Introductory Course on Non-smooth Optimisation

Lecture 03 - Krasnosel'skiĭ-Mann iteration

Outline

1. Method of alternating projection
2. Monotone and non-expansive mappings
3. Krasnosel'skiĭ-Mann iteration
4. "Accelerated" Krasnosel'skiĭ-Mann iteration

Recap

Recap of descent methods:

- include gradient descent, proximal gradient descent
- convergence (rate) properties
  - objective function value
    - $O(1/k)$ convergence rate
    - optimal $O(1/k^2)$ convergence rate
  - sequence
    - $O(1/\sqrt{k})$ convergence rate
    - optimal $O(1/k)$ convergence rate
  - linear convergence under *e.g.* strong convexity

Operator splitting

Consider the problem

$$\min_{x\in\mathbb{R}^n} \mu_1\|x\|_1 + \mu_2\|\nabla x\|_1 + \frac{1}{2}\|Ax - f\|^2.$$

In 1-D, both

$$\text{prox}_{\gamma\|\cdot\|_1}(\cdot) \quad \text{and} \quad \text{prox}_{\gamma\|\nabla\cdot\|_1}(\cdot)$$

have close form solution. However, not for

$$\text{prox}_{\gamma(\|\cdot\|_1+\|\nabla\cdot\|_1)}(\cdot).$$

Operator splitting design properly structured scheme such that

- the proximal mapping of non-smooth functions are evaluated separated
- gradient descent is applied to the smooth part

Outline

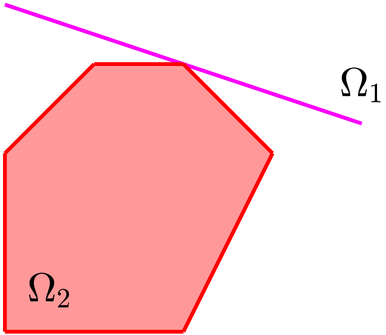
Feasibility problem

Problem (Feasibility problem)

Consider finding a common point

$$\text{find } \ x \in \Omega_1 \cap \Omega_2,$$

where $\Omega_1, \Omega_2 \in \mathbb{R}^n$ *are two closed and convex sets.*



Method of alternating projection

Equivalent formulation

$$\min_{x\in\mathbb{R}^n} \ \iota_{\Omega_1}(x) + \iota_{\Omega_2}(x).$$

Method of alternating projection (MAP)

initial: $x_0 \in \Omega_1$;

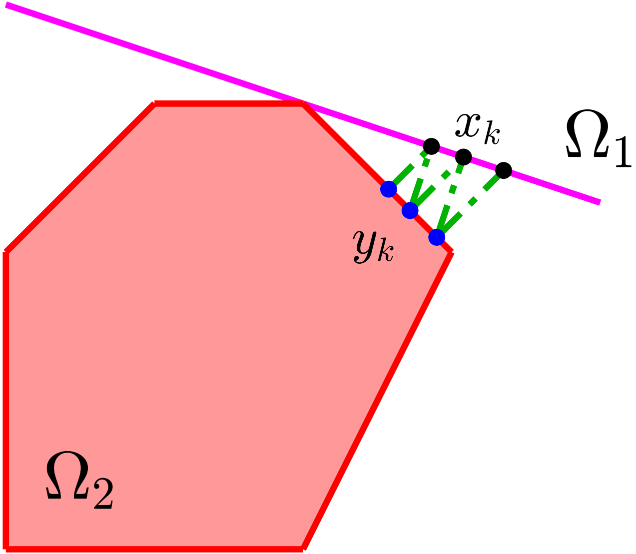
repeat:

1. Projection onto Ω_2: $y_k = \mathrm{proj}_{\Omega_2}(x_k)$
2. Projection onto Ω_1: $x_{k+1} = \mathrm{proj}_{\Omega_1}(y_k)$

until: stopping criterion is satisfied.

