Proximal Gradient Method

Pontus Giselsson

Today's lecture

- Proximal gradient method
- Key properties for convergence

Problem formulation

- *Empirical risk minimization* problems are of form

$$\underset{x}{\text{minimize}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} f_i(x)}_{f(x)} + g(x)$$

- We assume that:
  - all f_i and g are convex
  - all f_i are differentiable with L_i Lipschitz gradient

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$$

  - f has L-Lipschitz gradient
  - g is not necessarily differentiable (1-norm, indicator of set)
  - g is (typically) separable (often with $g_1 = \ldots = g_n$)

$$g(x) = \sum_{i=1}^{n} g_i(x_i)$$

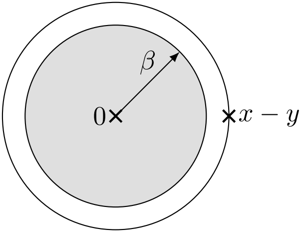
- (Note x and y are variables here, not data! Also x_i is ith element)

Lipschitz continuity

- The gradient ∇f is β-Lipschitz continuous if

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$$

  holds for all x, y
- Graphical representation ($\nabla f(x) - \nabla f(y)$ in gray area)



- 1-Lipschitz is called *nonexpansive*

Gradient method with Lipschitz gradient

- Let γ be such that $\gamma\nabla f$ is $\frac{1}{2}$-Lipschitz
- Gradient method

$$x_{k+1} := (I - \gamma\nabla f)x_k$$

tries to solve problem (which is special case with $g \equiv 0$):

$$\underset{x}{\text{minimize}}\, f(x)$$

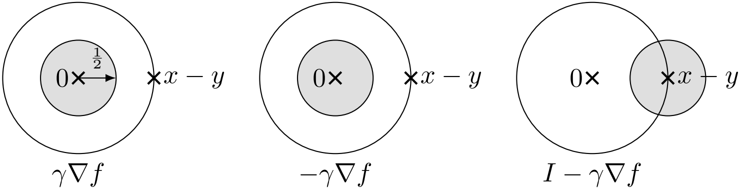
- Problem solved if $\nabla f(x^\star) = 0$, i.e., if

$$x^\star = (I - \gamma\nabla f)x^\star,$$

i.e., $x^\star$ is fixed-point of forward (gradient) step $(I - \gamma\nabla f)$

Gradient mapping properties

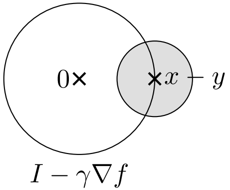
- The gradient mapping $G := I - \gamma\nabla f$ satisfies
- (recall $\gamma\nabla f$ is $\frac{1}{2}$-Lipschitz)



- Rightmost figure shows where $G(x) - G(y)$ can end up

Gradient mapping properties

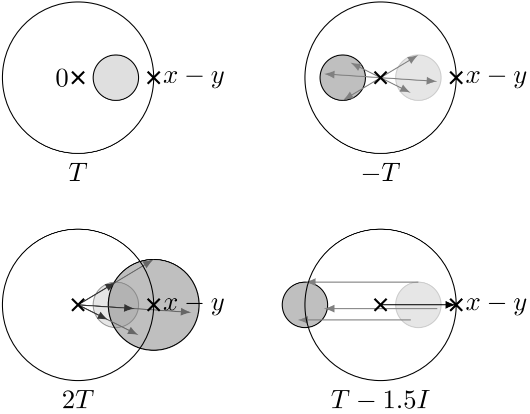
- Let $y = x^\star$ with $x^\star$ fixed-point of $I - \gamma\nabla f$ and shift figure by $x^\star$:



- Right figure shows where $G(x) - G(x^\star) + x^\star = G(x)$ can end up
- Gradient step $G(x) = I - \gamma\nabla f$ can take you further away from $x^\star$
- Gradient method does not work?
- (Recall: fixed-point $x^\star$ of $G = I - \gamma\nabla f$ is solution to problem)

