

CME 307 / MS&E 311: Optimization

Operators

Professor Udell

Management Science and Engineering
Stanford

May 7, 2023

Announcements



Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

Basic inequality

recall basic inequality for convex differentiable f :

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

- ▶ first-order approximation of f at x is global underestimator
- ▶ $(\nabla f(x), -1)$ supports **epi** f at $(x, f(x))$

what if f is not differentiable?

Non-differentiable functions

are these functions differentiable?

- ▶ $|t|$ for $t \in \mathbf{R}$
- ▶ $\|x\|_1$ for $x \in \mathbf{R}^n$
- ▶ $\|X\|_*$ for $X \in \mathbf{R}^{n \times n}$
- ▶ $\max_i a_i^T x + b_i$ for $x \in \mathbf{R}^n$
- ▶ $\lambda_{\max}(X)$ for $X \in \mathbf{R}^{n \times n}$
- ▶ indicators of convex sets \mathcal{C}

if not, where? can we find underestimators for them?

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

picture

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

picture

Q: Can a function f have > 1 subgradient at a point x ?

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

picture

Q: Can a function f have > 1 subgradient at a point x ?

A: Yes, if f is nonsmooth at x

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

picture

Q: Can a function f have > 1 subgradient at a point x ?

A: Yes, if f is nonsmooth at x

Q: Can a function f have no subgradient at a point x ?

Subgradient of a function

g is a **subgradient** of f (not necessarily convex) at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

picture

Q: Can a function f have > 1 subgradient at a point x ?

A: Yes, if f is nonsmooth at x

Q: Can a function f have no subgradient at a point x ?

A: Yes, if x does not lie on convex hull of f

Subgradients and convexity

- ▶ g is a subgradient of f at x iff $(g, -1)$ supports **epi** f at $(x, f(x))$
- ▶ g is a subgradient iff $f(x) + g^T(y - x)$ is a global (affine) underestimator of f
- ▶ if f is convex and differentiable, $\nabla f(x)$ is a subgradient of f at x

subgradients come up in several contexts:

- ▶ algorithms for nondifferentiable convex optimization
- ▶ convex analysis, e.g., optimality conditions, duality for nondifferentiable problems

(if $f(y) \leq f(x) + g^T(y - x)$ for all y , then g is a **supergradient**)

Subdifferential

set of all subgradients of f at x is called the **subdifferential** of f at x , denoted $\partial f(x)$

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x) \quad \forall y\}$$

for any f ,

- ▶ $\partial f(x)$ is a closed convex set (can be empty)
- ▶ $\partial f(x) = \emptyset$ if $f(x) = \infty$

proof: use the definition

Subdifferential

set of all subgradients of f at x is called the **subdifferential** of f at x , denoted $\partial f(x)$

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x) \quad \forall y\}$$

for any f ,

- ▶ $\partial f(x)$ is a closed convex set (can be empty)
- ▶ $\partial f(x) = \emptyset$ if $f(x) = \infty$

proof: use the definition

if f is convex,

- ▶ $\partial f(x)$ is nonempty, for $x \in \text{relint dom } f$
- ▶ $\partial f(x) = \{\nabla f(x)\}$, if f is differentiable at x
- ▶ if $\partial f(x) = \{g\}$, then f is differentiable at x and $g = \nabla f(x)$

Compute subgradient via definition

$g \in \partial f(x)$ iff

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(x) = |x|$ for $x \in \mathbf{R}$. suppose $s \in \text{sign}(x)$, where

$$\text{sign}(x) = \begin{cases} \{1\} & x > 0 \\ [-1, 1] & x = 0 \\ -\{1\} & x < 0. \end{cases}$$

then

$$f(y) = \max(y, -y) \geq sy = s(x + y - x) = |x| + s(y - x)$$

Compute subgradient via definition

$g \in \partial f(x)$ iff

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(x) = |x|$ for $x \in \mathbf{R}$. suppose $s \in \text{sign}(x)$, where

$$\text{sign}(x) = \begin{cases} \{1\} & x > 0 \\ [-1, 1] & x = 0 \\ -\{1\} & x < 0. \end{cases}$$

then

$$f(y) = \max(y, -y) \geq sy = s(x + y - x) = |x| + s(y - x)$$

so $\text{sign}(x) \subseteq \partial f(x)$ (in fact, holds with equality)

