Part 3. Projected and Proximal Gradient Methods

Math 126 Winter 18

Date of current version: February 5, 2018

Abstract This note studies projected subgradient methods, mirror descent methods, (accelerated) proximal gradient methods, and proximal point methods. Many parts of this note are based on the chapters [1, Chapter 6,8-10] and lecture notes and slides for EE364b course by S. Boyd and J. Duchi [4].

Please email me if you find any typos or errors.

- Projection operator (onto a set $\mathcal{C}$):

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \frac{1}{2}||\boldsymbol{v}-\boldsymbol{x}||_2^2 = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}||\boldsymbol{v}-\boldsymbol{x}||_2^2 + I_{\mathcal{C}}(\boldsymbol{x}) \right\} \tag{0.1}$$

  where $I_{\mathcal{C}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{C}$ and ∞ otherwise.
- Proximal operator:

$$\operatorname{prox}_f(\boldsymbol{v}) = \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}||\boldsymbol{v}-\boldsymbol{x}||_2^2 + f(\boldsymbol{x}) \right\} \tag{0.2}$$

1 Projected Subgradient Methods (see [1, Chapter 8] [4])

Consider the following convex constrained problem:

$$\min_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}) \tag{1.1}$$

where we assume that

- $f : \mathbb{R}^d \to (-\infty, \infty]$ is proper, closed and convex.
- $\mathcal{C} \subseteq \mathbb{R}^d$ is nonempty closed and convex.
- $\mathcal{C} \subseteq \operatorname{int}(\operatorname{dom} f)$.
- The optimal set is nonempty.

A projected subgradient method is given by

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ f(\boldsymbol{x}_k) + \langle \boldsymbol{g}_k, \boldsymbol{x}-\boldsymbol{x}_k\rangle + \frac{1}{2s_k}||\boldsymbol{x}-\boldsymbol{x}_k||_2^2 \right\} \text{ for } \boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k) \\ &= \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_k - s_k\boldsymbol{g}_k). \end{aligned} \tag{1.2}$$

Lemma 1.1 *Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the projected subgradient method. Then*

$$||\boldsymbol{x}_{i+1}-\boldsymbol{x}_*||_2^2 \leq ||\boldsymbol{x}_i-\boldsymbol{x}_*||_2^2 - 2s_i(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_*)) + s_i^2||\boldsymbol{g}_i||_2^2. \tag{1.3}$$

Donghwan Kim
Dartmouth College
E-mail: donghwan.kim@dartmouth.edu

Proof We have

$$\begin{aligned}||\boldsymbol{x}_{i+1} - \boldsymbol{x}_*||_2^2 &= ||\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_i - s_i\boldsymbol{g}_i) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_*)||_2^2 \\ &\leq ||\boldsymbol{x}_i - s_i\boldsymbol{g}_i - \boldsymbol{x}_*||_2^2 \\ &= ||\boldsymbol{x}_i - \boldsymbol{x}_*||_2^2 - 2s_i \langle \boldsymbol{g}_i, \boldsymbol{x}_i - \boldsymbol{x}_* \rangle + s_i^2||\boldsymbol{g}_i||_2^2 \\ &\leq ||\boldsymbol{x}_i - \boldsymbol{x}_*||_2^2 - 2s_i(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*)) + s_i^2||\boldsymbol{g}_i||_2^2,\end{aligned}$$

where the first inequality is due to the nonexpansiveness of the projection operator ($||\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{y})||_2 \leq ||\boldsymbol{x} - \boldsymbol{y}||_2$), and the second inequality follows by the definition of subgradient.

We used the inequality (1.3) for the convergence analysis of the subgradient method (without projection) in the previous class, so we can claim that the projected subgradient method has the same convergence rate results.

Example 1.1 Linear equality constraints. Consider

$$\begin{aligned}&\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.\end{aligned}$$

A projection of $\boldsymbol{y}$ onto a convex set $\mathcal{C} = \{\boldsymbol{x} \ : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}) = (\boldsymbol{I} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1})\boldsymbol{A})\boldsymbol{y} + \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{b}.$$

A projected subgradient update is

$$\boldsymbol{x}_{k+1} = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_k - s_k\boldsymbol{g}_k) = \boldsymbol{x}_k - s_k(\boldsymbol{I} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A})\boldsymbol{g}_k$$

where we used $\boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{b}$.

