

# Proximal Methods in Numerical Optimization

Lecture I – Motivation and Background

**Alberto DE MARCHI**

University of the Bundeswehr Munich

`alberto.demarchi@unibw.de`

`aldma.github.io`

Povo, UniTN – March 3, 2026



these slides are under development: please email me for corrections and suggestions



# Alberto De Marchi

Post-Doc at Institute of Applied Mathematics & Scientific Computing  
University of the Bundeswehr Munich (UniBw M) – Germany

## Background & Education

- ▶ Visiting Research Associate,  
Curtin University, Perth, AU      2022+2025
- ▶ Dr. rer. nat., Applied Mathematics  
UniBw M, Munich, DE                      2021
- ▶ M.Sc., Mechatronics Engineering  
Univ. of Trento, IT                      2016
- ▶ B.Sc., Industrial Engineering  
Univ. of Trento, IT                      2014

## Research Interests

- ▶ Nonsmooth & nonconvex optimization
- ▶ Constrained structured optimization
- ▶ Optimal control & model predictive control
- ▶ Numerical analysis, scientific computing

**PhD thesis:** Augmented Lagrangian and Proximal Methods for Constrained Structured Optimization

`alberto.demarchi@unibw.de`

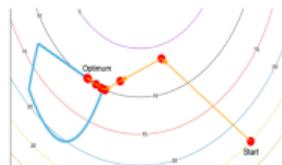
`aldma.github.io`

# Engineering Mathematics Group at UniBw Munich

Prof. Matthias Gerds, 12 PhD students, 6 Post-Doc

Institute within the **Department of Aerospace Engineering**

- ▶ **Optimal control** of ODEs, DAEs & PDEs
- ▶ **Model predictive control**  
real-time & online optimization
- ▶ **Nonsmooth & constrained optimization**  
proximal methods, augmented Lagrangians
- ▶ **Sensitivity analysis & parameter identification**
- ▶ **Trajectory optimization**  
aerospace, UAVs, autonomous vehicles
- ▶ **Multi-agent & cooperative control**  
generalized Nash equilibria, scheduling
- ▶ **Machine learning** for control  
imitation NMPC, neural networks



Optimization and Control



Problem-informed Machine Learning



Automated Driving



Flight Path Optimization



Space Robotics



Mobile Robots Lab

- ▶ Vehicle-in-the-Loop
- ▶ Flight Lab / Fixed-wing UAVs
- ▶ Multi-domain unmanned vehicles
- ▶ Industrial process control

## What is this seminar all about?

Adapted from Parikh & Boyd's **Proximal Algorithms** [PB14]:

*This [seminar] is about a class of **optimization** algorithms called **proximal algorithms**. Much like Newton's method is a standard tool for solving unconstrained smooth optimization problems of modest size, proximal algorithms can be viewed as an analogous **tool for nonsmooth, constrained, large-scale, or distributed** versions of these problems. They are very generally applicable, but are especially well-suited to problems of substantial recent interest involving large or high-dimensional datasets. **Proximal methods sit at a higher level of abstraction than classical algorithms** like Newton's method: **the base operation is evaluating the proximal operator** of a function, which itself involves solving a small [possibly non]convex optimization problem. [...]*

**Part I** Motivation and Background

**Part II** Proximal Point Methods

**Part III** Proximal Gradient Methods

**Part IV** Nonsmooth Problems with Constraints

**Part V** Convex Problems

# Outline

Motivating Example

Regularization

Applications

First-Order Optimization Methods

Proximal Operator

## Least Square Regression

### Setup

Let  $m \in \mathbb{N}_{++}$ . For  $i = 1, \dots, m$ , given  $x_i \in \mathbb{R}$ :

$$y_i = ax_i + b + \epsilon_i$$

with  $\epsilon_i$  random noise.

**Matrix-vector form:**

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}, \quad x = \begin{bmatrix} a \\ b \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{bmatrix}.$$

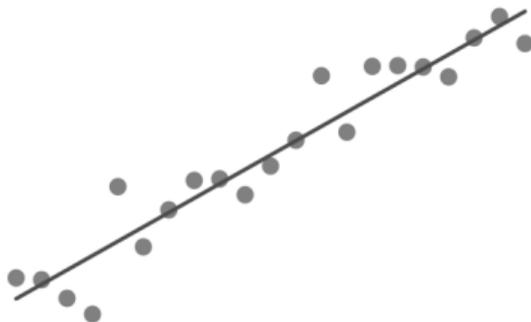
$$\implies y = Ax + \epsilon.$$

## Least Square Regression – Solution

Least square regression: estimating  $x$  from  $y$

$$\text{Model } y = Ax + \epsilon \quad \Longrightarrow \quad \underset{x}{\text{minimize}} \quad \|Ax - y\|^2$$

$$- \quad y = 0.23x - 0.08$$



**Full column rank case:**

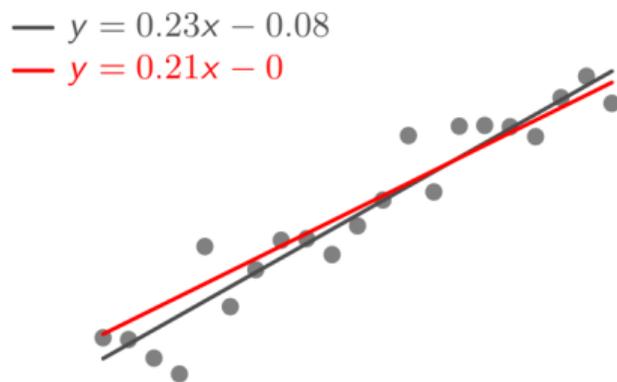
$$\underset{x}{\text{minimize}} \|Ax - y\|^2 \iff A^T Ax = A^T y \iff x = (A^T A)^{-1} A^T y.$$

$$A^T A \text{ not invertible} \implies \text{gradient descent: } x^{(k+1)} = x^{(k)} - \gamma_k A^T (Ax^{(k)} - y) \rightarrow x^*.$$

## Non-negative Least Square

### Non-negative least square regression

$$\underset{x}{\text{minimize}} \quad \|Ax - y\|^2 \quad \text{such that} \quad x \geq 0.$$



Smooth problem becomes **non-smooth** due to  $x \geq 0$ .