Circle exercise

- If $Tx - Ty$ ends up in gray area, given $x - y$, how about $\alpha T + \beta I$?
- Take every possible $v = Tx - Ty$ and compute $\alpha v + \beta(x - y)$



Convexity

- We have not exploited that f is convex
- A differentiable function is convex if and only if for all x, y

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

- Gradient satisfies (add two copies with x and y swapped):

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

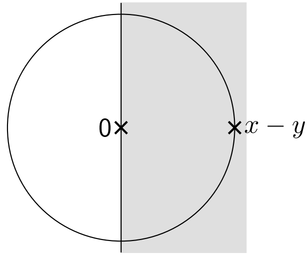
which is referred to that ∇f is *monotone*

Monotone operator

- Monotonicity of ∇f:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

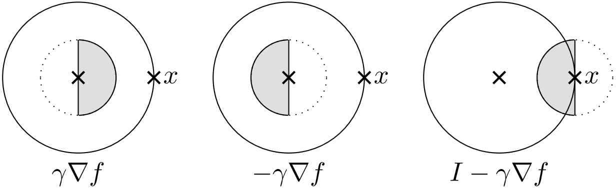
- Graphical representation



then $\nabla f(x) - \nabla f(y)$ in gray area (since scalar product positive)

Lipschitz and monotone

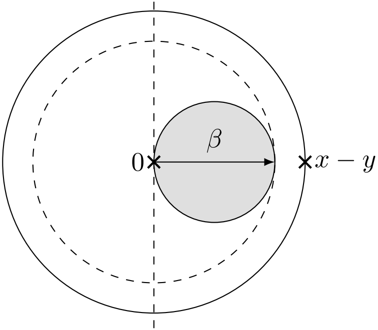
- $\gamma \nabla f$ monotone and 0.5-Lipschitz:



- May still become further away from fixed-point after iteration

Baillon-Haddad theorem — Cocoercivity

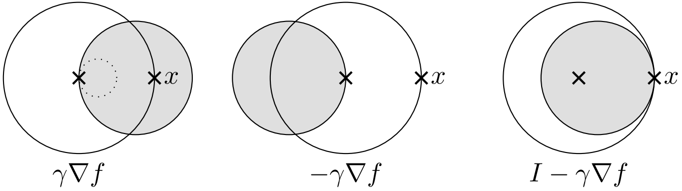
- If f is convex (∇f monotone) and ∇f β-Lipschitz
- Then ∇f is $\frac{1}{\beta}$-*cocoercive*: $\nabla f = \frac{\beta}{2}(I - N)$ with N nonexpansive
- Graphical representation



- Always: $\frac{1}{\beta}$-cocoercive implies β-Lipschitz
- For gradient of convex functions, converse implication holds
- This is known as Baillon-Haddad theorem

Gradient mapping properties

- ∇f is $\frac{1}{\beta}$-cocoercive with $\beta = \frac{1}{2}$
- $I - \gamma\nabla f$ with $\gamma = 3$:



- We have

$$I - \gamma\nabla f = I - \gamma\tfrac{\beta}{2}(I - N) = (1 - \tfrac{\gamma\beta}{2})I + \tfrac{\gamma\beta}{2}N$$

Averaged operators

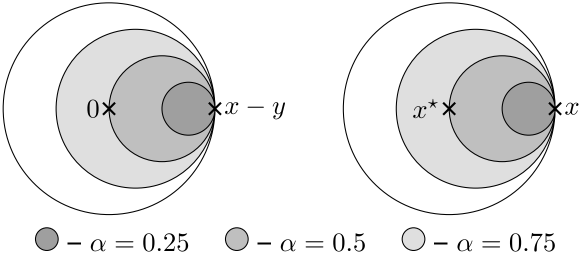
- For cocoercive ∇f, gradient mapping satisfies

$$I - \gamma\nabla f = (1 - \tfrac{\gamma\beta}{2})I + \tfrac{\gamma\beta}{2}N$$

  with N nonexpansive
- Operators T of the form

$$T = (1-\alpha)I + \alpha N$$

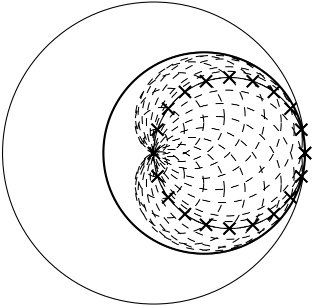
  with $\alpha \in (0,1)$ are called averaged
- Graphical representation (right: $x^\star = Tx^\star$ and shifted by $x^\star$)



