Convex Optimization for Machine Learning and Computer Vision

Tutorial

30.01.2019

Contents

Lipschitz fixed-point iteration

Fixed-point iteration

$\Phi : C \to C$, where C is a nonempty, closed, convex subset of $\mathbb{E}$

$$u^{k+1} = \Phi(u^k). \tag{1}$$

Lipschitz fixed-point iteration

Fixed-point iteration

$\Phi : C \to C$, where C is a nonempty, closed, convex subset of $\mathbb{E}$

$$u^{k+1} = \Phi(u^k). \tag{1}$$

Lipschitz operator

Φ μ-**Lipschitz** ($\mu \geq 0$): $\forall u, v \in C : \|\Phi(u) - \Phi(v)\| \leq \mu \|u - v\|$.

Φ **contractive**: $\mu < 1$; Φ **nonexpansive**: $\mu = 1$.

Lipschitz fixed-point iteration

Fixed-point iteration

$\Phi : C \to C$, where C is a nonempty, closed, convex subset of $\mathbb{E}$

$$u^{k+1} = \Phi(u^k). \tag{1}$$

Lipschitz operator

Φ μ-**Lipschitz** ($\mu \geq 0$): $\forall u, v \in C : \|\Phi(u) - \Phi(v)\| \leq \mu \|u - v\|$.
Φ **contractive**: $\mu < 1$; Φ **nonexpansive**: $\mu = 1$.

contractive $\Rightarrow$ linear convergence

Using *Banach fixed point theorem*, we have that when Φ is contractive:
1. Eq.(1) converges to u^*; 2. linearly: $\|u^{k+1} - u^*\| \leq \mu \|u^k - u^*\|$.

Lipschitz fixed-point iteration

Fixed-point iteration

$\Phi : C \to C$, where C is a nonempty, closed, convex subset of $\mathbb{E}$

$$u^{k+1} = \Phi(u^k). \tag{1}$$

Lipschitz operator

Φ μ-**Lipschitz** ($\mu \geq 0$): $\forall u, v \in C : \|\Phi(u) - \Phi(v)\| \leq \mu \|u - v\|$.
Φ **contractive**: $\mu < 1$; Φ **nonexpansive**: $\mu = 1$.

contractive $\Rightarrow$ linear convergence

Using *Banach fixed point theorem*, we have that when Φ is contractive:
1. Eq.(1) converges to u^*; 2. linearly: $\|u^{k+1} - u^*\| \leq \mu \|u^k - u^*\|$.

Problem

Most operator we encounter are only *nonexpansive*: proj, prox, refl ...
$\rightsquigarrow$ we need more refined analysis of nonexpansive operators!

Averaged operators: "damped" nonexpansive operators

Definition

Φ is α-**averaged** with $0 < \alpha < 1$ if

$$\Phi = (1-\alpha)I + \alpha\Psi, \text{ with } \Psi : C \to C \text{ nonexpansive.}$$

Φ **firmly nonexpansive**: $\alpha = \frac{1}{2}$

Averaged operators: "damped" nonexpansive operators

Definition

Φ is α-**averaged** with $0 < \alpha < 1$ if

$$\Phi = (1-\alpha)I + \alpha\Psi, \text{ with } \Psi : C \to C \text{ nonexpansive.}$$

Φ **firmly nonexpansive**: $\alpha = \frac{1}{2}$

Equivalent definitions of α-averaged

1. $(1 - \frac{1}{\alpha})I + \frac{1}{\alpha}\Phi$ is nonexpansive.
2. $\forall u, v \in C : \|\Phi(u) - \Phi(v)\|^2 \leq \|u - v\|^2 - \frac{1-\alpha}{\alpha}\|(I - \Phi)(u) - (I - \Phi)(v)\|^2$.
3. $\forall u, v \in C : \|\Phi(u) - \Phi(v)\|^2 + (1 - 2\alpha)\|u - v\|^2 \leq 2(1 - \alpha)\langle u - v, \Phi(u) - \Phi(v)\rangle$.

Remark (2), (3) can be more practical for proving Φ is α-averaged.