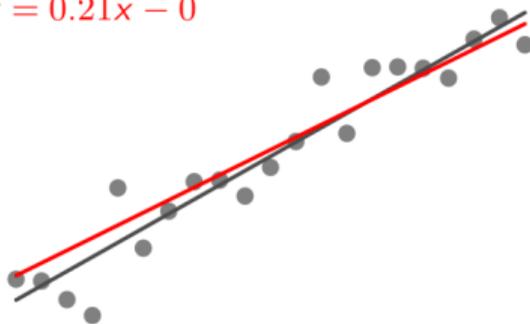
Jingwei Liang's "Selected Topics in Non-smooth Optimization"

# Sparse Least Square

## Sparse least square regression I

minimize  $\|Ax - y\|^2$  such that  $x$  has few nonzero components.

—  $y = 0.23x - 0.08$   
—  $y = 0.21x - 0$



## Sparse least square regression II

minimize  $\|Ax - y\|^2 + \text{penalty}(\text{nonzeros of } x)$ .

# Regularization

---

## Regularized Least Square

Let  $R(x)$  be a function promoting prior information, e.g. non-negativity or other simple constraints.

### Regularized least square

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|f(x)\|^2 + \mu R(x)$$

- ▶  $\mu > 0$  provides a balance between *simplicity* and *fidelity*.
- ▶ The choice of  $R(x)$  depends on the prior information.

### Regularization

$R(x)$  forces the solution to be “simpler”, “more robust”, “more expressive”:

- ▶ Obtain results for ill-posed problems (e.g. image processing).
- ▶ Prevent overfitting (e.g. machine learning).

Problems  $\min f + g$  are often called “composite” in the literature. . .

## A More General Formulation

### General regularized problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad F(x) + R(x) \quad \text{subject to} \quad \ell \leq c(x) \leq u$$

**Choices of  $F$ :** quadratic loss, logistic loss, squared hinge loss, ...

**Choices of  $R$ :**  $\ell_1$ -norm,  $\ell_2$ -norm,  $\ell_\infty$ -norm, MCP, SCAD, rank, nuclear norm, ...

- ▶ Without  $R$ , back to nonlinear programming.
- ▶ Without explicit constraints, back to structured problems.

## Penalties and Constraints

“It’s convenient for many purposes to consider functions  $f$  that are allowed to be **extended-real-valued**, i.e., to take values in  $\overline{\mathbb{R}} := [-\infty, \infty]$  instead of just  $\mathbb{R} := (-\infty, \infty)$ ”

Example: **indicator** of a set  $C \subset \mathbb{R}^n$ :

$$\delta_C: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}, \quad \delta_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{otherwise.} \end{cases}$$

From minimizing over  $C$

$$\underset{x \in C}{\text{minimize}} \quad f(x) \quad (\text{explicit constraint})$$

to minimizing over all of  $\mathbb{R}^n$

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \hat{f}(x) := f(x) + \delta_C(x) \quad (\text{implicit constraint})$$

“This helps in thinking abstractly about minimization and in achieving a single framework for the development of properties and results”

# Applications

---

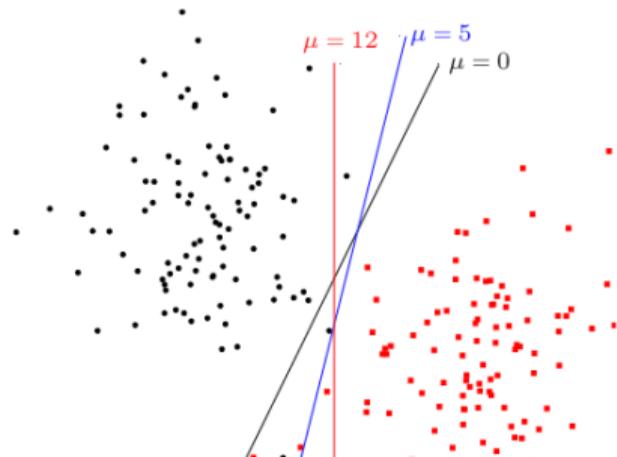
## Example: Machine Learning – Sparse Logistic Regression

### Sparse logistic regression

Let  $(A_i, b_i) \in \mathbb{R}^n \times \{\pm 1\}$ ,  $i = 1, \dots, m$ :

$$\underset{(x,y) \in \mathbb{R}^n \times \mathbb{R}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m f(x^T A_i + y; b_i) + \mu \|x\|_1$$

where  $f(u_i; b_i) = \log(1 + e^{-u_i b_i})$ .



Effect of  $\mu$ :

$\mu = 0$ : original logistic fit

$\mu = 5$ : moderate sparsity

$\mu = 12$ : stronger regularization

**sparser  $\implies$  less memory, faster inference**

## Examples: Total Variation

### Total Variation (TV) [ROF92]

Let  $\nabla$  be the discrete gradient operator:

$$\text{TV}(x) := \|\nabla x\|_1$$



Original image



Horizontal gradient



Vertical gradient

## Medical Imaging – MRI Reconstruction

Observation model:

$$f = \mathcal{F}x + \varepsilon$$

where  $x$  is piecewise constant/smooth and  $\mathcal{F}$  is partial Fourier transform.

