Introductory Lectures on Optimization

Beyond The Black-box Model (1)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

December 4, 2024

Outline

1 Proximal Gradient Method
- Proximal Operator
- Properties of Proximal Operator
- Analysis for Proximal Gradient Method
- Accelearted Proximal Gradient Method
- Special case: Proximal Point Method

2 Douglas-Rachford Splitting
- Different Settings for Convex Problem
- Fixed Point for Nonsmooth Composition
- Splitting Algorithm

3 Reference

Part I
Proximal Gradient Method

Proximal Operator

The proximal operator (or proximal mapping) of a convex function h is defined as

$$\operatorname{prox}_h(\boldsymbol{x}) = \underset{\boldsymbol{u}}{\operatorname{argmin}} \left(h(\boldsymbol{u}) + \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right).$$

Example 1

1. If $h(\boldsymbol{x}) = 0$, we have $\operatorname{prox}_h(\boldsymbol{x}) = \boldsymbol{x}$.
2. If $h(\boldsymbol{x})$ is an indicator function of a closed convex set $\mathcal{C}$, prox_h is projection on $\mathcal{C}$, that is
$$\operatorname{prox}_h(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathcal{C}}{\operatorname{argmin}} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 = \pi_{\mathcal{C}}(\boldsymbol{x}).$$

Proximal Operator

3 If $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, prox_h is the "soft-threshold" (shrinkage) operation:

$$\text{prox}_h(\boldsymbol{x})_i = \begin{cases} \boldsymbol{x}_i - 1, & \boldsymbol{x}_i \geq 1, \\ 0, & |\boldsymbol{x}_i| \leq 1, \\ \boldsymbol{x}_i + 1, & \boldsymbol{x}_i \leq -1. \end{cases}$$

Remark. The problem

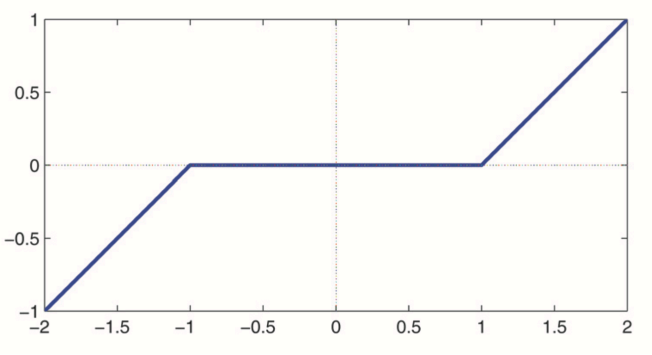
$$\underset{\boldsymbol{u}}{\text{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 + \|\boldsymbol{u}\|_1 \right\}$$

is separable with respect to both $\boldsymbol{u}$ and $\boldsymbol{x}$ hence one could solve the following problem:

$$\underset{\boldsymbol{u}_i}{\text{argmin}} \left\{ \frac{1}{2} (\boldsymbol{u}_i - \boldsymbol{x}_i)^2 + |\boldsymbol{u}_i| \right\}.$$

Proximal Operator

1. If $\boldsymbol{u}_i > 0$, we have $\boldsymbol{u}_i^* - \boldsymbol{x}_i + 1 = 0$, then $\boldsymbol{u}_i^* = \boldsymbol{x}_i - 1$ with $\boldsymbol{x}_i > 1$.
2. If $\boldsymbol{u}_i < 0$, we have $\boldsymbol{u}_i^* - \boldsymbol{x}_i - 1 = 0$, then $\boldsymbol{u}_i^* = \boldsymbol{x}_i + 1$ with $\boldsymbol{x}_i < 1$.
3. If $\boldsymbol{u}_i = 0$, we have $0 \in 0 - \boldsymbol{x}_i + \partial|0|$, then $\boldsymbol{x}_i \in \partial|0| = [-1, 1]$.



Proximal Gradient Method

Unconstrained optimization with objective split into two components:

$$\min\left\{f(\boldsymbol{x}) \triangleq g(\boldsymbol{x}) + h(\boldsymbol{x})\right\}.$$

where g is convex and differentiable, and $\operatorname{dom} g = \mathbb{R}^n$; h is convex with *inexpensive* proximal operator.

Proximal Gradient algorithm:

$$\boldsymbol{x}_{k+1} = \operatorname{prox}_{t_k h}(\boldsymbol{x}_k - t_k \nabla g(\boldsymbol{x}_k)).$$

1. $t_k > 0$ is step size, constant or determined by line search.
2. can start at infeasible $\boldsymbol{x}_0$ (However $\boldsymbol{x}_k \in \operatorname{dom} f = \operatorname{dom} h$ for $k \geq 1$.)

Proximal Gradient Method

How to interpret proximal algorithm? From the definition

$$
\begin{aligned}
\boldsymbol{x}_{k+1} &= \operatorname{prox}_{th}(\boldsymbol{x}_k - t\nabla g(\boldsymbol{x}_k)) \\
&= \operatorname*{argmin}_{\boldsymbol{u}} \left(h(\boldsymbol{u}) + \frac{1}{2t} \|\boldsymbol{u} - \boldsymbol{x}_k + t\nabla g(\boldsymbol{x}_k)\|_2^2 \right) \\
&= \operatorname*{argmin}_{\boldsymbol{u}} \left(h(\boldsymbol{u}) + g(\boldsymbol{x}_k) + \nabla g(\boldsymbol{x}_k)^\top (\boldsymbol{u} - \boldsymbol{x}_k) + \frac{1}{2t} \|\boldsymbol{u} - \boldsymbol{x}_k\|_2^2 \right).
\end{aligned}
$$

$\boldsymbol{x}_{k+1}$ minimize $h(\boldsymbol{u})$ plus a simple quadratic local model of $g(\boldsymbol{u})$ around $\boldsymbol{x}_k$.

Examples

Example 2 (Gradient Method:)

Special case with $h(\boldsymbol{x}) = 0$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla g(\boldsymbol{x}).$$

If $h(\boldsymbol{x}) = 0$, then $\text{prox}_h(\boldsymbol{x}) = \boldsymbol{x}$.