Compute subgradient via definition

$g \in \partial f(x)$ iff

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(x) = |x|$ for $x \in \mathbf{R}$. suppose $s \in \text{sign}(x)$, where

$$\text{sign}(x) = \begin{cases} \{1\} & x > 0 \\ [-1, 1] & x = 0 \\ -\{1\} & x < 0. \end{cases}$$

then

$$f(y) = \max(y, -y) \geq sy = s(x + y - x) = |x| + s(y - x)$$

so $\text{sign}(x) \subseteq \partial f(x)$ (in fact, holds with equality)

picture

Compute subgradient via definition

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(x) = \max_i a_i^T x + b_i$.

Compute subgradient via definition

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(x) = \max_i a_i^T x + b_i$. then for any i ,

$$\begin{aligned} f(y) &= \max_i a_i^T y + b_i \\ &\geq a_i^T y + b_i \\ &= a_i^T (x + y - x) + b_i \\ &= a_i^T x + b_i + a_i^T (y - x) \\ &= f(x) + a_i^T (y - x), \end{aligned}$$

where the last line holds for $i \in \operatorname{argmax}_j a_j^T x + b_j$. so

- ▶ $a_i \in \partial f(x)$ for each $i \in \operatorname{argmax}_j a_j^T x + b_j$
- ▶ $\partial f(x)$ is convex, so

$$\operatorname{Co}\{a_i : i \in \operatorname{argmax}_j a_j^T x + b_j\} \subseteq \partial f(x)$$

Compute subgradient via definition

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(X) = \lambda_{\max}(X)$.

Compute subgradient via definition

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

example. let $f(X) = \lambda_{\max}(X)$. then

$$\begin{aligned} f(Y) &= \sup_{\|v\| \leq 1} v^T Y v \\ &= \sup_{\|v\| \leq 1} v^T (X + Y - X) v, \quad \|v\| \leq 1 \\ &= \sup_{\|v\| \leq 1} \left(v^T X v + v^T (Y - X) v \right), \quad \|v\| \leq 1 \\ &= v^T X v + \text{tr}(v v^T (Y - X)), \quad v \in \underset{\|v\| \leq 1}{\text{argmax}} v^T X v \\ &= \lambda_{\max}(X) + \text{tr}(v v^T (Y - X)), \quad v \in \underset{\|v\| \leq 1}{\text{argmax}} v^T X v \end{aligned}$$

- ▶ $v v^T \in \partial f(X)$ for each $v \in \underset{\|v\| \leq 1}{\text{argmax}} v^T X v$
- ▶ $\partial f(x)$ is convex, so

$$\text{Co}\{v v^T : v \in \underset{\|v\| \leq 1}{\text{argmax}} v^T X v\} \subseteq \partial f(x)$$

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

Properties of subgradients

subgradient inequality:

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

for convex f , we'll show

- ▶ subgradients are monotone: for any $x, y \in \text{dom } f$, $g_y \in \partial f(y)$, and $g_x \in \partial f(x)$,

$$(g_y - g_x)^T(y - x) \geq 0$$

- ▶ $\partial f(x)$ is continuous: if f is (lower semi-)continuous, $x^{(k)} \rightarrow x$, $g^{(k)} \rightarrow g$, and $g^{(k)} \in \partial f(x^{(k)})$ for each k , then $g \in \partial f(x)$
- ▶ $\partial f(x) = \text{argmax } g^T x - f(x)$

these will help us compute subgradients

Subgradients are monotone

fact. for any $x, y \in \text{dom } f$, $g_y \in \partial f(y)$, and $g_x \in \partial f(x)$,

$$(g_y - g_x)^T (y - x) \geq 0$$

proof. same as for differentiable case:

$$f(y) \geq f(x) + g_x^T (y - x) \quad f(x) \geq f(y) + g_y^T (x - y)$$

add these to get

$$(g_y - g_x)^T (y - x) \geq 0$$

Subgradients are preserved under limits

subgradient inequality:

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

fact. if f is (lower semi-)continuous, $x^{(k)} \rightarrow x$, $g^{(k)} \rightarrow g$, and $g^{(k)} \in \partial f(x^{(k)})$ for each k , then $g \in \partial f(x)$

proof.