When $f(\boldsymbol{x}) = ||\boldsymbol{x}||_1$, this reduces to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - s_k(\boldsymbol{I} - \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A})\,\mathrm{sign}(\boldsymbol{x}_k)$$

by using the subgradient $\boldsymbol{g}_k = \mathrm{sign}(\boldsymbol{x}_k) \in \partial(||\boldsymbol{x}_k||_1)$.

2 Mirror Descent Methods (see [1, Chapter 9] [4])

This mirror descent methods generalize the projected subgradient methods to the non-Euclidean setting, using the following Bregman distance.

Definition 2.1 Let $h \ : \ \mathbb{R}^d \to (-\infty, \infty]$ be a proper closed and convex function that is differentiable over $\mathrm{dom}(\partial h)$. The *Bregman distance* associated with h is the function $D_h \ : \ \mathrm{dom}(h) \times \mathrm{dom}(\partial h) \to \mathbb{R}$ given by

$$D_h(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x}) - h(\boldsymbol{y}) - \langle \nabla h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle . \tag{2.1}$$

We further assume that

- $\mathcal{C} \subseteq \mathrm{dom}(h)$,
- $h + I_{\mathcal{C}}$ is μ-strongly convex ($\mu > 0$).

A Bregman distance is not necessarily a distance. This is because in general it is not symmetric and does not satisfy the triangle inequality. The properties of Bregman distances with additional assumption on h above are

- $D_h(\boldsymbol{x}, \boldsymbol{y}) \geq \frac{\mu}{2}||\boldsymbol{x} - \boldsymbol{y}||^2$ for all $\boldsymbol{x} \in \mathcal{C}$, $\boldsymbol{y} \in \mathcal{C} \cap \mathrm{dom}(\partial h)$,
- Let $\boldsymbol{x} \in \mathcal{C}$ and $\boldsymbol{y} \in \mathcal{C} \cap \mathrm{dom}(\partial h)$. Then $D_h(\boldsymbol{x}, \boldsymbol{y}) \geq 0$, $D_h(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$.

Using the Bregman distance D_h, a mirror descent (or non-linear subgradient) method is updated as

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ f(\boldsymbol{x}_k) + \langle \boldsymbol{g}_k, \boldsymbol{x} - \boldsymbol{x}_k\rangle + \frac{1}{s_k} D_h(\boldsymbol{x}, \boldsymbol{x}_k) \right\} \text{ for } \boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k) \\ &= \arg\min_{\boldsymbol{x}\in\mathcal{C}} \left\{ \langle s_k \boldsymbol{g}_k - \nabla h(\boldsymbol{x}_k), \boldsymbol{x}\rangle + h(\boldsymbol{x}) \right\}.\end{aligned} \tag{2.2}$$

This reduces to the projected subgradient methods when $h(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2^2$ that leads to $D_h(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}||\boldsymbol{x} - \boldsymbol{y}||_2^2$. Similar extension is possible for the (accelerated) gradient methods (see *e.g.*, [8]), which is not covered in this course.

The optimality condition of the mirror descent update is

$$\begin{aligned}\boldsymbol{0} &\in s_k \boldsymbol{g}_k - \nabla h(\boldsymbol{x}_k) + \nabla h(\boldsymbol{x}_{k+1}) + N_{\mathcal{C}}(\boldsymbol{x}_{k+1}) \\ \boldsymbol{x}_{k+1} &\in (\nabla h + N_{\mathcal{C}})^{-1}(\nabla h(\boldsymbol{x}_k) - s_k \boldsymbol{g}_k)\end{aligned} \tag{2.3}$$

where the normal cone of $\mathcal{C}$ at $\boldsymbol{x} \in \mathcal{C}$ is defined as $N_{\mathcal{C}}(\boldsymbol{x}) = \partial I_{\mathcal{C}}(\boldsymbol{x}) = \{\boldsymbol{g} \in \mathbb{R}^d : \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle \leq 0 \text{ for any } y \in \mathcal{C}\}$. Then, we can simplify the mirror descent update as

$$\boldsymbol{x}_{k+1} = \nabla h^*(\nabla h(\boldsymbol{x}_k) - s_k \boldsymbol{g}_k) \tag{2.4}$$

where a conjugate function is defined as $h^*(\boldsymbol{z}) = \max_{\boldsymbol{x}\in\mathcal{C}}\{\langle \boldsymbol{x}, \boldsymbol{z}\rangle - h(\boldsymbol{x})\}$ with

$$\nabla^* h(\boldsymbol{z}) = (\nabla h + N_{\mathcal{C}})^{-1}(\boldsymbol{z}) = \arg\min_{\boldsymbol{x}} \{-\langle \boldsymbol{z}, \boldsymbol{x}\rangle + h(\boldsymbol{x}) + I_{\mathcal{C}}(\boldsymbol{x})\}. \tag{2.5}$$

Now, we are left to choosing h to simplify the projection-like computation ∇h^*.

