Proximal Methods in Numerical Optimization

Lecture V – Convex Problems

Alberto DE MARCHI
University of the Bundeswehr Munich
`alberto.demarchi@unibw.de`
`aldma.github.io`

Povo, UniTN – March 5, 2026

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

Outline

Problem Class

We study **structured minimization** problems:

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

where $f, g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper, lsc, **convex**.

Motivating examples

- f smooth (least-squares, log-likelihood)
- $\ldots + g$ nonsmooth regularizer (e.g. ℓ_1, nuclear norm, indicator)
- Feasibility problems: $f = \delta_C$, $g = \delta_D$ (find $x \in C \cap D$)

Why not just gradient descent?

- f, g may not be differentiable (e.g. $\|\cdot\|_1$)
- Subgradient methods converge slowly: $\mathcal{O}(1/\sqrt{k})$
- Proximal methods exploit structure and attain $\mathcal{O}(1/k)$ or linear rates

Convex Analysis Recap

Subdifferential

$\partial f(x) = \{g : f(y) \geq f(x) + \langle g, y - x\rangle \ \forall y\}.$

Optimality: $\quad 0 \in \partial f(x^\star) + \partial g(x^\star).$

Conjugate function

$f^\star(y) = \sup_x \langle y, x\rangle - f(x).$

Fenchel: $f^{\star\star} = f$,

$y \in \partial f(x) \iff x \in \partial f^\star(y).$

Every convex proximal algorithm is a *special case* of a monotone operator splitting or a *fixed-point iteration* of a nonexpansive map.

Proximal mapping

$$\operatorname{prox}_{\lambda f}(v) := \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\lambda} \|x - v\|^2 \right\}$$

Properties

- Single-valued
- **Firmly nonexpansive**:
  $\|\operatorname{prox}_{\lambda f}(x) - \operatorname{prox}_{\lambda f}(y)\|^2$
  $\leq \langle x - y, \operatorname{prox}_{\lambda f}(x) - \operatorname{prox}_{\lambda f}(y)\rangle$
- Fixed points:
  $\operatorname{prox}_{\lambda f}(x^\star) = x^\star \iff 0 \in \partial f(x^\star)$
- **Resolvent** interpretation:
  $\operatorname{prox}_{\lambda f} = \mathcal{J}_{\lambda \partial f} := (\mathbb{I} + \lambda \partial f)^{-1}$

Operator Splitting: A Big Picture

Goal: solve $0 \in (A + B)(x)$ using A and B *separately*.

Method	**Setting**	**Requirements**
Forward-Backward	$A = \partial f$, $B = \nabla g$	A cocoercive
Douglas–Rachford	$A = \partial f$, $B = \partial g$	both maximal monotone
Peaceman–Rachford	$A = \partial f$, $B = \partial g$	both strongly monotone
ADMM	structured splitting	DR on dual
Davis–Yin	$A + B + C$	C cocoercive

Key operators:

- **Resolvent** of A: $\mathcal{J}_{\lambda A} = (\mathbb{I} + \lambda A)^{-1}$ (= $\text{prox}_{\lambda f}$ when $A = \partial f$)
- **Reflected resolvent**: $\mathcal{R}_{\lambda A} = 2\mathcal{J}_{\lambda A} - \mathbb{I}$ (nonexpansive when A is maximally monotone)

Alternating Projections and Douglas–Rachford

Alternating Projections

Feasibility problem: find $x \in C \cap D$, $C, D \subseteq \mathbb{R}^n$ closed convex sets.

Method of Alternating Projections (MAP)

$$x^{k+1} = \text{proj}_C\big(\text{proj}_D(x^k)\big)$$

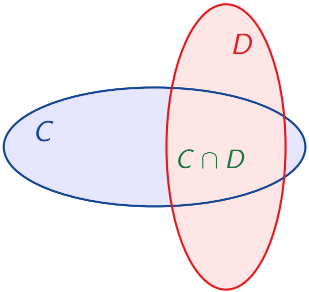
- **von Neumann (1950):** proved convergence for closed *subspaces*; linear rate $\cos^2\theta$ where θ is the *Friedrichs angle* between C and D
- **General convex sets:** converges, but rate can be arbitrarily slow (angle may be 0)

Alternating Projections: Geometry and Rates

Friedrichs angle θ (between subspaces)
Rate of MAP: $\left\|x^k - x^\star\right\| = O(\cos^k \theta)$

Pathology: two convex sets tangent at a point
$\implies \theta = 0$, sublinear convergence

Remedy: Douglas–Rachford [DR56], circumcentered reflections, RAAR (relaxed averaged alternating reflections).



