ORIE 6326: Convex Optimization

Operators

Professor Udell

Operations Research and Information Engineering
Cornell

May 15, 2017

Beating the lower bound

problem: lower bound for subgradient method is **slow**!

Beating the lower bound

problem: lower bound for subgradient method is **slow**!
solution: find a more powerful oracle

Proximal operator

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

Proximal operator

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $\mathbf{prox}_f : \mathbf{R}^d \to \mathbf{R}^d$

Proximal operator

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \underset{z}{\operatorname{argmin}}(f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $\mathbf{prox}_f : \mathbf{R}^d \to \mathbf{R}^d$
- **generalized projection:** if $\mathbf{1}_C$ is the indicator of set C,

$$\mathbf{prox}_{\mathbf{1}_C}(w) = \Pi_C(w)$$

Proximal operator

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \operatorname*{argmin}_z (f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $\mathbf{prox}_f : \mathbf{R}^d \to \mathbf{R}^d$
- **generalized projection:** if $\mathbf{1}_C$ is the indicator of set C,

$$\mathbf{prox}_{\mathbf{1}_C}(w) = \Pi_C(w)$$

- **implicit gradient step:** if $z = \mathbf{prox}_f(x)$

$$\begin{aligned} \partial f(z) + z - x &= 0 \\ z &= x - \partial f(z) \end{aligned}$$

Maps from functions to functions

no consistent notation for map from functions to functions.

for a function $f : \mathbf{R}^d \to \mathbf{R}$,

- **prox** maps f to a new function $\mathbf{prox}_f : \mathbf{R}^d \to \mathbf{R}^d$
  - $\mathbf{prox}_f(x)$ evaluates this function at the point x
- ∇ maps f to a new function $\nabla f : \mathbf{R}^d \to \mathbf{R}^d$
  - $\nabla f(x)$ evaluates this function at the point x

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)
- $f(x) = x^2$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \operatorname*{argmin}_z (f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \underset{z}{\operatorname{argmin}}(f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \underset{z}{\operatorname{argmin}}(f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)
- $f(x) = \sum_{i=1}^d f_i(x_i)$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)
- $f(x) = \sum_{i=1}^d f_i(x_i)$ (separable)

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \operatorname*{argmin}_z (f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)
- $f(x) = \sum_{i=1}^d f_i(x_i)$ (separable)
- $f(x) = \|x\|_1$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \underset{z}{\operatorname{argmin}}(f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)
- $f(x) = \sum_{i=1}^{d} f_i(x_i)$ (separable)
- $f(x) = \|x\|_1$ (soft-thresholding on each index)

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

$$\mathbf{prox}_f(x) = \operatorname*{argmin}_z (f(z) + \frac{1}{2}\|z - x\|_2^2)$$

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)
- $f(x) = \sum_{i=1}^d f_i(x_i)$ (separable)
- $f(x) = \|x\|_1$ (soft-thresholding on each index)
- $f(X) = \|X\|_*$

Let's evaluate some proximal operators!

define the **proximal operator** of the function $f : \mathbf{R}^d \to \mathbf{R}$

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

- $f(x) = 0$ (identity)
- $f(x) = x^2$ (shrinkage)
- $f(x) = |x|$ (soft-thresholding)
- $f(x) = \mathbf{1}(x \geq 0)$ (projection)
- $f(x) = \sum_{i=1}^d f_i(x_i)$ (separable)
- $f(x) = \|x\|_1$ (soft-thresholding on each index)
- $f(X) = \|X\|_*$ (soft-thresholding on singular values)

Proxable functions

we say a function f is **proxable** if it's easy to evaluate $\mathbf{prox}_f(x)$

all examples from previous slide are proxable

Roadmap

suppose f is smooth, g is non-smooth. solve

$$\text{minimize} \quad f(x) + g(x)$$

using proximal operators together with gradient steps?

- the proximal operator gives a **fast method** to step towards the minimum of g
- gradient method works well to step towards minimum of f
- put it together with gradients to make **fast optimization algorithms**

to do this elegantly, we will need more theory...