Example 2.1 For $h(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2^2$, we have $\nabla h(\boldsymbol{x}) = \boldsymbol{x}$ and $h^*(\boldsymbol{z}) = \frac{1}{2}(||\boldsymbol{z}||_2^2 - ||\boldsymbol{z} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{z})||_2^2)$ with $\nabla h^*(\boldsymbol{z}) = \mathcal{P}_{\mathcal{C}}(\boldsymbol{z})$ [2].

Example 2.2 $\mathcal{C}$ *is a simplex* Δ_d. We choose the following negative entropy

$$h(\boldsymbol{x}) = \begin{cases} \sum_{i=1}^d x_i \log x_i, & \boldsymbol{x} \in \mathbb{R}_+^d, \\ \infty, & \text{otherwise}, \end{cases} \tag{2.6}$$

which is 1-strongly convex with respect to the ℓ_1-norm. The associated Bregman distance is

$$D_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^d x_i \log(x_i/y_i), \tag{2.7}$$

which is the so-called Kullback-Leibler divergence distance. The resulting mirror descent update is

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}\in\Delta_d} \left\{ \sum_{i=1}^d (s_k [\boldsymbol{g}_k]_i - 1 - \log([\boldsymbol{x}_k]_i)) x_i + \sum_{i=1}^d x_i \log x_i \right\},$$

which can be simplified as

$$[\boldsymbol{x}_{k+1}]_i = \frac{\exp(\log([\boldsymbol{x}_k]_i) + 1 - s_k [\boldsymbol{g}_k]_i)}{\sum_{j=1}^d \exp(\log([\boldsymbol{x}_k]_j) + 1 - s_k [\boldsymbol{g}_k]_j)} = \frac{[\boldsymbol{x}_k]_i \exp(-s_k [\boldsymbol{g}_k]_i)}{\sum_{j=1}^d [\boldsymbol{x}_k]_j \exp(-s_k [\boldsymbol{g}_k]_j)},$$

using $h^*(\boldsymbol{z}) = \log\left(\sum_{i=1}^d \exp(z_i)\right)$ and $\nabla h^*(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z})}{\sum_{i=1}^d \exp(z_i)}$ [2].

Lemma 2.1 *Let* $\{\boldsymbol{x}_k\}_{k\geq 0}$ *be the sequence generated by the mirror descent method. Then*

$$2D_h(\boldsymbol{x}_*, \boldsymbol{x}_{i+1}) \leq 2D_h(\boldsymbol{x}_*, \boldsymbol{x}_i) - 2s_i(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*)) + \frac{s_i^2}{\mu}||\boldsymbol{g}_i||_*^2. \tag{2.8}$$

Theorem 2.1 *Suppose that there exists a constant* M *such that* $||\boldsymbol{g}||_* \leq M$ *for all* $\boldsymbol{g} \in \partial f(\boldsymbol{x})$ *and* $\boldsymbol{x} \in \mathcal{C}$. *Let* $\{\boldsymbol{x}_k\}_{k\geq 0}$ *be the sequence generated by the mirror descent method. Then*

$$\min_{0\leq i\leq k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*) \leq \frac{2D_h(\boldsymbol{x}_*, \boldsymbol{x}_0) + \frac{M^2}{\mu}\sum_{i=0}^k s_i^2}{2\sum_{i=0}^k s_i}. \tag{2.9}$$

3 Proximal Operators (see [1, Chapter 6] [4])

Definition 3.1 Given a function $\phi : \mathbb{R}^d \to (-\infty, \infty]$, the *proximal mapping* of ϕ is given by

$$\text{prox}_\phi(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{z}||_2^2 + \phi(\boldsymbol{z}) \right\} \tag{3.1}$$

for any $\boldsymbol{x}$.

Theorem 3.1 *Let $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper closed and convex function. Then $\text{prox}_\phi(\boldsymbol{x})$ is a singleton for any $\boldsymbol{x} \in \mathbb{R}^d$.*

Example 3.1 Examples of proximal mappings.

