SYS 6003: Optimization | **Fall 2016**

Lecture 20

Instructor: Quanquan Gu | *Date: Nov 7^{th}*

We are going to introduce the properties of proximal mapping (proximal operator) and the calculation rules of proximal mapping.

Definition 1 (Proximal Mapping) *The proximal mapping of $h(\mathbf{x})$ is*

$$\mathrm{Prox}_h(\mathbf{x}) = \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + h(\mathbf{u}).$$

Lemma 1 (Non-expansiveness) *If $\mathbf{u} = \mathrm{Prox}_h(\mathbf{x})$, $\mathbf{v} = \mathrm{Prox}_h(\mathbf{y})$, we have*

$$(\mathbf{u} - \mathbf{v})^\top(\mathbf{x} - \mathbf{y}) \geq \|\mathbf{u} - \mathbf{v}\|_2^2. \quad (1)$$

In addition,

$$\|\mathbf{u} - \mathbf{v}\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2. \quad (2)$$

Remark 1 *(2) implies that $\|\mathrm{Prox}_h(\mathbf{x}) - \mathrm{Prox}_h(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$. It means that the proximal function is 1-Lipschitz function.*

Proof: Since $\mathbf{u} = \mathrm{Prox}_h(\mathbf{x})$, we have $\mathbf{x} - \mathbf{u} \in \partial h(\mathbf{u})$, which implies

$$h(\mathbf{v}) \geq h(\mathbf{u}) + (\mathbf{x} - \mathbf{u})^\top(\mathbf{v} - \mathbf{u}).$$

Similarly we have $\mathbf{y} - \mathbf{v} \in \partial h(\mathbf{v})$, which implies

$$h(\mathbf{u}) \geq h(\mathbf{v}) + (\mathbf{y} - \mathbf{v})^\top(\mathbf{u} - \mathbf{v}).$$

Thus we have

$$(\mathbf{x} - \mathbf{u} + \mathbf{v} - \mathbf{y})^\top(\mathbf{u} - \mathbf{v}) \geq 0.$$

Rearrange the inequality, we have

$$(\mathbf{x} - \mathbf{y})^\top(\mathbf{u} - \mathbf{v}) \geq \|\mathbf{u} - \mathbf{v}\|_2^2. \quad (3)$$

By (3) we have

$$\|\mathbf{u} - \mathbf{v}\|_2^2 \leq (\mathbf{x} - \mathbf{y})^\top(\mathbf{u} - \mathbf{v}) \leq \|\mathbf{x} - \mathbf{y}\|_2 \cdot \|\mathbf{u} - \mathbf{v}\|_2.$$

Thus

$$\|\mathbf{u} - \mathbf{v}\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2.$$

■

Now we focus on the calculation rules of proximal mapping.

1. **Separable sum**

$$f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + h(\mathbf{y}),$$

we have

$$\text{Prox}_f\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \begin{bmatrix}\text{Prox}_g(\mathbf{x})\\ \text{Prox}_g(\mathbf{y})\end{bmatrix}.$$

Example 1

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |\mathbf{x}_i| = |x_1| + |x_2| + \cdots + |x_d|.$$

2. **Scaling and translation of argument**
For $a \neq 0, x \in \mathbb{R}, f(x) = g(ax+b)$, we have

$$\text{Prox}_f(x) = \frac{1}{a}\big[\text{Prox}_{a^2 g}(ax+b) - b\big]$$

Proof:

$$\hat{u} = \text{Prox}_f(x) = \arg\min_u \frac{1}{2}(u-x)^2 + g(au+b)$$

Let $t = au + b$, thus $u = (t-b)/a$, we have

$$\begin{aligned}\hat{t} &= \arg\min_t \frac{1}{2}\left(\frac{t-b}{a} - x\right)^2 + g(t)\\ &= \arg\min_t \frac{1}{2a^2}(t-b-ax)^2 + g(t)\\ &= \arg\min_t \frac{1}{2}(t-b-ax)^2 + a^2 g(t)\\ &= \text{Prox}_{a^2 g}(ax+b)\end{aligned}$$

Thus we have

$$\text{Prox}_f(x) = \hat{u} = \frac{\hat{t}-b}{a} = \frac{1}{a}\big[\text{Prox}_{a^2 g}(ax+b) - b\big].$$

■

3. **"Right" scalar multiplication**

$$f(x) = \lambda g(x/\lambda),$$

we have

$$\mathrm{Prox}_f(x) = \lambda \cdot \mathrm{Prox}_{\lambda^{-1}g}(x/\lambda)$$

Proof:

$$\begin{aligned}\hat{u} = \mathrm{Prox}_f(x) &= \arg\min_u \frac{1}{2}(u-x)^2 + f(u) \\ &= \arg\min_u \frac{1}{2}(u-x)^2 + \lambda f(u/\lambda)\end{aligned}$$