Outline

Functions

in much of what follows, we'll need to assume functions are

- closed: $\textbf{epi}(f)$ is a closed set
- convex: f is convex
- proper: $\textbf{dom}\, f$ is non-empty

which we abbreviate as CCP

Relations

$(x, \partial f(x))$ and $(x, \mathbf{prox}_f(x))$ define **relations** on $\mathbf{R}^n$

- a **relation** R on $\mathbf{R}^n$ is a subset of $\mathbf{R}^n \times \mathbf{R}^n$
- $\mathbf{dom}\, R = \{x : (x, y) \in R\}$
- let $R(x) = \{y : (x, y) \in R\}$
- if $R(x)$ is always empty or a singleton, we say R is a function
- any function $f : \mathbf{R} \to \mathbf{R}$ is a relation

Relations: examples

- empty relation: $\emptyset$
- full relation: $\mathbf{R}^n \times \mathbf{R}^n$
- identity: $\{(x, x) : x \in \mathbf{R}^n\}$
- zero: $\{(x, 0) : x \in \mathbf{R}^n\}$
- subdifferential: $\{(x, g : x \in \mathbf{dom}\, f, g \in \partial f(x)\}$

Operations on relations

if R and S are relations, define

- composition: $RS = \{(x, z) : (x, y) \in R, (y, z) \in S\}$
- addition: $R + S = \{(x, y + z) : (x, y) \in R, (x, z) \in S\}$
- inverses: $R^{-1} = \{(y, x) : (x, y) \in R\}$

use inequality on sets to mean the inequality holds for any element in the set, *e.g.*,

$$f(y) \geq f(x) + \partial f^T (y - x)$$

Example: fenchel conjugates and the subdifferential

if f is CPP, $(f^*)^* = f^{**} = f$, so

$$
\begin{aligned}
(u, v) \in (\partial f)^{-1} \quad &\iff \quad (v, u) \in \partial f \\
&\iff \quad u \in \partial f(v) \\
&\iff \quad 0 \in \partial f(v) - u \\
&\iff \quad v \in \operatorname*{argmin}_x (f(x) - u^T x) \\
&\iff \quad v \in \operatorname*{argmax}_x (u^T x - f(x)) \\
&\iff \quad f(v) + f^*(u) = u^T v \\
&\iff \quad u \in \operatorname*{argmax}_y (y^T v - f^*(y)) \\
&\iff \quad 0 \in v - \partial f^*(u)) \\
&\iff \quad (u, v) \in \partial f^*
\end{aligned}
$$

this shows $\partial f^* = \partial f^{-1}$

Outline

Zeros of a relation

- x is a **zero** of R if $0 \in R(x)$
- the **zero set** of R is $R^{-1}(0) = \{x : (x, 0) \in R\}$

Zeros of a relation

- x is a **zero** of R if $0 \in R(x)$
- the **zero set** of R is $R^{-1}(0) = \{x : (x, 0) \in R\}$

x is a zero of ∂f iff x solves minimize $\quad f(x)$

Lipschitz operators

relation F has Lipschitz constant L if for all $(x, u) \in F$ and $(y, v) \in F$,

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

fact: if F is Lipschitz, then F is a function.
proof:

Lipschitz operators

relation F has Lipschitz constant L if for all $(x, u) \in F$ and $(y, v) \in F$,

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

fact: if F is Lipschitz, then F is a function.
proof: if $(x, u) \in F$ and $(x, v) \in F$,

$$\|u - v\| \leq L\|x - x\| = 0$$

- the relation F is **nonexpansive** if $L \leq 1$
- the relation F is **contractive** if $L < 1$

Fixed points

x is a **fixed point** of F if $x = F(x)$

examples:

- $F(x) = x$: every point is a fixed point
- $F(x) = 0$: only 0 is a fixed point

Fixed points

x is a **fixed point** of F if $x = F(x)$

examples:

- $F(x) = x$: every point is a fixed point
- $F(x) = 0$: only 0 is a fixed point
- a contractive operator on $\mathbf{R}^n$ can have at most one FP