- Constant $\phi(\boldsymbol{x}) = c$:

$$\text{prox}_\phi(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{z}||_2^2 + c \right\} = \boldsymbol{x}$$

- Affine $\phi(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle + b$:

$$\text{prox}_\phi(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{z}||_2^2 + \langle \boldsymbol{a}, \boldsymbol{z} \rangle + b \right\} = \boldsymbol{x} - \boldsymbol{a}$$

- Convex Quadratic $\phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{p}^\top \boldsymbol{x} + r$, where $\boldsymbol{Q} \in \mathbb{S}_+^d$:

$$\text{prox}_\phi(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}||\boldsymbol{x} - \boldsymbol{z}||_2^2 + \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{Q}\boldsymbol{z} + \boldsymbol{p}^\top \boldsymbol{z} + r \right\} = (\boldsymbol{Q} + \boldsymbol{I})^{-1}(\boldsymbol{x} - \boldsymbol{p})$$

Example 3.2 One-dimensional examples of proximal mappings.

- $\phi(x) = \lambda|x|$ for $\lambda > 0$: If the minimizer of $\frac{1}{2}(x - z)^2 + \lambda|z|$ is attained at $z > 0$, then $0 = z - x + \lambda$, meaning that $z = x - \lambda$. Therefore, if $x > \lambda$, $\text{prox}_\phi(x) = x - \lambda$. Conversely, if $x < -\lambda$, $\text{prox}_\phi(x) = x + \lambda$. If $|x| \leq \lambda$, then $\text{prox}_\phi(x) = 0$. Therefore, we have

$$\text{prox}_\phi(x) = [|x| - \lambda]_+ \,\text{sign}(x).$$

- $\phi(x) = \begin{cases} -\lambda \log x, & x > 0, \\ \infty, & x \leq 0. \end{cases}$ for $\lambda > 0$: The minimizer of $\frac{1}{2}(x - z)^2 - \lambda \log z$ over $\mathbb{R}_{++}$ is a positive root of $0 = z - x - \frac{\lambda}{z}$, which yields:

$$\text{prox}_\phi(x) = \frac{x + \sqrt{x^2 + 4\lambda}}{2}.$$

Theorem 3.2 *Supposed that $\phi : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_m} \to (-\infty, \infty]$ is given by*

$$\phi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \sum_{i=1}^m \phi_i(\boldsymbol{x}_i) \tag{3.2}$$

for any $\boldsymbol{x}_i \in \mathbb{R}^{d_i}$, $i = 1, \ldots, m$. Then,

$$\text{prox}_\phi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \text{prox}_{\phi_1}(\boldsymbol{x}_1) \times \ldots \times \text{prox}_{\phi_m}(\boldsymbol{x}_m). \tag{3.3}$$

Example 3.3 Examples of proximal mappings of separable functions.

- $\phi(\boldsymbol{x}) = ||\boldsymbol{x}||_1$: $\text{prox}_\phi(\boldsymbol{x}) = [|\boldsymbol{x}| - \lambda \boldsymbol{1}]_+ \odot \text{sign}(\boldsymbol{x})$, where $\odot$ is an element-wise multiplicaton.
- $\phi(\boldsymbol{x}) = \begin{cases} -\lambda \sum_{i=1}^d \log x_i, & \boldsymbol{x} \succ \boldsymbol{0}, \\ \infty, & \text{otherwise.} \end{cases}$: $[\text{prox}_\phi(\boldsymbol{x})]_i = \frac{x_i + \sqrt{x_i^2 + 4\lambda}}{2}$

Theorem 3.3 *Let $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper closed and convex function. Then the following three claims are equivalent:*

– $\boldsymbol{z} = \mathrm{prox}_{\phi}(\boldsymbol{x})$
– $\boldsymbol{x} - \boldsymbol{z} \in \partial\phi(\boldsymbol{z})$
– $\phi(\boldsymbol{y}) \geq \phi(\boldsymbol{z}) + \langle \boldsymbol{x} - \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{z} \rangle$ *for any* $\boldsymbol{y}$

Corollary 3.1 *Let ϕ be a proper closed and convex function. Then $\boldsymbol{x}$ is a minimizer of ϕ if and only $\boldsymbol{x} = \mathrm{prox}_{\phi}(\boldsymbol{x})$.*

For $\boldsymbol{z} = \mathrm{prox}_{s\phi}(\boldsymbol{x})$, we have

$$\begin{aligned} \boldsymbol{0} \in \boldsymbol{z} - \boldsymbol{x} + s\partial\phi \quad &\Leftrightarrow \quad (I + s\partial\phi)\boldsymbol{z} \ni \boldsymbol{x} \\ &\Leftrightarrow \quad \boldsymbol{z} = (I + s\partial\phi)^{-1}(\boldsymbol{x}) \end{aligned} \tag{3.4}$$

A proximity operator $(I + s\partial\phi)^{-1}$ is also known as a resolvent (or backward) operator of subdifferential operator $\partial\phi$.