- Gradient mapping $I - \gamma\nabla f$ averaged if $\frac{\gamma\beta}{2} \in (0,1)$

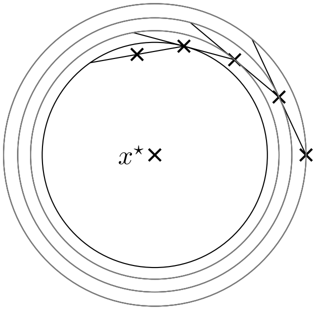
Composition of averaged operators

- composition of averaged operators is averaged
- assume that T_1 is α_1-averaged and T_2 is α_2-averaged, $\alpha_i \in (0, 1)$
- then $T_2 T_1$ is $\frac{\alpha}{\alpha+1}$-averaged with $\alpha = \frac{\alpha_1}{1-\alpha_1} + \frac{\alpha_2}{1-\alpha_2}$
- example $\alpha_1 = \alpha_2 = 0.5 \Rightarrow T_1 T_2$ is $\frac{2}{3}$-averaged



Iteration example - $\alpha = 0.5$

- rotation operator R_θ with $\theta = 50^\circ$ (nonexpansive)
- fixed-point $x^\star$ at origin
- iterate 0.5-averaged operator



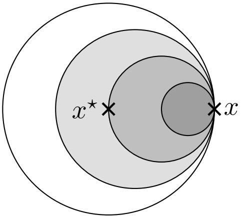
Convergence – Intuition from figures

- Let T be α-averaged with $\alpha \in (0,1)$
- Then

$$x_{k+1} = Tx_k$$

converges to fixed-point of T (provided it exists)

- Intuition: sufficiently much closer to fixed-point in every iteration



Convergence – Theory

1. Let T be α-averaged and R be 2-cocoercive, then

$$T = I - \alpha R$$

   (fixed-point of $T \Leftrightarrow$ zero of R)

2. R is β-cocoercive if and only if

$$\langle Rx - Ry, x - y \rangle \geq \beta \|Rx - Ry\|^2$$

3. Derive algorithm inequality and use: $x_k \to \mathrm{fix} T$ if (and only if)
   - $Rx_k \to 0$ as $k \to \infty$
   - $\|x_k - x^\star\|$ converges for all $x^\star \in \mathrm{fix} T$

Part 1

- Recall α-averaged T

$$T = (1 - \alpha)I + \alpha N$$

- Recall β-cocoercive R

$$R = \frac{1}{2\beta}(I - N)$$

- Therefore

$$T = (1 - \alpha)I + \alpha N = I - \alpha(I - N) = I - \alpha R$$

for $\frac{1}{2}$-cocoercive R

Part 2

- Recall β-cocoercive R

$$R = \tfrac{1}{2\beta}(I - N)$$

- Therefore

$$\begin{aligned}
\|Rx - Ry\|^2 &= \|\tfrac{1}{2\beta}(x - Nx) - \tfrac{1}{2\beta}(y - Ny)\|^2 \\
&= \tfrac{1}{4\beta^2}(\|x - y\|^2 + \|Nx - Ny\|^2 - 2\langle x - y, Nx - Ny\rangle \\
&= \tfrac{1}{4\beta^2}(-\|x - y\|^2 + \|Nx - Ny\|^2 \\
&\quad + 2\langle x - y, x - y - (Nx - Ny)\rangle) \\
&\le \tfrac{1}{2\beta^2}\langle x - y, x - y - (Nx - Ny)\rangle \\
&= \tfrac{1}{\beta}\langle x - y, Rx - Ry\rangle
\end{aligned}$$