Fixed points

x is a **fixed point** of F if $x = F(x)$

examples:

- $F(x) = x$: every point is a fixed point
- $F(x) = 0$: only 0 is a fixed point
- a contractive operator on $\mathbf{R}^n$ can have at most one FP
  proof: if x and y are FPs, $\|x - y\| = \|F(x) - F(y)\| < \|x - y\|$
  contradiction

Fixed points

x is a **fixed point** of F if $x = F(x)$

examples:

- $F(x) = x$: every point is a fixed point
- $F(x) = 0$: only 0 is a fixed point
- a contractive operator on $\mathbf{R}^n$ can have at most one FP
  proof: if x and y are FPs, $\|x - y\| = \|F(x) - F(y)\| < \|x - y\|$
  contradiction
- a nonexpansive operator F need not have a fixed point

Fixed points

x is a **fixed point** of F if $x = F(x)$

examples:

- $F(x) = x$: every point is a fixed point
- $F(x) = 0$: only 0 is a fixed point
- a contractive operator on $\mathbf{R}^n$ can have at most one FP
  proof: if x and y are FPs, $\|x - y\| = \|F(x) - F(y)\| < \|x - y\|$
  contradiction
- a nonexpansive operator F need not have a fixed point
  proof: translation

Fixed point iteration

to find a fixed point of F, try the fixed point iteration

$$x^{(k+1)} = F(x^{(k)})$$

Fixed point iteration

to find a fixed point of F, try the fixed point iteration

$$x^{(k+1)} = F(x^{(k)})$$

Q: when does this converge?

Fixed point iteration: contractive

Banach fixed point theorem: if F is a contraction, the iteration

$$x^{(k+1)} = F(x^{(k)})$$

converges to the unique fixed point of F

properties: if L is the Lipschitz constant of F,

- distance to fixed point decreases monotonically:

$$\|x^{(k+1)} - x^\star\| = \|F(x^{(k)}) - F(x^\star)\| \leq L\|x^{(k)} - x^\star\|$$

  (iteration is **Fejer-monotone**)
- linear convergence with rate L

Proof

if F is a contraction, $x^{(k+1)} = F(x^{(k)})$ converges to unique fixed point

proof:

Proof

if F is a contraction, $x^{(k+1)} = F(x^{(k)})$ converges to unique fixed point

proof: if F has Lipschitz constant $L < 1$,

- sequence $x^{(k)}$ is Cauchy:

$$\begin{array}{rcl}\|x^{(k+\ell)} - x^{(k)}\| & \leq & \|x^{(k+\ell)} - x^{(k+\ell-1)}\| + \cdots + \|x^{(k+1)} - x^{(k)}\| \\ & \leq & (L^{\ell-1} + \ldots + 1)\|x^{(k+1)} - x^{(k)}\| \\ & \leq & \dfrac{1}{1-L}\|x^{(k+1)} - x^{(k)}\| \\ & \leq & \dfrac{L^k}{1-L}\|x^{(1)} - x^{(0)}\|\end{array}$$

- so it converges to a point $x^\star$, which must be the (unique) FP
- converges to $x^\star$ linearly with rate L

$$\|x^{(k)} - x^\star\| = \|F(x^{(k-1)}) - F(x^\star)\| \leq L\|x^{(k-1)} - x^\star\| \leq L^k\|x^{(0)} - x^\star\|$$

Outline

Fixed point iteration: nonexpansive

if F is nonexpansive, the iteration

$$x^{(k+1)} = F(x^{(k)})$$

need not converge to a fixed point even if one exists.

proof:

Fixed point iteration: nonexpansive

if F is nonexpansive, the iteration

$$x^{(k+1)} = F(x^{(k)})$$

need not converge to a fixed point even if one exists.

proof:

- let F rotate its argument by θ degrees around the origin.
- then F is nonexpansive and has a fixed point at $x^\star = 0$.
- but if $\|x^{(0)}\| = r$, then $\|F(x^{(k)})\| = r$ for all k.