Theorem 3.4 *Let ϕ be a proper closed and convex function. Then*

– *Firm nonexpansivity (*i.e.*, Co-coercivity with constant* 1*):* $\langle \boldsymbol{x} - \boldsymbol{y}, \mathrm{prox}_{\phi}(\boldsymbol{x}) - \mathrm{prox}_{\phi}(\boldsymbol{y}) \rangle \geq ||\,\mathrm{prox}_{\phi}(\boldsymbol{x}) - \mathrm{prox}_{\phi}(\boldsymbol{y})||_2^2$
– *Nonexpansivity (*i.e.*, Lipschitz continuity with constant* 1*):* $||\,\mathrm{prox}_{\phi}(\boldsymbol{x}) - \mathrm{prox}_{\phi}(\boldsymbol{y})||_2 \leq ||\boldsymbol{x} - \boldsymbol{y}||_2$

Proof Denote $\boldsymbol{u} = \mathrm{prox}_{\phi}(\boldsymbol{x})$ and $\boldsymbol{v} = \mathrm{prox}_{\phi}(\boldsymbol{y})$. Then we have

$$\begin{aligned} \phi(\boldsymbol{v}) &\geq \phi(\boldsymbol{u}) + \langle \boldsymbol{x} - \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u} \rangle \\ \phi(\boldsymbol{u}) &\geq \phi(\boldsymbol{v}) + \langle \boldsymbol{y} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v} \rangle . \end{aligned}$$

Summing the above two yields

$$\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{u} - \boldsymbol{v} \rangle \geq ||\boldsymbol{u} - \boldsymbol{v}||_2^2.$$

Using the Cauchy-Schwarz inequality, we have

$$||\boldsymbol{u} - \boldsymbol{v}||_2^2 \leq ||\boldsymbol{x} - \boldsymbol{y}||_2 ||\boldsymbol{u} - \boldsymbol{v}||_2.$$

4 Proximal Gradient Methods (see [1, Chapter 10])

4.1 Composite Nonsmooth Problem

Consider the following composite problem:

$$\min_{\boldsymbol{x}}\{F(\boldsymbol{x}) \equiv f(\boldsymbol{x}) + \phi(\boldsymbol{x})\}, \tag{4.1}$$

where we assume the following.

Assumption 4.1
- $f : \mathbb{R}^d \to (-\infty, \infty]$ is proper and closed, $\mathrm{dom}\, f$ is convex, $\mathrm{dom}(\phi) \subseteq \mathrm{int}(\mathrm{dom}\, f)$, and f is L-smooth over $\mathrm{int}(\mathrm{dom}\, f)$.
- $\phi : \mathbb{R}^d \to (-\infty, \infty]$ is proper, closed and convex.
- The optimal set is nonempty.

Example 4.1 Examples of composite problems.

- $\phi(\boldsymbol{x}) = 0$: Unconstrained smooth minimization
- $\phi(\boldsymbol{x}) = I_{\mathcal{C}}(\boldsymbol{x})$ where $\mathcal{C} \subseteq \mathbb{R}^d$ is convex: Constrained smooth minimization
- $\phi(\boldsymbol{x}) = \lambda||\boldsymbol{x}||_1$: ℓ_1-regularized minimization

4.2 Proximal Gradient Methods

Proximal gradient update is given by

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2s_k}||\boldsymbol{x} - \boldsymbol{x}_k||_2^2 + \phi(\boldsymbol{x}) \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}||\boldsymbol{x} - (\boldsymbol{x}_k - s_k \nabla f(\boldsymbol{x}_k))||_2^2 + s_k \phi(\boldsymbol{x}) \right\} \\
&= \mathrm{prox}_{s_k \phi}(\boldsymbol{x}_k - s_k \nabla f(\boldsymbol{x}_k)) \\
&= \mathrm{T}_{1/s_k}(\boldsymbol{x}_k)
\end{aligned} \tag{4.2}$$

where we define the so-called prox-grad operator $\mathrm{T}_{1/s}(\boldsymbol{x}) \equiv \mathrm{prox}_{s\phi}(\boldsymbol{x} - s\nabla f(\boldsymbol{x}))$. This method is effective when the proximal operator $\mathrm{prox}_{s_k \phi}$ is efficient!