Part 3

- Algorithm $x_{k+1} = Tx_k = x_k - \alpha R x_k$ satisfies

$$
\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|x_k - \alpha R x_k - x^\star\|^2 \\
&= \|x_k - x^\star\|^2 - 2\alpha\langle Rx_k, x_k - x^\star\rangle + \alpha^2\|Rx_k\|^2 \\
&= \|x_k - x^\star\|^2 - 2\alpha\langle Rx_k - Rx^\star, x_k - x^\star\rangle + \alpha^2\|Rx_k - Rx^\star \\
&\leq \|x_k - x^\star\|^2 - \alpha(1-\alpha)\|Rx_k - Rx^\star\|^2 \\
&= \|x_k - x^\star\|^2 - \alpha(1-\alpha)\|Rx_k\|^2
\end{aligned}
$$

- Let $\alpha \in (0, 1)$ to conclude that
  - $Rx_k \to 0$ as $k \to \infty$
  - $\|x_k - x^\star\|$ converges for all $x^\star \in \mathrm{fix} T$

  and apply result to get $x_k \to x^\star \in \mathrm{fix} T$

Summary so far

- Have considered gradient method for

$$\underset{x}{\text{minimize}}\, f(x)$$

  where f is convex and differentiable with Lipschitz gradient
- Important property for convergence; Cocoercivity of gradient
- Follows from Baillon-Haddad theorem
- Implies that gradient mapping is iteration of averaged map

Composite form

- Next, consider composite form

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

  where f as before and g convex and nonsmooth

- Handle f via gradient as before
- Handle g via *proximal operator*

$$\text{prox}_{\gamma g}(z) = \underset{x}{\text{argmin}}(g(x) + \tfrac{1}{2\gamma}\|x - z\|_2^2)$$

  where $\gamma > 0$ is a parameter

Prox is generalization of projection

- Introduce the indicator function of a set C

$$\iota_C(x) := \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} \end{cases}$$

(we can use extended valued functions that take value ∞)

- Then

$$\begin{aligned} \Pi_C(z) &= \underset{x}{\operatorname{argmin}}(\|x - z\|_2 : x \in C) \\ &= \underset{x}{\operatorname{argmin}}(\tfrac{1}{2}\|x - z\|_2^2 : x \in C) \\ &= \underset{x}{\operatorname{argmin}}(\tfrac{1}{2}\|x - z\|_2^2 + \iota_C(x)) \\ &= \operatorname{prox}_{\iota_C}(z) \end{aligned}$$

projection onto C equals prox of indicator function of C

Examples of proximal operators

- Quadratic function, $g(x) = \frac{1}{2}x^T H x + h^T x$:

$$\mathrm{prox}_{\gamma g}(z) = (I + \gamma H)^{-1}(z - \gamma h)$$

- The squared 2-norm, $g(x) = \frac{1}{2}\|x\|_2^2$:

$$\mathrm{prox}_{\gamma g}(z) = (1 + \gamma)^{-1} z$$

- The 2-norm, $g(x) = \|x\|_2$:

$$\mathrm{prox}_{\gamma g}(z) = \begin{cases} (1 - \gamma/\|z\|_2)z & \text{if } \|z\|_2 \geq \gamma \\ 0 & \text{otherwise} \end{cases}$$

- Affine subspace, $V = \{x : Ax = b\}$:

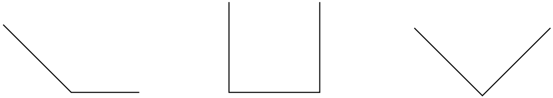
$$\mathrm{prox}_{\iota_V}(z) = \Pi_V(z) = z - A^T(AA^T)^{-1}(Az - b)$$