Averaged operators

an operator F is **averaged** if

$$F = \theta G + (1 - \theta) I$$

for $\theta \in (0, 1)$, G nonexpansive

Averaged operators

an operator F is **averaged** if

$$F = \theta G + (1 - \theta) I$$

for $\theta \in (0, 1)$, G nonexpansive

fact: if F is averaged, then x if FP of $F \iff x$ is FP of G
proof:

Averaged operators

an operator F is **averaged** if

$$F = \theta G + (1-\theta) I$$

for $\theta \in (0,1)$, G nonexpansive

fact: if F is averaged, then x if FP of $F \iff x$ is FP of G
proof:

$$\begin{aligned} x &= Fx = \theta Gx + (1-\theta) Ix = \theta Gx + (1-\theta)x \\ \theta x &= \theta Gx \\ x &= Gx \end{aligned}$$

$\Longrightarrow$ if G is nonexpansive, $F = \frac{1}{2} I + \frac{1}{2} G$ is averaged with same FPs

Fixed point iteration: averaged

if $F = \theta G + (1-\theta)I$ is averaged ($\theta \in (0,1)$, G nonexpansive), the iteration

$$x^{(k+1)} = F(x^{(k)})$$

converges to a fixed point if one exists.

(also called the damped, averaged, or Mann-Krasnosel'skii iteration.)

properties:

- distance to fixed point decreases monotonically (Fejer-monotone)
- sublinear convergence of fixed point residual

$$\|Gx^{(k)} - x^{(k)}\|^2 \leq \frac{1}{(k+1)\theta(1-\theta)}\|x^{(0)} - x^\star\|^2$$

Proof

proof follows [Ryu and Boyd, 2015]

use $\|(1-\theta)a+\theta b\|^2 = (1-\theta)\|a\|^2 + \theta\|b\|^2 - \theta(1-\theta)\|a-b\|^2$ (proof by expanding), and set

$$x^{(k+1)} - x^\star = (1-\theta)(x^{(k)} - x^\star) + \theta(Gx^{(k)} - x^\star) = (1-\theta)a + \theta b$$

so $a - b = x^{(k)} - x^\star - (Gx^{(k)} - x^\star) = x^{(k)} - Gx^{(k)}$. then

$$\begin{aligned}
& \|x^{(k+1)} - x^\star\|^2 \\
= \;& (1-\theta)\|x^{(k)} - x^\star\|^2 + \theta\|Gx^{(k)} - x^\star\|^2 - \theta(1-\theta)\|x^{(k)} - Gx^{(k)}\|^2 \\
\le \;& (1-\theta)\|x^{(k)} - x^\star\|^2 + \theta\|x^{(k)} - x^\star\|^2 - \theta(1-\theta)\|x^{(k)} - Gx^{(k)}\|^2 \\
= \;& \|x^{(k)} - x^\star\|^2 - \theta(1-\theta)\|x^{(k)} - Gx^{(k)}\|^2,
\end{aligned}$$

(using the fact that G is nonexpansive for the first inequality)

this shows the iteration is Fejer monotone

Proof (II)

sum the last inequality over iterations k. it telescopes:

$$
\begin{aligned}
\|x^{(k+1)} - x^\star\|^2 &\leq \|x^{(0)} - x^\star\|^2 - \theta(1-\theta)\sum_{i=0}^{k}\|x^{(i)} - Gx^{(i)}\|^2 \\
\sum_{i=0}^{k}\|x^{(i)} - Gx^{(i)}\|^2 &\leq \frac{1}{\theta(1-\theta)}\|x^{(0)} - x^\star\|^2 \\
\min_{i=0,\ldots,k}\|x^{(i)} - Gx^{(i)}\|^2 &\leq \frac{1}{(k+1)\theta(1-\theta)}\|x^{(0)} - x^\star\|^2 \\
\|x^{(k)} - Gx^{(k)}\|^2 &\leq \frac{1}{(k+1)\theta(1-\theta)}\|x^{(0)} - x^\star\|^2
\end{aligned}
$$

where the last inequality uses the fact that G is nonexpansive.