> *Amongst many ways to improve the convergence rate of the method of alternating projections, one can 'overproject' by pushing the projection beyond the boundary of the closed convex set. An extreme version of such an overprojection is a reflection in the hyperplane perpendicular to the direction of the projection.*
>
> *from "Douglas–Rachford is the best projection method" [DDLR23]*

Douglas–Rachford Splitting: Derivation

Problem: $\text{minimize}_x\, f(x) + g(x)$ (no smoothness required on either term).

Optimality condition: $0 \in \partial f(x^\star) + \partial g(x^\star)$.

Algorithm: Douglas–Rachford (DR)

$$x^{k+1} = \text{prox}_{\lambda f}(z^k)$$
$$y^{k+1} = \text{prox}_{\lambda g}(2x^{k+1} - z^k)$$
$$z^{k+1} = z^k + y^{k+1} - x^{k+1}$$

Both x^k and y^k converge to the *same* primal solution $x^\star$.
DR operator: introduce auxiliary variable z and iterate

$$z^{k+1} = \underbrace{\frac{1}{2}\left(\mathbb{I} + \mathcal{R}_{\lambda\partial g} \circ \mathcal{R}_{\lambda\partial f}\right)}_{\mathcal{T}_{\mathrm{DR}}} z^k$$

where $\mathcal{R}_\square := 2\mathcal{J}_\square - \mathbb{I}$.

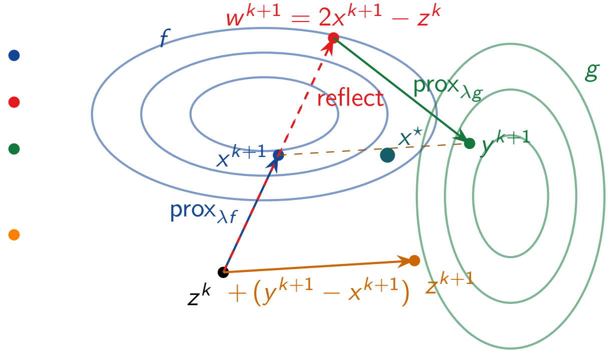
Douglas–Rachford: Geometric Illustration

One DR step from z^k:

1. $x^{k+1} = \text{prox}_{\lambda f}(z^k)$
2. $w^{k+1} = 2x^{k+1} - z^k$
3. $y^{k+1} = \text{prox}_{\lambda g}(w^{k+1})$
4. $z^{k+1} = z^k + (y^{k+1} - x^{k+1})$ (midpoint update)

At convergence: $x^\infty = y^\infty = x^\star$.

Fixed point: shadow z^∞ satisfies $\mathcal{R}_{\lambda\partial g}\mathcal{R}_{\lambda\partial f} z^\infty = z^\infty$.



Douglas–Rachford: Convergence Analysis

DR does *not* require f or g to be differentiable—both are accessed only through $\mathrm{prox}_{\lambda f}$ and $\mathrm{prox}_{\lambda g}$.

Operator viewpoint: $\mathcal{T}_{\mathrm{DR}} := \frac{1}{2}(\mathbb{I} + \mathcal{R}_{\lambda\partial g}\mathcal{R}_{\lambda\partial f})$

- $\mathcal{R}_{\lambda\partial f}$, $\mathcal{R}_{\lambda\partial g}$ nonexpansive $\implies$ $\mathcal{T}_{\mathrm{DR}}$ is $\frac{1}{2}$-**averaged nonexpansive**.
- Krasnosel'skiĭ–Mann theorem: z^k converges to a fixed point z^∞.
- At z^∞: $x^\infty := \mathrm{prox}_{\lambda f}(z^\infty)$ solves the original problem.

Convergence rates

- General convex: $\mathcal{O}(1/k)$ on the fixed-point residual $\left\|z^{k+1} - z^k\right\|^2$
- f or g strongly convex: **linear** (geometric) convergence

DR vs. Peaceman–Rachford

Peaceman–Rachford (PR) uses the full reflection without averaging:

$$z^{k+1} = \mathcal{R}_{\lambda\partial g}\mathcal{R}_{\lambda\partial f} z^k$$

	Douglas–Rachford	Peaceman–Rachford
Operator	$\frac{1}{2}(\mathbb{I} + \mathcal{R}_{\lambda\partial g}\mathcal{R}_{\lambda\partial f})$	$\mathcal{R}_{\lambda\partial g}\mathcal{R}_{\lambda\partial f}$
Convergence (general)	Guaranteed (averaged)	Not guaranteed in general
Convergence (strongly convex)	Linear	Linear, faster constant
Oscillation behaviour	Damped	Can oscillate

Relaxed DR: Replace the z-update by $z^{k+1} = z^k + 2\alpha(y^{k+1} - x^{k+1})$, $\alpha \in (0, 1)$.