Examples

Example 3 (Projected Gradient Method)

Special case with $h(\boldsymbol{x}) = I_{\mathcal{C}}(\boldsymbol{x})$

$$\boldsymbol{x}_{k+1} = \pi_{\mathcal{C}}(\boldsymbol{x}_k - t\nabla g(\boldsymbol{x}_k)).$$

Examples

Example 4 (Soft-thresholding Method)

Special case with $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$

$$\boldsymbol{x}_{k+1} = \text{prox}_{th}(\boldsymbol{x} - t\nabla g(\boldsymbol{x})),$$

where

$$\text{prox}_{th}(\boldsymbol{u})_i = \begin{cases} \boldsymbol{u}_i - t, & \boldsymbol{u}_i \geq t, \\ 0, & |\boldsymbol{u}_i| \leq t, \\ \boldsymbol{u}_i + t, & \boldsymbol{u}_i \leq -t. \end{cases}$$

Properties

Propsition 5

If h is convex and closed (has a closed epigraph), then

$$\operatorname{prox}_h(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{u}} \left(h(\boldsymbol{u}) + \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right)$$

exists and is unique for all $\boldsymbol{x}$.

Proof.

$h(\boldsymbol{u}) + \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2$ is strongly convex. □

Properties

Propsition 6

$\boldsymbol{u} = \mathrm{prox}_h(\boldsymbol{x})$ is equivalent to the following

1. $\boldsymbol{x} - \boldsymbol{u} \in \partial h(\boldsymbol{u})$,
2. $h(\boldsymbol{z}) \geq h(\boldsymbol{u}) + (\boldsymbol{x} - \boldsymbol{u})^\top (\boldsymbol{z} - \boldsymbol{u})$ for all $\boldsymbol{z}$.

Proof. $0 \in \partial\{h(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2\}$. Thus

$$0 \in \partial h(\boldsymbol{u}) + \boldsymbol{u} - \boldsymbol{x}.$$

Also we have for $g \in \partial h(\boldsymbol{u})$,

$$h(\boldsymbol{z}) \geq h(\boldsymbol{u}) + g^\top (\boldsymbol{z} - \boldsymbol{u}).$$

□

Properties

Propsition 7

Proximal mapping of indicator function $I_{\mathcal{C}}$ is Euclidean projection on $\mathcal{C}$

$$\boldsymbol{u}^* = \operatorname{prox}_{I_{\mathcal{C}}}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{C}} \|\boldsymbol{u}-\boldsymbol{x}\|_2^2 = \pi_{\mathcal{C}}(\boldsymbol{x}),$$

and

$$(\boldsymbol{x}-\boldsymbol{u}^*)^\top(\boldsymbol{z}-\boldsymbol{u}^*) \leq 0, \forall \boldsymbol{z}\in\mathcal{C},$$

for all $\boldsymbol{z}\in\mathcal{C}$.

Proof. First of all, $h(\boldsymbol{z}) \geq h(\boldsymbol{u}^*) + (\boldsymbol{x}-\boldsymbol{u}^*)^\top(\boldsymbol{z}-\boldsymbol{u}^*)$ for all $\boldsymbol{z}$. And from the definition of indicator function, we have $h(\boldsymbol{z}) = h(\boldsymbol{u}^*) = 0$. □

Properties

Propsition 8 (Fixed Point)

Let f be a convex function, we have that a point $\boldsymbol{x}_*$ minimizes $f(\boldsymbol{x})$ if and only if $\boldsymbol{x}_* = \mathrm{prox}_f(\boldsymbol{x}_*)$.

proof. First, if $\boldsymbol{x}_*$ minimize $f(\boldsymbol{x})$, we have $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_*)$. Hence,

$$f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 \geq f(\boldsymbol{x}_*) + \frac{1}{2}\|\boldsymbol{x}_* - \boldsymbol{x}_*\|_2^2 .$$

This is implies that

$$\boldsymbol{x}_* = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 \right\} = \mathrm{prox}_f(\boldsymbol{x}_*).$$

Properties

Proof. (Continued.)

To prove the converse, consider if

$$\boldsymbol{x}_* = \operatorname{prox}_f(\boldsymbol{x}_*) = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 \right\}.$$

By the optimality condition, this implies that

$$0 \in \partial f(\boldsymbol{x}_*) + (\boldsymbol{x}_* - \boldsymbol{x}_*) \Rightarrow 0 \in \partial f(\boldsymbol{x}_*).$$

Therefore, $\boldsymbol{x}_*$ minimizes f. $\square$

Properties

Propsition 9 (Non-expansive)

$$\left\|\operatorname{prox}_f(\boldsymbol{x}) - \operatorname{prox}_f(\boldsymbol{y})\right\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2 .$$

Proof. Let us denote $\boldsymbol{u} = \operatorname{prox}_f(\boldsymbol{x})$ and $\boldsymbol{v} = \operatorname{prox}_f(\boldsymbol{y})$, then in view of Prop. 6,

$$\boldsymbol{x} - \boldsymbol{u} \in \partial f(\boldsymbol{u}), \qquad \boldsymbol{y} - \boldsymbol{v} \in \partial f(\boldsymbol{v}).$$

Combining this with monotonicity of subdifferential gvies $\langle \boldsymbol{x} - \boldsymbol{u} - \boldsymbol{y} + \boldsymbol{v},\ \boldsymbol{u} - \boldsymbol{v} \rangle \geq 0$. Hence, we have

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

By Cauchy-Schwarts inequality, this also leads to $\left\|\operatorname{prox}_f(\boldsymbol{x}) - \operatorname{prox}_f(\boldsymbol{y})\right\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2$. □

Properties

Remark.
The subdifferential of a convex function si a monotone operator:

$$\langle g_{\boldsymbol{u}} - g_{\boldsymbol{v}},\ \boldsymbol{u} - \boldsymbol{v} \rangle \geq 0$$

for all $\boldsymbol{u}, \boldsymbol{v}, g_{\boldsymbol{u}} \in \partial f(\boldsymbol{u})$, and $g_{\boldsymbol{v}} \in \partial f(\boldsymbol{v})$.

The proof is very straightforward. Combining the following two inequalities shows it:

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + g_{\boldsymbol{u}}^\top (\boldsymbol{v} - \boldsymbol{u}), \text{ and } f(\boldsymbol{u}) \geq f(\boldsymbol{v}) + g_{\boldsymbol{v}}^\top (\boldsymbol{u} - \boldsymbol{v}).$$

Problem

For problem

$$\min \left\{ f(\boldsymbol{x}) \triangleq g(\boldsymbol{x}) + h(\boldsymbol{x}) \right\},$$

we assume that

1. h is closed and convex (so that prox_{th} is well-defined).
2. g is differentiable with dom $g = \mathbb{R}^n$, and belong to $\mathcal{S}_{1,1}^{L,m}(\mathbb{R}^n)$.
3. The optimal value f^* is finite and attained at $\boldsymbol{x}^*$ (not necessarily unique).