- The projection onto two sets are computed separately
- Stopping criterion: $\|x_k - x_{k-1}\| \le \epsilon$

Method of alternating projection



Convergence analysis

MAP:

$$x_{k+1} = \mathrm{proj}_{\Omega_1} \circ \mathrm{proj}_{\Omega_2}(x_k).$$

How to:

- analyse the convergence proerties
- not convergence result for the objective function value
- how about the sequence $\{x_k\}_{k\in\mathbb{N}}$

Outline

Notations

Given two non-empty sets $\mathcal{X},\mathcal{U} \subseteq \mathbb{R}^n$, $A : \mathcal{X} \rightrightarrows \mathcal{U}$ is called set-valued operator if A maps every point in $\mathcal{X}$ to a subset of $\mathcal{U}$, *i.e.*

$$A : \mathcal{X} \rightrightarrows \mathcal{U},\ x \in \mathcal{X} \mapsto A(x) \subseteq \mathcal{U}.$$

- The graph of A is defined by

$$\mathrm{gra}(A) \overset{\text{def}}{=} \big\{(x, u) \in \mathcal{X} \times \mathcal{U} : u \in A(x)\big\}.$$

- The domain and range of A are

$$\mathrm{dom}(A) \overset{\text{def}}{=} \big\{x \in \mathcal{X} : A(x) \neq \emptyset\big\},\ \mathrm{ran}(A) \overset{\text{def}}{=} A(\mathcal{X}).$$

- The inverse of A defined through its graph

$$\mathrm{gra}(A^{-1}) \overset{\text{def}}{=} \big\{(u, x) \in \mathcal{U} \times \mathcal{X} : u \in A(x)\big\}.$$

- The set of zeros of A are the points such that

$$\mathrm{zer}(A) \overset{\text{def}}{=} A^{-1}(0) = \big\{x \in \mathcal{X} : 0 \in A(x)\big\}.$$

Monotone operator

Definition (Monotone operator)

Let $\mathcal{X},\mathcal{U} \subseteq \mathbb{R}^n$ be two non-empty convex sets, $A : \mathcal{X} \rightrightarrows \mathcal{U}$ is monotone if

$$\langle x - y,\, u - v\rangle \geq 0,\ \forall (x, u), (y, v) \in \mathrm{gra}(A).$$

It is moreover maximal monotone if $\mathrm{gra}(A)$ is not strictly contained in the graph of any other monotone operators.

A is called α-strongly monotone for some $\kappa > 0$ if

$$\langle x - y,\, u - v\rangle \geq \kappa \|x - y\|^2.$$

Lemma

Let $R \in \Gamma_0$, then ∂R is maximal monotone.

Definition (Resolvent and reflection)

The resolvent and reflection of $A : \mathcal{X} \rightrightarrows \mathcal{U}$ are defined by

$$\mathcal{J}_A \overset{\text{def}}{=} (\mathrm{Id} + A)^{-1} \quad \text{and} \quad \mathcal{R}_A \overset{\text{def}}{=} 2\mathcal{J}_A - \mathrm{Id}.$$

Cocoersive operator

Definition (Cocoercive operator)

An operator $B : \mathbb{R}^n \to \mathbb{R}^n$ is called β-cocoercive if there exists $\beta > 0$ such that

$$\beta \|B(x) - B(y)\|^2 \leq \langle B(x) - B(y), x - y \rangle, \ \ \forall x, y \in \mathbb{R}^n.$$

The above equation implies that B is $(1/\beta)$-Lipschitz continuous.

Theorem (Baillon-Haddad)

Let $F \in C_L^1$, then ∇F is β-cocoercive.

Lemma

Let $C : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be β-strongly monotone, then its inverse C^{-1} is β-cocoercive.