### TV-based MRI reconstruction

Let  $p \in \{1, 2\}$ :

$$\underset{x \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \frac{1}{2} \|\mathcal{F}x - f\|_p^p + \mu \|\nabla x\|_1$$

## Video Decomposition

Observation model:

$$y_{\text{obs}} = \bar{B} + \bar{F} + \varepsilon$$

where  $\bar{B}$  is the background (prior: low rank  $\rightsquigarrow$  nuclear norm, rank constraint),  $\bar{F}$  is the foreground (prior: sparse  $\rightsquigarrow$   $\ell_1$ -norm).

### Principal component pursuit [CLMW11]

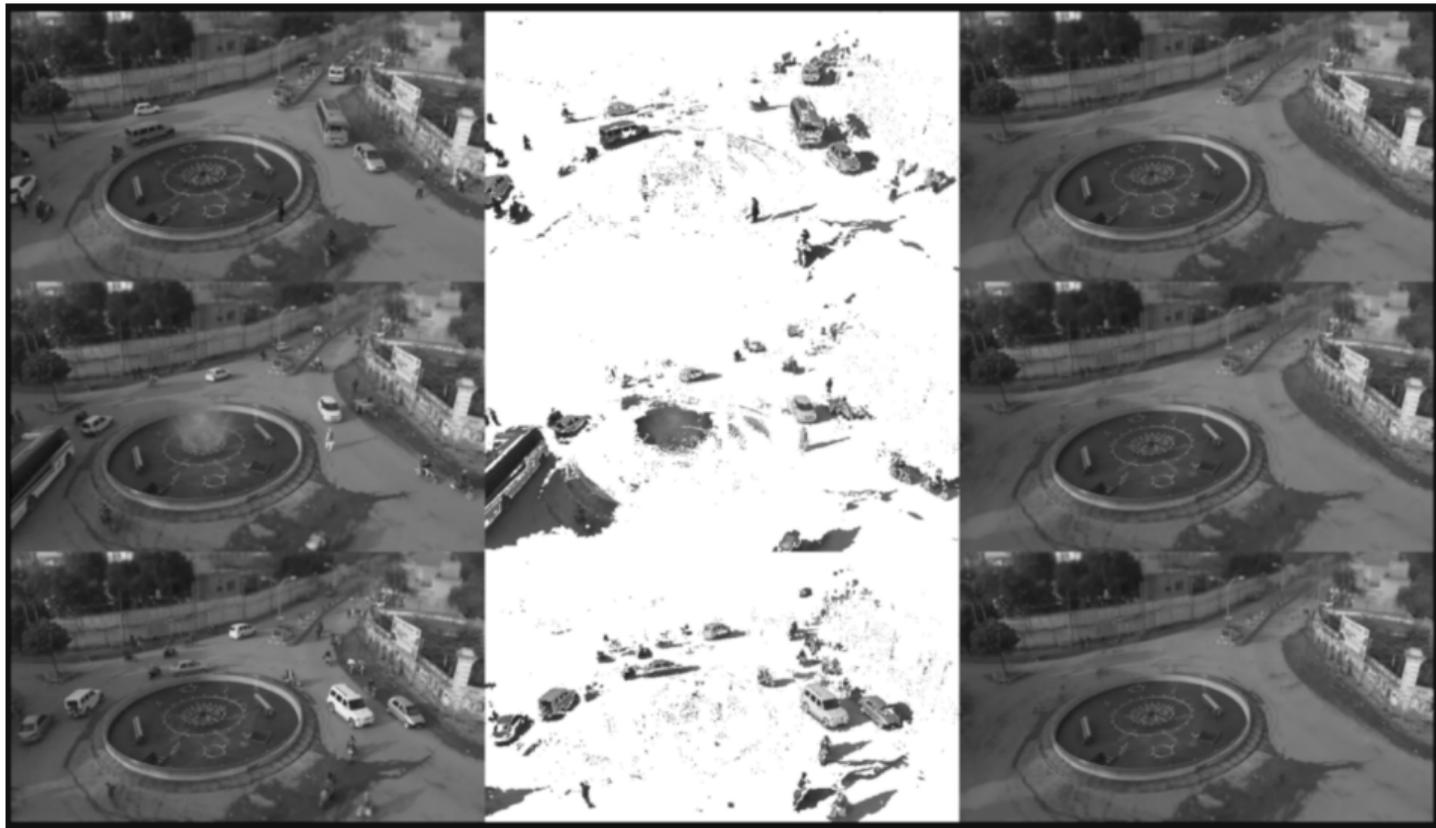
$$\underset{B, F \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \frac{1}{2} \|B + F - y_{\text{obs}}\|^2 + \mu \|B\|_* + \nu \|F\|_1$$

### Rank-constrained decomposition [ASPvW20]

$$\underset{B, F \in \mathbb{R}^{m \times n}}{\text{minimize}} \quad \frac{1}{2} \|B + F - y_{\text{obs}}\|^2 + \nu \|F\|_1 \quad \text{subject to} \quad \text{rank}(B) \leq r.$$