From Lemma 1.2.3 and definition 2.1.2, for all $\boldsymbol{x}$, $\boldsymbol{y}$,

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{1}$$

and

$$g(\boldsymbol{y}) \leq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{2}$$

Update rule for proximal gradient

The update rule for proximal gradient:

$$\begin{aligned} \boldsymbol{x}' &= \operatorname{prox}_{th}(\boldsymbol{x} - t\nabla g(\boldsymbol{x})) \\ &= \boldsymbol{x} - tG_t(\boldsymbol{x}), \end{aligned}$$

where $G_t(\boldsymbol{x})$ is the gradient map in this settings, defined as

$$G_t(\boldsymbol{x}) = \frac{1}{t}\left(\boldsymbol{x} - \operatorname{prox}_{th}(\boldsymbol{x} - t\nabla g(\boldsymbol{x}))\right).$$

(The following properties of $G_t(\boldsymbol{x})$ will be left as excercises)

1. $G_t(\boldsymbol{x}) \in \nabla g(\boldsymbol{x}) + \partial h(\boldsymbol{x} - tG_t(\boldsymbol{x}))$.
2. $G_t(\boldsymbol{x}_*) = 0$ if and only if $\boldsymbol{x}_*$ minimizes $f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x})$.

Analysis for Fixed Step Size

For $0 < t \leq 1/L$,

1. The upper bound (2) implies (substitute $\boldsymbol{y} = \boldsymbol{x} - tG_t(\boldsymbol{x})$ into bound (2))

$$
\begin{aligned}
g(\boldsymbol{x} - tG_t(\boldsymbol{x})) &\leq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^\top(-tG_t(\boldsymbol{x})) + \frac{L}{2}\left\|tG_t(\boldsymbol{x})\right\|_2^2; \\
&= g(\boldsymbol{x}) - t\nabla g(\boldsymbol{x})^\top(G_t(\boldsymbol{x})) + \frac{t \cdot t \cdot L}{2}\left\|G_t(\boldsymbol{x})\right\|_2^2; \\
\Rightarrow \quad g(\boldsymbol{x} - tG_t(\boldsymbol{x})) &\leq g(\boldsymbol{x}) - t\nabla g(\boldsymbol{x})^\top(G_t(\boldsymbol{x})) + \frac{t}{2}\left\|G_t(\boldsymbol{x})\right\|_2^2. \qquad (3)
\end{aligned}
$$

The last inequality comes from $t \cdot L \leq 1$ (since $0 < t \leq 1/L$).

Analysis for Fixed Step Size

For $0 < t \leq 1/L$,

2. If the inequality (1) is also satisfied, we have $\boxed{mt \leq 1}$, since

$$\frac{mt^2}{2}\|G_t(\boldsymbol{x})\|_2^2 \leq \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2 \text{ and } t \neq 0.$$

3. We also establish an upper bound for $f(\boldsymbol{x} - tG_t(\boldsymbol{x}))$, that is

$$f(\boldsymbol{x} - tG_t(\boldsymbol{x})) \leq f(\boldsymbol{z}) + G_t(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{z}) - \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2 - \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2. \quad (4)$$

(Please refer to the remark below for the establishing process.)

Analysis for Fixed Step Size

From inequality(3), we have

$$f(\boldsymbol{x} - tG_t(\boldsymbol{x})) = g(\boldsymbol{x} - tG_t(\boldsymbol{x})) + h(\boldsymbol{x} - tG_t(\boldsymbol{x}))$$
$$\le \underbrace{g(\boldsymbol{x}) - t\nabla g(\boldsymbol{x})^\top G_t(\boldsymbol{x}) + \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2}_{\text{using bound (3) for } g(\boldsymbol{x} - tG_t(\boldsymbol{x}))} + h(\boldsymbol{x} - tG_t(\boldsymbol{x})).$$

Using upper bound of g, we arrive at

$$f(\boldsymbol{x} - tG_t(\boldsymbol{x})) = g(\boldsymbol{x} - tG_t(\boldsymbol{x})) + h(\boldsymbol{x} - tG_t(\boldsymbol{x}))$$
$$\le \underbrace{g(\boldsymbol{z}) - \nabla g(\boldsymbol{x})^\top (\boldsymbol{z} - \boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2}_{\text{using bound (1) for } g(\boldsymbol{z})} - t\nabla g(\boldsymbol{x})^\top G_t(\boldsymbol{x}) + \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2$$
$$+ h(\boldsymbol{x} - tG_t(\boldsymbol{x})).$$

Analysis for Fixed Step Size

Recall that

$$G_t(\boldsymbol{x}) = \frac{1}{t}\left(\boldsymbol{x} - \mathrm{prox}_{th}(\boldsymbol{x} - t\nabla g(\boldsymbol{x}))\right).$$

In view of properties of gradient mapping, we have

$$\begin{aligned} \boldsymbol{x} - tG_t(\boldsymbol{x}) &= \mathrm{prox}_{th}(\boldsymbol{x} - t\nabla g(\boldsymbol{x})), && (5)\\ \Rightarrow (\boldsymbol{x} - t\nabla g(\boldsymbol{x})) - (\boldsymbol{x} - tG_t(\boldsymbol{x})) &\in \partial h(\boldsymbol{x} - tG_t(\boldsymbol{x})), && (6)\\ \Rightarrow t(G_t(\boldsymbol{x}) - \nabla g(\boldsymbol{x})) &\in \partial th(\boldsymbol{x} - tG_t(\boldsymbol{x})), && (7)\\ \Rightarrow G_t(\boldsymbol{x}) - \nabla g(\boldsymbol{x}) &\in \partial h(\boldsymbol{x} - tG_t(\boldsymbol{x})). && (8) \end{aligned}$$

Analysis for Fixed Step Size

Thus,

$$
\begin{aligned}
f(\boldsymbol{x} - tG_t(\boldsymbol{x})) &= g(\boldsymbol{x} - tG_t(\boldsymbol{x})) + \boxed{h(\boldsymbol{x} - tG_t(\boldsymbol{x}))} \\
&\le g(\boldsymbol{z}) - \nabla g(\boldsymbol{x})^\top (\boldsymbol{z} - \boldsymbol{x}) - \frac{m}{2} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 \\
&\quad - t\nabla g(\boldsymbol{x})^\top G_t(\boldsymbol{x}) + \frac{t}{2} \|G_t(\boldsymbol{x})\|_2^2 \\
&\quad + \boxed{h(\boldsymbol{z}) - (G_t(\boldsymbol{x}) - \nabla g(\boldsymbol{x}))^\top (\boldsymbol{z} - \boldsymbol{x} + tG_t(\boldsymbol{x}))} \\
&= \underbrace{g(\boldsymbol{z}) + h(\boldsymbol{z})}_{f(\boldsymbol{z})} + G_t(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{z}) - \frac{t}{2} \|G_t(\boldsymbol{x})\|_2^2 - \frac{m}{2} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 .
\end{aligned}
$$