How to design an algorithm

- look for an operator F
  - whose fixed points solve optimization problems
  - which is contractive or averaged
- iterate $x^{(k+1)} = F(x^{(k)})$
- get convergence
  - if F is contractive, get linear convergence
  - if F is averaged, get sublinear convergence

Outline

Properties of operators

we'll need these properties of operators:

- Lipschitz
  - contractive
  - nonexpansive
  - averaged
- monotone
  - maximal
  - strongly monotone == coercive
- cocoercive

Properties of optimization operators

we'll show connection to optimization:

- gradient of convex function is monotone
- gradient of strongly convex function is strongly monotone
- gradient of smooth function is cocoercive
- prox of convex function is $\frac{1}{2}$-averaged
- gradient step $I - \frac{2}{\beta}\nabla f$ of smooth function is nonexpansive (and so smaller stepsizes are averaged)
- prox of strongly convex function is contractive
- gradient step of SSC function is contractive

Monotone operators

- a relation F is **monotone** if, for all $(x, u) \in F$ and $(y, v) \in F$

$$(u - v)^T (x - y) \geq 0.$$

example: ∂f is monotone

Monotone operators

- a relation F is **monotone** if, for all $(x, u) \in F$ and $(y, v) \in F$

  $$(u - v)^T (x - y) \geq 0.$$

  example: ∂f is monotone
- a relation F is α-**strongly monotone** if, for all $(x, u) \in F$ and $(y, v) \in F$

  $$(u - v)^T (x - y) \geq \alpha \|x - y\|^2.$$

  (also called α-**coercive**)
  example: ∂f is strongly monotone iff f is strongly convex

Monotone operators

- a relation F is **monotone** if, for all $(x,u) \in F$ and $(y,v) \in F$
$$(u-v)^T(x-y) \geq 0.$$
example: ∂f is monotone
- a relation F is α-**strongly monotone** if, for all $(x,u) \in F$ and $(y,v) \in F$
$$(u-v)^T(x-y) \geq \alpha\|x-y\|^2.$$
(also called α-**coercive**)
example: ∂f is strongly monotone iff f is strongly convex
- a relation F is **maximal monotone** if there is no monotone relation that properly contains it (as a subset of $\mathbf{R}^n \times \mathbf{R}^n$)
examples:
  - if $F : \mathbf{R}^n \to \mathbf{R}^n$ is continuous and monotone, then F is maximal monotone
  - if f is CCP then ∂f is maximal monotone

  useful: makes sure FP iteration doesn't leave domain of F

Cocoercive operators

an operator F is $\frac{1}{\beta}$-**cocoercive** if

$$(F(x) - F(y))^T(x - y) \geq \frac{1}{\beta}\|F(x) - F(y)\|^2.$$

example: we've already seen that ∇f is cocoercive if f is β-smooth

Cocoercive operators

an operator F is $\frac{1}{\beta}$-**cocoercive** if

$$(F(x) - F(y))^T(x - y) \geq \frac{1}{\beta}\|F(x) - F(y)\|^2.$$

example: we've already seen that ∇f is cocoercive if f is β-smooth

- F is $\frac{1}{\beta}$-cocoercive $\implies$ F is β-Lipschitz

Cocoercive operators

an operator F is $\frac{1}{\beta}$-**cocoercive** if

$$(F(x) - F(y))^T(x - y) \geq \frac{1}{\beta}\|F(x) - F(y)\|^2.$$

example: we've already seen that ∇f is cocoercive if f is β-smooth

- F is $\frac{1}{\beta}$-cocoercive $\implies$ F is β-Lipschitz
  (proof by Cauchy-Schwarz)

Cocoercive operators

an operator F is $\frac{1}{\beta}$-**cocoercive** if

$$(F(x) - F(y))^T(x - y) \geq \frac{1}{\beta}\|F(x) - F(y)\|^2.$$

example: we've already seen that ∇f is cocoercive if f is β-smooth

- F is $\frac{1}{\beta}$-cocoercive $\implies$ F is β-Lipschitz
  (proof by Cauchy-Schwarz)
- F is α-coercive $\iff$ F^{-1} is α-cocoercive

