divergence

8. Extension of Pythagorean:

Lemma 23 (Extension of Pythagorean)

Suppose L is a proper convex function whose domain is an open set containing C. L is not necessarily differentiable. Let x^* be

$$x^* = \operatorname*{argmin}_{x \in C} \left\{ L(x) + \triangle_\psi(x, x_0) \right\}.$$

Then for any $y \in C$ we have

$$L(y) + \triangle_\psi(y, x_0) \geq L(x^*) + \triangle_\psi(x^*, x_0) + \triangle_\psi(y, x^*).$$

Properties of Bregman Divergence

$$\boxed{L(y) + \triangle_\psi(y, x_0) \geq L(x^*) + \triangle_\psi(x^*, x_0) + \triangle_\psi(y, x^*).}$$

Proof. Denote $J(x) = L(x) + \triangle_\psi(x, x_0)$. Since x^* minimizes J over C, there must exist a subgradient $d \in \partial J(x^*)$ such that

$$\langle d,\ x - x^* \rangle \geq 0, \quad \forall x \in C.$$

Since $\partial J(x^*) = \{g + \nabla_{x=x^*} \triangle_\psi(x, x_0) : g \in \partial L(x^*)\}$, we have $\partial J(x^*) = \{g + \nabla\psi(x^*) - \nabla\psi(x_0) : g \in \partial L(x^*)\}$. So there must be a subgradient $g \in L(x^*)$ such that

$$\begin{aligned} &\langle g + \nabla\psi(x^*) - \nabla\psi(x_0),\ x - x^* \rangle \geq 0, \quad \forall x \in C \\ &\Rightarrow \langle g,\ x - x^* \rangle \geq \langle \nabla\psi(x_0) - \nabla\psi(x^*),\ x - x^* \rangle. \end{aligned} \tag{5}$$

Properties of Bregman Divergence

$$\boxed{L(y) + \triangle_\psi(y, x_0) \geq L(x^*) + \triangle_\psi(x^*, x_0) + \triangle_\psi(y, x^*).}$$

Proof. (continued.) Therefore using the property of subgradient, we have for all $y \in C$ that

$$\begin{aligned} L(y) &\geq L(x^*) + \langle g,\ y - x^* \rangle \\ &\geq L(x^*) + \langle \nabla\psi(x_0) - \nabla\psi(x^*),\ y - x^* \rangle \quad \text{by (5)} \\ &\geq L(x^*) - \langle \nabla\psi(x_0),\ x^* - x_0 \rangle + \psi(x^*) - \psi(x_0) \\ &\qquad + \langle \nabla\psi(x_0),\ y - x_0 \rangle - \psi(y) + \psi(x_0) \\ &\qquad - \langle \nabla\psi(x^*),\ y - x^* \rangle + \psi(y) - \psi(x^*) \\ &= L(x^*) + \triangle_\psi(x^*, x_0) - \triangle_\psi(y, x_0) + \triangle_\psi(y, x^*). \end{aligned}$$

□

Mirror Descent

If we use the Bregman divergence as a measure of displacement:

$$\begin{aligned} x_{k+1} &= \underset{x \in C}{\operatorname{argmin}} \left\{ f(x_k) + \langle g_k,\ x - x_k \rangle + \frac{1}{\alpha_k} \triangle_\psi (x, x_k) \right\} \\ &= \underset{x \in C}{\operatorname{argmin}} \{ \alpha_k f(x_k) + \alpha_k \langle g_k,\ x - x_k \rangle + \triangle_\psi (x, x_k) \} \end{aligned}$$

Suppose the constraint set C is the whole space (i.e. no constraint). Then we can take gradient with respect to x and find the optimality condition.

$$\begin{aligned} & g_k + \frac{1}{\alpha_k} (\nabla \psi(x_{k+1}) - \nabla \psi(x_k)) = 0 \\ \Leftrightarrow\ & \nabla \psi(x_{k+1}) = \nabla \psi(x_k) - \alpha_k g_k \\ \Leftrightarrow\ & x_{k+1} = (\nabla \psi)^{-1} (\nabla \psi(x_k) - \alpha_k g_k) = (\nabla \psi^*)(\nabla \psi(x_k) - \alpha_k g_k). \end{aligned}$$