Subgradients are preserved under limits

subgradient inequality:

$$g \in \partial f(x) \iff f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \text{dom}(f)$$

fact. if f is (lower semi-)continuous, $x^{(k)} \rightarrow x$, $g^{(k)} \rightarrow g$, and $g^{(k)} \in \partial f(x^{(k)})$ for each k , then $g \in \partial f(x)$

proof. For each k and for every y ,

$$\begin{aligned} f(y) &\geq f(x^{(k)}) + (g^{(k)})^T(y - x^{(k)}) \\ \lim_{k \rightarrow \infty} f(y) &\geq \lim_{k \rightarrow \infty} f(x^{(k)}) + (g^{(k)})^T(y - x^{(k)}) \\ f(y) &\geq f(x) + g^T(y - x) \end{aligned}$$

moral. To find a subgradient $g \in \partial f(x)$, find points $x^{(k)} \rightarrow x$ where f is differentiable, and let $g = \lim_{k \rightarrow \infty} \nabla f(x^{(k)})$.

Subgradients are preserved under limits: example

consider $f(x) = |x|$. we know

$$\partial f(x) = \begin{cases} \{-1\} & x < 0 \\ ? & x = 0 \\ \{1\} & x > 0 \end{cases}$$

so

- ▶ $\lim_{x \rightarrow 0^+} \nabla(x) = 1$
- ▶ $\lim_{x \rightarrow 0^-} \nabla(x) = -1$

hence

Subgradients are preserved under limits: example

consider $f(x) = |x|$. we know

$$\partial f(x) = \begin{cases} \{-1\} & x < 0 \\ ? & x = 0 \\ \{1\} & x > 0 \end{cases}$$

so

- ▶ $\lim_{x \rightarrow 0^+} \nabla(x) = 1$
- ▶ $\lim_{x \rightarrow 0^-} \nabla(x) = -1$

hence

- ▶ $-1 \in \partial f(0)$ and $1 \in \partial f(0)$
- ▶ $\partial f(0)$ is convex, so $[-1, 1] \subseteq \partial f(0)$
- ▶ and $\partial f(0)$ is monotone, so $[-1, 1] = \partial f(0)$

Convex functions can't be very non-differentiable

Theorem

Rockafellar, Convex Analysis, Thm 25.5 a convex function f is differentiable almost everywhere on the interior of its domain.

Convex functions can't be very non-differentiable

Theorem

Rockafellar, Convex Analysis, Thm 25.5 a convex function f is differentiable almost everywhere on the interior of its domain.

corollary: pick $x \in \text{dom } f$ uniformly at random. then f is differentiable at x w/prob 1.

Convex functions can't be very non-differentiable

Theorem

Rockafellar, Convex Analysis, Thm 25.5 a convex function f is differentiable almost everywhere on the interior of its domain.

corollary: pick $x \in \text{dom } f$ uniformly at random. then f is differentiable at x w/prob 1.

corollary: For a convex function f and any x , there is a sequence of points $x^{(k)} \rightarrow x$ where f is differentiable.

Subgradients and fenchel conjugates

fact. $g \in \partial f(x) \iff f^*(g) + f(x) = g^T x$

(recall the conjugate function $f^*(g) = \sup_x g^T x - f(x)$.)

Subgradients and fenchel conjugates

proof. if $f^*(g) + f(x) = g^T x$,

$$\begin{aligned} f^*(g) &= \sup_y g^T y - f(y) \\ &\geq g^T y - f(y) \quad \forall y \\ f(y) &\geq g^T y - f^*(g) \quad \forall y \\ &= g^T y - g^T x + f(x) \quad \forall y \\ &= g^T (y - x) + f(x) \quad \forall y \end{aligned}$$

so $g \in \partial f(x)$. conversely, if $g \in \partial f(x)$,

Subgradients and fenchel conjugates

proof. if $f^*(g) + f(x) = g^T x$,

$$\begin{aligned}f^*(g) &= \sup_y g^T y - f(y) \\&\geq g^T y - f(y) \quad \forall y \\f(y) &\geq g^T y - f^*(g) \quad \forall y \\&= g^T y - g^T x + f(x) \quad \forall y \\&= g^T (y - x) + f(x) \quad \forall y\end{aligned}$$

so $g \in \partial f(x)$. conversely, if $g \in \partial f(x)$,

$$\begin{aligned}f(y) &\geq g^T (y - x) + f(x) \\g^T x - f(x) &\geq g^T y - f(y) \\\sup_y g^T x - f(x) &\geq \sup_y g^T y - f(y) \\g^T x - f(x) &\geq f^*(g)\end{aligned}$$