$\alpha = \frac{1}{2}$ gives DR; $\alpha \to 1$ recovers PR.
Over-relaxation ($\alpha > 0.5$) often accelerates in practice.

Augmented Lagrangian and ADMM

Augmented Lagrangian Method

Constrained problem:

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad Ax = b$$

Standard Lagrangian: $\mathscr{L}(x, y) := f(x) + \langle y, Ax - b \rangle$

Augmented Lagrangian ($\rho > 0$):

$$\mathscr{L}_\varrho(x, y) := f(x) + \langle y, Ax - b \rangle + \frac{\varrho}{2} \|Ax - b\|^2$$

Method of Multipliers [Hes69, Pow69]:

$$x^{k+1} = \arg\min_x \mathscr{L}_\varrho(x, y^k)$$
$$y^{k+1} = y^k + \varrho(Ax^{k+1} - b)$$

- Dual update = gradient ascent step on the augmented dual function
- Converges even for fixed finite ϱ (no need $\varrho \to \infty$)
- **Limitation:** x-minimization is a *joint* problem—no decomposition

From Augmented Lagrangian to ADMM

Two-block separable problem:

$$\underset{x,z}{\text{minimize}} \quad f(x) + g(z) \qquad \text{subject to} \quad Ax + Bz = c$$

Augmented Lagrangian:

$$\mathscr{L}_\varrho(x, z, y) := f(x) + g(z) + \langle y, Ax + Bz - c\rangle + \frac{\varrho}{2}\|Ax + Bz - c\|^2$$

Key idea: instead of jointly minimizing over (x, z), minimize block-wise and only once.

ADMM (Alternating Direction Method of Multipliers) [GM75, BPC+11]:

$$x^{k+1} = \arg\min_x \; f(x) + \frac{\varrho}{2}\left\|Ax + Bz^k - c + \frac{y^k}{\varrho}\right\|^2$$

$$z^{k+1} = \arg\min_z \; g(z) + \frac{\varrho}{2}\left\|Ax^{k+1} + Bz - c + \frac{y^k}{\varrho}\right\|^2$$

$$y^{k+1} = y^k + \varrho[Ax^{k+1} + Bz^{k+1} - c]$$

ADMM as Douglas–Rachford on the Dual

Theorem [GM76, EB92]: ADMM is *equivalent* to DR applied to the dual problem.

Simplified case: $A = \mathbb{I}$, $B = -\mathbb{I}$, $c = 0$: minimize$_{x,z}$ $f(x) + g(z)$ s.t. $x - z = 0$.

DR on f and $\tilde{g}(x) := g(-x)$ gives exactly the ADMM updates:

$$x^{k+1} = \text{prox}_{\varrho^{-1}f}(z^k - \varrho^{-1}y^k)$$
$$z^{k+1} = \text{prox}_{\varrho^{-1}g}(x^{k+1} + \varrho^{-1}y^k)$$
$$y^{k+1} = y^k + \varrho[x^{k+1} - z^{k+1}]$$

Consequences:

- All convergence guarantees of DR carry over to ADMM
- The dual variable y^k/ϱ plays the role of the DR shadow variable z^k
- Over-relaxation on x^{k+1} (with $\alpha x^{k+1} + (1-\alpha)z^k$) corresponds to relaxed DR

ADMM: Convergence

Assumptions: f, g proper closed convex; A full column rank (or suitable constraint qualification); problem has a solution.

Primal and dual residuals:

$$r_p^k := Ax^k + Bz^k - c \quad \text{(primal)}, \qquad r_d^k := \varrho A^\top B(z^k - z^{k-1}) \quad \text{(dual)}$$

Stopping criterion: $\left\| r_p^k \right\| \le \varepsilon_{\mathrm{p}}$ and $\left\| r_d^k \right\| \le \varepsilon_{\mathrm{d}}$.