Piece-wise linear function

- Define $h_i \; : \; \mathbb{R} \to \overline{\mathbb{R}}$ is

$$h_i(x) = \begin{cases} c_l(l - x) & \text{if } x \le l \\ 0 & \text{if } l \le x \le u \\ c_u(x - u) & \text{if } x \ge u \end{cases}$$

  where $c_l, c_u \in (0, \infty]$ (∞ included) and $l \le u$
- graphical representations of different h_i



$c_l = 1,\ c_u = 0$, $l = 1,\ u = \infty$ $c_l = c_u = \infty$, $l = -1,\ u = 1$ $c_l = c_u = 1$, $l = u = 0$

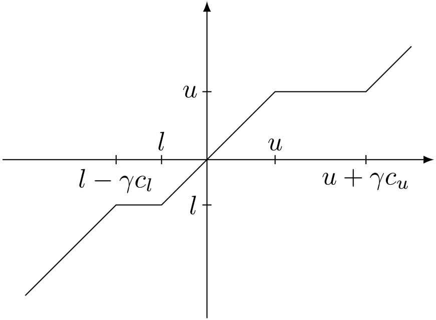
- special cases of h_i
  - hinge loss (SVM)
  - upper and lower bounds
  - "soft" upper and lower bounds
  - absolute value

Prox of h_i

- Prox of h_i:

$$\text{prox}_{\gamma h_i}(z) = \begin{cases} z + \gamma c_l & \text{if } z \leq l - \gamma c_l \\ l & \text{if } l - \gamma c_l \leq z \leq l \\ z & \text{if } l \leq z \leq u \\ u & \text{if } u \leq z \leq u + \gamma c_u \\ z - \gamma c_u & \text{if } z \geq u + \gamma c_u \end{cases}$$

- Graphical representation ($l = -1, u = 1.5, \gamma c_l = 1, \gamma c_u = 2$):



Examples prox h_i

- Hinge loss, $g = h_i$ with $l = 1$, $u = \infty$, $c_l = 1$, $c_u = 0$:

$$\mathrm{prox}_{\gamma g}(z) = \begin{cases} z + \gamma & \text{if } z \leq 1 - \gamma \\ 1 & \text{if } 1 - \gamma \leq z \leq 1 \\ z & \text{if } z \geq 1 \end{cases}$$

- Absolute value, $g = h_i$ with $l = u = 0$ and $c_l = c_u = 1$:

$$\mathrm{prox}_{\gamma g}(z) = \begin{cases} z + \gamma & \text{if } z \leq -\gamma \\ 0 & \text{if } -\gamma \leq z \leq \gamma \\ z - \gamma & \text{if } z \geq \gamma \end{cases}$$

- Upper and lower bounds, $g = h_i$ with $l < u$ and $c_l = c_u = \infty$:

$$\mathrm{prox}_{\gamma g}(z) = \begin{cases} l & \text{if } z \leq l \\ z & \text{if } l \leq z \leq u \\ u & \text{if } u \leq z \end{cases}$$

Computational cost

- Computing prox requires solving optimization problem

$$\mathrm{prox}_{\gamma g}(z) = \underset{x}{\mathrm{argmin}}(g(x) + \tfrac{1}{2\gamma}\|x - z\|_2^2)$$

- Prox typically more expensive to evaluate than gradient
- Example: Quadratic $g(x) = \frac{1}{2}x^T Hx + h^T x$:

$$\mathrm{prox}_{\gamma g}(z) = (I + \gamma H)^{-1}(z - \gamma h), \qquad \nabla g(z) = Hz - h$$

- Often use prox for nondifferentiable and separable functions

Prox for separable functions

- Separable function

$$g(x) = \sum_{i=1}^{n} g_i(x_i)$$

where $x = (x_1, \dots, x_n)$:

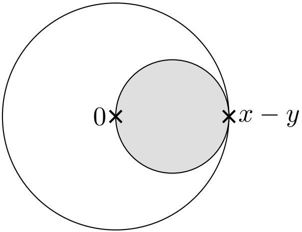
$$\operatorname{prox}_{\gamma g}(z) = \begin{bmatrix} \operatorname{prox}_{\gamma g_1}(z_1) \\ \vdots \\ \operatorname{prox}_{\gamma g_n}(z_n) \end{bmatrix}$$