Averaged operators: "damped" nonexpansive operators

Definition

Φ is α-**averaged** with $0 < \alpha < 1$ if

$$\Phi = (1-\alpha)I + \alpha\Psi, \text{ with } \Psi : C \to C \text{ nonexpansive.}$$

Φ **firmly nonexpansive**: $\alpha = \frac{1}{2}$

Equivalent definitions of α-averaged

1. $(1 - \frac{1}{\alpha})I + \frac{1}{\alpha}\Phi$ is nonexpansive.
2. $\forall u, v \in C : \|\Phi(u) - \Phi(v)\|^2 \leq \|u - v\|^2 - \frac{1-\alpha}{\alpha}\|(I - \Phi)(u) - (I - \Phi)(v)\|^2$.
3. $\forall u, v \in C : \|\Phi(u) - \Phi(v)\|^2 + (1 - 2\alpha)\|u - v\|^2 \leq 2(1-\alpha)\langle u - v, \Phi(u) - \Phi(v)\rangle$.

Remark (2), (3) can be more practical for proving Φ is α-averaged.

Properties for constructing more averaged operators

The composition and convex combination conserve the "averagedness".

Convergence with averaged operators

Theorem (Krasnoselskii)

1. Φ is α-averaged for some $\alpha \in (0, 1)$.
2. Φ has at least one fixed point.

Then $\{u^k\}$ converges to a fixed point of Φ.

Convergence with averaged operators

Theorem (Krasnoselskii)

1. Φ is α-averaged for some $\alpha \in (0, 1)$.
2. Φ has at least one fixed point.

Then $\{u^k\}$ converges to a fixed point of Φ.

Theorem (Krasnoselskii-Mann)

$u^{k+1} = (1 - \tau^k)u^k + \tau^k \Psi(u^k)$ where

1. $\Psi : C \to C$ is nonexpansive and has at least one fixed point.
2. $\{\tau^k\} \subset [0, 1]$ s.t. $\sum_{k=0}^{\infty} \tau^k (1 - \tau^k) = \infty$

Then $\{u^k\}$ converges to a fixed point of Ψ.

Convergence with averaged operators

Theorem (Krasnoselskii)

1. Φ is α-averaged for some $\alpha \in (0,1)$.
2. Φ has at least one fixed point.

Then $\{u^k\}$ converges to a fixed point of Φ.

Theorem (Krasnoselskii-Mann)

$u^{k+1} = (1-\tau^k)u^k + \tau^k \Psi(u^k)$ where

1. $\Psi : C \to C$ is nonexpansive and has at least one fixed point.
2. $\{\tau^k\} \subset [0,1]$ s.t. $\sum_{k=0}^{\infty} \tau^k(1-\tau^k) = \infty$

Then $\{u^k\}$ converges to a fixed point of Ψ.

Remarks on Krasnoselskii-Mann Theorem

"Customized damping", Fulfilled if $\{\tau^k\} \subset [\epsilon, 1-\epsilon]$ for some $\epsilon \in (0, 1/2]$.

What operators are averaged?

Overview

proj firmly nonexpansive (see exercise sheet 11 ex2 or below)

prox firmly nonexpansive (use e.g. last equivalent definition)

refl nonexpansive only (refl = 2 * prox - Id)

Gradient descent if G is μ-Lipschitz differentiable and $\tau \in (0, 2/\mu)$, then $I - \tau \nabla G$ is $(\tau\mu/2)$-averaged (Corollary of Baillon-Haddad)

CPI if M is spd and R is maximal monotone, then $\Phi^{(\text{cpi})} = (M + R)^{-1} M$ is firmly nonexpansive in the Euclidean space with scaled inner-product $< \cdot, \cdot >_M$

What operators are averaged?

Overview

- **proj**: firmly nonexpansive (see exercise sheet 11 ex2 or below)
- **prox**: firmly nonexpansive (use e.g. last equivalent definition)
- **refl**: nonexpansive only (refl = 2 * prox - Id)
- **Gradient descent**: if G is μ-Lipschitz differentiable and $\tau \in (0, 2/\mu)$, then $I - \tau \nabla G$ is $(\tau\mu/2)$-averaged (Corollary of Baillon-Haddad)
- **CPI**: if M is spd and R is maximal monotone, then $\Phi^{(\text{cpi})} = (M+R)^{-1}M$ is firmly nonexpansive in the Euclidean space with scaled inner-product $< \cdot, \cdot >_M$

Remark on CPI

1. The notions of *Lipschitz*, *nonexpansive*, *averaged* etc. all depend on the choice of inner product (which induce the norm).
2. The *convergence* doesn't depend on the choice of inner product, Since all norms are equivalent in Euclidean space.