Subgradients and fenchel conjugates

Conclusion.

$$\begin{aligned}g \in \partial f(x) &\iff f^*(g) + f(x) = g^T x \\ &\iff x \in \operatorname{argmax}_x g^T x - f(x)\end{aligned}$$

consider the same implications for the function f^* :

$$\begin{aligned}x \in \partial f^*(g) &\iff f(x) + f^*(g) = x^T g \\ &\iff g \in \operatorname{argmax}_g g^T x - f^*(g)\end{aligned}$$

so all these conditions are equivalent, and

$$g \in \partial f(x) \iff x \in \partial f^*(g)!$$

Compute subgradient via fenchel conjugate

$$\partial f(x) = \operatorname{argmax}_g g^T x - f^*(g)$$

example. let $f(x) = \|x\|_1$. compute

$$f^*(g) = \sup_x g^T x - \|x\|_1$$

=

Compute subgradient via fenchel conjugate

$$\partial f(x) = \operatorname{argmax}_g g^T x - f^*(g)$$

example. let $f(x) = \|x\|_1$. compute

$$\begin{aligned} f^*(g) &= \sup_x g^T x - \|x\|_1 \\ &= \begin{cases} 0 & \|g\|_\infty \leq 1 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

Compute subgradient via fenchel conjugate

$$\partial f(x) = \operatorname{argmax}_g g^T x - f^*(g)$$

example. let $f(x) = \|x\|_1$. compute

$$\begin{aligned} f^*(g) &= \sup_x g^T x - \|x\|_1 \\ &= \begin{cases} 0 & \|g\|_\infty \leq 1 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

given x ,

$$\begin{aligned} \partial f(x) &= \operatorname{argmax}_g g^T x - f^*(g) \\ &= \operatorname{argmax}_{\|g\|_\infty \leq 1} g^T x \\ &= \mathbf{sign}(x) \end{aligned}$$

where **sign** is computed elementwise.

Compute subgradient via fenchel conjugate

$$\partial f(x) = \operatorname{argmax}_g g^T x - f^*(g)$$

example. let $f(X) = \|X\|_*$. compute

$$f^*(G) = \sup_X \operatorname{tr}(G, X) - \|X\|_*$$

=

Compute subgradient via fenchel conjugate

$$\partial f(x) = \operatorname{argmax}_g g^T x - f^*(g)$$

example. let $f(X) = \|X\|_*$. compute

$$\begin{aligned} f^*(G) &= \sup_X \operatorname{tr}(G, X) - \|X\|_* \\ &= \begin{cases} 0 & \|G\| \leq 1 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

where $\|G\| = \sigma_1(G)$ is the operator norm of G .

Compute subgradient via fenchel conjugate

$$\partial f(x) = \operatorname{argmax}_g g^T x - f^*(g)$$

example. let $f(X) = \|X\|_*$. compute

$$\begin{aligned} f^*(G) &= \sup_X \operatorname{tr}(G, X) - \|X\|_* \\ &= \begin{cases} 0 & \|G\| \leq 1 \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

where $\|G\| = \sigma_1(G)$ is the operator norm of G .

given $X = U \mathbf{diag}(\sigma) V^T$,

$$\begin{aligned} \partial f(x) &= \operatorname{argmax}_G \operatorname{tr}(G, X) - f^*(G) \\ &= \operatorname{argmax}_{\|G\| \leq 1} \operatorname{tr}(G, X) \\ &= U \mathbf{diag}(\mathbf{sign}(\sigma)) V^T \end{aligned}$$

where **sign** is computed elementwise.