The optimality condition of the proximal gradient step (4.2) is

$$\begin{aligned}
\mathbf{0} \in \boldsymbol{x}_{k+1} - (\boldsymbol{x}_k - s_k \nabla f(\boldsymbol{x}_k)) + s_k \partial\phi(x_{k+1}) \quad &\Leftrightarrow \quad \boldsymbol{x}_{k+1} + s_k \partial\phi(\boldsymbol{x}_{k+1}) \ni \boldsymbol{x}_k - s_k \nabla f(\boldsymbol{x}_k) \\
&\Leftrightarrow \quad (I + s_k \partial\phi)(\boldsymbol{x}_{k+1}) \ni (I - s_k \nabla f)(\boldsymbol{x}_k) \\
&\Leftrightarrow \quad \boldsymbol{x}_{k+1} = (I + s_k \partial\phi)^{-1}(I - s_k \nabla f)(\boldsymbol{x}_k).
\end{aligned} \tag{4.3}$$

The proximal gradient method is an instance of operator splitting approaches, becauses the method exploits the fact that ∂F can be splitted as $\partial F = \nabla f + \partial\phi$. In specific, the proximal gradient method is an instance of forward-backward splitting method (among various operator splitting approaches).

An optimal point $\boldsymbol{x}_*$ is a fixed point of forward-backward operator, because

$$\begin{aligned}
\mathbf{0} \in \partial F(\boldsymbol{x}_*) = \nabla f(\boldsymbol{x}_*) + \partial\phi(\boldsymbol{x}_*) \quad &\Leftrightarrow \quad \mathbf{0} \in (\boldsymbol{x}_* + s\partial\phi(\boldsymbol{x}_*)) - (\boldsymbol{x}_* - s\nabla f(\boldsymbol{x}_*)) \\
&\Leftrightarrow \quad \mathbf{0} \in (I + s\partial\phi)(\boldsymbol{x}_*) - (I - s\nabla f)(\boldsymbol{x}_*) \\
&\Leftrightarrow \quad (I + s\partial\phi)(\boldsymbol{x}_*) \ni (I - s\nabla f)(\boldsymbol{x}_*) \\
&\Leftrightarrow \quad \boldsymbol{x}_* = (I + s\partial\phi)^{-1}(I - s\nabla f)(\boldsymbol{x}_*).
\end{aligned}$$

4.3 Convergence Analysis of Proximal Gradient Methods

Definition 4.1 The *gradient mapping* is the operator $G_{1/s} : \text{int}(\text{dom}\, f) \to \mathbb{R}$ defined by

$$G_{1/s}(\boldsymbol{x}) \equiv \frac{1}{s}(\boldsymbol{x} - \text{T}_{1/s}(\boldsymbol{x})) \tag{4.4}$$

for any $\boldsymbol{x} \in \text{int}(\text{dom}\, f)$.

Using the terminology of the gradient mapping, the proximal gradient update step can be simply rewritten as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - s_k G_{1/s_k}(\boldsymbol{x}_k). \tag{4.5}$$

Theorem 4.1 *Suppose that Assumption 4.1 holds and let $s > 0$. Then*

- *When $\phi(\boldsymbol{x}) = 0$, $G_{1/s}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \text{int}(\text{dom}\, f)$.*
- *For $\boldsymbol{x}_* \in \text{int}(\text{dom}\, f)$, $G_{1/s}(\boldsymbol{x}_*) = \mathbf{0}$ if and only $\boldsymbol{x}_*$ is a stationary point.*

Corollary 4.1 *Suppose that Assumption 4.1 holds and f is convex, and let $s > 0$. Then for $\boldsymbol{x}_* \in \text{dom}\, \phi$, $G_{1/s}(\boldsymbol{x}_*) = \mathbf{0}$ if and only if $\boldsymbol{x}_*$ is an optimal point.*

Lemma 4.1 *(Sufficient decrease lemma) Suppose that Assumption 4.1 holds. For any $\boldsymbol{x} \in \text{int}(\text{dom}\, f)$ and $s > 0$*

$$F(\boldsymbol{x}) - F(\text{T}_{1/s}(\boldsymbol{x})) \geq s\left(1 - \frac{sL}{2}\right)||G_{1/s}(\boldsymbol{x})||_2^2. \tag{4.6}$$

Theorem 4.2 *Suppose that Assumption 4.1 holds. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with a constant step size $s \in \left(0, \frac{2}{L}\right)$. Then,*

- *the sequence $\{F(\boldsymbol{x}_k)\}_{k\geq 0}$ is nonincreasing. In addition, $F(\boldsymbol{x}_{k+1}) < F(\boldsymbol{x}_k)$ unless $G_{1/s}(\boldsymbol{x}_k) = \mathbf{0}$.*
- *$G_{1/s}(\boldsymbol{x}_k) \to \mathbf{0}$ as $k \to \infty$*
- $\min_{k=0,\ldots,N} ||G_{1/s}(\boldsymbol{x}_k)||_2 \leq \sqrt{\frac{F(\boldsymbol{x}_0) - F(\boldsymbol{x}_*)}{s\left(1-\frac{sL}{2}\right)(N+1)}}$

Similar analysis is given in [1, Theorem 10.15] for the proximal gradient method with a backtracking line search.

