Optimization for data science

Non-smooth optimization: Proximal methods

R. Flamary

Master Data Science, Institut Polytechnique de Paris

October 29, 2024

Full course overview

1. **Introduction to optimization for data science**
   1.1 ML optimization problems and linear algebra recap
   1.2 Optimization problems and their properties (Convexity, smoothness)
2. **Smooth optimization : Gradient descent**
   2.1 First order algorithms, convergence for smooth and strongly convex functions
3. **Smooth Optimization : Quadratic problems**
   3.1 Solvers for quadratic problems, conjugate gradient
   3.2 Linesearch methods
4. **Non-smooth Optimization : Proximal methods**
   4.1 Proximal operator and proximal algorithms
   4.2 Lab 1: Lasso and group Lasso
5. **Stochastic Gradient Descent**
   5.1 SGD and variance reduction techniques
   5.2 Lab 2: SGD for Logistic regression
6. **Standard formulation of constrained optimization problems**
   6.1 LP, QP and Mixed Integer Programming
7. **Coordinate descent**
   7.1 Algorithms and Labs
8. **Newton and quasi-newton methods**
   8.1 Second order methods and Labs
9. **Beyond convex optimization**
   9.1 Nonconvex reg., Frank-Wolfe, DC programming, autodiff

Current course overview

1. Introduction to optimization 4

2. Smooth optimization : Gradient descent 4

3. Smooth Optimization : Quadratic problems 4

4. Non-smooth optimization : Proximal methods 4

4.1 Non-smooth optimization and definitions 4

- 4.1.1 Non-smooth Machine Learning problems
- 4.1.2 Optimality and subgradient
- 4.1.3 Subgradient methods

4.2 Proximal Gradient descent 23

- 4.2.1 Majorization Minimization and proximal operator
- 4.2.2 Proximal Gradient Descent and application to Lasso
- 4.2.3 Accelerated Proximal Gradient Descent (APGD)

4.3 Other proximal methods and Primal Dual Algorithms 49

- 4.3.1 Primal-Dual Algorithms
- 4.3.2 Alternating Direction Method of Multipliers (ADMM)

4.4 Conclusion 53

5. Stochastic Gradient Descent 54

6. Standard formulation of constrained optimization problems 54

7. Coordinate descent 54

8. Newton and quasi-newton methods 54

9. Beyond convex optimization 54

Nonsmooth optimization

Optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; F(\mathbf{x}), \tag{1}$$

- F is convex, proper, lower semi-continuous can be non smooth, non continuous.
- Can be constrained optimization with $F(\mathbf{x}) = f(\mathbf{x}) + \chi_{\mathcal{C}}(\mathbf{x})$.
- General strategy : use the structure of F, find fast iterations.

Optimization strategies

- Subgradient descent: slower than GD, used for training NN.
- Proximal Splitting : divide an conquer strategy, can be accelerated.
- Projected Gradient Descent : special case of proximal splitting.
- Conditional Gradient : Use a linearization of F (see last course).

Constraints VS non-smooth

Characteristic function

Let A be a subset of $\mathbb{R}^n$, the **characteristic function** χ_A of A is the function

$$\chi_A(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \\ +\infty, & \text{if } \mathbf{x} \notin A \end{cases} \tag{2}$$

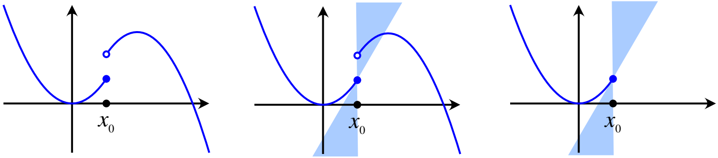
- If A is a closed set, χ_A is lower semi-continuous.
- If A is a closed convex set, χ_A is convex.

Equivalent optimization problems

$$\min_{\mathbf{x} \in \mathcal{C}} F(\mathbf{x}) \qquad \equiv \qquad \min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x}) + \chi_{\mathcal{C}}(\mathbf{x})$$

- Constrained OP can be reformulated as a non-smooth unconstrained OP.
- The new objective function is a sum of two functions (splitting algorithms).

Semicontinuity



Lower semi-continuous function

A function F is **lower semi-continuous (l.s.c.)** if for any point $\mathbf{x}_0 \in \mathcal{C}$ we have

$$F(\mathbf{x}_0) \leq \lim_{\mathbf{x} \to \mathbf{x}_0} \inf F(\mathbf{x}) \tag{3}$$

- Continuous functions are l.s.c. since it implies the equality above.
- If the function is l.s.c., there exists a local affine minorant.
- If the function is l.s.c. and convex it means that the sub-differential is never empty and the minorant is global : well defined problem.

Optimization problem in machine learning

Regularized supervised learning

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad f(\mathbf{x}) + g(\mathbf{x}) \tag{4}$$

- f is the data fitting term, g the regularization term.
- Usually f is smooth (K Lipschitz gradient).
- g can be non-smooth for instance Lasso regularization.
- This course will focus on the optimization of this type of non-smooth problem.

Data fiting examples

- Least square:

$$f(\mathbf{x}) = \sum_i (y_i - \mathbf{h}_i^T\mathbf{x})^2$$

- Logistic regression:

$$f(\mathbf{x}) = \sum_i \log(1+\exp(-y_i\mathbf{h}_i^T\mathbf{x}))$$

Regularization examples

- Ridge

$$g(\mathbf{x}) = \frac{\lambda}{2}\sum_k x_k^2$$

- Lasso

$$g(\mathbf{x}) = \lambda\sum_k |x_k|$$

Non-smooth ML problems

Linear SVM [Vapnik, 2013]

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \max(0, 1 - y_i \mathbf{w}^T \mathbf{h}_i) \tag{5}$$

Lasso regression [Tibshirani, 1996]

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1 \tag{6}$$

Multi-task learning (MTL)

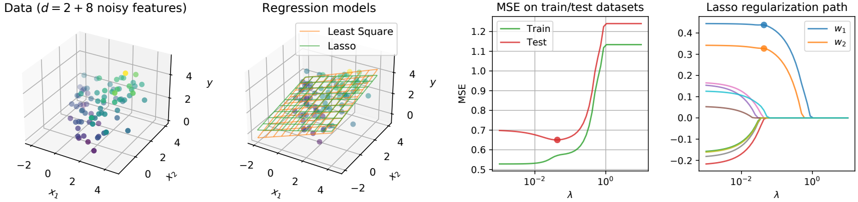
- **Low rank MTL [Argyriou et al., 2008]**:

$$\min_{\mathbf{W}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{W} - \mathbf{Y}\|^2 + \lambda\|\mathbf{W}\|_* \tag{7}$$

- **Group Lasso MTL [Argyriou et al., 2008, Obozinski et al., 2010]**:

$$\min_{\mathbf{W}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{W} - \mathbf{Y}\|^2 + \lambda\sum_{k=1}^{d}\|\mathbf{W}_{k,:}\|_2 \tag{8}$$

Lasso regression



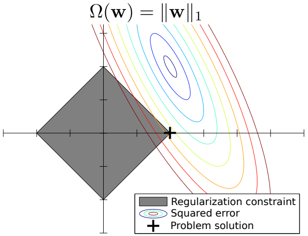
Principle [Tibshirani, 1996]

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$$

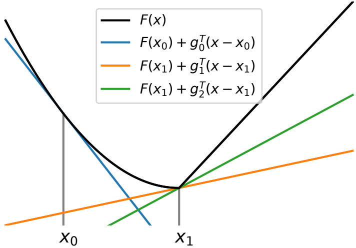
- For a large enough λ the solution of the problem is sparse.
- Under some conditions, support of true $\mathbf{w}$ can be recovered [Zhao and Yu, 2006].
- L1 regularization creates attraction points in 0 (see optimality condition).
- Lasso Problem is also equivalent to

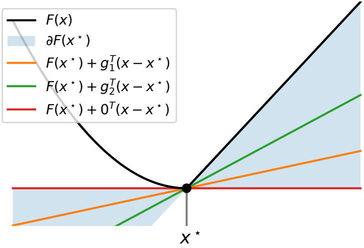
$$\min_{\mathbf{w},, \|\mathbf{w}\|_1 \leq \tau} \quad \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \qquad (9)$$

- The geometrical constraints promotes sparse $\mathbf{w}$ on the axis.