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

Subgradient method

the **subgradient method** is a simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- ▶ $x^{(k)}$ is the k th iterate
- ▶ $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- ▶ $\alpha_k > 0$ is the k th step size

Subgradient method

the **subgradient method** is a simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- ▶ $x^{(k)}$ is the k th iterate
- ▶ $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- ▶ $\alpha_k > 0$ is the k th step size

warning: subgradient method is **not** a descent method.

Subgradient method

the **subgradient method** is a simple algorithm to minimize nondifferentiable convex function f

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

- ▶ $x^{(k)}$ is the k th iterate
- ▶ $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- ▶ $\alpha_k > 0$ is the k th step size

warning: subgradient method is **not** a descent method.
instead, keep track of best point so far

$$f_{\text{best}}^{(k)} = \min_{i=1,\dots,k} f(x^{(i)})$$

How to avoid slow convergence

don't use subgradient method for very high accuracy!

instead,

- ▶ for high accuracy: rewrite problem as LP or SDP; use IPM
- ▶ for medium accuracy:

- ▶ regularize your objective (so it's strongly convex)

$$\tilde{f}(x) = f(x) + \alpha \|x - x^0\|^2$$

- ▶ smooth your objective (so it's smooth)

$$\tilde{f}(x) = \mathbb{E}_{y: \|y-x\| \leq \delta} f(y)$$

- ▶ infimal convolution (so it's smooth and strongly convex):

$$\tilde{f}(x) = \inf_y f(y) + \frac{\rho}{2} \|y - x\|^2$$

- ▶ more on these later...
 - ▶ for low accuracy: use a constant step size; terminate when you stop improving much or get bored

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

Proximal operator

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

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

Proximal operator

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

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

► $\mathbf{prox}_f : \mathbf{R}^d \rightarrow \mathbf{R}^d$

Proximal operator

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

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

- ▶ $\mathbf{prox}_f : \mathbf{R}^d \rightarrow \mathbf{R}^d$
- ▶ **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 $f : \mathbf{R}^d \rightarrow \mathbf{R}$

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

- ▶ $\mathbf{prox}_f : \mathbf{R}^d \rightarrow \mathbf{R}^d$
- ▶ **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 $f : \mathbf{R}^d \rightarrow \mathbf{R}$,

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

Let's evaluate some proximal operators!

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

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

► $f(x) = 0$

Let's evaluate some proximal operators!

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

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

- ▶ $f(x) = 0$ (identity)

Let's evaluate some proximal operators!

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

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

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

Let's evaluate some proximal operators!

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

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

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

Let's evaluate some proximal operators!

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

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

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

Let's evaluate some proximal operators!

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

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

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

Let's evaluate some proximal operators!

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

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

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

Let's evaluate some proximal operators!

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

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

- ▶ $f(x) = 0$ (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 $f : \mathbf{R}^d \rightarrow \mathbf{R}$

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

- ▶ $f(x) = 0$ (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 $f : \mathbf{R}^d \rightarrow \mathbf{R}$

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

- ▶ $f(x) = 0$ (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)

Let's evaluate some proximal operators!

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

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

- ▶ $f(x) = 0$ (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)
- ▶ $f(x) = \|x\|_1$

Let's evaluate some proximal operators!

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

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

- ▶ $f(x) = 0$ (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 $f : \mathbf{R}^d \rightarrow \mathbf{R}$

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

- ▶ $f(x) = 0$ (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)
- ▶ $f(X) = \|X\|_*$

Let's evaluate some proximal operators!

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

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

- ▶ $f(x) = 0$ (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)
- ▶ $f(X) = \|X\|_*$ (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

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

Proximal gradient method

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

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

using proximal operators together with gradient steps?

Proximal gradient method

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

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

using proximal operators together with gradient steps? idea:

$$x^+ = \mathbf{prox}_{tg}(x - t\nabla f(x))$$

- ▶ the proximal operator gives a **fast method** to step towards the minimum of g
- ▶ gradient method works well to step towards minimum of f

Proximal gradient: examples

- ▶ projected gradient $g(x) = \mathbf{1}(\Omega)(x)$
- ▶ nonnegative least squares: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$,
 $g(x) = \lambda\|x\|_1$
- ▶ lasso: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, $g(x) = \lambda\|x\|_1$
- ▶ ...