Non-expansive operator

Definition (Non-expansive operator)

An operator $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ is called non-expansive if it is 1-Lipschitz continuous, *i.e.*

$$\|\mathcal{T}(x) - \mathcal{T}(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

For any $\alpha \in]0, 1[$, $\mathcal{T}$ is α-averaged if there exists a non-expansive operator $\mathcal{T}'$ such that

$$\mathcal{T} = \alpha \mathcal{T}' + (1 - \alpha)\mathrm{Id}.$$

- $\mathcal{A}(\alpha)$ denotes the class of α-averaged operators on $\mathbb{R}^n$
- $\mathcal{A}(\frac{1}{2})$ is the class of firmly non-expansive operators

Properties: α-averaged operators

$\mathcal{A}(\alpha)$ is closed under relaxations, convex combinations and compositions.

Lemma

Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be non-expansive and $\alpha \in]0,1[$. The following statements are equivalent:

- *$\mathcal{T}$ is α-averaged non-expansive.*
- *The operator*
$$\big(1 - \frac{1}{\alpha}\big)\mathrm{Id} + \frac{1}{\alpha}\mathcal{T}$$
is non-expansive.
- *For any $x, y \in \mathbb{R}^n$,*
$$\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 \leq \|x - y\|^2 - \frac{1-\alpha}{\alpha}\|(\mathrm{Id} - \mathcal{T})(x) - (\mathrm{Id} - \mathcal{T})(y)\|^2.$$

Properties: firmly non-expansive operators

Lemma

Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be non-expansive. The following statements are equivalent:

- *$\mathcal{T}$ is firmly non-expansive.*
- $2\mathcal{T} - \mathrm{Id}$ *is non-expansive.*
- $\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 \leq \langle \mathcal{T}(x) - \mathcal{T}(y), \, x - y \rangle, \forall x, y \in \mathbb{R}^n.$
- *For any $x, y \in \mathbb{R}^n$,*
$$\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 + \|(\mathrm{Id} - \mathcal{T})(x) - (\mathrm{Id} - \mathcal{T})(y)\|^2 \leq \|x - y\|^2.$$
- *$\mathcal{T}$ is the resolvent of a maximal monotone operator A, i.e. $\mathcal{T} = \mathcal{J}_A$.*

Lemma

Let operator $B : \mathbb{R}^n \to \mathbb{R}^n$ be β-cocoercive for some $\beta > 0$. Then

- $\beta B \in \mathcal{A}(\frac{1}{2})$, *i.e. is firmly non-expansive.*
- $\mathrm{Id} - \gamma B \in \mathcal{A}(\frac{\gamma}{2\beta})$ *for $\gamma \in]0, 2\beta[$.*

Outline

Fixed point

Definition (Fixed point)

Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a non-expansive operator, $x \in \mathbb{R}^n$ is called the fixed point of $\mathcal{T}$ if

$$x = \mathcal{T}(x).$$

The set of fixed points of $\mathcal{T}$ is denoted as $\mathrm{fix}(\mathcal{T})$.

$\mathrm{fix}(\mathcal{T})$ may be empty, *e.g.* translation by a non-zero vector.

Theorem

Let $\mathcal{X}$ be a non-empty bounded closed convex subset of $\mathbb{R}^n$ and $\mathcal{T} : \mathcal{X} \to \mathcal{X}$ be a non-expansive operator, then $\mathrm{fix}(\mathcal{T}) \neq \emptyset$.

Lemma

Let $\mathcal{X}$ be a non-empty closed convex subset of $\mathbb{R}^n$ and $\mathcal{T} : \mathcal{X} \to \mathbb{R}^n$ be a non-expansive operator, then $\mathrm{fix}(\mathcal{T})$ is closed and convex.

Krasnosel'skiĭ-Mann iteration

Fixed-point iteration

$$x_{k+1} = \mathcal{T}(x_k).$$

However, it may not converge, *e.g.* $\mathcal{T} = -\mathrm{Id}$...