Analysis for Fixed Step Size

$$\boxed{f(\boldsymbol{x} - tG_t(\boldsymbol{x})) \le f(\boldsymbol{z}) + G_t(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{z}) - \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2 - \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2}. \tag{4}$$

- inequality (4) with $\boldsymbol{z} = \boldsymbol{x}$ shows that the algorithm is a descent method:

$$\left\{f(\boldsymbol{x}^+) \triangleq f(\boldsymbol{x} - tG_t(\boldsymbol{x}))\right\} \le f(\boldsymbol{x}) - \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2 .$$

Analysis for Fixed Step Size

$$\boxed{f(\boldsymbol{x} - tG_t(\boldsymbol{x})) \le f(\boldsymbol{z}) + G_t(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{z}) - \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2 - \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2}.$$

- inequality (4) with $\boldsymbol{z} = \boldsymbol{x}_*$ shows that

$$\begin{aligned}
f(\boldsymbol{x}^+) - f(\boldsymbol{x}_*) &\le G_t(\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{x}_*) - \frac{t}{2}\|G_t(\boldsymbol{x})\|_2^2 - \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 \\
&= \frac{1}{2t}\left(\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x} - \boldsymbol{x}_* - tG_t(\boldsymbol{x})\|_2^2\right) - \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 \\
&= \frac{1}{2t}\left((1 - mt)\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 - \left\|\boldsymbol{x}^+ - \boldsymbol{x}_*\right\|_2^2\right) && (9) \\
&\le \frac{1}{2t}\left(\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 - \left\|\boldsymbol{x}^+ - \boldsymbol{x}_*\right\|_2^2\right) \quad (\text{ since } mt \le 1) && (10)
\end{aligned}$$

Analysis for Fixed Step Size

Add inequalites (10) with $\boldsymbol{x} = \boldsymbol{x}_i$, $\boldsymbol{x}^+ = \boldsymbol{x}_{i+1}$, $t = t_i = 1/L$, from $i = 0, \ldots, k-1$,

$$
\begin{aligned}
\sum_{i=1}^{k} (f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*)) &\leq \frac{1}{2t} \sum_{i=1}^{k} \left(\|\boldsymbol{x}_i - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_*\|_2^2 \right) \\
&= \frac{1}{2t} \left(\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}_k - \boldsymbol{x}_*\|_2^2 \right) \\
&\leq \frac{1}{2t} \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .
\end{aligned}
$$

Since $f(\boldsymbol{x}_i)$ is nonincreasing,

$$
f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*) \leq \frac{1}{k} \sum_{i=1}^{k} (f(\boldsymbol{x}_i) - f(\boldsymbol{x}_*)) \leq \frac{1}{2kt} \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 \tag{11}
$$

Analysis for Fixed Step Size

Distance to Optimal Set: from (9) and $f(\boldsymbol{x}^+) \geq f(\boldsymbol{x}_*)$, we have

$$\begin{aligned} 0 \leq f(\boldsymbol{x}^+) - f(\boldsymbol{x}_*) &\leq \frac{1}{2t}\left((1-mt)\|\boldsymbol{x}-\boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}^+-\boldsymbol{x}_*\|_2^2\right) \\ \Rightarrow \|\boldsymbol{x}^+-\boldsymbol{x}_*\|_2^2 &\leq (1-mt)\|\boldsymbol{x}-\boldsymbol{x}_*\|_2^2 \\ &\leq \|\boldsymbol{x}-\boldsymbol{x}_*\|_2^2 \quad (\text{since } 0 < mt \leq 1). \end{aligned}$$

For fixed step size $t_k = 1/L$,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_*\|_2^2 \leq \left(1 - \frac{m}{L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .$$

This implies the linear convergence if g is strongly convex (m > 0).

Analysis with line search

From inequality (10), if (3) holds in iteration i, then $f(\boldsymbol{x}_{i+1}) < f(\boldsymbol{x}_i)$ and

$$t_i(f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_*)) \leq \frac{1}{2}\left(\|\boldsymbol{x}_i - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_*\|_2^2\right)$$

Adding inequalities for $i = 0$ to $k-1$ gives

$$\left(\sum_{i=0}^{k-1} t_i\right)(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)) \leq \sum_{i=0}^{k-1} t_i(f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_*)) \leq \frac{1}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .$$

First inequality holds beacause $f(\boldsymbol{x}_i)$ is nonincreasing. Since $t_i \geq t_{min}$, we obtain a similar $1/k$ bound as for fixed step size

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*) \leq \frac{1}{2kt_{min}}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .$$

Analysis with line search

Distance to Optimal Set: from inequality (10), if (3) holds in iteration i, then

$$\begin{aligned} 0 \le f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_*) &\le \frac{1}{2t}\left((1 - mt_i)\|\boldsymbol{x}_i - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_*\|_2^2\right) \\ \Rightarrow \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_*\|_2^2 &\le (1 - mt_i)\|\boldsymbol{x}_i - \boldsymbol{x}_*\|_2^2 \\ &\le (1 - mt_{min})\|\boldsymbol{x}_i - \boldsymbol{x}_*\|_2^2 \quad (\text{if } 0 < mt_{min} \le 1). \end{aligned}$$

Thus, for $c = 1 - mt_{min}$,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_*\|_2^2 \le c^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .$$

This implies the linear convergence if g is strongly convex (m > 0).

Analysis with line search

Backtracking line-search for Lipschitz constant:

(1) We initialize $L_0 = 1$ and some $\alpha > 1$.

(2) At each iteration i, we find the smallest integer t such that $L = \alpha^t L_{i-1}$, specifically:

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}_i) + \nabla f(\boldsymbol{x}_i)(\boldsymbol{x}^+ - \boldsymbol{x}_i) + \frac{L}{2}\left\|\boldsymbol{x}^+ - \boldsymbol{x}_i\right\|_2^2,$$

where $\boldsymbol{x}^+ = \text{prox}_h(\boldsymbol{x}_i - \frac{1}{L}\nabla f(\boldsymbol{x}_i))$.