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

Functions

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

- ▶ closed: **epi**(f) is a closed set
- ▶ convex: f is convex
- ▶ proper: **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 $f : \mathbf{R}^n \rightarrow \mathbf{R}^n$ defines a relation $\{(x, f(x)) : x \in \mathbf{dom} f\}$

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

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

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

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

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 $f(x)$

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 $L < 1$

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 $t \in (0, \frac{2\alpha}{\beta^2})$.

proof:

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

Gradient update is contractive for SSC functions

suppose f is α -strongly convex and β -smooth
the relation

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

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

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 shows $I - t\nabla f$ is Lipschitz with parameter $L = \max\{|1 - t\alpha|, |1 - t\beta|\}$. if $t = \frac{2}{\alpha + \beta}$, $L = \frac{\kappa - 1}{\kappa + 1}$
(see Ryu and Yin exercise 2.16)

Proximal map is nonexpansive

the proximal map of any convex function f is nonexpansive:

$$\|\mathbf{prox}_f(y) - \mathbf{prox}_f(x)\| \leq \|y - x\|$$

Proximal map is nonexpansive

the proximal map of any convex function f is nonexpansive:

$$\|\mathbf{prox}_f(y) - \mathbf{prox}_f(x)\| \leq \|y - x\|$$

proof: let $u = \mathbf{prox}_f(x)$ and $v = \mathbf{prox}_f(y)$, so

$$x - u \in \partial f(u), \quad y - v \in \partial f(v)$$

then by the subgradient inequality,

$$f(v) \geq f(u) + \langle x - u, v - u \rangle \quad \text{and} \quad f(u) \geq f(v) + \langle y - v, u - v \rangle$$

add these to show

$$\begin{aligned} 0 &\geq \langle y - x + u - v, u - v \rangle \\ \langle x - y, u - v \rangle &\geq \|u - v\|^2 \end{aligned}$$

Proximal map is nonexpansive

the proximal map of any convex function f is nonexpansive:

$$\|\mathbf{prox}_f(y) - \mathbf{prox}_f(x)\| \leq \|y - x\|$$

proof: let $u = \mathbf{prox}_f(x)$ and $v = \mathbf{prox}_f(y)$, so

$$x - u \in \partial f(u), \quad y - v \in \partial f(v)$$

then by the subgradient inequality,

$$f(v) \geq f(u) + \langle x - u, v - u \rangle \quad \text{and} \quad f(u) \geq f(v) + \langle y - v, u - v \rangle$$

add these to show

$$\begin{aligned} 0 &\geq \langle y - x + u - v, u - v \rangle \\ \langle x - y, u - v \rangle &\geq \|u - v\|^2 \end{aligned}$$

shows **firm nonexpansiveness** of \mathbf{prox}_f .

now use Cauchy-Schwarz to show \mathbf{prox}_f is nonexpansive

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,

$$\|x - y\| = \|F(x) - F(y)\| < \|x - y\| \text{ contradiction}$$

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:

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

(iteration is **Fejer-monotone**)

- ▶ linear convergence with rate L

Proof

proof:

Proof

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

► sequence $x^{(k)}$ is Cauchy:

$$\begin{aligned}\|x^{(k+\ell)} - x^{(k)}\| &\leq \|x^{(k+\ell)} - x^{(k+\ell-1)}\| + \dots + \|x^{(k+1)} - x^{(k)}\| \\ &\leq (L^{\ell-1} + \dots + 1)\|x^{(k+1)} - x^{(k)}\| \\ &\leq \frac{1}{1-L}\|x^{(k+1)} - x^{(k)}\| \\ &\leq \frac{L^k}{1-L}\|x^{(1)} - x^{(0)}\|\end{aligned}$$

► so it converges to a point x^* . must be the (unique) FP!

► converges to x^* linearly with rate L

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

Outline

Subgradients

Subgradient properties

Subgradient method

Proximal operators

Proximal gradient method

Relations

Fixed points

Averaged operators

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 $x^* = 0$.
- ▶ 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 is 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 is FP of $F \iff x$ is FP of G

proof:

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

\implies 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: **Ryu and Boyd; 2015**

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

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