Theorem 4.3 *Suppose that Assumption 4.1 holds and f is convex. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with a constant step size $s = \frac{1}{L}$. Then, for $k \geq 0$*

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_*) \leq \frac{L||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2}{2k}. \tag{4.7}$$

Similar analysis is given in [1, Theorem 10.21] for the proximal gradient method with a backtracking line search.

Theorem 4.4 *Suppose that Assumption 4.1 holds and f is μ-strongly convex. Let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with a constant step size $s = \frac{1}{L}$. Then, for $k \geq 0$*

$$\begin{aligned} ||\boldsymbol{x}_{k+1} - \boldsymbol{x}_*||_2^2 &\leq \left(1 - \frac{\mu}{L}\right)||\boldsymbol{x}_k - \boldsymbol{x}_*||_2^2 \\ ||\boldsymbol{x}_k - \boldsymbol{x}_*||_2^2 &\leq \left(1 - \frac{\mu}{L}\right)||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2 \\ F(\boldsymbol{x}_k) - F(\boldsymbol{x}_*) &\leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^k ||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2. \end{aligned} \tag{4.8}$$

Similar analysis is given in [1, Theorem 10.29] for the proximal gradient method with a backtracking line search.

4.4 Accelerated Proximal Gradient Methods

Fast proximal gradient method that is widely popularized under the name FISTA [3] uses Nesterov's acceleration technique to accelerate the convergence rate of the proximal gradient method.

Algorithm 1 Fast Proximal Gradient Method [3]

1: **Input:** $f \in \mathcal{F}_L^{1,1}$, ϕ is convex, $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^d$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \text{prox}_{\frac{1}{L}\phi}\left(\boldsymbol{y}_k - \frac{1}{L}\nabla f(\boldsymbol{y}_k)\right)$
4: $\quad \boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$

There is a backtracking version of this accelerated method [3], [1, Chapter 10.7].
Two widely used choices of β_k are as follows.

- Choice 1: $\beta_k = \frac{t_k - 1}{t_{k+1}}$ with $t_k = \begin{cases} 1, & k = 0, \\ \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, & k = 1, 2, \ldots \end{cases}$
- Choice 2: $\beta_k = \frac{t_k - 1}{t_{k+1}} = \frac{k}{k+3}$ with $t_k = \frac{k+2}{2}$.

Theorem 4.5 *Let $f \in \mathcal{F}_L^{1,1}$ and ϕ is convex, and let $\{\boldsymbol{x}_k\}_{k\geq 0}$ be the sequence generated by the fast proximal gradient method (using either choice of the parameters above). Then, for $k \geq 0$*

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_*) \leq \frac{L||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2}{2t_{k-1}^2} \leq \frac{2L||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2}{(k+1)^2}. \tag{4.9}$$

Remark 4.1 Fast proximal gradient method is widely known as fast iterative shrinkage/thresholding algorithm (FISTA), because the paper [3] specifically considered the choice $\phi(\boldsymbol{x}) = \lambda||\boldsymbol{x}||_1$ involving shrinkage/thresholding operator (or equivalently soft-thresholding operator), while the proof was for general ϕ. FISTA reduces to iterative shrinkage/thresholding algorithm (ISTA) [5,6] without the acceleration, which is an instance of a proximal gradient method.

For an L-smooth and μ-strongly convex function f, fast proximal gradient method with the following coefficient:

$$\beta_k = \frac{1 - \sqrt{q}}{1 + \sqrt{q}} \tag{4.10}$$

for $q = \frac{\mu}{L}$ satisfies

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_*) \leq 2(1 - \sqrt{q})^q (F(\boldsymbol{x}_0) - F(\boldsymbol{x}_*)). \tag{4.11}$$

5 Proximal Point Methods (see [1, Chapter 10])

Consider the nonsmooth convex problem:

$$\min_{\boldsymbol{x}} \phi(\boldsymbol{x}) \tag{5.1}$$

where we assume that $\phi : \mathbb{R}^d \to (-\infty, \infty]$ is a proper, closed and convex function. This is an instance of the composite problem (4.1), where $f(\boldsymbol{x}) \equiv 0$ is 0-smooth or L-smooth with any constant $L \geq 0$.