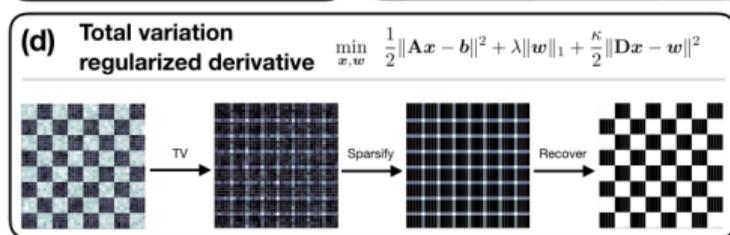
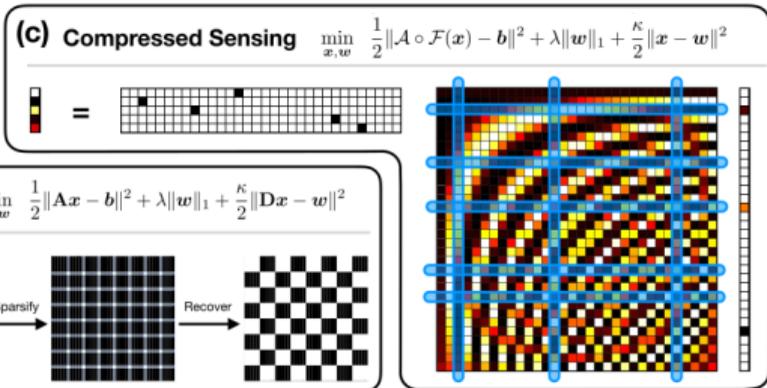
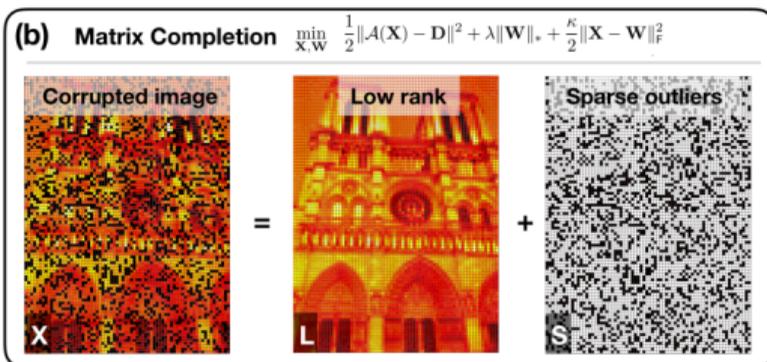
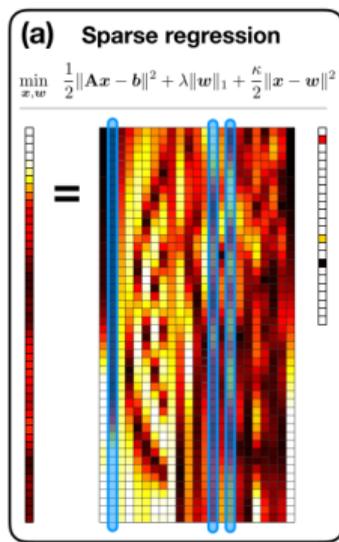
VideoBackgroundRemoval.ipynb in StructuredOptimization.jl

## Video decomposition



VideoBackgroundRemoval.ipynb in StructuredOptimization.jl

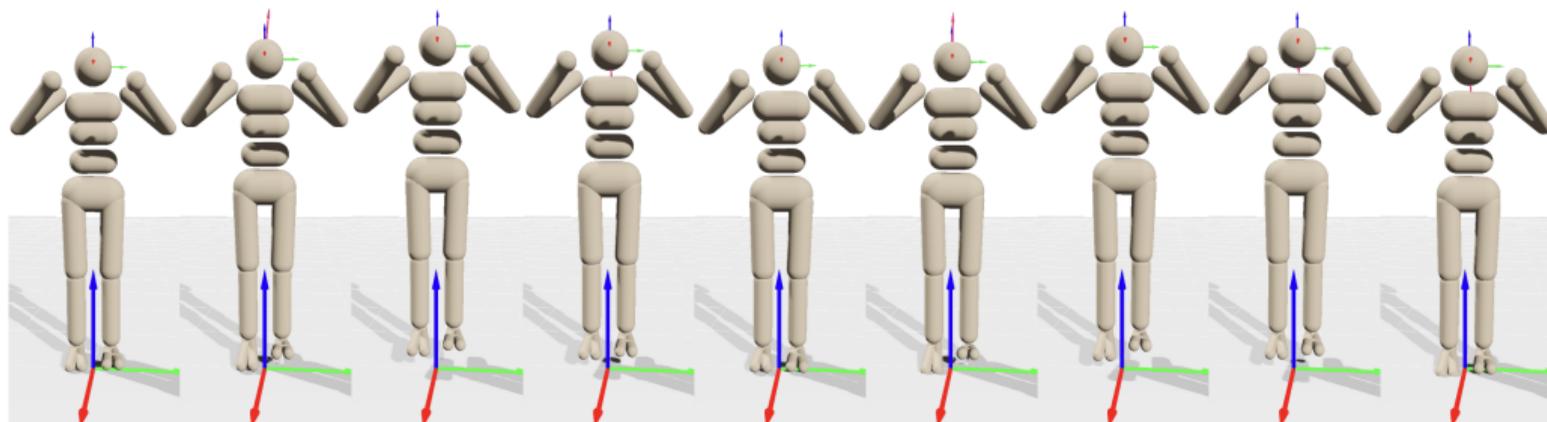
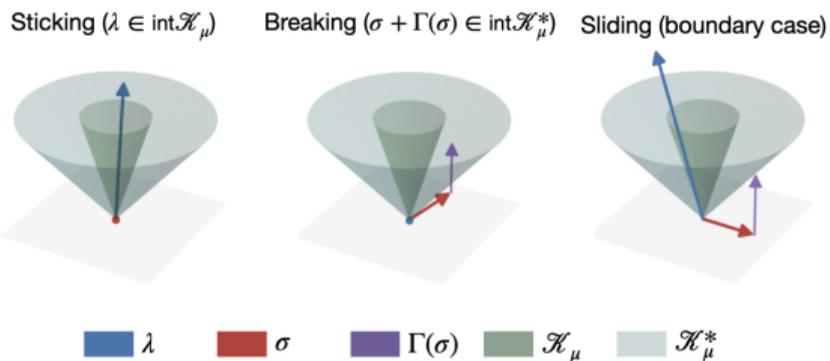
# ... and many more examples