Mirror Descent

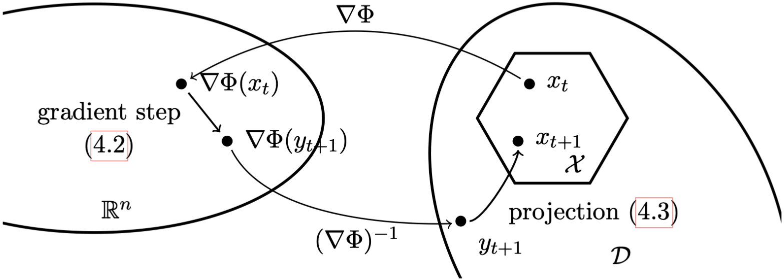


Illustration of mirror descent (from Bubeck et al. [2015])

Mirror Descent

For example, in KL-divergence over simplex, since we have

$$\psi(x) = \sum_i x^{(i)} \log x^{(i)},$$

the $\nabla \psi(x)^{(i)} = \log x^{(i)} + 1$. Thus the update rule becomes

$$\text{for } i: \qquad \log x_{k+1}^{(i)} = \log x_k^{(i)} - \alpha_k g_k \Leftrightarrow x_{k+1}^{(i)} = x_k^{(i)} \exp(-\alpha_k g_k).$$

For the simplex constraint, we have $x_0^{(i)} = 1/n$, and in each iteration we set,

$$\text{for } i: \qquad x_{k+1}^{(i)} = x_{k+1}^{(i)} / \sum_i x_{k+1}^{(i)}$$

Analysis of Mirror Descent

$$\boxed{L(y) + \triangle_\psi(y, x_0) \geq L(x^*) + \triangle_\psi(x^*, x_0) + \triangle_\psi(y, x^*).}$$

We further assume ψ is μ strongly convex. for

$$L(x) = \alpha_k \left(f(x_k) + \langle g_k,\ x - x_k \rangle \right),$$

in view of Lemma 23, we have

$$\begin{aligned} L(x^*) + \triangle_\psi(x^*, x_k) &\left(= \alpha_k(f(x_k) + \langle g_k,\ x^* - x_k\rangle) + \triangle_\psi(x^*, x_k)\right) \\ &\geq \underbrace{L(x_{k+1})}_{\alpha_k(f(x_k)+\langle g_k,\ x_{k+1}-x_k\rangle)} + \triangle_\psi(x_{k+1}, x_k) + \triangle_\psi(x^*, x_{k+1}). \end{aligned}$$

For Extension of Pythagorean, we use $y \leftarrow x^*$, $x_0 \leftarrow x_k$, and $x^* \leftarrow x_{k+1}$.

Analysis of Mirror Descent

$$\alpha_k(f(x_k) + \langle g_k,\ x^* - x_k\rangle) + \triangle_\psi(x^*, x_k) \\ \geq \alpha_k(f(x_k) + \langle g_k,\ x_{k+1} - x_k\rangle) + \triangle_\psi(x_{k+1}, x_k) + \triangle_\psi(x^*, x_{k+1}).$$

Some terms can be canceled. Thus, we have

$$\triangle_\psi(x^*, x_{k+1}) \leq \triangle_\psi(x^*, x_k) + \alpha_k \boxed{\langle g_k,\ x^* - x_k\rangle} \\ + \alpha_k\langle g_k,\ x_k - x_{k+1}\rangle \boxed{- \triangle_\psi\ (x_{k+1}, x_k)}.$$

Analysis of Mirror Descent

Since $\psi(\cdot)$ is strongly convex, we have

$$\begin{aligned}\triangle_\psi(x_{k+1}, x_k) &= \psi(x_{k+1}) - \psi(x_k) - \langle \nabla\psi(x_k),\ x_{k+1} - x_k\rangle \\ &\geq \frac{\mu}{2}\left\|x_{k+1} - x_k\right\|^2.\end{aligned}$$

This implies

$$\boxed{-\triangle_\psi\left(x_{k+1}, x_k\right)} \leq -\frac{\mu}{2}\left\|x_k - x_{k+1}\right\|^2.$$