Cocoercive operators

an operator F is $\frac{1}{\beta}$-**cocoercive** if

$$(F(x) - F(y))^T(x - y) \geq \frac{1}{\beta}\|F(x) - F(y)\|^2.$$

example: we've already seen that ∇f is cocoercive if f is β-smooth

- F is $\frac{1}{\beta}$-cocoercive $\implies$ F is β-Lipschitz
  (proof by Cauchy-Schwarz)
- F is α-coercive $\iff$ F^{-1} is α-cocoercive
  (proof by interchanging x and $F(x)$)

Gradient step is nonexpansive (or averaged)

F is $\frac{1}{\beta}$-**cocoercive** $\iff$ $I - \frac{2}{\beta}F$ is nonexpansive
(and smaller step sizes are averaged)

proof:

Gradient step is nonexpansive (or averaged)

F is $\frac{1}{\beta}$-**cocoercive** $\iff$ $I - \frac{2}{\beta}F$ is nonexpansive
(and smaller step sizes are averaged)

proof:

$$
\begin{aligned}
& \|(I - \frac{2}{\beta}F)y - (I - \frac{2}{\beta}F)x\| \\
= \quad & \|y - \frac{2}{\beta}Fy - x - \frac{2}{\beta}Fx\| \\
= \quad & \|y - x\|^2 - \frac{4}{\beta}(Fy - Fx)^T(y - x) + \frac{4}{\beta^2}\|Fy - Fx\|^2 \\
= \quad & \|y - x\|^2 - \frac{4}{\beta}\left((Fy - Fx)^T(y - x) - \frac{1}{\beta}\|Fy - Fx\|^2\right) \\
\leq \quad & \|y - x\|^2
\end{aligned}
$$

inequality is definition of $\frac{1}{\beta}$-cocoercivity

Outline

∂f as an operator

if f is convex, ∂f is maximal monotone

if f is β-smooth,

- $\partial f = \nabla f$ is $\frac{1}{\beta}$-cocoercive
- $\partial f = \nabla f$ is β-Lipschitz
-

if f is α-strongly convex

- ∂f is α-strongly monotone

Gradient method converges

if f is convex and β-smooth,

- $I - \frac{2}{\beta}\nabla f$ is nonexpansive,
- gradient mapping $I - t\nabla f$ is averaged for $t \in (0, \frac{2}{\beta})$
- so fixed point iteration

$$x^{(k+1)} = (I - \frac{2}{\beta}\nabla f)x^{(k)}$$

  converges

- from convergence guarantee for damped iteration,

$$\begin{aligned} \frac{1}{(k+1)\theta(1-\theta)}\|x^{(0)} - x^\star\|^2 &\geq \min_{i=0,\ldots,k} \|(I - \frac{2}{\beta}\nabla f)x^{(k)} - x^{(k)}\|^2 \\ &\geq \min_{i=0,\ldots,k} (\frac{2}{\beta})^2 \|\nabla f(x^{(k)})\|^2 \end{aligned}$$

Strong convexity and smoothness

f is α-strongly convex iff f^* is $\frac{1}{\alpha}$-smooth

proof:

Strong convexity and smoothness

f is α-strongly convex iff f^* is $\frac{1}{\alpha}$-smooth

proof: $\partial f^{-1} = \partial f^*$.
∂f α-coercive $\iff$ ∂f^* α-cocoercive $\iff$ f^* $\frac{1}{\alpha}$-smooth.

moral: strong convexity and (strong) smoothness are dual

Gradient update is contractive for SSC functions

suppose f is α-strongly convex and β-smooth

the relation

$$I - t\nabla f = \{(x, x - t\nabla f(x)) : x \in \mathbf{dom}\, f\}$$

is contractive if $t \in (0, \frac{2\alpha}{\beta^2})$. best contraction factor if $t = \frac{\alpha}{\beta^2}$

proof:

$$\begin{aligned}
& \|x - t\nabla f(x) - (y - t\nabla f(y))\|^2 \\
\leq\ & \|x - y\|^2 + t^2\|\nabla f(x) - \nabla f(y)\|^2 - 2t(\nabla f(x) - \nabla f(y))^T(x - y) \\
\leq\ & \|x - y\|^2 + t^2\beta^2\|x - y\|^2 - 2t\alpha\|x - y\|^2 \\
\leq\ & (1 - 2t\alpha + t^2\beta^2)\|x - y\|^2
\end{aligned}$$

note: stronger proof (using co+cocoercive inequality from slide 28 of GD lecture) shows $I - t\nabla f$ is Lipschitz with parameter $L = \max\{|1 - t\alpha|, |1 - t\beta|\}$. if $t = \frac{2}{\alpha+\beta}$, $L = \frac{\kappa-1}{\kappa+1}$

Outline

Resolvent operator

for relation F, define the **resolvent** of F

$$R_F = (I + F)^{-1}$$

consider resolvent of F

- $(I + F) = \{(x, x + y) : (x, y) \in F\}$
- $R_F = (I + F)^{-1} = \{(x + y, x) : (x, y) \in F\}$
- $R_F = \{(u, v) : (u - v) \in F(v)\}$

Prox is the resolvent of ∂f

- $\mathbf{prox}_f = R_{\partial f} = (I + \partial f)^{-1}$

Prox is the resolvent of ∂f

- $\mathbf{prox}_f = R_{\partial f} = (I + \partial f)^{-1}$
  proof: let $z \in \mathbf{prox}_f(x)$,

$$\begin{aligned} z &= \operatorname*{argmin}_z f(z) + \frac{1}{2}\|z - x\|^2 \\ 0 &\in \partial f(z) + z - x \\ (x - z) &\in \partial f(z) \end{aligned}$$

- $\mathbf{prox}_f = \nabla h^*$ where $h(x) = f(x) + \frac{1}{2}\|x\|^2$

Prox is the resolvent of ∂f

- $\mathbf{prox}_f = R_{\partial f} = (I + \partial f)^{-1}$
  proof: let $z \in \mathbf{prox}_f(x)$,

$$
\begin{aligned}
z &= \operatorname*{argmin}_z f(z) + \frac{1}{2}\|z - x\|^2 \\
0 &\in \partial f(z) + z - x \\
(x - z) &\in \partial f(z)
\end{aligned}
$$

- $\mathbf{prox}_f = \nabla h^*$ where $h(x) = f(x) + \frac{1}{2}\|x\|^2$
  proof: h is CCP and $\partial h = \partial f + I$, so

$$\nabla h^* = (\partial h)^{-1} = (I + \partial f)^{-1}$$

- $\mathbf{prox}_f$ is a function

Prox is the resolvent of ∂f

- $\mathbf{prox}_f = R_{\partial f} = (I + \partial f)^{-1}$
  proof: let $z \in \mathbf{prox}_f(x)$,

$$
\begin{aligned}
z &= \operatorname*{argmin}_z f(z) + \frac{1}{2}\|z - x\|^2 \\
0 &\in \partial f(z) + z - x \\
(x - z) &\in \partial f(z)
\end{aligned}
$$

- $\mathbf{prox}_f = \nabla h^*$ where $h(x) = f(x) + \frac{1}{2}\|x\|^2$
  proof: h is CCP and $\partial h = \partial f + I$, so

$$\nabla h^* = (\partial h)^{-1} = (I + \partial f)^{-1}$$

- $\mathbf{prox}_f$ is a function
  proof: h is strongly convex, so h^* is smooth

Projection is resolvent of indicator function

let $f = I_C$, the indicator function of the convex set C

- ∂f is the **normal cone operator**

$$\partial f(x) = N_C(x) = \begin{cases} \emptyset & x \notin C \\ \{w : w^T(z - x) \le 0, \quad \forall z \in C\} & x \in C \end{cases}$$

- $R_{\partial f} = \mathbf{prox}_f$ is

$$(I + \partial I_C)^{-1}(x) = \operatorname*{argmin}_u (I_C(u) + \frac{1}{2}\|u - x\|^2 = \Pi_C(x)$$