(3)Update $L_i = L$ and $\boldsymbol{x}_{i+1} = \boldsymbol{x}^+$

Accelearted Proximal Gradient Method

Originally developed by [Nesterov, 2013a] and [Beck and Teboulle, 2009], we can accelerate the prixmal gradient method simply as follows

$$\begin{aligned} \boldsymbol{x}_{i+1} &= \operatorname{prox}_{th}(\boldsymbol{y}_i - t\nabla g(\boldsymbol{y}_i)), \\ \boldsymbol{y}_{i+1} &= \boldsymbol{x}_{i+1} + \beta_i(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i). \end{aligned}$$

Some simple choice for β:

1. NESTEROV07[Nesterov, 2013a]: $\beta_i = \frac{i}{i+3}$.
2. FISTA[Beck and Teboulle, 2009]: $\beta_i = \frac{\lambda_i - 1}{\lambda_{i+1}}$, where $\lambda_0 = 0$, $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_i^2}}{2}$.

Accelearted Proximal Gradient Method

It was shown in [Beck and Teboulle, 2009] that

Theorem 10

The sequences $\boldsymbol{x}_k, f(\boldsymbol{x}_k)$ generated via FISTA with either a constant or backtracking (with a ratio $\alpha \geq 1$) stepseize rule satisfy

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

Proximal Point Method

Proximal Point Method is an algorithm for minimizing a closed convex function f:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \operatorname{prox}_{tf}(\boldsymbol{x}_k) \\ &= \operatorname*{argmin}_{\boldsymbol{u}} \left(f(\boldsymbol{u}) + \frac{1}{2t} \|\boldsymbol{u} - \boldsymbol{x}_k\|_2 \right). \end{aligned}$$

1. PPM can be viewed as proximal gradient method with $g(\boldsymbol{x}) = 0$.
2. PPM is basis of the augmented Lagrangian method (coming sections).
3. PPM is related to Moreau-Yosida smoothing (coming sections).

Proximal Point Method

Theorem 11

For PPM, if f is closed and convex and optimal value f^* is finite and attained at $\boldsymbol{x}_*$. We have

$$f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2}{2\sum_{i=0}^{k-1} t_i}, \quad \text{for } k \geq 1.$$

Remark.

(1) Implies convergence if $\sum_i t_i \to \infty$

(2) Convergence rate is $1/k$ if t_i is fixed, or variable but bounded, away from zero.

(3) t_i is arbitrary; however cost of PPM evaluations will depend on t_i

Part II
Splitting Method

Different Settings for Convex Problem

1. Nonsmooth Minimization:
   $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, where $f(\boldsymbol{x})$ is convex and nonsmooth. We show the optimal solution has the following fixed point property:

$$\boldsymbol{x}_* \text{ is optimal.} \Leftrightarrow 0 \in \partial f(\boldsymbol{x}_*) \Leftrightarrow \forall \lambda > 0, \boldsymbol{x}_* = \operatorname{prox}_{\lambda f}(\boldsymbol{x}_*)$$

   The fixed point iteration gives rise to the Proximal Point Algorithm:

$$\boldsymbol{x}_{t+1} = \operatorname{prox}_{\lambda_t f}(\boldsymbol{x}_t).$$

Different Settings for Convex Problem

2. Smooth + Nonsmooth Minimization:
   $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \triangleq g(\boldsymbol{x}) + h(\boldsymbol{x})$, where $g(\boldsymbol{x})$ is convex and smooth and $h(\boldsymbol{x})$ is comvex and nonsmooth. We show the optimal solution has the following fixed point property:

$$\boldsymbol{x}_* \text{ is optimal.} \Leftrightarrow 0 \in \nabla g(\boldsymbol{x}_*) + \partial h(\boldsymbol{x}_*) \Leftrightarrow \forall \lambda > 0, \boldsymbol{x}_* = \mathrm{prox}_{\lambda h}(\boldsymbol{x}_* - \lambda \nabla g(\boldsymbol{x}_*))$$

   The fixed point iteration gives rise to the Proximal Gradient Algorithm:

$$\boldsymbol{x}_{t+1} = \mathrm{prox}_{\lambda_t h}(\boldsymbol{x}_t - \lambda_t \nabla g(\boldsymbol{x}_t)).$$

Different Settings for Convex Problem

3. Nonsmooth + Nonsmooth Minimization:
   $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \triangleq g(\boldsymbol{x}) + h(\boldsymbol{x})$, where both $g(\boldsymbol{x})$ and $h(\boldsymbol{x})$ are convex and nonsmooth.
   We show the optimal solution has the following fixed point property:

$$\boldsymbol{x}_* \text{ is optimal.} \Leftrightarrow 0 \in \partial g(\boldsymbol{x}_*) + \partial h(\boldsymbol{x}_*) \Leftrightarrow \forall \lambda > 0, \boldsymbol{x}_* = \mathrm{prox}_{\lambda(g+h)}(\boldsymbol{x}_*)$$

   The fixed point iteration gives rise to the Splitting Algorithms:

$$\boldsymbol{x}_{t+1} = \mathrm{prox}_{\lambda_t(g+h)}(\boldsymbol{x}_t).$$

However, this would require the proximal operator of the sum of two convex function, which is not alway easy to compute, even if the proximal operators of both functions separately may be easy to compute.

Different Settings for Convex Problem

Example 12

Let $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, and $h(\boldsymbol{x}) = \|A\boldsymbol{x}\|_2^2$. The proximal operator of g is given by

$$\operatorname{prox}_g(\boldsymbol{y}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{x}\|_1 \right\}.$$

The proximal operator of h (see 6.2.3 Convex Quadratic of [Beck, 2017])is given by

$$\operatorname{prox}_h(\boldsymbol{y}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|A\boldsymbol{x}\|_2^2 \right\}.$$

Both are easy to compute. However, the proximal operator of the sum of g and h is gvien by

$$\operatorname{prox}_{g+h}(\boldsymbol{y}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \|\boldsymbol{x}\|_1 + \|A\boldsymbol{x}\|_2^2 \right\}.$$

Therefore, the above fixed point property is not really useful.