Subgradients and subdifferential





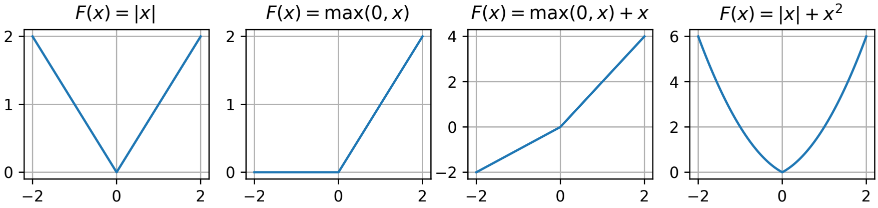
Non differentiable function

- For a convex function $F(\mathbf{x})$, $\mathbf{g}$ is a subgradient of F in $\mathbf{x}_0$ if

$$F(\mathbf{x}) \geq F(\mathbf{x}_0) + \mathbf{g}^\top(\mathbf{x} - \mathbf{x}_0) \tag{10}$$

- The set of all subgradients at $\mathbf{x}_0$ is the subdifferential $\partial f(\mathbf{x}_0)$.
- If F is differentiable in $\mathbf{x}_0$ there is a unique subgradient: $\partial f(\mathbf{x}_0) = \{\nabla_\mathbf{x} F(\mathbf{x})\}$
- **Optimality** : $\mathbf{x}^\star$ is a minimum of the convex function F if $\mathbf{0} \in \partial F(\mathbf{x}^\star)$.

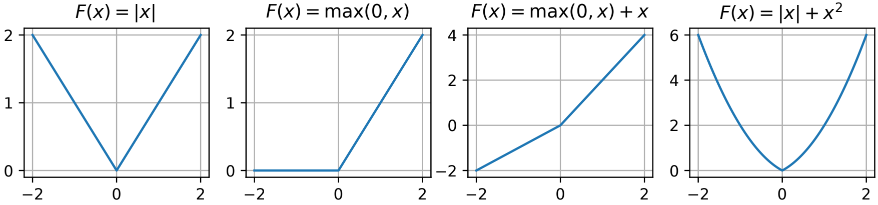
Exercise 1: Subgradients and subdifferential



Find the subdifferential $\partial F(\mathbf{x})$ for the following 1D functions:

1. $F(x) = |x|$, at $x \in \{-1, 0, 1\}$
2. $F(x) = \max(x, 0)$, at $x \in \{-1, 0, 1\}$
3. $F(x) = \max(x, 0) + x$, at $x \in \{-1, 0, 1\}$
4. $F(x) = |x| + x^2$, at $x \in \{-1, 0, 1\}$

Exercise 1: Subgradients and subdifferential



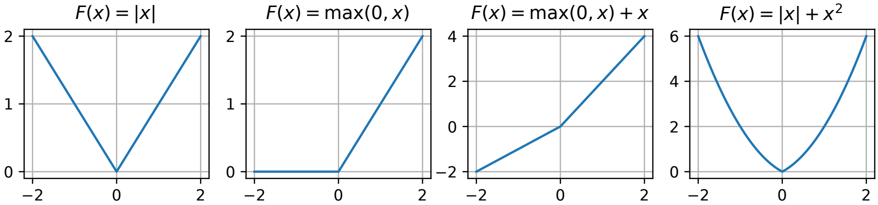
Find the subdifferential $\partial F(\mathbf{x})$ for the following 1D functions:

1. $F(x) = |x|$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{-1\}, \qquad \partial F(0) = \{g | -1 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

2. $F(x) = \max(x, 0)$, at $x \in \{-1, 0, 1\}$
3. $F(x) = \max(x, 0) + x$, at $x \in \{-1, 0, 1\}$
4. $F(x) = |x| + x^2$, at $x \in \{-1, 0, 1\}$

Exercise 1: Subgradients and subdifferential



Find the subdifferential $\partial F(\mathbf{x})$ for the following 1D functions:

1. $F(x) = |x|$, at $x \in \{-1, 0, 1\}$

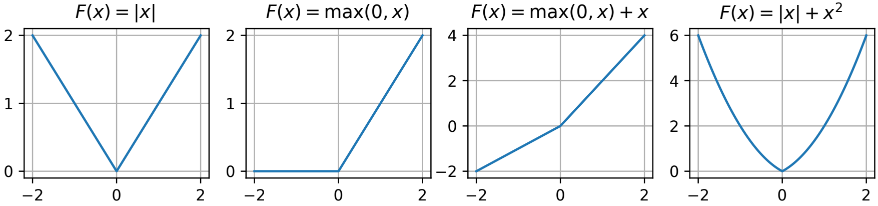
$$\partial F(-1) = \{-1\}, \qquad \partial F(0) = \{g | -1 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

2. $F(x) = \max(x, 0)$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{0\}, \qquad \partial F(0) = \{g | 0 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

3. $F(x) = \max(x, 0) + x$, at $x \in \{-1, 0, 1\}$
4. $F(x) = |x| + x^2$, at $x \in \{-1, 0, 1\}$

Exercise 1: Subgradients and subdifferential



Find the subdifferential $\partial F(\mathbf{x})$ for the following 1D functions:

1. $F(x) = |x|$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{-1\}, \qquad \partial F(0) = \{g | -1 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

2. $F(x) = \max(x, 0)$, at $x \in \{-1, 0, 1\}$

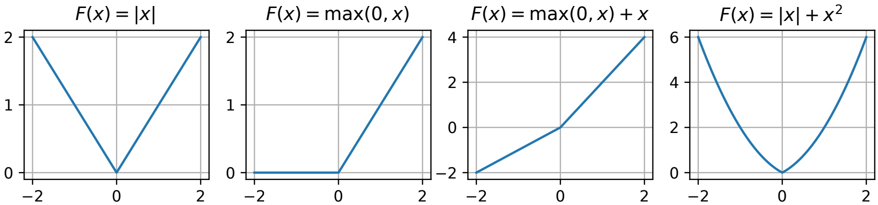
$$\partial F(-1) = \{0\}, \qquad \partial F(0) = \{g | 0 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

3. $F(x) = \max(x, 0) + x$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{0\}, \qquad \partial F(0) = \{g | 1 \leq g \leq 2\}, \qquad \partial F(1) = \{2\}$$

4. $F(x) = |x| + x^2$, at $x \in \{-1, 0, 1\}$

Exercise 1: Subgradients and subdifferential



Find the subdifferential $\partial F(\mathbf{x})$ for the following 1D functions:

1. $F(x) = |x|$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{-1\}, \qquad \partial F(0) = \{g| -1 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

2. $F(x) = \max(x, 0)$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{0\}, \qquad \partial F(0) = \{g|0 \leq g \leq 1\}, \qquad \partial F(1) = \{1\}$$

3. $F(x) = \max(x, 0) + x$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{0\}, \qquad \partial F(0) = \{g|1 \leq g \leq 2\}, \qquad \partial F(1) = \{2\}$$

4. $F(x) = |x| + x^2$, at $x \in \{-1, 0, 1\}$

$$\partial F(-1) = \{-3\}, \qquad \partial F(0) = \{g| -1 \leq g \leq 1\}, \qquad \partial F(1) = \{3\}$$

Optimal solution for the Lasso

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$$

Optimality for Least Square ($\lambda = 0$)

$$\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\mathbf{w}^\star_{LS}) = \mathbf{0}$$

Orthogonality between the columns of $\mathbf{X}$ and the residuals $\mathbf{y}-\mathbf{X}\mathbf{w}^\star_{LS}$.