Definition (Krasnosel'skiĭ-Mann iteration)

Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a non-expansive operator such that $\mathrm{fix}(\mathcal{T}) \neq \emptyset$. Let $\lambda_k \in [0, 1]$ and choose x_0 arbitrarily from $\mathbb{R}^n$, then the Krasnosel'skiĭ-Mann iteration of $\mathcal{T}$ reads

$$x_{k+1} = x_k + \lambda_k(\mathcal{T}(x_k) - x_k). \tag{3.1}$$

- If $\mathcal{T} \in \mathcal{A}(\alpha)$, then $\lambda_k \in [0, 1/\alpha]$

Fejér monotonicity

Definition (Fejér monotonicity)

Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a non-empty set and $\{x_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Then

(i) $\{x_k\}_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathcal{S}$ if

$$\|x_{k+1} - x\| \leq \|x_k - x\|, \quad \forall x \in \mathcal{S}, \forall k \in \mathbb{N}.$$

(ii) $\{x_k\}_{k\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathcal{S}$, if there exists a summable sequence $\{\epsilon_k\}_{k\in\mathbb{N}} \in \ell^1_+$ such that

$$\forall k \in \mathbb{N}, \quad \|x_{k+1} - x\| \leq \|x_k - x\| + \epsilon_k, \quad \forall x \in \mathcal{S}.$$

Example

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a non-empty convex set, and $\mathcal{T} : \mathcal{X} \to \mathcal{X}$ be a non-expansive operator such that $\mathrm{fix}(\mathcal{T}) \neq \emptyset$. The sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by $x_{k+1} = \mathcal{T}(x_k)$ is Fejér monotone with respect to $\mathrm{fix}(\mathcal{T})$.

Convergence

Lemma

Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a non-empty set and $\{x_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Assume the $\{x_k\}_{k\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathcal{S}$, then the following holds

- *$\{x_k\}_{k\in\mathbb{N}}$ is bounded.*
- *$\|x_k - x\|$ is bounded for any $x \in \mathcal{S}$.*
- *$\{\mathrm{dist}(x_k, \mathcal{S})\}_{k\in\mathbb{N}}$ is decreasing and convergent.*

If every sequential cluster point of $\{x_k\}_{k\in\mathbb{N}}$ belongs to $\mathcal{S}$, then $\{x_k\}_{k\in\mathbb{N}}$ converges to a point in $\mathcal{S}$.

- Weak convergence in general real Hilbert space

Convergence

Theorem

Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a non-expansive operator such that $\mathrm{fix}(\mathcal{T}) \neq \emptyset$. *Consider the Krasnosel'skiĭ-Mann iteration of $\mathcal{T}$, and choose $\lambda_k \in [0, 1]$ such that*

$$\sum_{k\in\mathbb{N}} \lambda_k(1 - \lambda_k) = +\infty,$$

then the following holds

- $\{x_k\}_{k\in\mathbb{N}}$ *is Fejér monotone with respect to* $\mathrm{fix}(\mathcal{T})$.
- $\{x_k - \mathcal{T}(x_k)\}_{k\in\mathbb{N}}$ *converges strongly to* 0.
- $\{x_k\}_{k\in\mathbb{N}}$ *converges to a point in* $\mathrm{fix}(\mathcal{T})$.

When $\mathcal{T}$ is α-averaged, then

$$\lambda_k \in [0, 1/\alpha] \ \text{ such that } \ \textstyle\sum_{k\in\mathbb{N}} \lambda_k(1/\alpha - \lambda_k) = +\infty.$$

Preliminiary

Krasnosel'skiĭ-Mann iteration with constant relaxation

$$\begin{aligned} x_{k+1} &= x_k + \lambda(\mathcal{T}(x_k) - x_k) \\ &= \big((1-\lambda)\mathrm{Id} + \lambda\mathcal{T}\big)(x_k). \end{aligned}$$