Fixed Point for Nonsmooth Composition

Theorem 13 (Fixed Point)

Consider $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \triangleq g(\boldsymbol{x}) + h(\boldsymbol{x})$, where both $g(\boldsymbol{x})$ and $h(\boldsymbol{x})$ are convex and nonsmooth. If $\boldsymbol{x}_*$ is optimal if and only if for any $\lambda > 0$ and $\rho \in \mathbb{R}$,

$$\boldsymbol{x}_* = \mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) \quad \text{and} \quad \boldsymbol{y}_* = \boldsymbol{y}_* + \rho\left[\mathrm{prox}_{\lambda h}\left(\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*}\right) - \mathrm{prox}_{\lambda h}(\boldsymbol{y}_*)\right]$$

Proof. Suppose that $\boldsymbol{x}_*$ is optimal.

1. $\Leftrightarrow 0 \in \partial g(\boldsymbol{x}_*) + \partial h(\boldsymbol{x}_*)$.
2. $\Leftrightarrow \forall\lambda > 0$, there exists $\boldsymbol{z}$ such that $\boldsymbol{z} \in \partial(\lambda g)(\boldsymbol{x}_*)$ and $-\boldsymbol{z} \in \partial(\lambda h)(\boldsymbol{x}_*)$.
3. $\Leftrightarrow \forall\lambda > 0$, there exists $\boldsymbol{y}$ such that $\boldsymbol{y} - \boldsymbol{x}_* \in \partial(\lambda g)(\boldsymbol{x}_*)$ and $\boldsymbol{x}_* - \boldsymbol{y} \in \partial(\lambda h)(\boldsymbol{x}_*)$.

Fixed Point for Nonsmooth Composition

4 $\Leftrightarrow \forall \lambda > 0$, there exists y such that $\boldsymbol{y} - \boldsymbol{x}_* \in \partial(\lambda g)(\boldsymbol{x}_*)$ and $2\boldsymbol{x}_* - \boldsymbol{y} \in \boldsymbol{x}_* + \partial(\lambda h)(\boldsymbol{x}_*)$.

5 $\Leftrightarrow \boldsymbol{x}_* = \mathrm{prox}_{\lambda g}(\boldsymbol{y})$, thus we have $\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}) - \boldsymbol{y}} \in \boldsymbol{x}_* + \partial(\lambda h)(\boldsymbol{x}_*)$.

6 $\Leftrightarrow \boldsymbol{x}_* = \mathrm{prox}_{\lambda g}(\boldsymbol{y})$, and $\boldsymbol{x}_* = \mathrm{prox}_{\lambda h}(\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}) - \boldsymbol{y}})$.

7 $\Leftrightarrow \boldsymbol{x}_* = \mathrm{prox}_{\lambda g}(\boldsymbol{y})$, and $\forall \rho$,

$$\boldsymbol{y} = \boldsymbol{y} + \rho[\underbrace{\mathrm{prox}_{\lambda h}(\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}) - \boldsymbol{y}})}_{\boldsymbol{x}_*} - \underbrace{\mathrm{prox}_{\lambda g}(\boldsymbol{y})}_{\boldsymbol{x}_*}].$$

All the statements are equivalent. $\square$

Fixed Point for Nonsmooth Composition

Define $\mathrm{refl}_f(\boldsymbol{x}) \triangleq 2\mathrm{prox}_f(\boldsymbol{x}) - \boldsymbol{x}$.

1. When $\rho = 1$, we have

$$\begin{aligned}\boldsymbol{y}_* &= \boldsymbol{y}_* + \left[\mathrm{prox}_{\lambda h}(2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*) - \mathrm{prox}_{\lambda g}(\boldsymbol{y}_*)\right] \\ &= \frac{1}{2}[\boldsymbol{y}_* + 2\mathrm{prox}_{\lambda h}(\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*}) - (\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*})] \\ &= \frac{1}{2}[\boldsymbol{y}_* + \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{y}_*)].\end{aligned}$$

Hence, $\boldsymbol{y}_* = \mathcal{T}_{\lambda,g,h}(\boldsymbol{y}_*)$ with operator

$$\mathcal{T}_{\lambda,g,h} = \frac{1}{2}[I + \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}].$$

This is known as the Douglas-Rachford operator[Lions and Mercier, 1979].

Fixed Point for Nonsmooth Composition

2 When $\rho = 2$, we have

$$\begin{aligned} \boldsymbol{y}_* &= \boldsymbol{y}_* + 2\left[\mathrm{prox}_{\lambda h}(2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*) - \mathrm{prox}_{\lambda g}(\boldsymbol{y}_*)\right] \\ &= 2\mathrm{prox}_{\lambda h}(\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*}) - (\boxed{2\mathrm{prox}_{\lambda g}(\boldsymbol{y}_*) - \boldsymbol{y}_*}) \\ &= \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{y}_*). \end{aligned}$$

Hence, $\boldsymbol{y}_* = \mathcal{T}_{\lambda,g,h}(\boldsymbol{y}_*)$ with operator

$$\mathcal{T}_{\lambda,g,h} = \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}.$$

This is known as the Peaceman-Rachford operator[Lions and Mercier, 1979].

Non-expansive

Lemma 14

The reflection operator $\mathrm{refl}_{\lambda f}(\cdot)$ is non-expansive for any $\lambda > 0$.

Proof. This is beacause

$$
\begin{aligned}
\|\mathrm{refl}_{\lambda f}(\boldsymbol{x}) - \mathrm{refl}_{\lambda f}(\boldsymbol{y})\|_2^2 &= \left\|2\mathrm{prox}_{\lambda f}(\boldsymbol{x}) - 2\mathrm{prox}_{\lambda f}(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})\right\|_2^2 \\
&= 4\left\|\mathrm{prox}_{\lambda f}(\boldsymbol{x}) - \mathrm{prox}_{\lambda f}(\boldsymbol{y})\right\|_2^2 \\
&\quad - 4\langle \mathrm{prox}_{\lambda f}(\boldsymbol{x}) - \mathrm{prox}_{\lambda f}(\boldsymbol{y}),\ \boldsymbol{x} - \boldsymbol{y}\rangle + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\
&\leq 4\left\|\mathrm{prox}_{\lambda f}(\boldsymbol{x}) - \mathrm{prox}_{\lambda f}(\boldsymbol{y})\right\|_2^2 - 4\left\|\mathrm{prox}_{\lambda f}(\boldsymbol{x}) - \mathrm{prox}_{\lambda f}(\boldsymbol{y})\right\|_2^2 \\
&\quad + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\
&= \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 .
\end{aligned}
$$

Non-expansive

The proximal operator is firmly nonexpansiv, i.e.,

$$\left\|\operatorname{prox}_{\lambda f}(\boldsymbol{x}) - \operatorname{prox}_{\lambda f}(\boldsymbol{y})\right\|_2^2 \leq \langle \operatorname{prox}_{\lambda f}(\boldsymbol{x}) - \operatorname{prox}_{\lambda f}(\boldsymbol{y}),\ \boldsymbol{x} - \boldsymbol{y} \rangle.$$

Non-expansive

Lemma 15

Both the Peaceman-Rachford operator $\mathcal{T}_1$ and the Douglas-Rachford operator $\mathcal{T}_2$ are non-expansive.