SR3 approach [ZAB<sup>+</sup>19]

[link to pysr3]

## Inverse dynamics and trajectory optimization with **frictional contact**



# First-Order Optimization Methods

---

## Non-smooth Optimization

### Non-smooth optimization problem

Let  $r \in \mathbb{N}_{++}$ :

$$\underset{x_1, \dots, x_r}{\text{minimize}} \left\{ \Phi(x) := F(x_1, \dots, x_r) + \sum_{i=1}^r R_i(K_i x_i) \right\}$$

where  $F$  is smooth (objective/loss/data fidelity),  
 $R_i$  is prox-friendly (regularization),  
and  $K_i$  are linear/nonlinear operators.

Non-smooth, (non-convex), composite, high dimension

## First-Order Methods: Basic Ingredients

First-Order Methods: Numerical schemes that use *at most* first-order oracles of the objective (gradient, subgradient, prox, ...)

### Gradient Descent [Cauchy 1847]

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad F(x)$$

$F$  convex smooth,  $\nabla F$   $L$ -Lipschitz.

**GD update:**

$$x^{(k+1)} = x^{(k)} - \gamma_k \nabla F(x^{(k)}), \quad \gamma_k \in \left(0, \frac{2}{L}\right).$$

where  $\nabla F$  is the **gradient** of  $F$ .

### Proximal Point Algorithm [Roc76b]

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad R(x)$$

$R$  proper, lsc, convex.

**PPA update:**

$$x^{(k+1)} = \text{prox}_{\gamma_k R}(x^{(k)}), \quad \gamma_k > 0.$$

where the **proximal operator** is

$$\text{prox}_R(v) := \arg \min_x \left\{ R(x) + \frac{1}{2} \|x - v\|^2 \right\}.$$

## Problem Classes

---

| Problem structure                                 | Method (convex case)              |
|---|-----------------------------------|
| $\min R_1 + R_2$                                  | Douglas–Rachford splitting [DR56] |
| $\min_{x,z} \{R_1(x) + R_2(z) \mid Ax + Bz = c\}$ | ADMM [GM75, GM76]                 |
| $\min F + R$                                      | Forward–Backward splitting [LM79] |
| $\min F + R \circ K$                              | Primal–Dual splitting [CP11]      |
| $\min F + \sum_i R_i$                             | Generalized FB splitting [RFP13]  |

---

### Part V: Convex Problems on Friday

More difficult to exploit structure in nonconvex settings, or so it seems.

## Settings and History: A New Hope

Douglas & Rachford [DR56], Moreau [Mor65], Rockafellar [Roc76a], Gabay & Mercier [GM76]

$$\underset{x}{\text{minimize}} \quad R(x), \quad \underset{x}{\text{minimize}} \quad R_1(x) + R_2(x)$$

Lions & Mercier [LM79], Fukushima & Mine [FM81]

$$\underset{x}{\text{minimize}} \quad F(x) + R(x)$$

... forgotten for decades, until Beck & Teboulle's **accelerated prox-grad** [BT09]

Primal-dual: Chambolle & Pock [CP11]

$$\underset{x}{\text{minimize}} \quad F(x) + R(Kx)$$

Parih & Boyd's FnT on "Proximal Algorithms" [PB14] **still convex**

## Settings and History: Nonconvexity Awakens

Attouch, Bolte, Teboulle et al. [ABS13, BST14, BBT17, BSTV18]

Beck's book *First-Order Methods in Optimization* [Bec17]

↪ Winter school on **“Modern methods in nonsmooth optimization”** in March 2018

Recent and fast-growing field [ABO22, DMT22, KM22, HT23, DMJKM23, DM23, DMT25, EPLP25, DFZ<sup>+</sup>25, DHO26]

Relaxing assumptions, improving rates/complexity, weakening oracles, broader settings. . .

**Convexity strikes back!** Adaptive variants [MM24, LTSP25] for ML

## Not in this course, but topics of active research

- ▶ Accelerated: same oracles, faster rates
- ▶ Decentralized / distributed
- ▶ Inexact evaluations
- ▶ Stochastic / randomized
- ▶ Riemannian / non-Euclidean / manifolds

# Proximal Operator

---

## Proximal Mapping – Definition

### Definition – Proximal Mapping

The proximal mapping (proximity operator) of  $R$  with parameter  $\gamma > 0$  is

$$\text{prox}_{\gamma R}(y) := \arg \min_{x \in \mathbb{R}^n} \left\{ R(x) + \frac{1}{2\gamma} \|x - y\|^2 \right\}.$$

Function  $R: \mathbb{R}^n \rightarrow \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$  should be proper and lsc.

Proximal methods adopt the proximal mapping as oracle to access  $R$

- ▶ for **convex**  $R$ ,  $\text{prox}_{\gamma R}(y)$  is nonempty and unique for all  $y \in \mathbb{R}^n$  and  $\gamma > 0$ .
- ▶ in general, the stepsize  $\gamma$  has to be small enough for  $\text{prox}_{\gamma R}$  to be well-defined.
- ▶ **prox-boundedness** of  $R$  at  $y$ : there exists  $\gamma_{R,y} > 0$  such that  $\text{prox}_{\gamma R}(y)$  is nonempty for all  $\gamma \in (0, \gamma_{R,y})$ .
- ▶ **uniform prox-boundedness** of  $R$ : there exists  $\gamma_R > 0$  such that  $\text{prox}_{\gamma R}(y)$  is nonempty for all  $y \in \mathbb{R}^n$  and  $\gamma \in (0, \gamma_R)$ .