- Decomposes into n individual proxes $\Rightarrow$ cheap to evaluate
- 1-norm $\|x\|_1$, upper/lower bounds, hinge loss constructed from h_i

Property of proximal operator

- Proximal operator is 1-Lipschitz, i.e., nonexpansive
- It is also gradient of convex function
- Hence, it is 1-cocoercive, i.e., $\frac{1}{2}$-averaged

$$\mathrm{prox}_{\gamma f} = \underbrace{\tfrac{1}{2}(I+N)}_{1-\text{cocoercive}} = \underbrace{(1-\tfrac{1}{2})I + \tfrac{1}{2}N}_{\frac{1}{2}-\text{averaged}}$$



- This property makes it useful for algorithms

Proximal gradient method

- Applicable to models

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

- The method iterates

$$x_{k+1} = \text{prox}_{\gamma g}(I - \gamma \nabla f) x_k$$

- Prox generalizes projection $\Rightarrow$ generalizes projected gradient
- Easily implemented using `ProximalOperators` package in Julia

Why does it work?

- The point $x^\star$ solves

$$\operatorname*{minimize}_{x} f(x) + g(x)$$

if and only if fixed-point to proximal gradient mapping

$$x^\star = \mathrm{prox}_{\gamma g}(I - \gamma \nabla f)x^\star$$

- Iteration of $\mathrm{prox}_{\gamma g}(I - \gamma \nabla f)x^\star$ converges to fixed-point – why?

Convergence

- Know gradient mapping $\frac{\gamma\beta}{2}$-averaged if $\gamma \in (0, \frac{2}{\beta})$
- Know that $\mathrm{prox}_{\gamma f}$ is $\frac{1}{2}$-averaged for all $\gamma > 0$
- Composition $\mathrm{prox}_{\gamma g}(I - \gamma \nabla f)$ is therefore also averaged
- Iteration of averaged map converges to fixed-point, i.e., solution

Another way to prove convergence

- Can prove convergence in similar but different way
- Use nonexpansiveness of $\mathrm{prox}_{\gamma g}$ and $\frac{1}{\beta}$-cocoercivity of ∇f

$$
\begin{aligned}
\|x_{k+1} - x^\star\|^2 &= \|\mathrm{prox}_{\gamma g}(I - \gamma\nabla f)x_k - \mathrm{prox}_{\gamma g}(I - \nabla f)x^\star\|^2 \\
&\leq \|x_k - \gamma\nabla f(x_k) - (x^\star - \gamma\nabla f(x^\star))\|^2 \\
&= \|x_k - x^\star\|^2 - 2\gamma\langle \nabla f(x_k) - \nabla f(x^\star), x_k - x^\star\rangle \\
&\quad + \gamma^2\|\nabla f(x_k) - \nabla f(x^\star)\|^2 \\
&= \|x_k - x^\star\|^2 - \gamma(\tfrac{2}{\beta} - \gamma)\|\nabla f(x_k) - \nabla f(x^\star)\|^2
\end{aligned}
$$

- Sufficient decrease if $\gamma \in (0, \frac{2}{\beta})$, just like gradient method

Summary

- f convex and ∇f Lipschitz $\Rightarrow$ ∇f cocoercive (Baillon-Haddad)
- ∇f cocoercive implies $I - \gamma\nabla f$ averaged (for small γ)
- $\mathrm{prox}_{\gamma g}$ is $\frac{1}{2}$-averaged
- Composition of averaged is averaged; $\mathrm{prox}_{\gamma g}(I - \gamma\nabla f)$ averaged
- Iteration of averaged operator converges to fixed-point
- Fixed-point of $\mathrm{prox}_{\gamma g}(I - \gamma\nabla f)$ is solution to problem

Next lecture

- Apply method to formulations from Lecture 1
- Modify method to exploit structure
  - Stochastic gradients
  - Coordinate-wise updates