Convergence of the proximal algorithms: an overview

Convergence that we can proof:

FBS composition pf prox and gradient descent, so averaged if G is μ-Lipschitz differentiable and $\tau \in (0, 2/\mu)$.

DRS $\Phi^{(\text{drs})} = \frac{1}{2}I + \frac{1}{2}\text{refl}_{\tau F} \circ \text{refl}_{\tau G}$, so firmly nonexpansive

PDHG direct from CPI: firmly nonexpansive if $s \times t > \|K\|^2_{\text{spec}}$

Convergence of the proximal algorithms: an overview

Convergence that we can proof:

- **FBS** composition pf prox and gradient descent, so averaged if G is μ-Lipschitz differentiable and $\tau \in (0, 2/\mu)$.
- **DRS** $\Phi^{(\text{drs})} = \frac{1}{2}I + \frac{1}{2}\text{refl}_{\tau F} \circ \text{refl}_{\tau G}$, so firmly nonexpansive
- **PDHG** direct from CPI: firmly nonexpansive if $s \times t > \|K\|^2_{\text{spec}}$

About ADMM

- We can not use CPI conditions: M is only spsd, not spd; However:
- We can make M spd using e.g. the variant in exercise sheet 10 ex1. Then we can prove convergence of the variant!

Convergence of the proximal algorithms: an overview

Convergence that we can proof:

FBS composition pf prox and gradient descent, so averaged if G is μ-Lipschitz differentiable and $\tau \in (0, 2/\mu)$.

DRS $\Phi^{(\text{drs})} = \frac{1}{2}I + \frac{1}{2}\text{refl}_{\tau F} \circ \text{refl}_{\tau G}$, so firmly nonexpansive

PDHG direct from CPI: firmly nonexpansive if $s \times t > \|K\|^2_{\text{spec}}$

About ADMM

- We can not use CPI conditions: M is only spsd, not spd; However:
- We can make M spd using e.g. the variant in exercise sheet 10 ex1. Then we can prove convergence of the variant!

About PRS

Doesn't have convergence guarantee, since composition of nonexpansive operators are just nonexpansive.

Linear convergence under strong monotonicity

Strongly monotone operator

- R is said μ-**strongly monotone** if $R - \mu I$ is monotone.
- For proper, convex, lsc function J, ∂J is μ-strongly monotone iff J is μ-strongly convex, i.e., $J - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Linear convergence under strong monotonicity

Strongly monotone operator

- R is said μ-**strongly monotone** if $R - \mu I$ is monotone.
- For proper, convex, lsc function J, ∂J is μ-strongly monotone iff J is μ-strongly convex, i.e., $J - \frac{\mu}{2}\|\cdot\|^2$ is convex.

- R is μ-strongly monotone

$$\Rightarrow \left\langle u^{k+1} - u^*, \xi^{k+1} - 0 \right\rangle \geq \mu \|u^{k+1} - u^*\|^2 \tag{2}$$

$$\Rightarrow \|u^{k+1} - u^*\|_M \leq \sqrt{\frac{1}{1 + 2\mu/\lambda_{\max}(M)}} \|u^k - u^*\|_M. \tag{3}$$

Linear convergence under strong monotonicity

Strongly monotone operator

- R is said μ-**strongly monotone** if $R - \mu I$ is monotone.
- For proper, convex, lsc function J, ∂J is μ-strongly monotone iff J is μ-strongly convex, i.e., $J - \frac{\mu}{2}\|\cdot\|^2$ is convex.

- R is μ-strongly monotone

$$\Rightarrow \quad \langle u^{k+1} - u^*, \xi^{k+1} - 0 \rangle \geq \mu \|u^{k+1} - u^*\|^2 \qquad (2)$$

$$\Rightarrow \quad \|u^{k+1} - u^*\|_M \leq \sqrt{\frac{1}{1 + 2\mu/\lambda_{\max}(M)}} \|u^k - u^*\|_M. \qquad (3)$$

- Recall in PDHG/ADMM:

$$R = \begin{bmatrix} \partial G & K^\top \\ -K & \partial F^* \end{bmatrix}.$$

R is μ-strongly monotone $\Leftrightarrow$ G, F^* are μ-strongly convex;
F^* is μ-strongly convex $\Leftrightarrow$ F is $\frac{1}{\mu}$-Lipschitz differentiable.