Optimality for Lasso ($\lambda > 0$)

$$\mathbf{0} \in \mathbf{X}^\top(\mathbf{y}-\mathbf{X}\mathbf{w}^\star) + \lambda\partial\|\mathbf{w}^\star\|_1$$

Which is equivalent to

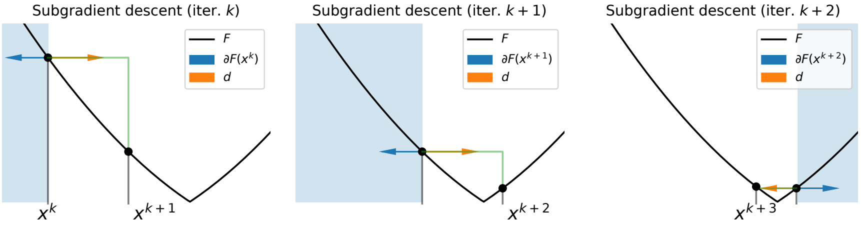
$$-\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\mathbf{w}^\star) \in \lambda\partial\|\mathbf{w}^\star\|_1$$

Using the subdifferential of the absolute value we can get $\forall i$

$$\mathbf{X}_{:,i}^\top(\mathbf{y}-\mathbf{X}\mathbf{w}^\star) \in \begin{cases}\{\lambda\} & \text{if } w_i^\star > 0\\ [-\lambda,\lambda] & \text{if } w_i^\star = 0\\ \{-\lambda\} & \text{if } w_i^\star < 0\end{cases} \quad = \quad \begin{cases}\{\lambda\text{sign}(w_i^\star)\} & \text{if } w_i^\star \neq 0\\ [-\lambda,\lambda] & \text{if } w_i^\star = 0\end{cases}$$

What happens when $\max_i |\mathbf{X}_{:,i}^\top\mathbf{y}| < \lambda$?

Subgradient methods

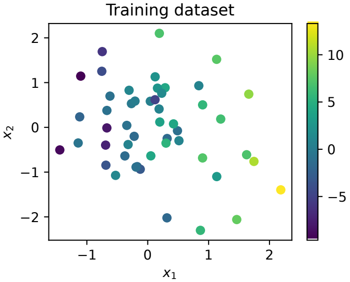


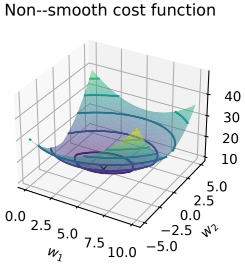
Subgradient descent

Initialize $\mathbf{x}^{(0)}$
for $k = 0, 1, 2, \dots$ **do**
  $\mathbf{g}^{(k)} \in \partial F(\mathbf{x}^{(k)})$
  $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} - \rho^{(k)}\mathbf{g}^{(k)}$
end for

- No convergence guarantee to a minimum with fixed step size $\rho^{(k)} = \rho$.
- For fixed step on L Lipschitz F reaches an $\epsilon = \frac{L^2\rho}{2}$ approx. solution.
- Convergence for a Lischitz function is $O(\frac{1}{\sqrt{k}})$ with decreasing step $\rho^{(k)} = \frac{1}{\sqrt{n}}$.
- Subgradient descent is slower than gradient descent.

Example dataset for the Lasso



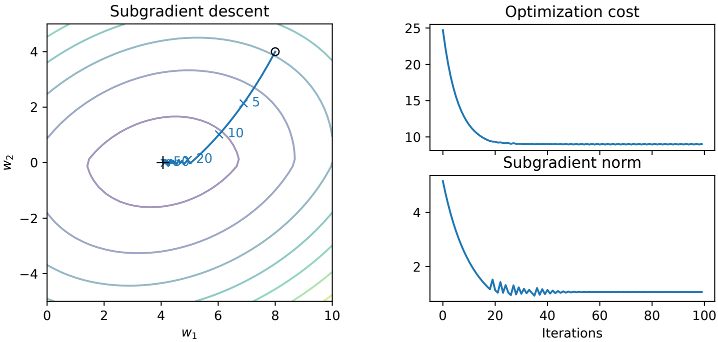


2D Lasso optimization problem

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$$

- $\mathbf{X}$ is a $n \times 2$ matrix, $\mathbf{y}$ is a n vector with $n = 50$
- True model is $\mathbf{w}^\star = [5, 0]$ and additive noise is added to the data.
- Least square solution is not sparse $\mathbf{w}_{LS} = [5.32, 0.30]$.
- λ selected to have a sparse solution (only the relevant variable) with solution $\mathbf{w}_{Lasso} = [4.064, 0]$.

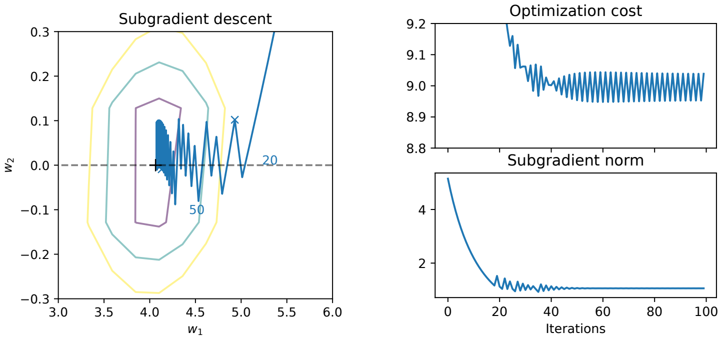
Exemple of Subgradient Descent for the Lasso



Discussion

- Subgradient descent fixed step $\rho^{(k)} = \rho$ does not converge.
- Oscillation around optimal value 0 for w_2.
- Convergence with decreasing step size $\rho^{(k)} = \frac{1}{\sqrt{k}}$.
- But slow convergence in $O(\frac{1}{\sqrt{k}})$.

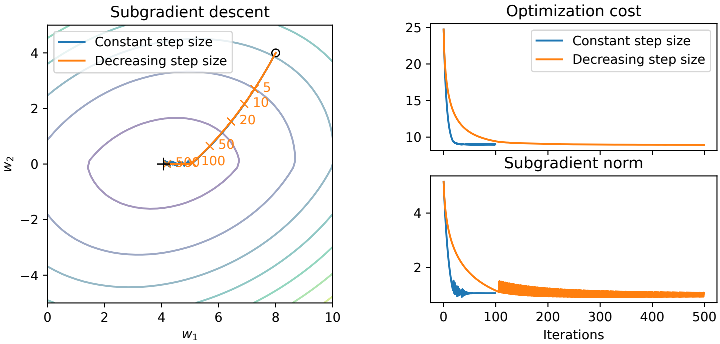
Exemple of Subgradient Descent for the Lasso



Discussion

- Subgradient descent fixed step $\rho^{(k)} = \rho$ does not converge.
- Oscillation around optimal value 0 for w_2.
- Convergence with decreasing step size $\rho^{(k)} = \frac{1}{\sqrt{k}}$.
- But slow convergence in $O(\frac{1}{\sqrt{k}})$.

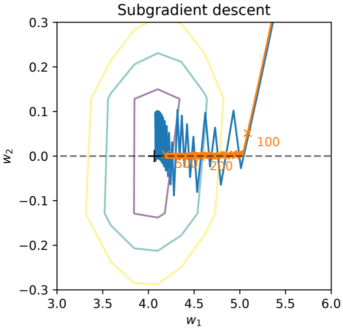
Exemple of Subgradient Descent for the Lasso

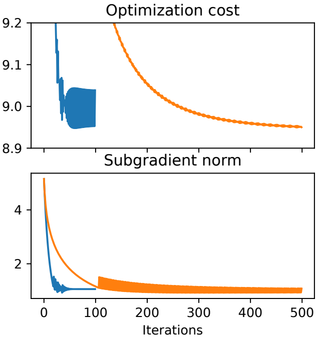


Discussion

- Subgradient descent fixed step $\rho^{(k)} = \rho$ does not converge.
- Oscillation around optimal value 0 for w_2.
- Convergence with decreasing step size $\rho^{(k)} = \frac{1}{\sqrt{k}}$.
- But slow convergence in $O(\frac{1}{\sqrt{k}})$.

Exemple of Subgradient Descent for the Lasso





Discussion

- Subgradient descent fixed step $\rho^{(k)} = \rho$ does not converge.
- Oscillation around optimal value 0 for w_2.
- Convergence with decreasing step size $\rho^{(k)} = \frac{1}{\sqrt{k}}$.
- But slow convergence in $O(\frac{1}{\sqrt{k}})$.

Majorization Minimization of non-smooth functions

Assumptions (separable F)

$$F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$$

- f is L-smooth and convex.
- g is convex and lower semi-continuous (can be smooth but not necessary).