Proof. For $\mathcal{T}_1$,

$$\begin{aligned}
\|\mathcal{T}_1 \boldsymbol{x} - \mathcal{T}_1 \boldsymbol{y}\|_2^2 &= \|\mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{x}) - \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{y})\|_2^2 \\
&\leq \|\mathrm{refl}_{\lambda g}(\boldsymbol{x}) - \mathrm{refl}_{\lambda g}(\boldsymbol{y})\|_2^2 \\
&\leq \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 .
\end{aligned}$$

For $\mathcal{T}_2$,

$$\|\mathcal{T}_2 \boldsymbol{x} - \mathcal{T}_2 \boldsymbol{y}\|_2^2 \leq \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{2} \|\mathcal{T}_1 \boldsymbol{x} - \mathcal{T}_2 \boldsymbol{y}\|_2^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 .$$

$\square$

Splitting Algorithm

The fixed point iteration correspoinding to these non-expansive operators lead to the following algorithm:

1. Douglas-Rachford Splitting Algorithm:

$$\boldsymbol{y}_{t+1} = \frac{1}{2}[\boldsymbol{y}_t + \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{y}_t)].$$

2. Peaceman-Rachford Splitting Algorithm:

$$\boldsymbol{y}_{t+1} = \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{y}_t).$$

3. relaxed Peaceman-Rachford Splitting Algorithm: let $\gamma_t \in [0, 1]$,

$$\boldsymbol{y}_{t+1} = (1 - \gamma_t)\boldsymbol{y}_t + \gamma_t \mathrm{refl}_{\lambda h} \circ \mathrm{refl}_{\lambda g}(\boldsymbol{y}_t).$$

When $\gamma_t = 1$, Relaxed PR is PR. And if $\gamma_t = 1/2$, Relaxed PR is DR.

Splitting Algorithm

Let us initialize $\boldsymbol{y}_1$ and $\boldsymbol{x}_1 = \mathrm{prox}_{\lambda g}(\boldsymbol{y}_1)$, the DR algorithm can be rewritten as

$$\begin{aligned}
\boldsymbol{y}_{t+1} &= \boldsymbol{y}_t + \mathrm{prox}_{\lambda h}(2\boldsymbol{x}_t - \boldsymbol{y}_t) - \boldsymbol{x}_t, \\
\boldsymbol{x}_{t+1} &= \mathrm{prox}_{\lambda g}(\boldsymbol{y}_{t+1}).
\end{aligned}$$

Let $\boldsymbol{z}_{t+1} = \mathrm{prox}_{\lambda h}(2\boldsymbol{x}_t - \boldsymbol{y}_t)$ and switch $\boldsymbol{x}$ and $\boldsymbol{y}$ updates, this can be further formulated as

$$\begin{aligned}
\boldsymbol{z}_{t+1} &= \mathrm{prox}_{\lambda h}(2\boldsymbol{x}_t - \boldsymbol{y}_t), \\
\boldsymbol{x}_{t+1} &= \mathrm{prox}_{\lambda g}(\boldsymbol{y}_t + \boldsymbol{z}_{t+1} - \boldsymbol{x}_t), \\
\boldsymbol{y}_{t+1} &= \boldsymbol{y}_t + \boldsymbol{z}_{t+1} - \boldsymbol{x}_t.
\end{aligned}$$

Splitting Algorithm

Let $\boldsymbol{u}_t = \boldsymbol{x}_t - \boldsymbol{y}_t$. We have

$$\begin{aligned}
\boldsymbol{z}_{t+1} &= \operatorname{prox}_{\lambda h}(\boldsymbol{x}_t + \boldsymbol{u}_t), \\
\boldsymbol{x}_{t+1} &= \operatorname{prox}_{\lambda g}(\boldsymbol{z}_{t+1} - \boldsymbol{u}_t), \\
\boldsymbol{u}_{t+1} &= \boldsymbol{u}_t + (\boldsymbol{x}_{t+1} - \boldsymbol{z}_{t+1}),
\end{aligned}$$

The above is a special case of the Alternating Direction Methods of Multipliers (ADMM).

Related to ADMM

We consider the following optimization problem

$$\begin{aligned} &\min g(\boldsymbol{x}) + h(\boldsymbol{z}) \\ \text{s.t.} \quad &A\boldsymbol{x} + B\boldsymbol{z} = c. \end{aligned}$$

The iteration of ADMM is

$$\begin{aligned} \boldsymbol{z}_{t+1} &= \operatorname*{argmin}_{\boldsymbol{z}} \{h(\boldsymbol{z}) + \frac{\rho}{2}\|A\boldsymbol{x}_t + B\boldsymbol{z}_t - c + \boldsymbol{u}_t\|_2^2\}, \\ \boldsymbol{x}_{t+1} &= \operatorname*{argmin}_{\boldsymbol{x}} \{h(\boldsymbol{z}) + \frac{\rho}{2}\|A\boldsymbol{x}_t + B\boldsymbol{z}_{t+1} - c + \boldsymbol{u}_t\|_2^2\}, \\ \boldsymbol{u}_{t+1} &= \boldsymbol{u}_t + (A\boldsymbol{x}_{t+1} + B\boldsymbol{z}_{t+1} - c). \end{aligned}$$

Let $A = I, B = -I, c = 0$. One can see that the above problem is exactly the same as the Douglas-Rachford splitting algorthm in this case.

Convergence Analysis

Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a nonexpansive operator. We consider the relaxed fixed point algorithm

$$\boldsymbol{x}_{t+1} = (1-\gamma_t)\boldsymbol{x}_t + \gamma_t \cdot \mathcal{T}\boldsymbol{x}_t, \quad \text{for all } t \geq 0$$

where $\gamma_t \in (0,1]$. This is known as the Krasnosel'skii-Mann (KM) algorithm. Note that when $\gamma_t = 1$, this reduce to the usual fixed point algorithm.