Also, we have

$$f(x^*) \geq f(x_k) + \langle g_k,\ x^* - x_k\rangle.$$

Thus,

$$\boxed{\langle g_k,\ x^* - x_k\rangle} \leq -(f(x_k) - f(x^*)).$$

Analysis of Mirror Descent

Thus, we have

$$
\begin{aligned}
\triangle_\psi(x^*, x_{k+1}) &= \triangle_\psi(x^*, x_k) + \alpha_k \boxed{\langle g_k,\ x^* - x_k\rangle} \\
&\quad + \alpha_k\langle g_k,\ x_k - x_{k+1}\rangle \boxed{- \triangle_\psi\ (x_{k+1}, x_k)} \\
&\le \triangle_\psi(x^*, x_k) - \alpha_k(f(x_k) - f(x^*)) + \underbrace{\alpha_k\langle g_k,\ x_k - x_{k+1}\rangle}_{u^\top v \le \frac{1}{2\alpha}\|u\|_*^2 + \frac{\alpha}{2}\|v\|} \\
&\quad - \frac{\mu}{2}\|x_k - x_{k+1}\|^2 \\
&\le \triangle_\psi(x^*, x_k) - \alpha_k(f(x_k) - f(x^*)) + \boxed{\frac{\alpha_k^2}{2\mu}\|g_k\|_*^2 + \frac{\mu}{2}\|x_k - x_{k+1}\|} \\
&\quad - \frac{\mu}{2}\|x_k - x_{k+1}\|^2 .
\end{aligned}
$$

Analysis of Mirror Descent

Thus, we have

$$\triangle_\psi(x^*, x_{k+1}) \leq \triangle_\psi(x^*, x_k) - \alpha_k(f(x_k) - f(x^*)) + \frac{\alpha_k^2}{2\mu} \|g_k\|_*^2 .$$

Then, we arrive at ($\alpha_k = \alpha$)

$$\min_{k\in\{1,\dots,T\}} (f(x_k) - f(x^*)) \leq \frac{1}{T}\left(\frac{\triangle_\psi(x^*, x_1)}{\alpha} + \frac{\alpha}{2\mu}\sum_{k=1}^{T} \|g_k\|_*^2\right) = c\frac{RM}{\sqrt{T}},$$

where M bounds the $\|g_k\|_*$, R^2 bounds $\triangle_\psi(x^*, x_1)$, and $\alpha = \frac{R}{M}\sqrt{\frac{2\mu}{T}}$. This is the same as the bound of sub-GD.

Analysis of Mirror Descent

The advantage of using mirror descent over sub-gradient descent is that it takes into account the geometry of the problem through the potential function ψ. Consider the following problem.

$$\min_{x \in \mathcal{X}} f(x),$$

where $\mathcal{X} = \{x \in \mathbb{R}^n : x \geq 0, \sum x_i = 1\}$.

For sub-GD, assume that f is 1-Lipchitz with norm $\|\cdot\|_1$, equivalently, $\|g\|_\infty \leq 1$. Recall that implies $\|g\|_2 = \sqrt{n}\,\|g\|_\infty \leq \sqrt{n} \triangleq M$. Thus, the bound is

$$\|x_1 - x^*\|_2 \cdot \frac{M}{\sqrt{T}} = \|x_1 - x^*\|_2 \cdot \sqrt{\frac{n}{T}}.$$

Analysis of Mirror Descent

For MD, set $\psi(x) = \sum x_i \log x_i$. The fact is that if ψ is 1-strongly convex on $\mathcal{X}$, with $\|\cdot\|_1$, we have

$$\begin{aligned} \triangle_\psi(x^*, x_1) &\leq \log n \\ &\triangleq R^2, \end{aligned}$$

if x_1 is $(\frac{1}{n}, \ldots, \frac{1}{n})^\top$. Thus, we have the bound is

$$\sqrt{2}\sqrt{\frac{\log n}{T}}.$$

Remark. $\sqrt{\log n}$ is smaller than $\sqrt{n}$. This is crucial when n is very large.

References I

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

A Ben-Tal and M Teboulle. A smoothing technique for nondifferentiable optimization problems. In *Optimization*, pages 1–11. Springer, 1989.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Niao He. Big data optimizaton course, ie598. URL `https://github.com/niaohe/Big-Data-Optimization-Course/blob/main/lecture_scribe/IE598-lecture16-smoothing-techniques-I.pdf`.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

References II

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Xinhua Zhang. Bregman divergence and mirror descent. URL `https://www2.cs.uic.edu/~zhangx/teaching/bregman.pdf`.

Thank You!

Email:qianhui@zju.edu.cn