Majorization Minimization of the smooth part

- Since f is L gradient Lipschitz F can be upper bounded around $\mathbf{x}^{(0)}$ by:

$$F(\mathbf{x}) \leq f(\mathbf{x}^{(0)}) + \boldsymbol{\nabla} f(\mathbf{x}^{(0)})^\top (\mathbf{x} - \mathbf{x}^{(0)}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^{(0)}\|^2 + g(\mathbf{x}), \tag{11}$$

- Minimizing the upper bound above is equivalent to minimize:

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{L} g(\mathbf{x}) \tag{12}$$

with

$$\mathbf{y} =$$

Majorization Minimization of non-smooth functions

Assumptions (separable F)

$$F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$$

- f is L-smooth and convex.
- g is convex and lower semi-continuous (can be smooth but not necessary).

Majorization Minimization of the smooth part

- Since f is L gradient Lipschitz F can be upper bounded around $\mathbf{x}^{(0)}$ by:

$$F(\mathbf{x}) \leq f(\mathbf{x}^{(0)}) + \boldsymbol{\nabla} f(\mathbf{x}^{(0)})^\top (\mathbf{x} - \mathbf{x}^{(0)}) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^{(0)}\|^2 + g(\mathbf{x}), \tag{11}$$

- Minimizing the upper bound above is equivalent to minimize:

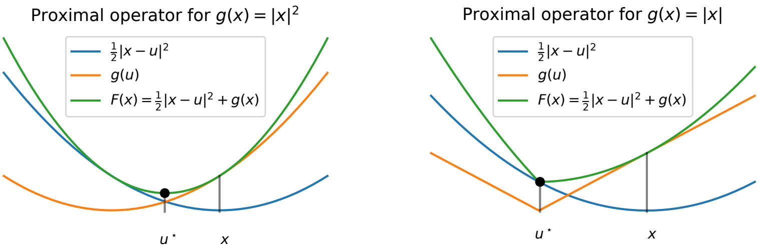
$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{L} g(\mathbf{x}) \tag{12}$$

  with

$$\mathbf{y} = \mathbf{x}^{(0)} - \frac{1}{L}\nabla f(\mathbf{x}^{(0)})$$

- The solution of (12) is the proximal operator of g.
- Minimizing the upper bound iteratively corresponds to the Forward Backward Splitting or Proximal Gradient Descent algorithm.

Proximal operator



Definition [Bauschke et al., 2011]

The Proximity (or proximal) operator of a function g is:

$$\mathbf{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{u}\in\mathbb{R}^n} \; g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2.$$

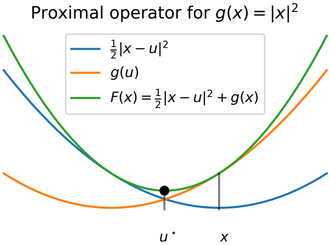
- Returns a vector minimizing g but close to $\mathbf{x}$ in the quadratic sense.
- **Fixed point:** $\mathbf{prox}_g(\mathbf{x}) = \mathbf{x}$ if $\mathbf{x}$ if an only if $\mathbf{0} \in \partial g(\mathbf{x})$ (i.e. $\mathbf{x}$ is minimizer).
- **Non expansiveness**: $\|\mathbf{prox}_g(\mathbf{x}) - \mathbf{prox}_g(\mathbf{y})\| \leq \|\mathbf{x}-\mathbf{y}\|$.

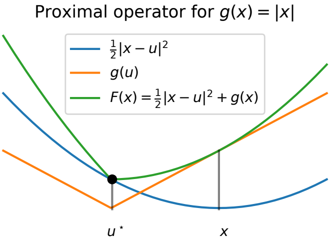
Exercise 2: Proximal operator for L2 norm

Compute the proximal operator for $g(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|^2$ with $\lambda \geq 0$

Solution :

Proximal operator





Definition [Bauschke et al., 2011]

The Proximity (or proximal) operator of a function g is:

$$\mathbf{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{u}\in\mathbb{R}^n} \; g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2.$$

- Returns a vector minimizing g but close to $\mathbf{x}$ in the quadratic sense.
- **Fixed point:** $\mathbf{prox}_g(\mathbf{x}) = \mathbf{x}$ if $\mathbf{x}$ if an only if $\mathbf{0} \in \partial g(\mathbf{x})$ (i.e. $\mathbf{x}$ is minimizer).
- **Non expansiveness**: $\|\mathbf{prox}_g(\mathbf{x}) - \mathbf{prox}_g(\mathbf{y})\| \leq \|\mathbf{x}-\mathbf{y}\|$.

Exercise 2: Proximal operator for L2 norm

Compute the proximal operator for $g(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|^2$ with $\lambda \geq 0$

Solution : $\mathbf{prox}_{\frac{\lambda}{2}\|\cdot\|^2}(\mathbf{x}) = \frac{1}{1+\lambda}\mathbf{x}$

Properties of proximal operator

Exercise 3: Separable function g

If $g(\mathbf{x}) = \sum_k g_k(x_k)$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Exercise 4: Characteristic function of set A

If $g(\mathbf{x}) = \chi_A(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \\ +\infty, & \text{if } \mathbf{x} \notin A \end{cases}$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Exercise 5: Linear function

If $g(\mathbf{x}) = \mathbf{b}^\top \mathbf{x} + c$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Exercise 6: Quadratic function

If $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Properties of proximal operator

Exercise 3: Separable function g

If $g(\mathbf{x}) = \sum_k g_k(x_k)$ then

$$\mathbf{prox}_g(\mathbf{x}) = [\mathbf{prox}_{g_1}(x_1), \ldots, \mathbf{prox}_{g_d}(x_d)]^\top$$

Exercise 4: Characteristic function of set A

If $g(\mathbf{x}) = \chi_A(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \\ +\infty, & \text{if } \mathbf{x} \notin A \end{cases}$ then

$$\mathbf{prox}_g(\mathbf{x}) = \text{proj}_A(\mathbf{x}) \quad \text{(projection operator)}$$

Exercise 5: Linear function

If $g(\mathbf{x}) = \mathbf{b}^\top \mathbf{x} + c$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Exercise 6: Quadratic function

If $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Properties of proximal operator

Exercise 3: Separable function g

If $g(\mathbf{x}) = \sum_k g_k(x_k)$ then

$$\mathbf{prox}_g(\mathbf{x}) = [\mathbf{prox}_{g_1}(x_1), \ldots, \mathbf{prox}_{g_d}(x_d)]^\top$$

Exercise 4: Characteristic function of set A

If $g(\mathbf{x}) = \chi_A(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \\ +\infty, & \text{if } \mathbf{x} \notin A \end{cases}$ then

$$\mathbf{prox}_g(\mathbf{x}) = \text{proj}_A(\mathbf{x}) \quad \text{(projection operator)}$$

Exercise 5: Linear function

If $g(\mathbf{x}) = \mathbf{b}^\top \mathbf{x} + c$ then

$$\mathbf{prox}_g(\mathbf{x}) = \mathbf{x} - \mathbf{b}$$

Exercise 6: Quadratic function

If $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$ then

$$\mathbf{prox}_g(\mathbf{x}) =$$

Properties of proximal operator

Exercise 3: Separable function g

If $g(\mathbf{x}) = \sum_k g_k(x_k)$ then

$$\mathbf{prox}_g(\mathbf{x}) = [\mathbf{prox}_{g_1}(x_1), \ldots, \mathbf{prox}_{g_d}(x_d)]^\top$$

Exercise 4: Characteristic function of set A

If $g(\mathbf{x}) = \chi_A(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in A \\ +\infty, & \text{if } \mathbf{x} \notin A \end{cases}$ then

$$\mathbf{prox}_g(\mathbf{x}) = \mathsf{proj}_A(\mathbf{x}) \quad \text{(projection operator)}$$