Denote $\mathcal{T}_\lambda = (1-\lambda)\mathrm{Id} + \lambda\mathcal{T}$, and define residual

$$e_k = (\mathrm{Id} - \mathcal{T})(x_k) = \frac{1}{\lambda}(x_k - x_{k+1}).$$

Preliminiary

Krasnosel'skiĭ-Mann iteration with constant relaxation

$$\begin{aligned} x_{k+1} &= x_k + \lambda(\mathcal{T}(x_k) - x_k) \\ &= \big((1-\lambda)\mathrm{Id} + \lambda\mathcal{T}\big)(x_k). \end{aligned}$$

Denote $\mathcal{T}_\lambda = (1-\lambda)\mathrm{Id} + \lambda\mathcal{T}$, and define residual

$$e_k = (\mathrm{Id} - \mathcal{T})(x_k) = \frac{1}{\lambda}(x_k - x_{k+1}).$$

- $\mathcal{T}_\lambda \in \mathcal{A}(\lambda)$ if $\lambda \in]0,1[$. If $\mathcal{T} \in \mathcal{A}(\alpha)$, then $\mathcal{T}_\lambda \in \mathcal{A}(\lambda\alpha)$
- For any $x^\star \in \mathrm{fix}(\mathcal{T})$,

$$x^\star \in \mathrm{fix}(\mathcal{T}) \Leftrightarrow x^\star \in \mathrm{fix}(\mathcal{T}_\lambda) \Leftrightarrow x^\star \in \mathrm{zer}(\mathrm{Id} - \mathcal{T})$$

- If $\lambda \in [\epsilon, 1-\epsilon], \epsilon \in]0, 1/2]$,
  - e_k converges to 0.
  - $\{x_k\}_{k\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathrm{fix}(\mathcal{T})$, and converges to a $x^\star \in \mathrm{fix}(\mathcal{T})$.

Pointwise convergence rate

Rate of $\|e_k\|^2$:

- For residual

$$\|e_{k+1}\|^2 \le \|e_k\|^2 - \frac{1-\lambda}{\lambda}\|e_k - e_{k+1}\|^2.$$

- $\mathcal{T}_\lambda \in \mathcal{A}(\lambda)$, $\tau = \lambda(1-\lambda)$

$$\|x_{k+1} - x^\star\|^2 \le \|x_k - x^\star\|^2 - \tau\|e_k\|^2.$$

- Summation

$$(k+1)\|e_k\|^2 \le \tau \sum\nolimits_{i=0}^{k} \|e_i\|^2 \le \|x_0 - x^\star\|^2 - \|x_{k+1} - x^\star\|^2.$$

- Rate

$$\|e_k\|^2 \le \frac{\|x_0 - x^\star\|^2}{k+1}.$$

If $T \in \mathcal{A}(\alpha)$, then the above holds for $\lambda \in [\epsilon, 1/\alpha - \epsilon]$.

Ergodic convergence rate

Define $\bar{e}_k = \frac{1}{k+1}\sum_{i=0}^{k} e_i$.

- Boundedness
$$\begin{aligned}\|x_{k+1} - x^\star\| = \|\mathcal{T}_\lambda(x_k) - \mathcal{T}_\lambda(x^\star)\| &\le \|x_k - x^\star\| \\ &\le \|x_0 - x^\star\|\end{aligned}$$

- $\lambda e_k = x_k - x_{k+1}$
$$\begin{aligned}\|\bar{e}_k\| = \frac{1}{k+1}\|\sum_{i=0}^{k} e_i\| &= \frac{1}{\lambda(k+1)}\|\sum_{i=0}^{k}(x_i - x_{i+1})\| \\ &= \frac{1}{\lambda(k+1)}\|x_0 - x_{k+1}\| \\ &\le \frac{1}{\lambda(k+1)}(\|x_0 - x^\star\| + \|x_{k+1} - x^\star\|) \\ &\le \frac{2\|x_0 - x^\star\|}{\lambda(k+1)}\end{aligned}$$

Both rates (pointwise and ergodic) can be extended to the inexact case...

