1 Outline

In this lecture, we study

- set-valued operators and their properties,
- Banach's fixed-point theorem for contractive operators, and
- Krasnosel'skii-Mann iteration for averaged operators.

2 Fixed-point Iteration

Recall that the proximal point algorithm is given by

$$x_{t+1} = \mathrm{prox}_{\eta f}(x_t) = \operatorname*{argmin}_x \left\{ f(x) + \frac{1}{2\eta}\|x - x_t\|^2 \right\}.$$

Moreover, we learned that the following statements are equivalent:

- $\hat{x} = \mathrm{prox}_{\eta f}(\hat{x})$.
- $\hat{x} \in \mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$.

This is because $\hat{x} = \mathrm{prox}_{\eta f}(\hat{x})$ holds if and only if $0 \in \partial f(\hat{x})$, which is the optimality condition for $\hat{x}$ to be a minimizer of f. We also learned that the proximal point algorithm converges to a minimizer of f at a rate of $O(1/T)$. In this lecture, we will see that the proximal point algorithm can be viewed as a so-called **fixed-point iteration**, and we will analyze its convergence using this perspective.

2.1 Set-Valued Operators

A set-valued operator, or relation, is a mapping $\mathbb{T} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ that maps each point $x \in \mathbb{R}^d$ to a set $\mathbb{T}(x) \subseteq \mathbb{R}^d$. An operator $\mathbb{T}$ is single-valued if $\mathbb{T}(x)$ is a singleton or empty for every $x \in \mathbb{R}^d$. Let us define some relevant concepts for set-valued operators.

- **Domain**: $\mathrm{domain}(\mathbb{T}) = \{x \in \mathbb{R}^d : \mathbb{T}(x) \neq \emptyset\}$.
- **Range**: $\mathrm{range}(\mathbb{T}) = \{y \in \mathbb{R}^d : y \in \mathbb{T}(x), x \in \mathbb{R}^d\}$.
- **Graph**: $\mathrm{gra}(\mathbb{T}) = \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : y \in \mathbb{T}(x)\}$.

Moreover, given two operators $\mathbb{T}$ and $\mathbb{S}$, we can define some operations on them as follows:

- **Sum**: $(\mathbb{T} + \mathbb{S})(x) = \{y + z : y \in \mathbb{T}(x), z \in \mathbb{S}(x)\}$.
- **Composition**: $(\mathbb{T} \circ \mathbb{S})(x) = \{y : z \in \mathbb{S}(x), y \in \mathbb{T}(z)\}$.

Based on these, we may define the identity and zero operators as follows:

- **Identity operator**: $\mathbb{I}(x) = \{x\}$.
- **Zero operator**: $\mathbb{O}(x) = \{0\}$.

Hence, we have $\mathbb{T} + \mathbb{O} = \mathbb{T}$, $\mathbb{T} \circ \mathbb{I} = \mathbb{T}$, and $\mathbb{I} \circ \mathbb{T} = \mathbb{T}$ for any operator $\mathbb{T}$. Lastly, the inverse of an operator $\mathbb{T}$ is defined as

$$\mathbb{T}^{-1}(y) = \{x : y \in \mathbb{T}(x)\}.$$

As $x \in T^{-1}(y)$ if and only if $y \in \mathbb{T}(x)$, we have

$$\mathrm{gra}(\mathbb{T}^{-1}) = \{(y, x) : (x, y) \in \mathrm{gra}(\mathbb{T})\}.$$

Moreover, we have $(T^{-1})^{-1} = \mathbb{T}$ and $\mathrm{domain}(\mathbb{T}^{-1}) = \mathrm{range}(\mathbb{T})$.

2.2 Fixed Points and Banach's Fixed-Point Theorem

For $L > 0$, we say that an operator $\mathbb{T}$ is **L-Lipschitz** if for any $(x, u), (y, v) \in \mathrm{gra}(\mathbb{T})$,

$$\|u - v\| \leq L\|x - y\|.$$

We also write it as

$$\|T(x) - T(y)\| \leq L\|x - y\|, \forall x, y \in \mathrm{domain}(\mathbb{T}).$$

Note that if $\mathbb{T}$ is L-Lipschitz, then $\mathbb{T}$ is single-valued because for any $(x, u), (x, v) \in \mathrm{gra}(\mathbb{T})$, we have $\|u - v\| \leq L\|x - x\| = 0$, which implies $u = v$.

We say that $\mathbb{T}$ is **non-expansive** if $\mathbb{T}$ is 1-Lipschitz, and $\mathbb{T}$ is **contractive** if $\mathbb{T}$ is L-Lipschitz for some $L \in (0, 1)$.