## Interpretations: Smoothing

### Definition – Moreau Envelope

Given  $\gamma > 0$ , the Moreau envelope (or Moreau–Yosida regularization)  $R^\gamma$  of  $R$  is

$$R^\gamma(y) := \inf_{x \in \mathbb{R}^n} \left\{ R(x) + \frac{1}{2\gamma} \|x - y\|^2 \right\}.$$

Also denoted  $e_\gamma R$  or  $\text{env}_\gamma R$  in the literature.

The Moreau envelope  $R^\gamma$  is a convolution of  $R$  with  $\frac{1}{2\gamma} \|\cdot\|^2$ : essentially a smoothed or regularized form of  $R$ .

- ▶ It has full domain  $\mathbb{R}^n$ , even when  $R$  does not.
- ▶ The sets of minima and minimizers of  $R$  and  $R^\gamma$  are the same.
- ▶ For **convex**  $R$ , it is continuously differentiable, even when  $R$  is not.

## Moreau envelope: Example I

### Absolute value

Let  $R(x) := |x|$ . Then

$$R^\gamma(x) = \inf_{y \in \mathbb{R}} \left\{ |y| + \frac{1}{2\gamma}(y - x)^2 \right\} = \begin{cases} \frac{x^2}{2\gamma} & \text{if } |x| \leq \gamma, \\ |x| - \frac{\gamma}{2} & \text{otherwise.} \end{cases}$$

Since  $R$  is convex,  $R^\gamma$  is differentiable:

$$\nabla R^\gamma(x) = \begin{cases} \frac{x}{\gamma} & \text{if } |x| \leq \gamma, \\ \text{sign}(x) & \text{otherwise.} \end{cases}$$

[link to desmos]

## Moreau envelope: Example II

### Indicator of a convex set

Given  $l, u \in \mathbb{R}$ ,  $l \leq u$ , let  $R(x) := \delta_{[l,u]}(x) := \begin{cases} 0 & \text{if } x \in [l, u], \\ \infty & \text{otherwise} \end{cases}$ . Then

$$R^\gamma(x) = \inf_{y \in \mathbb{R}} \left\{ \delta_{[l,u]}(y) + \frac{1}{2\gamma}(y - x)^2 \right\} = \frac{1}{2\gamma} \begin{cases} (u - x)^2 & \text{if } x \geq u, \\ (\ell - x)^2 & \text{if } x \leq \ell, \\ 0 & \text{otherwise} \end{cases} = \frac{1}{2\gamma} \text{dist}_{[l,u]}^2(x).$$

Since  $R$  is convex,  $R^\gamma$  is differentiable:

$$\nabla R^\gamma(x) = \frac{1}{\gamma} \begin{cases} x - u & \text{if } x \geq u, \\ x - \ell & \text{if } x \leq \ell, \\ 0 & \text{otherwise} \end{cases} = \frac{x - \text{proj}_{[l,u]}(x)}{\gamma}.$$

[link to desmos]

## Proximal operator and Moreau envelope

Given  $R$  and  $\gamma > 0$ ,

$$\text{prox}_{\gamma R}(x) := \arg \min_{z \in \mathbb{R}^n} \left\{ R(z) + \frac{1}{2\gamma} \|z - x\|^2 \right\}, \quad R^\gamma(x) := \inf_{z \in \mathbb{R}^n} \left\{ R(z) + \frac{1}{2\gamma} \|z - x\|^2 \right\}.$$

$\text{prox}_{\gamma R}$  returns a point that actually achieves the infimum that defines  $R^\gamma$

$$\forall x \in \mathbb{R}^n: \quad R^\gamma(x) = R(z) + \frac{1}{2\gamma} \|z - x\|^2 \quad \text{for all suff. small } \gamma > 0 \text{ and } z \in \text{prox}_{\gamma R}(x).$$

and (for **convex**  $R$ ) the gradient of the Moreau envelope is given by

$$\nabla R^\gamma(x) = \frac{x - \text{prox}_{\gamma R}(x)}{\gamma}.$$

## Interpretations: Gradient step on the Moreau envelope

(for **convex**  $R$ ) the gradient of the Moreau envelope is given by

$$\nabla R^\gamma(x) = \frac{x - \text{prox}_{\gamma R}(x)}{\gamma}.$$

Rewriting as

$$\text{prox}_{\gamma R}(x) = x - \gamma \nabla R^\gamma(x)$$

shows that  $\text{prox}_{\gamma R}$  **can be viewed as a gradient step**, with stepsize  $\gamma$ , **for minimizing**  $R^\gamma$  (which has the same minimizers as  $R$ ).

## Interpretations: Generalized Projection

### Definition: Euclidean/Orthogonal projection

Given a closed nonempty set  $S$ ,

$$\text{proj}_S(y) := \arg \min_{x \in S} \|x - y\|.$$

Let  $\delta_S$  be the indicator of  $S$ . Then

$$\begin{aligned} \text{proj}_S(y) &:= \arg \min_{x \in S} \|x - y\| = \arg \min_{x \in S} \frac{1}{2} \|x - y\|^2 = \arg \min_{x \in S} \frac{1}{2\gamma} \|x - y\|^2 \\ &= \arg \min_{x \in \mathbb{R}^n} \left\{ \delta_S(x) + \frac{1}{2\gamma} \|x - y\|^2 \right\} = \text{prox}_{\gamma \delta_S}(y) \end{aligned}$$

for all  $\gamma > 0$ .