The update of a proximal point method is given by

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2s_k} ||\boldsymbol{x} - \boldsymbol{x}_k||_2^2 + \phi(\boldsymbol{x}) \right\} \\ &= \mathrm{prox}_{s_k \phi}(\boldsymbol{x}_k). \end{aligned} \tag{5.2}$$

Theorem 5.1 *Let $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper, closed and convex function. Assume that $\min_{\boldsymbol{x}} \phi(\boldsymbol{x})$ has a nonempty optimal set. Let $\{\boldsymbol{x}_k\}_{k \geq 0}$ be the sequence generated by the proximal point method with $s_k > 0$. Then,*

$$\phi(\boldsymbol{x}_k) - \phi(\boldsymbol{x}_*) \leq \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2}{2 \sum_{i=0}^{k-1} s_i}. \tag{5.3}$$

For a constant step $s_k = s > 0$, this reduces to

$$\phi(\boldsymbol{x}_k) - \phi(\boldsymbol{x}_*) \leq \frac{||\boldsymbol{x}_0 - \boldsymbol{x}_*||_2^2}{2sk}. \tag{5.4}$$

Choosing large s_k will provide fast convergence rate, while each update becomes almost equivalent to solving the original problem. On the other hand, small s_k yields slow convergence rate, whereas each update becomes easier to solve due to a larger quadratic term.

Example 5.1 Consider minimizing the quadratic function:

$$\phi(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{p}^\top \boldsymbol{x}, \tag{5.5}$$

where $\boldsymbol{Q} \in \mathbb{S}_+^d$. The proximal point method updates as

$$\boldsymbol{x}_{k+1} = \mathrm{prox}_{s\phi}(\boldsymbol{x}_k) = \left(\boldsymbol{Q} + \frac{1}{s} \boldsymbol{I} \right)^{-1} \left(\boldsymbol{p} + \frac{1}{s} \boldsymbol{x}_k \right), \tag{5.6}$$

which involves the regularized inverse $\left(\boldsymbol{Q} + \frac{1}{s} \boldsymbol{I} \right)^{-1}$.

Remark 5.1 The proximal point method seems not useful in general, but this becomes more useful when combined with duality, which leads to the augmented Lagrangian methods. We will learn this in few weeks.

It is also possible to accelerate the proximal point method.

Algorithm 2 Fast Proximal Point Method [7]

1: **Input:** ϕ is convex, $\boldsymbol{x}_0 = \boldsymbol{y}_0 \in \mathbb{R}^d$.
2: **for** $k \geq 0$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \mathrm{prox}_{s_k \phi}(\boldsymbol{x}_k)$
4: $\quad \boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1} + \beta_k (\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$

Theorem 5.2 *Let $\phi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper, closed and convex function. Assume that $\min_{\boldsymbol{x}} \phi(\boldsymbol{x})$ has a nonempty optimal set. Let $\{\boldsymbol{x}_k\}_{k \geq 0}$ be the sequence generated by the fast proximal point method with $s_k \geq s > 0$ (and either choice of parameters β_k for the fast proximal gradient method). Then,*

$$\phi(\boldsymbol{x}_k) - \phi(\boldsymbol{x}_*) \leq O\left(\frac{1}{sk^2} \right). \tag{5.7}$$

References

1. Beck, A.: First-Order Methods in Optimization. Soc. Indust. Appl. Math., Philadelphia (2017). URL http://epubs.siam.org/doi/book/10.1137/1.9781611974997
2. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters **3**(31), 167–75 (2003). DOI '10.1016/S0167-6377(02)00231-6'
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009). DOI 10.1137/080716542
4. Boyd, S., Duchi, J.: EE364b: Convex Optimization II. Lecture notes and slides
5. Chambolle, A., De Vore, R.A., Lee, N.Y., Lucier, B.J.: Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. IEEE Trans. Im. Proc. **7**(3), 319–35 (1998). DOI 10.1109/83.661182
6. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. **57**(11), 1413–57 (2004). DOI 10.1002/cpa.20042
7. Güler, O.: New proximal point algorithms for convex minimization. SIAM J. Optim. **2**(4), 649–64 (1992). DOI 10.1137/0802032
8. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming **103**(1), 127–52 (2005). DOI 10.1007/s10107-004-0552-5