Given a single-valued operator $\mathbb{T}$, we say that $x \in \mathbb{R}^d$ is a **fixed point** of $\mathbb{T}$ if $\mathbb{T}(x) = x$. We denote the set of fixed points of $\mathbb{T}$ as

$$\mathrm{fix}(\mathbb{T}) = \{x \in \mathbb{R}^d : x \in \mathbb{T}(x)\} = (\mathbb{I} - \mathbb{T})^{-1}(0).$$

Proposition 19.1. *If $\mathbb{T}$ is nonexpansive and $\mathrm{dom}(\mathbb{T}) = \mathbb{R}^d$, then $\mathrm{fix}(\mathbb{T})$ is closed and convex.*

Proof. Since $\mathbb{T}$ is nonexpansive, $\mathbb{T}$ is continuous, and thus, $\mathbb{I} - \mathbb{T}$ is continuous. Hence, $\mathrm{fix}(\mathbb{T}) = (\mathbb{I} - \mathbb{T})^{-1}(0)$ is closed. For convexity, take $x, y \in \mathrm{fix}(\mathbb{T})$ and $\lambda \in [0, 1]$. Consider $z = \lambda x + (1 - \lambda)y$. Note that

$$\|\mathbb{T}(z) - x\| \leq \|z - x\| = (1 - \lambda)\|y - x\|,$$

and

$$\|\mathbb{T}(z) - y\| \leq \|z - y\| = \lambda\|x - y\|.$$

Hence, we have

$$\|\mathbb{T}(z) - x\| + \|\mathbb{T}(z) - y\| \leq \|x - y\|.$$

On the other hand, by the triangle inequality, we have

$$\|\mathbb{T}(z) - x\| + \|\mathbb{T}(z) - y\| \geq \|x - y\|.$$

Combining the above two inequalities, we have $\|\mathbb{T}(z) - x\| + \|\mathbb{T}(z) - y\| = \|x - y\|$. Hence, $\mathbb{T}(z)$ lies on the line segment between x and y. This implies $\mathbb{T}(z) = \lambda x + (1 - \lambda)y = z$, and thus, $z \in \mathrm{fix}(\mathbb{T})$. □

Then we are interested in the following algorithm, called the **fixed-point iteration**:

$$x_{t+1} = \mathbb{T}(x_t).$$

The algorithm is also called the **Picard iteration**. In general, the fixed-point iteration may not converge, even if $\mathbb{T}$ is non-expansive. One example is the operator $\mathbb{T}(x) = -x$ for $x \in \mathbb{R}$. More generally, a rotation operator would not necessarily lead to covergence.

However, if $\mathbb{T}$ is contractive, then the fixed-point iteration converges to a unique fixed point of $\mathbb{T}$ at a linear rate. This is known as **Banach's fixed-point theorem**.

Theorem 19.2 (Banach's fixed-point theorem)**.** *If $\mathbb{T}$ is contractive, then* $\text{fix}(\mathbb{T})$ *is a singleton, and the fixed-point iteration converges to the unique fixed point of $\mathbb{T}$ at a linear rate.*

Proof. Since $\mathbb{T}$ is contractive, there exists $L \in (0, 1)$ such that $\|\mathbb{T}(x) - \mathbb{T}(y)\| \le L\|x - y\|$ for any $x, y \in \text{domain}(\mathbb{T})$.

For convergence, note that $\|x_{t+1} - x_t\| \le L^t\|x_1 - x_0\|$. Hence, for any $s > t$, we have

$$\|x_s - x_t\| \le \sum_{i=t}^{s-1} \|x_{i+1} - x_i\| \le \sum_{i=t}^{s-1} L^i\|x_1 - x_0\| \le \frac{L^t}{1-L}\|x_1 - x_0\|.$$

This implies that $\{x_t\}$ is a Cauchy sequence in $\mathbb{R}^d$, and thus, $\{x_t\}$ converges to some point x^*.

Note that $\mathbb{T}$ is continuous, and thus, we have $\mathbb{T}(x^*) = \lim_{t\to\infty} \mathbb{T}(x_t) = \lim_{t\to\infty} x_{t+1} = x^*$. Hence, x^* is a fixed point of $\mathbb{T}$.