Exercise 5: Linear function

If $g(\mathbf{x}) = \mathbf{b}^\top \mathbf{x} + c$ then

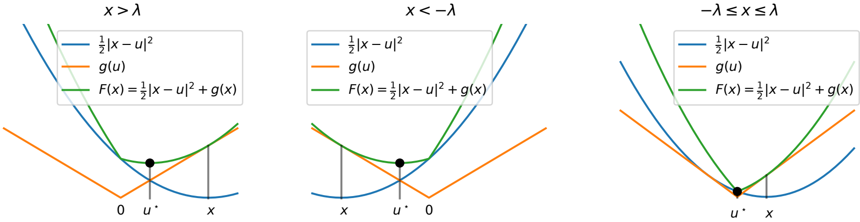
$$\mathbf{prox}_g(\mathbf{x}) = \mathbf{x} - \mathbf{b}$$

Exercise 6: Quadratic function

If $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}$ then

$$\mathbf{prox}_g(\mathbf{x}) = (I + \mathbf{A})^{-1}(\mathbf{x} - \mathbf{b})$$

Proximal operator for L1 norm: Soft Thresholding



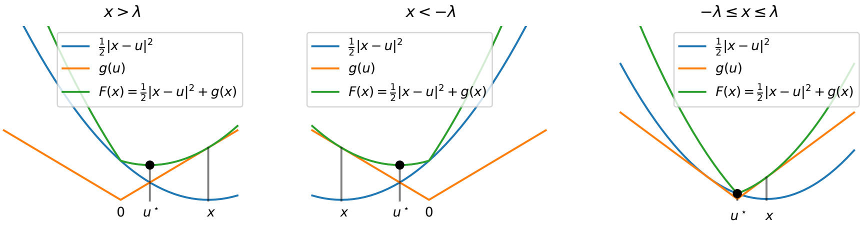
$$g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 = \lambda \sum_k |x_k|$$

Exercise 7: Soft Thresholding operator

L1 norm is separable so we can compute the proximal operator for each component:

1. Optimality condition for proximal operator: $\min_u \frac{1}{2}(u-x)^2 + \lambda|u|$
2. If $x > \lambda$ then
3. If $x < -\lambda$ then
4. If $-\lambda \le x \le \lambda$ then

Proximal operator for L1 norm: Soft Thresholding



$$g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 = \lambda \sum_k |x_k|$$

Exercise 7: Soft Thresholding operator

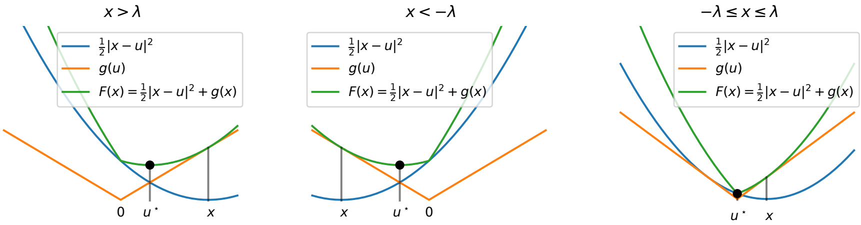
L1 norm is separable so we can compute the proximal operator for each component:

1. Optimality condition for proximal operator: $\min_u \frac{1}{2}(u-x)^2 + \lambda|u|$

$$u^\star \in x - \lambda\partial|u^\star|$$

2. If $x > \lambda$ then
3. If $x < -\lambda$ then
4. If $-\lambda \leq x \leq \lambda$ then

Proximal operator for L1 norm: Soft Thresholding



$$g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1 = \lambda \sum_k |x_k|$$

Exercise 7: Soft Thresholding operator

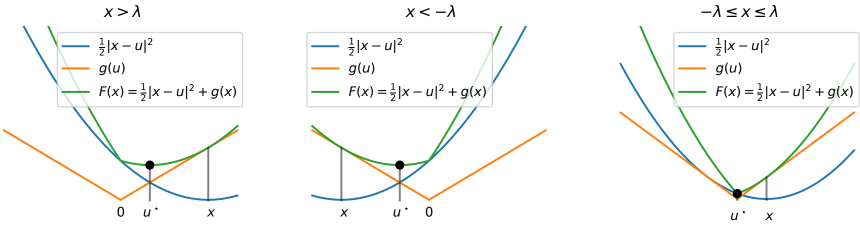
L1 norm is separable so we can compute the proximal operator for each component:

1. Optimality condition for proximal operator: $\min_u \frac{1}{2}(u-x)^2 + \lambda|u|$

$$u^\star \in x - \lambda \partial |u^\star|$$

2. If $x > \lambda$ then $u^\star = x - \lambda$ ($u \leq 0$ not possible)
3. If $x < -\lambda$ then
4. If $-\lambda \leq x \leq \lambda$ then

Proximal operator for L1 norm: Soft Thresholding



$$g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1 = \lambda \sum_k |x_k|$$

Exercise 7: Soft Thresholding operator

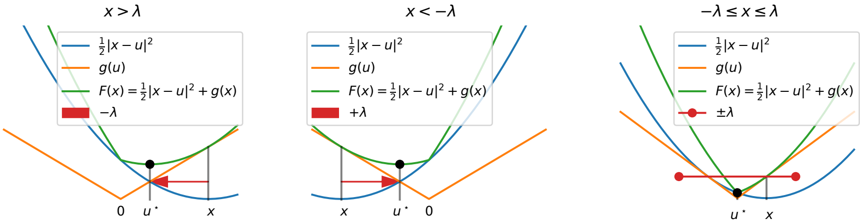
L1 norm is separable so we can compute the proximal operator for each component:

1. Optimality condition for proximal operator: $\min_u \frac{1}{2}(u - x)^2 + \lambda|u|$

$$u^\star \in x - \lambda \partial |u^\star|$$

2. If $x > \lambda$ then $u^\star = x - \lambda$ ($u \leq 0$ not possible)
3. If $x < -\lambda$ then $u^\star = x + \lambda$($u \geq 0$ not possible)
4. If $-\lambda \leq x \leq \lambda$ then

Proximal operator for L1 norm: Soft Thresholding



$$g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1 = \lambda \sum_k |x_k|$$

Exercise 7: Soft Thresholding operator

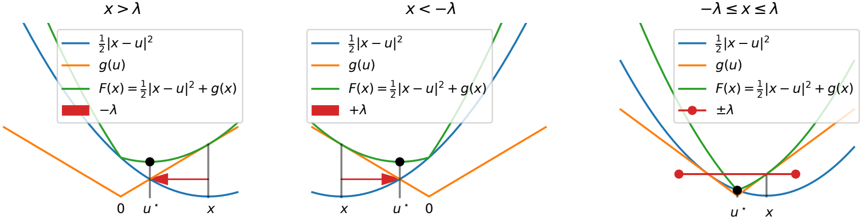
L1 norm is separable so we can compute the proximal operator for each component:

1. Optimality condition for proximal operator: $\min_u \frac{1}{2}(u - x)^2 + \lambda |u|$

$$u^\star \in x - \lambda \partial |u^\star|$$

2. If $x > \lambda$ then $u^\star = x - \lambda$ ($u \leq 0$ not possible)
3. If $x < -\lambda$ then $u^\star = x + \lambda$($u \geq 0$ not possible)
4. If $-\lambda \leq x \leq \lambda$ then $-\lambda \leq x - u^\star \leq \lambda$ only for $u^\star = 0$.

Proximal operator for L1 norm: Soft Thresholding



$$g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 = \lambda \sum_k |x_k|$$

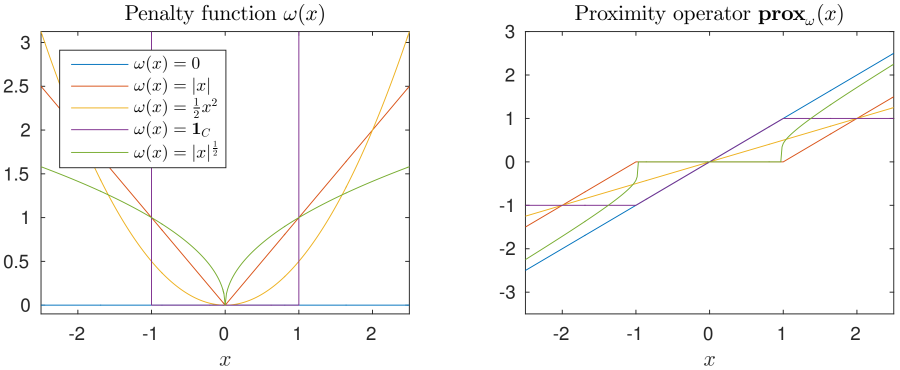
Exercise 7: Soft Thresholding operator

The proximal operator for $\lambda\|\cdot\|_1$ is the soft thresholding operator:

$$\mathbf{prox}_{\lambda\|\cdot\|_1}(\mathbf{x}) = \begin{cases} x - \lambda & \text{if } x > \lambda \\ 0 & \text{if } |x| \leq \lambda \\ x + \lambda & \text{if } x < -\lambda \end{cases} \quad = \quad \text{sign}(\mathbf{x}) \max(0, |\mathbf{x}| - \lambda)$$

The soft thresholding operator shrinks the values of $\mathbf{x}$ towards 0 and promotes sparsity.