Optimal relaxation parameter?

Consider $\lambda_k \in [0, 1]$ and $x_{k+1} = (1 - \lambda_k)x_k + \lambda_k \mathcal{T}(x_k)$. Then

$$\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|(1 - \lambda_k)(x_k - x^\star) + \lambda_k(\mathcal{T}(x_k) - x^\star)\|^2 \\
&= (1 - \lambda_k)\|x_k - x^\star\|^2 + \lambda_k\|\mathcal{T}(x_k) - x^\star\|^2 \\
&\qquad - \lambda_k(1 - \lambda_k)\|x_k - \mathcal{T}(x_k)\|^2 \\
&= \lambda_k^2\|x_k - \mathcal{T}(x_k)\|^2 \\
&\qquad - \lambda_k\big(\|x_k - x^\star\|^2 - \|\mathcal{T}(x_k) - x^\star\|^2 + \|x_k - \mathcal{T}(x_k)\|^2\big) \\
&\qquad + \|x_k - x^\star\|^2,
\end{aligned}$$

which is a quadratic fucntion of λ_k, and minimises at

$$\lambda = \frac{1}{2} + \frac{\|x_k - x^\star\|^2 - \|\mathcal{T}(x_k) - x^\star\|^2}{2\|x_k - \mathcal{T}(x_k)\|^2}.$$

Approximation:

$$\lambda = \frac{1}{2} + \frac{\|x_k - \mathcal{T}(x_k)\|^2 - \|\mathcal{T}(x_k) - \mathcal{T}^2(x_k)\|^2}{2\|(x_k - \mathcal{T}(x_k)) - (\mathcal{T}(x_k) - \mathcal{T}^2(x_k))\|^2}.$$

Outline

Inertial Krasnosel'skiĭ-Mann iteration

An inertial Krasnosel'skiĭ-Mann iteration

Initial: $x_0 \in \mathbb{R}^n, x_{-1} = x_0$;

$$y_k = x_k + a_k(x_k - x_{k-1}),\ a_k \in [0, 1],$$

$$z_k = x_k + b_k(x_k - x_{k-1}),\ b_k \in [0, 1],$$

$$x_{k+1} = (1 - \lambda_k)y_k + \lambda_k \mathcal{T}(z_k),\ \lambda_k \in [0, 1].$$

- Covers heavy-ball method, Nesterov's scheme, inertial FB and FISTA
- Convergence analysis is much harder than the inertial version of descent methods
- No convergence rate
- May perform very poorly in practice, slower than the original scheme

A multi-step inertial scheme

A multi-step inertial Krasnosel'skiĭ-Mann iteration

Initial: $x_0 \in \mathbb{R}^n, x_{-1} = x_0$ and $\gamma \in]0, 2/L[$;

$$y_k = x_k + a_{0,k}(x_k - x_{k-1}) + a_{1,k}(x_{k-1} - x_{k-2}) + \cdots \ ,$$
$$z_k = x_k + b_{0,k}(x_k - x_{k-1}) + b_{1,k}(x_{k-1} - x_{k-2}) + \cdots \ ,$$
$$x_{k+1} = (1 - \lambda_k)y_k + \lambda_k \mathcal{T}(z_k),\ \lambda_k \in [0, 1].$$

- Even harder to analyse convergence
- No rate
- However, can outperform the original scheme...

Convergence

- Conditional convergence, $i = 0, 1, \ldots$

$$\sum_{k \in \mathbb{N}} \max\left\{ \max_i |a_{i,k}|, \max_i |b_{i,k}| \right\} \sum_i \|x_{k-i} - x_{k-i-1}\| < +\infty.$$

- Online updating rule

$$a_{i,k} = \min\left\{a_i, c_{i,k}\right\}$$

where

$$c_{i,k} = \frac{c_i}{k^{1+\delta} \sum_i \|x_{k-i} - x_{k-i-1}\|}, \; \delta > 0.$$