For uniqueness, take any $x, y \in \text{fix}(\mathbb{T})$. We have

$$\|x - y\| = \|\mathbb{T}(x) - \mathbb{T}(y)\| \le L\|x - y\|.$$

Since $L < 1$, we must have $\|x - y\| = 0$, which implies $x = y$. Hence, $\text{fix}(\mathbb{T})$ is a singleton. $\square$

Example 19.3. When f is α-strongly convex and β-smooth,

$$\mathbb{T} = I - \eta\nabla f$$

is a contractive operator for any $\eta \in (0, 2/(\alpha + \beta))$. Moreover, the fixed-point iteration of $\mathbb{T}$ is exactly the gradient descent algorithm with step size η. Hence, Banach's fixed-point theorem implies that gradient descent converges to the unique minimizer of f at a linear rate for any step size $\eta \in (0, 2/(\alpha + \beta))$.

However, when f is smooth but not strongly convex, $\mathbb{T} = I - \eta\nabla f$ is not contractive for any $\eta > 0$. Hence, Banach's fixed-point theorem does not apply to the gradient descent algorithm in this case. However, we will see that the corresponding fixed-point iteration still converges.

2.3 Averaged Operators and Krasnosel'skii-Mann Iteration

We say that an opeator $\mathbb{T}$ is **θ-averaged** for some $\theta \in (0, 1)$ if $\mathbb{T} = (1 - \theta)\mathbb{I} + \theta\mathbb{S}$ for some non-expansive operator $\mathbb{S}$.

Example 19.4. When f is convex and β-smooth, $\mathbb{T} = I - \eta\nabla f$ is θ-averaged for any $\eta \in (0, 2/\beta)$ with $\theta = \eta\beta/2$. To see this, let $\mathbb{S} = I - \frac{2}{\beta}\nabla f$. Note that $\mathbb{S}$ is non-expansive because for any

$x, y \in \mathbb{R}^d$,

$$\begin{aligned}
\|\mathbb{S}(x) - \mathbb{S}(y)\|^2 &= \|x-y\|^2 - \frac{4}{\beta}\langle \nabla f(x) - \nabla f(y), x-y\rangle + \frac{4}{\beta^2}\|\nabla f(x) - \nabla f(y)\|^2 \\
&\le \|x-y\|^2 - \frac{4}{\beta}\langle \nabla f(x) - \nabla f(y), x-y\rangle + \frac{4}{\beta}\langle \nabla f(x) - \nabla f(y), x-y\rangle \\
&= \|x-y\|^2
\end{aligned}$$

where the inequality holds due to the co-coercivity of ∇f. Moreover, we have $\mathbb{T} = (1-\theta)\mathbb{I} + \theta\mathbb{S}$ for $\theta = \eta\beta/2$.

Theorem 19.5 (Krasnosel'skii-Mann Iteration)**.** *Let $\mathbb{T}$ be θ-averaged for some $\theta \in (0,1)$ with $\mathrm{fix}(\mathbb{T}) \neq \emptyset$. Then the fixed-point iteration $x_{t+1} = \mathbb{T}(x_t)$ with any starting point $x_0 \in \mathbb{R}^d$ converges to a fixed point of $\mathbb{T}$. Moreover, we have*

$$\mathrm{dist}(x_t, \mathrm{fix}(\mathbb{T})) \to 0,$$

and

$$\|x_{t+1} - x_t\|^2 \le \frac{\theta}{(t+1)(1-\theta)}\mathrm{dist}^2(x_0, \mathrm{fix}(\mathbb{T})).$$

Proof. Let $z \in \mathrm{fix}(\mathbb{T})$ be a fixed point of $\mathbb{T}$. Then $z = \mathbb{T}(z) = (1-\theta)z + \theta\mathbb{S}(z)$, which implies $\mathbb{S}(z) = z$. Hence, z is also a fixed point of $\mathbb{S}$. Note that

$$\begin{aligned}
\|x_{t+1} - z\|^2 &= \|(1-\theta)(x_t - z) + \theta(\mathbb{S}(x_t) - \mathbb{S}(z))\|^2 \\
&= (1-\theta)\|x_t - z\|^2 + \theta\|\mathbb{S}(x_t) - \mathbb{S}(z)\|^2 - \theta(1-\theta)\|x_t - \mathbb{S}(x_t) + z - \mathbb{S}(z)\|^2 \\
&\le (1-\theta)\|x_t - z\|^2 + \theta\|x_t - z\|^2 - \theta(1-\theta)\|x_t - \mathbb{S}(x_t)\|^2 \\
&= \|x_t - z\|^2 - \theta(1-\theta)\|x_t - \mathbb{S}(x_t)\|^2.
\end{aligned}$$