Let $t = u/\lambda$, thus $u = \lambda t$, we have

$$\begin{aligned}\hat{t} &= \arg\min_t \frac{1}{2}(\lambda t - x)^2 + \lambda g(t) \\ &= \arg\min_t \frac{\lambda^2}{2}\Big(t - \frac{x}{\lambda}\Big)^2 + \lambda g(t) \\ &= \arg\min_t \frac{1}{2}\Big(t - \frac{x}{\lambda}\Big)^2 + \frac{1}{\lambda} g(t) \\ &= \mathrm{Prox}_{\lambda^{-1}g}(x/\lambda)\end{aligned}$$

Thus we have

$$\mathrm{Prox}_f(x) = \hat{u} = \lambda\hat{t} = \lambda \cdot \mathrm{Prox}_{\lambda^{-1}g}(x/\lambda).$$

■

4. **Addition to linear function**
For $\mathbf{a}, \mathbf{x} \in \mathbb{R}^d$,

$$f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{a}^\top \mathbf{x},$$

we have

$$\mathrm{Prox}_f(\mathbf{x}) = \lambda \cdot \mathrm{Prox}_g(\mathbf{x} - \mathbf{a})$$

Proof:

$$\begin{aligned}\mathrm{Prox}_f(\mathbf{x}) &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + f(\mathbf{u}) \\ &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + \lambda g(\mathbf{u}) + \mathbf{a}^\top \mathbf{u} \\ &= \arg\min_{\mathbf{u}} \frac{1}{2}[\mathbf{u}^\top\mathbf{u} - 2\mathbf{u}^\top\mathbf{x} + \mathbf{x}^\top\mathbf{x} + 2\mathbf{a}^\top\mathbf{u}] + g(\mathbf{u}) \\ &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{x} + \mathbf{a}\|_2^2 + g(\mathbf{u}) \\ &= \mathrm{Prox}_g(\mathbf{x} - \mathbf{a})\end{aligned}$$

■

5. **Addition to quadratic function**
   For $\mu > 0$,
$$f(\mathbf{x}) = g(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{a}\|_2^2,$$
   we have
$$\text{Prox}_f(\mathbf{x}) = \text{Prox}_{\theta g}\big(\theta\mathbf{x} + (1-\theta)\mathbf{a}\big),$$
   where $\theta = 1/(1+\mu)$.

   Example 2 (Elastic Net)
$$\min_{\mathbf{x}} \frac{1}{2n}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 + \mu\|\mathbf{x}\|_2^2$$

   Proof:
$$\begin{aligned}
\text{Prox}_f(\mathbf{x}) &= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + f(\mathbf{u}) \\
&= \arg\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 + g(\mathbf{u}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{a}\|_2^2 \\
&= \arg\min_{\mathbf{u}} \frac{1}{2}[\mathbf{u}^\top\mathbf{u} - 2\mathbf{u}^\top\mathbf{x} + \mathbf{x}^\top\mathbf{x} + \mu\mathbf{u}^\top\mathbf{u} - 2\mu\mathbf{u}^\top\mathbf{a} + \mu\mathbf{a}^\top\mathbf{a}] + g(\mathbf{u}) \\
&= \arg\min_{\mathbf{u}} \frac{1}{2}[(1+\mu)\mathbf{u}^\top\mathbf{u} - 2(\mathbf{x} + \mu\mathbf{a})^\top\mathbf{u} + C] + g(\mathbf{u}) \\
&= \arg\min_{\mathbf{u}} \frac{1+\mu}{2}\left[\mathbf{u}^\top\mathbf{u} - 2\Big(\frac{\mathbf{x} + \mu\mathbf{a}}{1+\mu}\Big)^\top\mathbf{u} + C'\right] + g(\mathbf{u}) \\
&= \arg\min_{\mathbf{u}} \frac{1}{2}\left\|\mathbf{u} - \frac{\mathbf{x} + \mu\mathbf{a}}{1+\mu}\right\|_2^2 + \frac{1}{1+\mu}g(\mathbf{u}) \\
&= \text{Prox}_{\theta g}\big(\theta\mathbf{x} + (1-\theta)\mathbf{a}\big)
\end{aligned}$$

   ■

Examples:

1. **Quadratic function:** For $\mathbf{A} > 0, \mathbf{x}, \mathbf{b} \in \mathbb{R}^d, c \in \mathbb{R}$,
$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c,$$
   we have
$$\text{Prox}_{tf}(\mathbf{x}) = (\mathbf{I} + t\mathbf{A})^{-1}(\mathbf{x} - t\mathbf{b}).$$

2. **Logarithmic function:**
$$f(\mathbf{x}) = -\sum_{i=1}^{d}\log(x_i),$$
   we have
$$\big[\text{Prox}_{tf}(\mathbf{x})\big]_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}.$$

3. **Euclidean norm(ℓ_2 norm):**

$$f(\mathbf{x}) = \|\mathbf{x}\|_2,$$

we have

$$\text{Prox}_{tf}(\mathbf{x}) = \begin{cases} \left(1 - \frac{t}{\|\mathbf{x}\|_2}\right)\mathbf{x} & , \|\mathbf{x}\|_2 > t \\ 0 & , \text{otherwise} \end{cases}$$