## Calculus rules for prox

### Quadratic perturbation

$$H(x) := R(x) + \frac{\alpha}{2} \|x\|^2 + x^\top u + b, \alpha \geq 0 \implies \text{prox}_H(y) = \text{prox}_{R/(\alpha+1)}\left(\frac{y-u}{\alpha+1}\right).$$

### Translation

$$H(x) := R(x - z) \implies \text{prox}_H(y) = z + \text{prox}_R(y - z).$$

### Scaling

$$H(x) := R(x/\rho) \implies \text{prox}_H(y) = \rho \text{prox}_{R/\rho^2}(y/\rho).$$

### Reflection

$$H(x) := R(-x) \implies \text{prox}_H(y) = -\text{prox}_R(-y).$$

### Composition ( $K^{-1} = K^*$ , bijective bounded linear)

$$H := R \circ K \implies \text{prox}_H(y) = K^* \text{prox}_R(Ky).$$

## Examples of prox

**Projection**  $R(x) = \delta_S(x)$ . Then  $\text{prox}_{\gamma R}(x) = \text{proj}_S(x)$ .

- ▶ Hyperplane  $S = \{x : a^\top x = b\}$ ,  $a \neq 0$ :  $\text{proj}_S(y) = y + \frac{b - a^\top y}{\|a\|^2} a$ .
- ▶ Affine subspace  $S = \{x : Ax = b\}$ ,  $A \in \mathbb{R}^{m \times n}$ ,  $\text{rank}(A) = m$ :  
 $\text{proj}_S(y) = y + A^\top (AA^\top)^{-1} (b - Ay)$ .
- ▶ Nonnegative orthant  $S = \mathbb{R}_+^n$ :  $\text{proj}_S(y) = \max\{0, y\}$  (elementwise).

**Quadratic**  $R(x) = \frac{1}{2}x^\top Ax + b^\top x + c$ ,  $A \succeq 0$ :

$$\text{prox}_{\gamma R}(x) = (\mathbb{I} + \gamma A)^{-1}(x - \gamma b).$$

## Examples of prox

**Absolute value/ $\ell_1$  norm**  $R(x) = |x|$ :

$$\text{prox}_{\gamma R}(x) = \begin{cases} x - \gamma & \text{if } x \geq \gamma, \\ x + \gamma & \text{if } x \leq -\gamma, \\ 0 & \text{otherwise.} \end{cases}$$

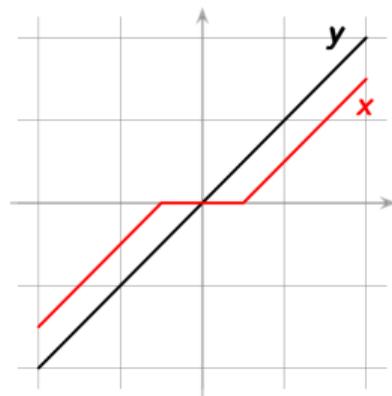
**Euclidean norm**  $R(x) = \|x\|_2$ :

$$\text{prox}_{\gamma R}(x) = \begin{cases} (1 - \frac{\gamma}{\|x\|})x & \|x\| > \gamma, \\ 0 & \text{otherwise.} \end{cases}$$

**Nuclear norm**  $R(X) = \sum_i \sigma_i(X)$ ,  $y = USV^T$ ,  $S = \text{diag } \sigma$ :

$$\text{prox}_{\gamma R}(y) = U \text{prox}_{\gamma \|\cdot\|_1}(S) V^T.$$

Soft-thresholding operator:



## Fancy some code?

### Proximal **operators**

- ▶ [proximity-operator.net](http://proximity-operator.net) (MATLAB/python)
- ▶ [cvxpy.org](http://cvxpy.org) (python)
- ▶ [JuliaFirstOrder/ProximalOperators.jl](https://github.com/JuliaFirstOrder/ProximalOperators.jl) (julia)

### Proximal **algorithms**

- ▶ [kul-optec/ForBES](https://github.com/kul-optec/ForBES) (MATLAB)
- ▶ [cvxpy.org](http://cvxpy.org) (python)
- ▶ [JuliaFirstOrder/ProximalAlgorithms.jl](https://github.com/JuliaFirstOrder/ProximalAlgorithms.jl) (julia)
- ▶ [JuliaSmoothOptimizers/RegularizedOptimization.jl](https://github.com/JuliaSmoothOptimizers/RegularizedOptimization.jl) (julia)

## Asking claude.ai (Sonnet 4.6)

### Prompt:

*Why are proximal methods important in nonsmooth optimization?*

### Reply:

**The core difficulty they address.** *Gradient descent breaks down when  $f$  is nonsmooth — the gradient may not exist, and even subgradient methods converge slowly and cannot be accelerated in general. Proximal methods sidestep this by replacing the gradient step with the **proximal operator**:*

$$\text{prox}_{\alpha f}(x) = \arg \min_u \left\{ f(u) + \frac{1}{2\alpha} \|u - x\|^2 \right\}$$

*This is **always well-defined (strongly convex subproblem, unique minimizer)** [🦉] even when  $f$  has no gradient. Geometrically it finds the point that best trades off reducing  $f$  against staying near  $x$ .*

**They encode structure implicitly.** *The power of the proximal framework is that for many nonsmooth functions that appear in practice,  $\text{prox}_{\alpha f}(x)$  has **closed-form solution**, so the "expensive" nonsmooth subproblem costs essentially nothing: [...]*

## References I

-  Aleksandr Y. Aravkin, Robert Baraldi, and Dominique Orban.  
A proximal quasi-Newton trust-region method for nonsmooth regularized optimization.  
*SIAM Journal on Optimization*, 32(2):900–929, 2022.
-  Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter.  
Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods.  
*Mathematical Programming*, 137(1):91–129, 2 2013.
-  Niccoló Antonello, Lorenzo Stella, Panagiotis Patrinos, and Toon van Waterschoot.  
Proximal gradient algorithms: Applications in signal processing, 2020.
-  Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle.  
A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications.  
*Mathematics of Operations Research*, 42(2):330–348, 2017.
-  Amir Beck.  
*First-Order Methods in Optimization*.  
Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.
-  Jérôme Bolte, Shoham Sabach, and Marc Teboulle.  
Proximal alternating linearized minimization for nonconvex and nonsmooth problems.  
*Mathematical Programming*, 146(1):459–494, 8 2014.