Resolvent and Cayley operators

for relation F, recall the **resolvent** of F

$$R_F = (I + F)^{-1}$$

and define the **Cayley** operator (sometimes called reflection operator)

$$C_F = 2R_F - I = 2(I + F)^{-1} - I$$

properties:

- if F is monotone, then R_F and C_F are nonexpansive (and hence are functions)
- if F is maximal monotone, then R_F and C_F have full domain
- zeros of F are FPs of R_F and C_F

we'll prove the first and last properties; second is Minty's theorem

Fixed points of R and C are zeros of F

fact: fixed points of R and C are zeros of F

proof:

Fixed points of R and C are zeros of F

fact: fixed points of R and C are zeros of F

proof:

$$\begin{aligned} 0 \in F(x) \quad &\iff \quad x \in (I + F)(x) \\ &\iff \quad (I + F)^{-1}(x) \ni x \\ &\iff \quad x = R(x) \end{aligned}$$

(using the fact that R is a function)

and if $x = R(x)$,

$$C(x) = 2R(x) - I(x) = 2x - x = x$$

Fixed points of R and C are zeros of F

fact: fixed points of R and C are zeros of F

proof:

$$\begin{aligned} 0 \in F(x) \quad &\iff \quad x \in (I+F)(x) \\ &\iff \quad (I+F)^{-1}(x) \ni x \\ &\iff \quad x = R(x) \end{aligned}$$

(using the fact that R is a function)

and if $x = R(x)$,

$$C(x) = 2R(x) - I(x) = 2x - x = x$$

in particular, this means fixed points of $\mathbf{prox}_f(x)$ are zeros of ∂f

Resolvent is nonexpansive

fact: $R = R_{\lambda F}$ is nonexpansive

proof:

Resolvent is nonexpansive

fact: $R = R_{\lambda F}$ is nonexpansive

proof: if $(x, u) \in R$ and $(y, v) \in R$,

$$x \in u + F(u), \quad y \in v + F(v)$$

so for some $f_u \in F(u)$, $f_v \in F(v)$,

$$\begin{aligned} x - y &= u - v + f_u - f_v \\ (u - v)^T(x - y) &= \|u - v\|^2 + (u - v)^T(f_u - f_v) \\ (u - v)^T(x - y) &\geq \|u - v\|^2 \end{aligned}$$

so R is 1-cocoercive. this implies R is nonexpansive, too:

$$\begin{aligned} \|u - v\|\|x - y\| &\geq \|u - v\|^2 \\ \|x - y\| &\geq \|u - v\| \end{aligned}$$

note: this also proves projections and proxs are nonexpansive

Cayley operator is nonexpansive

fact: $C = C_F$ is nonexpansive

proof:

Cayley operator is nonexpansive

fact: $C = C_F$ is nonexpansive

proof:

$$\begin{aligned} \|C(x) - C(y)\|^2 &= \|2(u-v) - (x-y)\|^2 \\ &= 4\|u-v\|^2 - 4(x-y)^T(u-v) + \|x-y\|^2 \\ &\leq \|x-y\|^2 \end{aligned}$$

using 1-cocoercivity of R (from above): $\|u-v\|^2 \leq (u-v)^T(x-y)$.

note:

- this also shows $R = \frac{1}{2}C + \frac{1}{2}I$ is $\frac{1}{2}$-averaged
- more generally, $\theta C + (1-\theta)I$ is θ-averaged for any $\theta \in (0,1)$

More contractions!

- if F is α-strongly monotone
- then $(I + F)$ is $1 + \alpha$-strongly monotone
- so $R_F = (I + F)^{-1}$ is $(1 + \alpha)$-cocoercive
- and hence R_F is $\frac{1}{1+\alpha}$-Lipschitz

More contractions!

- if F is α-strongly monotone
- then $(I + F)$ is $1 + \alpha$-strongly monotone
- so $R_F = (I + F)^{-1}$ is $(1 + \alpha)$-cocoercive
- and hence R_F is $\frac{1}{1+\alpha}$-Lipschitz

moral: we now have two ways to manufacture contractions:

- as a gradient update of an SSC function
- as a proximal update of a strongly convex function