Examples of separable proximal operators

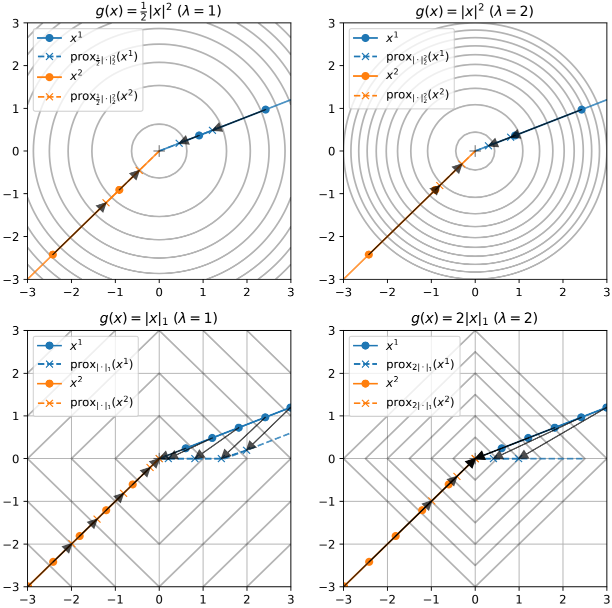


Common proximal operators

$g(\mathbf{x}) = 0$	$\mathbf{prox}_g(\mathbf{x}) = \mathbf{x}$	identity
$g(\mathbf{x}) = \lambda\|\mathbf{x}\|_2^2$	$\mathbf{prox}_g(\mathbf{x}) = \frac{1}{1+\lambda}\mathbf{x}$	scaling
$g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$	$\mathbf{prox}_g(\mathbf{x}) = \mathrm{sign}(\mathbf{x})\max(0, \lvert\mathbf{x}\rvert - \lambda)$	soft shrinkage
$g(\mathbf{x}) = \lambda\|\mathbf{x}\|_{1/2}^{1/2}$	[Xu et al., 2012, Equation 11]	power family
$g(\mathbf{x}) = \chi_C(\mathbf{x})$	$\mathbf{prox}_g(\mathbf{x}) = \operatorname{argmin}_{\mathbf{u}\in C} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2$	orthogonal projection.

- Both $|x|$ and $|x|^{\frac{1}{2}}$ promote sparsity (soft thresholds).

Proximal operator in 2D



Proximal Gradient Descent (PGD)

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$$

PGD algorithm [Combettes and Pesquet, 2011][Parikh and Boyd, 2014].

1: Initialize $\mathbf{x}^{(0)}$
2: **for** $k = 0, 1, 2, \dots$ **do**
3: $\quad \mathbf{d}^{(k)} \leftarrow -\nabla f(\mathbf{x}^{(k)})$
4: $\quad \mathbf{x}^{(k+1)} \leftarrow \mathbf{prox}_{\rho^{(k)} g}(\mathbf{x}^{(k)} + \rho^{(k)}\mathbf{d}^{(k)})$
5: **end for**

- One gradient step *w.r.t.* f and one proximal step *w.r.t.* g.
- Also known as Forward Backward Splitting (FBS) [Combettes and Pesquet, 2011]
- Efficient when the proximal operator is simple to compute (closed form).
- When g is a characteristic function, FBS/PGD is the projected Gradient Descent.
- **Optimal solution is a fixed point**: $\mathbf{x}^\star$ min of F implies that for $\rho \leq \frac{2}{L}$

$$-\nabla f(\mathbf{x}^\star) \in \partial g(\mathbf{x}^\star) \quad \Leftrightarrow \quad \mathbf{x}^\star = \mathbf{prox}_{\rho g}(\mathbf{x}^\star - \rho \nabla f(\mathbf{x}^\star)) \tag{13}$$

Convergence of PGD

Convergenge for L-smooth f [Beck and Teboulle, 2009]

For and L-smooth function f and a convex g the PGD with step size $\rho \le \frac{1}{L}$ converges to a minimum of F with the following speed:

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \le \frac{L}{2k}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2$$

Convergence for L-smooth and μ-convex f

For and L-smooth and μ-convex function f and a convex g the PGD with step size $\rho \le \frac{1}{L}$ converges to a minimum of F with the following speed:

$$\|\mathbf{x}^{(k)} - \mathbf{x}^\star\| \le \left(1 - \frac{\mu}{L}\right)^k \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2$$

Sketch of proof

$$\begin{aligned}
\|\mathbf{x}^{(k)} - \mathbf{x}^\star\| &= \|\mathbf{prox}_{\rho g}(\mathbf{x}^{(k)} - \rho\nabla f(\mathbf{x}^{(k)})) - \mathbf{x}^\star\| \\
&\underset{1}{=} \|\mathbf{prox}_{\rho g}(\mathbf{x}^{(k)} - \rho\nabla f(\mathbf{x}^{(k)})) - \mathbf{prox}_{\rho g}(\mathbf{x}^\star - \rho\nabla f(\mathbf{x}^\star))\| \\
&\underset{2}{\le} \|\mathbf{x}^{(k)} - \rho\nabla f(\mathbf{x}^{(k)}) - \mathbf{x}^\star - \rho\nabla f(\mathbf{x}^\star)\|
\end{aligned}$$

Next steps are similar to proof of Gradient descent convergence.

[1] Use fixed point property (13)

[2] Use non-expansiveness of proximal operator

Exercise 8: Solving the Lasso with PGD

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{H}\mathbf{x}-\mathbf{y}\|^2 + \lambda\sum_k |x_k|$$

Known as Iterative Soft Thresholding Algorithm (ISTA) [Beck and Teboulle, 2009].

1. Express the smooth function f and non-smooth functions g for the problem above

$$f(\mathbf{x}) = \qquad\qquad g(\mathbf{x}) =$$

2. Compute the gradient $\nabla f(\mathbf{x})$ and express the proximal of g.

$$\nabla f(\mathbf{x}) = \qquad\qquad \mathbf{prox}_g(\mathbf{x}) =$$

3. Express the FBS algorithm in Python/Numpy for solving the lasso with a fixed step rho :

```
def lasso(H,y,reg,rho,nbiter):
```

Exercise 8: Solving the Lasso with PGD

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{H}\mathbf{x}-\mathbf{y}\|^2 + \lambda\sum_k |x_k|$$

Known as Iterative Soft Thresholding Algorithm (ISTA) [Beck and Teboulle, 2009].

1. Express the smooth function f and non-smooth functions g for the problem above

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{H}\mathbf{x}-\mathbf{y}\|^2 \qquad g(\mathbf{x}) = \lambda\sum_k |x_k|$$

2. Compute the gradient $\nabla f(\mathbf{x})$ and express the proximal of g.

$$\nabla f(\mathbf{x}) = \qquad \mathbf{prox}_g(\mathbf{x}) =$$

3. Express the FBS algorithm in Python/Numpy for solving the lasso with a fixed step rho :

```
def lasso(H,y,reg,rho,nbiter):
```

Exercise 8: Solving the Lasso with PGD

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad \frac{1}{2}\|\mathbf{H}\mathbf{x}-\mathbf{y}\|^2 + \lambda \sum_k |x_k|$$

Known as Iterative Soft Thresholding Algorithm (ISTA) [Beck and Teboulle, 2009].