## References II

-  Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd.  
First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems.  
*SIAM Journal on Optimization*, 28(3):2131–2151, 2018.
-  Amir Beck and Marc Teboulle.  
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.  
*SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
-  Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright.  
Robust principal component analysis?  
*J. ACM*, 58(3), 2011.
-  Antonin Chambolle and Thomas Pock.  
A first-order primal-dual algorithm for convex problems with applications to imaging.  
*Journal of Mathematical Imaging and Vision*, 40(1):120–145, 5 2011.
-  T. Diamandis, Z. Frangella, S. Zhao, B. Stellato, and M. Udell.  
Genios: an (almost) second-order operator-splitting solver for large-scale convex optimization.  
*Mathematical Programming Computation (forthcoming)*, 2025.

## References III

-  Youssef Diouane, Mohamed Laghdaf Habiboullah, and Dominique Orban.  
A proximal modified quasi-newton method for nonsmooth regularized optimization.  
*SIAM Journal on Optimization*, 2026.
-  Alberto De Marchi.  
Proximal gradient methods beyond monotony.  
*Journal of Nonsmooth Analysis and Optimization*, 4, 2023.
-  Alberto De Marchi, Xiaoxi Jia, Christian Kanzow, and Patrick Mehlitz.  
Constrained composite optimization and augmented Lagrangian methods.  
*Mathematical Programming*, 201(1):863–896, 9 2023.
-  Alberto De Marchi and Andreas Themelis.  
Proximal gradient algorithms under local Lipschitz gradient continuity.  
*Journal of Optimization Theory and Applications*, 194(3):771–794, 2022.
-  Alberto De Marchi and Andreas Themelis.  
A penalty barrier framework for nonconvex constrained optimization.  
*Journal of Nonsmooth Analysis and Optimization*, 5(14585), 2025.
-  Jim Douglas and Henry H. Rachford.  
On the numerical solution of heat conduction problems in two and three space variables.  
*Transactions of the American Mathematical Society*, 82(2):421–439, 1956.

## References IV

-  **Brecht Evens, Pieter Pas, Puya Latafat, and Panagiotis Patrinos.**  
Convergence of the preconditioned proximal point method and douglas–rachford splitting in the absence of monotonicity.  
*Mathematical Programming*, 214(1):247–301, 2025.
-  **Masao Fukushima and Hisashi Mine.**  
A generalized proximal point algorithm for certain non-convex minimization problems.  
*International Journal of Systems Science*, 12(8):989–1000, 1981.
-  **R. Glowinski and A. Marrocco.**  
Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité, d’une classe de problèmes de dirichlet non linéaires.  
*Rev. Française Automat. Informat. Recherche Opérationnelle RAIRO Analyse Numérique*, 9:41–76, 1975.
-  **Daniel Gabay and Bertrand Mercier.**  
A dual algorithm for the solution of nonlinear variational problems via finite element approximation.  
*Computers & Mathematics with Applications*, 2(1):17–40, 1976.
-  **Nadav Hallak and Marc Teboulle.**  
An adaptive Lagrangian-based scheme for nonconvex composite optimization.  
*Mathematics of Operations Research*, 48(4):2337–2352, 2023.

## References V

-  **Christian Kanzow and Patrick Mehlitz.**  
Convergence properties of monotone and nonmonotone proximal gradient methods revisited.  
*Journal of Optimization Theory and Applications*, 195(2):624–646, 2022.
-  **Pierre-Louis Lions and Bertrand Mercier.**  
Splitting algorithms for the sum of two nonlinear operators.  
*SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.
-  **Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos.**  
Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient.  
*Mathematical Programming*, 213(1):433–471, 2025.
-  **Yura Malitsky and Konstantin Mishchenko.**  
Adaptive proximal gradient method for convex optimization, 2024.
-  **Jean Jacques Moreau.**  
Proximité et dualité dans un espace hilbertien.  
*Bulletin de la Société Mathématique de France*, 93:273–299, 1965.
-  **Neal Parikh and Stephen Boyd.**  
Proximal algorithms.  
*Foundations and Trends in Optimization*, 1(3):127–239, 2014.

## References VI

-  Hugo Raguét, Jalal Fadili, and Gabriel Peyré.  
A generalized forward-backward splitting.  
*SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.
-  Ralph Tyrrell Rockafellar.  
Augmented Lagrangians and applications of the proximal point algorithm in convex programming.  
*Mathematics of operations research*, 1(2):97–116, 5 1976.
-  Ralph Tyrrell Rockafellar.  
Monotone operators and the proximal point algorithm.  
*SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.
-  Leonid I. Rudin, Stanley Osher, and Emad Fatemi.  
Nonlinear total variation based noise removal algorithms.  
*Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
-  R. Tyrrell Rockafellar and Roger J.B. Wets.  
*Variational Analysis*, volume 317.  
Springer, 1998.
-  Peng Zheng, Travis Askham, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin.  
A unified framework for sparse relaxed regularized regression: SR3.  
*IEEE Access*, 7:1404–1423, 2019.