Therefore, for any fixed point z of $\mathbb{T}$, we have $\|x_{t+1} - z\| \le \|x_t - z\|$. This implies that $\mathrm{dist}(x_{t+1}, \mathrm{fix}(\mathbb{T})) \le \mathrm{dist}(x_t, \mathrm{fix}(\mathbb{T}))$. Moreover,

$$\|x_{t+1} - z\|^2 \le \|x_0 - z\|^2 - \theta(1-\theta)\sum_{i=0}^{t}\|x_i - \mathbb{S}(x_i)\|^2 = \|x_0 - z\|^2 - \frac{(1-\theta)}{\theta}\sum_{i=0}^{t}\|x_i - \mathbb{T}(x_i)\|^2.$$

Hence, we have

$$\sum_{i=0}^{t}\|x_i - \mathbb{T}(x_i)\|^2 \le \frac{\theta}{1-\theta}\|x_0 - z\|^2.$$

As $\mathbb{T}$ is non-expansive, $\|x_{t+1} - x_t\| \le \|x_t - x_{t-1}$, and moreover, we have

$$\|x_t - \mathbb{T}(x_t)\|^2 \le \frac{\theta}{(t+1)(1-\theta)}\|x_0 - z\|^2.$$

For convergence to a fixed point, note that $\{x_t\}$ is bounded because $\|x_t - z\| \le \|x_0 - z\|$ for any fixed point z of $\mathbb{T}$. Hence, there exists a subsequence $\{x_{t_i}\}$ that converges to some point x^*. Note that $\|x_{t_i} - \mathbb{T}(x_{t_i})\| \to 0$ as $i \to \infty$. Since $\mathbb{T}$ is continuous, we have $\|x^* - \mathbb{T}(x^*)\| = 0$, which implies $x^* \in \mathrm{fix}(\mathbb{T})$. Hence, the whole sequence $\{x_t\}$ converges to x^*. $\square$

Example 19.6. $\mathbb{T} = \mathrm{prox}_{\eta f}$ is 1/2-averaged. To see this, let $\mathbb{S} = 2\mathrm{prox}_{\eta f} - \mathbb{I}$. Note that $\mathbb{S}$ is non-expansive because for any $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
\|\mathbb{S}(x) - \mathbb{S}(y)\|^2 &= 4\|\mathrm{prox}_{\eta f}(x) - \mathrm{prox}_{\eta f}(y)\|^2 - 4\langle \mathrm{prox}_{\eta f}(x) - \mathrm{prox}_{\eta f}(y), x - y\rangle + \|x - y\|^2 \\
&\leq 4\|\mathrm{prox}_{\eta f}(x) - \mathrm{prox}_{\eta f}(y)\|^2 - 4\|\mathrm{prox}_{\eta f}(x) - \mathrm{prox}_{\eta f}(y)\|^2 + \|x - y\|^2 \\
&= \|x - y\|^2
\end{aligned}$$

where the inequality holds due to the co-coercivity of $\mathrm{prox}_{\eta f}$. The co-coercivity holds because $u = \mathrm{prox}_{\eta f}(x)$ and $v = \mathrm{prox}_{\eta f}(y)$ satisfy $u - x \in \eta \partial f(u)$ and $v - y \in \eta \partial f(v)$, which implies $\langle u - v, u - x - (v - y)\rangle = \eta\langle u - v, g_u - g_v\rangle \geq \|u - v\|^2$ for any $g_u \in \partial f(u)$ and $g_v \in \partial f(v)$. Moreover, we have $\mathbb{T} = (1/2)\mathbb{I} + (1/2)\mathbb{S}$.

When we have non-expansive operators $\mathbb{T}_1, \ldots, \mathbb{T}_k$, the composition $\mathbb{T}_1 \circ \cdots \circ \mathbb{T}_k$ is also non-expansive. Furthermore, if the operators are contractive, then their composition is also contractive. On the other hand, if $\mathbb{T}_i$ is θ_i-averaged for some $\theta_i \in (0, 1)$ for each $i \in [k]$, then $\mathbb{T}_1 \circ \cdots \circ \mathbb{T}_k$ is θ-averaged for

$$\theta = \frac{k}{k - 1 + \frac{1}{\max_{i \in [k]} \theta_i}}.$$

Example 19.7. Recall that $I - \eta\nabla g$ is θ-averaged for any $\eta \in (0, 2/\beta)$ with $\theta = \eta\beta/2$ when g is convex and β-smooth. Moreover, $\mathrm{prox}_{\eta h}$ is 1/2-averaged. Hence, the composition

$$\mathrm{prox}_{\eta h} \circ (I - \eta\nabla g)$$

is averaged. Here, the fixed-point iteration of $\mathrm{prox}_{\eta h} \circ (I - \eta\nabla g)$ is exactly the proximal gradient algorithm with step size η.