1. Express the smooth function f and non-smooth functions g for the problem above

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{H}\mathbf{x}-\mathbf{y}\|^2 \qquad g(\mathbf{x}) = \lambda \sum_k |x_k|$$

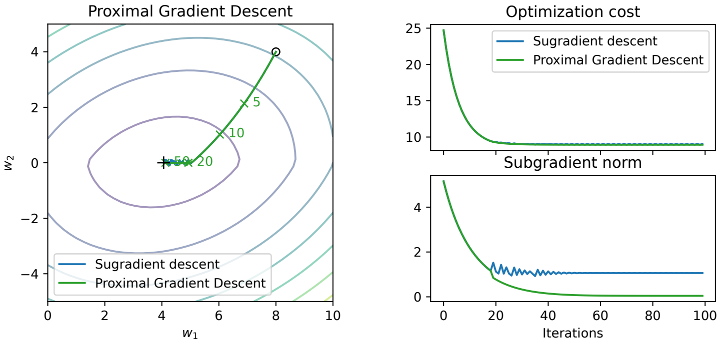
2. Compute the gradient $\nabla f(\mathbf{x})$ and express the proximal of g.

$$\nabla f(\mathbf{x}) = \mathbf{H}^T(\mathbf{H}\mathbf{x}-\mathbf{y}) \qquad \mathbf{prox}_g(\mathbf{x}) = \text{sign}(\mathbf{x})\max(0, |\mathbf{x}|-\lambda)$$

3. Express the FBS algorithm in Python/Numpy for solving the lasso with a fixed step rho :

```
def lasso(H,y,reg,rho,nbiter):
```

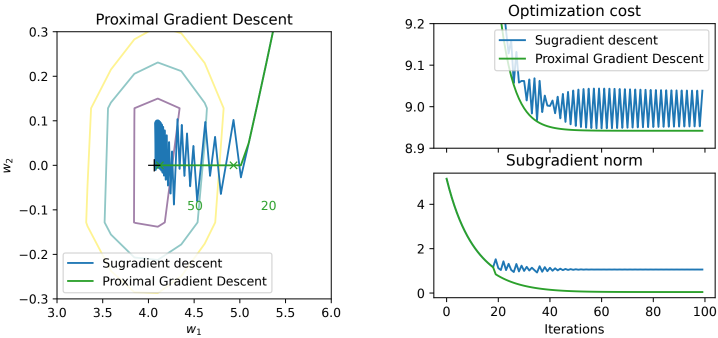
Example: PGD/ISTA for solving the Lasso



Discussion

- PGD with fixed step $\rho^{(k)} = \rho$ is more stable than subgradient descent.
- No oscillation and only monotonous decrease.
- One variable is exactly 0 after 20 iterations.
- 2 regimes: support selection and then optimization of the subset of non-zeros components (that can be strongly convex on the subset).

Example: PGD/ISTA for solving the Lasso



Discussion

- PGD with fixed step $\rho^{(k)} = \rho$ is more stable than subgradient descent.
- No oscillation and only monotonous decrease.
- One variable is exactly 0 after 20 iterations.
- 2 regimes: support selection and then optimization of the subset of non-zeros components (that can be strongly convex on the subset).

Accelerated Proximal Gradient Descent (APGD)

PGD with Nesterov acceleration [Beck and Teboulle, 2009]

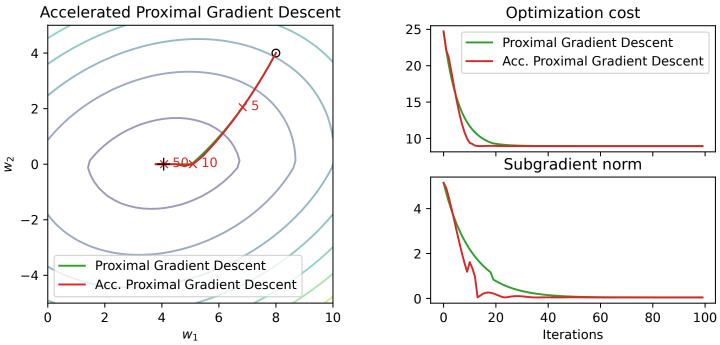
Initialize $\mathbf{y}^{(1)} = \mathbf{x}^{(0)}, t^{(1)} = 1$
for $k = 1, 2, \dots$ **do**
    $\mathbf{x}^{(k)} \leftarrow \mathbf{prox}_{\rho^{(k)} g}(\mathbf{y}^{(k)} - \rho^{(k)} \nabla f(\mathbf{y}^{(k)}))$
    $t^{(k+1)} \leftarrow \frac{1+\sqrt{1+4(t^{(k)})^2}}{2}$
    $\mathbf{y}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \frac{t^{(k)}-1}{t^{(k+1)}}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})$
end for

- Use a similar momentum to accelerated gradient.
- The function might not decrease at each iteration due to the momentum.
- Convergence for and L-smooth function f is :

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{2L\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{(k+1)^2}$$

- Also known as Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) when applied to the Lasso [Beck and Teboulle, 2009].

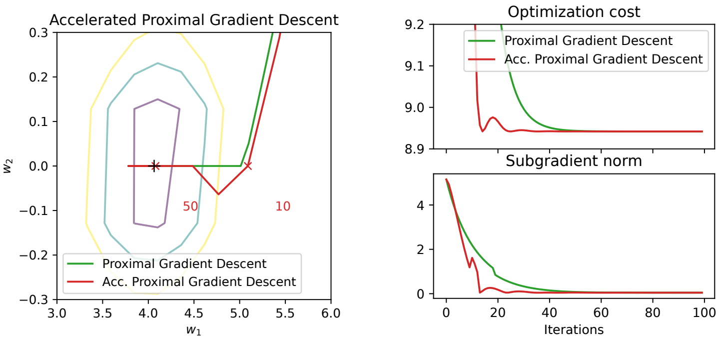
Example: Accelerated PGD/FISTA for the Lasso



Discussion

- Accelerated PGD with fixed step $\rho^{(k)} = \rho$ is faster than PGD.
- Inertia causes overshooting and oscillations but the algorithm converges faster.
- One variable is exactly 0 after 20 iterations.
- 2 regimes: support selection and then optimization of non-zeros components.

Example: Accelerated PGD/FISTA for the Lasso



Discussion

- Accelerated PGD with fixed step $\rho^{(k)} = \rho$ is faster than PGD.
- Inertia causes overshooting and oscillations but the algorithm converges faster.
- One variable is exactly 0 after 20 iterations.
- 2 regimes: support selection and then optimization of non-zeros components.

Chambolle-Pock Algorithm

Assumptions

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad F(\mathbf{x}) = f(\mathbf{A}\mathbf{x}) + g(\mathbf{x})$$

- Both f and g are convex (no smoothness necessary).
- $\mathbf{A}$ is a linear operator (not needed to be square or invertible).

Chambolle-Pock Algorithm [Chambolle and Pock, 2011]

1: Initialize $\mathbf{x}^{(0)} = \bar{\mathbf{x}}^{(0)}, \mathbf{y}^{(0)}, \rho_1, \rho_2 > 0, 0 \leq \theta \leq 1$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad \mathbf{y}^{(k+1)} \leftarrow \mathbf{prox}_{\rho_1 f}(\mathbf{y}^{(k)} + \rho_1 \mathbf{A}\bar{\mathbf{x}}^{(k)})$
4: $\quad \mathbf{x}^{(k+1)} \leftarrow \mathbf{prox}_{\rho_2 g}(\mathbf{x}^{(k)} - \rho_2 \mathbf{A}^\top \mathbf{y}^{(k+1)})$
5: $\quad \bar{\mathbf{x}}^{(k+1)} \leftarrow \mathbf{x}^{(k+1)} + \theta(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})$
6: **end for**

- Generalization of the Douglas-Rachford splitting (with a linear operator $\mathbf{A}$).
- θ allows to use a momentum when > 0.
- Interesting when the prox of f and g are efficient.

Vu-Condat Algorithm

Assumptions

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$$

- f convex and L-smooth, $\mathbf{A}$ is a linear operator.
- g and h are convex and have "simple" proximal operators.