Theorem 16

Let $\mathcal{T}$ be a nonexpansive operator and not a self-map. The KM algorithm satisfies that

$$\|\mathcal{T}\boldsymbol{x}_t - \boldsymbol{x}_t\|_2^2 \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2}{\sum_{\tau=0}^{t} \gamma_\tau(1-\gamma_\tau)}.$$

Convergence Analysis

Proof. First, $\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2$ is non-increasing, since

$$\begin{aligned}
\|\boldsymbol{x}_{t+1} - \mathcal{T}\boldsymbol{x}_{t+1}\|_2^2 &= \|(1-\gamma_t)\boldsymbol{x}_t + \gamma_t\mathcal{T}\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_{t+1}\|_2^2 \\
&= \|(1-\gamma_t)(\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t) + \mathcal{T}\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_{t+1}\|_2^2 \\
&\leq (1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 + \|\mathcal{T}\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_{t+1}\|_2^2 \quad \text{(Triangular inequality)} \\
&\leq (1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 + \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|_2^2 \quad (\mathcal{T} \text{ is a nonexpansive operator}) \\
&= (1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 + \gamma_t\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 \\
&= \|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 .
\end{aligned}$$

To be continued ...

Convergence Analysis

Proof. (continued) We now show that

$$
\begin{aligned}
\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_*\|_2^2 &= \|(1-\gamma_t)\boldsymbol{x}_t + \gamma_t \mathcal{T}\boldsymbol{x}_t - (1-\gamma_t)\boldsymbol{x}_* - \gamma_t \mathcal{T}\boldsymbol{x}_*\|_2^2 \\
&= \|(1-\gamma_t)(\boldsymbol{x}_t - \boldsymbol{x}_*) + \gamma_t(\mathcal{T}\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_*)\|_2^2 \\
&= (1-\gamma_t)\|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 + \gamma_t \|\mathcal{T}\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_*\|_2^2 - \gamma_t(1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 \\
&= \|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 - \underbrace{\gamma_t\left(\|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 - \|\mathcal{T}\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_*\|_2^2\right)}_{\geq 0 \text{ since } \mathcal{T} \text{ is non-expansive.}} - \gamma_t(1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 \\
&\leq \|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 - \gamma_t(1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 .
\end{aligned}
$$

The third equality is due to the fact that

$$
\|(1-\gamma_t)\boldsymbol{u} + \gamma_t \boldsymbol{v}\|_2^2 = (1-\gamma_t)\|\boldsymbol{u}\|_2^2 + \gamma_t\|\boldsymbol{v}\|_2^2 - \gamma_t(1-\gamma_t)\|\boldsymbol{u} - \boldsymbol{v}\|_2^2 .
$$

Convergence Analysis

Proof. (continued) We now have that

$$\boxed{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_*\|_2^2 \le \|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2^2 - \gamma_t(1-\gamma_t)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2}.$$

Taking summation over t leads to

$$\begin{aligned}\sum_{\tau=0}^{t} \gamma_\tau(1-\gamma_\tau)\|\boldsymbol{x}_\tau - \mathcal{T}\boldsymbol{x}_\tau\|_2^2 &\le \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_*\|_2^2 \\ &\le \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .\end{aligned}$$

Thus, we obtain

$$\left(\sum_{\tau=0}^{t} \gamma_\tau(1-\gamma_\tau)\right)\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 \le \sum_{\tau=0}^{t} \gamma_\tau(1-\gamma_\tau)\|\boldsymbol{x}_\tau - \mathcal{T}\boldsymbol{x}_\tau\|_2^2 \le \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2 .$$

Convergence Analysis

If we set $\gamma_t = \gamma \in (0, 1)$, we have

$$\|\boldsymbol{x}_t - \mathcal{T}\boldsymbol{x}_t\|_2^2 \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2}{\sum_{\tau=0}^{t} \gamma_\tau(1 - \gamma_\tau)} = \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2}{\gamma(1 - \gamma)(t + 1)}.$$

Appendix: Contraction Mapping

Let $(\mathcal{X}, D)$ be a metric space, and let $\mathcal{T} : \mathcal{X} \to \mathcal{X}$ be an operator. $\mathcal{T}$ is considered a contraction mapping if there exists a constant $0 \leq k < 1$ such that, for any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$, the distance between $\mathcal{T}(\boldsymbol{x})$ and $\mathcal{T}(\boldsymbol{y})$ is less than or equal to the contraction factor times the distance between and $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$D(\mathcal{T}(\boldsymbol{x}), \mathcal{T}(\boldsymbol{y})) \leq k \cdot D(\boldsymbol{x}, \boldsymbol{y}).$$

let $(\mathcal{X}, D)$ be a metric space, and let $\mathcal{T} : \mathcal{X} \to \mathcal{X}$ be an operator. $\mathcal{T}$ is considered non-expansive if, for any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$, the distance between $\mathcal{T}(\boldsymbol{x})$ and $\mathcal{T}(\boldsymbol{y})$ is less than or equal to the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$D(\mathcal{T}(\boldsymbol{x}), \mathcal{T}(\boldsymbol{y})) \leq D(\boldsymbol{x}, \boldsymbol{y}).$$

Appendix: Fixed-Point Theorem

The Banach Fixed-Point Theorem states that if a mapping $\mathcal{T}$ is a contraction mapping on a complete metric space, then it has a unique fixed point and any sequence generated by iteratively applying the mapping will converge to that fixed point.

In the case of a non-expansive operator, which is a weaker condition than contraction, the guarantee for convergence depends on the specific properties of the operator and the space it operates on. If the non-expansive operator is defined on a compact set or operates in a finite-dimensional space, convergence is guaranteed. This is because compact sets and finite-dimensional spaces have certain properties that ensure convergence for non-expansive mappings.

Appendix: Fixed-Point Theorem

However, in an infinite-dimensional space, the convergence of a non-expansive operator is NOT always guaranteed. Additional conditions, such as convexity or strong monotonicity, may be required to ensure convergence in such cases.

Therefore, while a non-expansive operator with a fixed point indicates the potential for convergence, the actual convergence depends on the specific properties of the operator and the space in which it operates.

References I

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013a.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

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

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

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

References II

L. Vandenberghe. Proximal gradient method, course ece236c. URL `http://www.seas.ucla.edu/~vandenbe/236C/lectures/proxgrad.pdf`.

Niao He. Big data optimizaton course ece236c. URL `http://niaohe.ise.illinois.edu/IE598_2016/index.html`.

Thank You!

Email:qianhui@zju.edu.cn