Vu-Conda Algorithm [Vũ, 2013, Condat, 2014]

1: Initialize $\mathbf{x}^{(0)} = \bar{\mathbf{x}}^{(0)}, \mathbf{y}^{(0)} = \bar{\mathbf{y}}^{(0)}, \rho_1, \rho_2 > 0, 0 \leq \theta \leq 1$
2: **for** $k = 1, 2, \dots$ **do**
3: $\quad \mathbf{x}^{(k+1)} \leftarrow \mathbf{prox}_{\rho_2 g}(\bar{\mathbf{x}}^{(k)} - \rho_2 \nabla f(\bar{\mathbf{x}}^{(k)}) - \rho_2 \mathbf{A}^\top \bar{\mathbf{y}}^{(k)})$
4: $\quad \bar{\mathbf{x}}^{(k+1)} \leftarrow \bar{\mathbf{x}}^{(k+1)} + \theta(\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k)})$
5: $\quad \mathbf{y}^{(k+1)} \leftarrow \mathbf{prox}_{\rho_1 h^*}(\bar{\mathbf{y}}^{(k)} + \rho_1 \mathbf{A}(2\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k)}))$
6: $\quad \bar{\mathbf{y}}^{(k+1)} \leftarrow \bar{\mathbf{y}}^{(k+1)} + \theta(\mathbf{y}^{(k+1)} - \bar{\mathbf{y}}^{(k)})$
7: **end for**

- $\mathbf{prox}_{\rho h^*}(\mathbf{x}) = \mathbf{x} - \rho \mathbf{prox}_{h/\rho}(\mathbf{x}/\rho)$ is the proximal operator of the Fenchel–Rockafellar conjugate of h also called convex conjugate.
- General formulation in parallel with $h(\mathbf{A}\mathbf{x}) = \sum_i h_i(\mathbf{A}_i \mathbf{x})$ in [Condat, 2014].

Alternating Direction Method of Multipliers (ADMM)

Optimization problem and augmented Lagrangian

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{z}\in\mathbb{R}^m} \quad f(\mathbf{x}) + g(\mathbf{z}) \qquad \text{s.t.} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}$$

The augmented Lagrangian of the problem is expressed as:

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2 \tag{14}$$

ADMM Algorithm [Boyd et al., 2011]

1: Initialize $\mathbf{x}^{(0)}, \mathbf{z}^{(0)}, \mathbf{y}^{(0)}, \rho > 0$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad \mathbf{x}^{(k+1)} \leftarrow \arg\min_{\mathbf{x}} \; L_\rho(\mathbf{x}, \mathbf{z}^{(k)}, \mathbf{y}^{(k)})$
4: $\quad \mathbf{z}^{(k+1)} \leftarrow \arg\min_{\mathbf{z}} \; L_\rho(\mathbf{x}^{(k+1)}, \mathbf{z}, \mathbf{y}^{(k)})$
5: $\quad \mathbf{y}^{(k+1)} \leftarrow \mathbf{y}^{(k)} + \rho(\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{B}\mathbf{z}^{(k+1)} - \mathbf{c})$
6: **end for**

- Updates 3 and 4 can often be expressed as proximal updates.
- When f or g is separable, the updates can be done in parallel.

Example: 2D Total Variation denoising

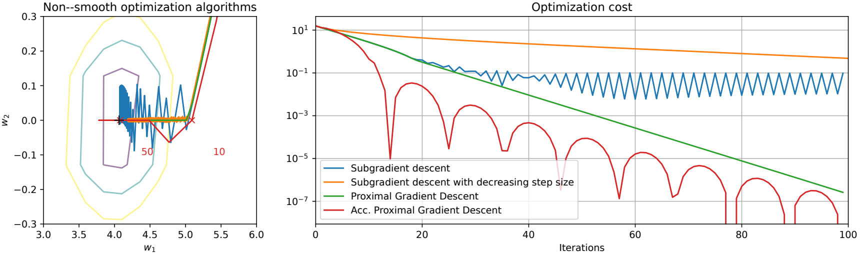
$$\min_{\mathbf{X}\in\mathbb{R}_+^{d\times d}} \quad \|\mathbf{Y}-\mathbf{X}\|_F^2 + \lambda\left(\sum_{i=1,j=1}^{d,d-1} |X_{i,j}-X_{i,j+1}| + \sum_{i=1,j=1}^{d-1,d} |X_{i,j}-X_{i+1,j}|\right)$$

- Image $\mathbf{Y}$ is noisy but a clean $\mathbf{X}$ that has piecewise constant parts.
- The regularization term measure the total variation (L1 norm of the gradients) of the image horizontally and vertically.

Exercise 9 (optional): Solve the problem

- For each algorithm: ADMM, Chambolle-Pock and Vu-Conda.
- Reformulate the problem with and without positivity constraints (recover f, g, h).
- Which algorithms can be used if the first term is $\|\mathbf{Y}-\mathbf{H}*\mathbf{X}\|_F^2$ (deconvolution)?

Conclusion



Proximal methods [Parikh and Boyd, 2014]

- General strategy of proximal splitting: divide and conquer the objective function.
- Search for a stationary point, avoid subgradients.
- PGD/APGD for simple problems, ADMM or other for more complex splitting.
- For sparse optimization, intermediate iterates are sparse and better conditioned.
- Works also for non-convex problems [Attouch et al., 2010].
- For deep learning non-convex problems subgradient descent is often used [Goodfellow, 2016].

Bibliography I

Convex Optimization [Boyd and Vandenberghe, 2004]

- Available freely online: `https://web.stanford.edu/~boyd/cvxbook/`.

Nonlinear Programming [Bertsekas, 1997]

- Reference optimization book, contains also most of the course.
- Unconstrained optimization (Ch. 1), duality and lagrangian (Ch. 3, 4 ,5).

Convex analysis and monotone operator theory in Hilbert spaces [Bauschke et al., 2011]

- Awesome book with lot's of algorithms, and convergence proofs.
- All definitions (convexity, lower semi continuity) in specific chapters.

Numerical optimization [Nocedal and Wright, 2006]

- Classic introduction to numerical optimization.

References I

Argyriou, A., Evgeniou, T., and Pontil, M. (2008).
Convex multi-task feature learning.
Machine learning, 73:243–272.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010).
Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-Lojasiewicz inequality.
Mathematics of Operations Research, 35(2):438–457.

Bauschke, H. H., Combettes, P. L., et al. (2011).
Convex analysis and monotone operator theory in Hilbert spaces, volume 408.
Springer.

Beck, A. and Teboulle, M. (2009).
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202.

Bertsekas, D. P. (1997).
Nonlinear programming.
Journal of the Operational Research Society, 48(3):334–334.

References II

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011).
Distributed optimization and statistical learning via the alternating direction method of multipliers.
Foundations and Trends® in Machine learning, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004).
Convex optimization.
Cambridge university press.

Chambolle, A. and Pock, T. (2011).
A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of mathematical imaging and vision, 40(1):120–145.

Combettes, P. L. and Pesquet, J.-C. (2011).
Proximal splitting methods in signal processing.
In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer.

Condat, L. (2014).
A generic proximal algorithm for convex optimization—application to total variation minimization.
IEEE Signal Processing Letters, 21(8):985–989.

References III

Goodfellow, I. (2016).
Deep learning.

Nocedal, J. and Wright, S. (2006).
Numerical optimization.
Springer Science & Business Media.

Obozinski, G., Taskar, B., and Jordan, M. I. (2010).
Joint covariate selection and joint subspace selection for multiple classification problems.
Statistics and Computing, 20:231–252.

Parikh, N. and Boyd, S. P. (2014).
Proximal algorithms.
Foundations and Trends in optimization, 1(3):127–239.

Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

Vapnik, V. (2013).
The nature of statistical learning theory.
Springer science & business media.

References IV

Vũ, B. C. (2013).
A splitting algorithm for dual monotone inclusions involving cocoercive operators.
Advances in Computational Mathematics, 38(3):667–681.

Xu, Z., Chang, X., Xu, F., and Zhang, H. (2012).
$L_{1/2}$ regularization: a thresholding representation theory and a fast solver.
Neural Networks and Learning Systems, IEEE Transactions on, 23(7):1013–1027.

Zhao, P. and Yu, B. (2006).
On model selection consistency of lasso.
The Journal of Machine Learning Research, 7:2541–2563.