Convergence results [BPC+11]:

- Primal / dual residuals: $r_p^k \to 0$, $r_d^k \to 0$.
- Objective value: $f(x^k) + g(z^k) \to$ optimal value.
- Dual variable: $y^k \to y^\star$.
- Rate: $\mathcal{O}(1/k)$ in general; **linear** if f or g is strongly convex.

Parameter ϱ: Adaptive scheme for practical speedup $\rightsquigarrow$ rescale when unbalanced:

$$\varrho \leftarrow 2\varrho \quad \text{if } \left\| r_p^k \right\| > 5 \left\| r_d^k \right\|, \qquad \varrho \leftarrow \varrho/2 \quad \text{if } \left\| r_d^k \right\| > 5 \left\| r_p^k \right\|.$$

ADMM: Distributed / Consensus Form

Global consensus:

$$\underset{x}{\text{minimize}} \sum_{i=1}^{N} f_i(x) \equiv \underset{x_i, z}{\text{minimize}} \sum_{i=1}^{N} f_i(x_i) \quad \text{subject to} \quad x_i = z \ \forall i$$

ADMM updates decompose over agents:

$$x_i^{k+1} = \text{prox}_{\rho^{-1} f_i}(z^k - u_i^k), \qquad z^{k+1} = \frac{1}{N} \sum_{i=1}^{N} (x_i^{k+1} + u_i^k), \qquad u_i^{k+1} = u_i^k + x_i^{k+1} - z^{k+1}$$

- Each agent minimises its own f_i *in parallel*
- Coordinator computes the mean (all-reduce)

General graph: replace the mean with message passing on the network graph; convergence follows from the DR equivalence on the edge-node incidence structure.

Example: Lasso via ADMM

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\left\|Cx - d\right\|^2 + \mu\left\|x\right\|_1$$

Introduce $z = x$, write as $\text{minimize}_{x,z}\, f(x) + g(z)$ s.t. $x - z = 0$:

ADMM updates:

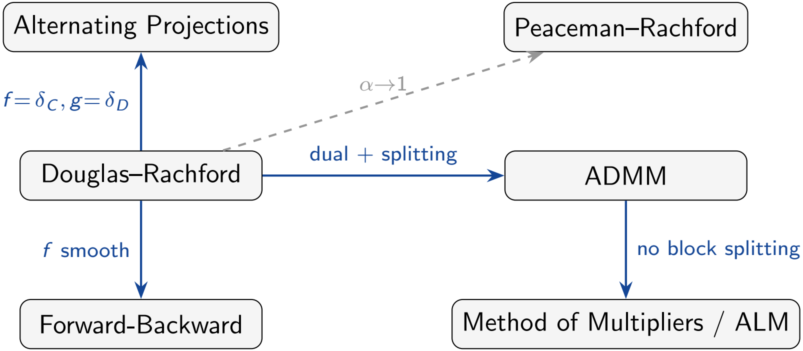
$$x^{k+1} = (C^\top C + \varrho\,\mathbb{I})^{-1}\big(C^\top d + \varrho z^k - y^k\big) \qquad \text{[linear system, pre-factor!]}$$

$$z^{k+1} = \text{prox}_{\mu/\varrho\|\cdot\|_1}(x^{k+1} + u^k) \qquad [\text{soft thresholding, } \mathcal{O}(n)]$$

$$y^{k+1} = y^k + \varrho[x^{k+1} - z^{k+1}]$$

- Factor $C^\top C + \varrho\,\mathbb{I}$ once; each iteration costs **one backsolve + soft thresholding**.
- Handles efficiently large-scale instances.
- Change g to get elastic net, group lasso, fused lasso: same x-update!

Connections Between Methods



Performance Estimation

Performance Estimation Problems (PEP)

Question: What is the *exact* worst-case convergence rate of an algorithm?

Drori & Teboulle [DT14], Taylor, Hendrickx & Glineur [THG17]

Performance Estimation Problems:

1. Fix algorithm $\mathcal{A}$ running N steps on function class. $\mathcal{F}_{L,\mu}$
2. Worst-case over all $f \in \mathcal{F}_{L,\mu}$ and initial conditions:
$$\sup_{f \in \mathcal{F},\, x^0} \Phi(x^0, x^1, \ldots, x^N).$$
3. Interpolation conditions for $\mathcal{F}_{L,\mu}$ are *linear* in the Gram matrix of inner products $G_{ij} = \langle x^i - x^j, g^i - g^j \rangle$.
4. The supremum becomes a **finite-dimensional SDP**—solvable numerically.

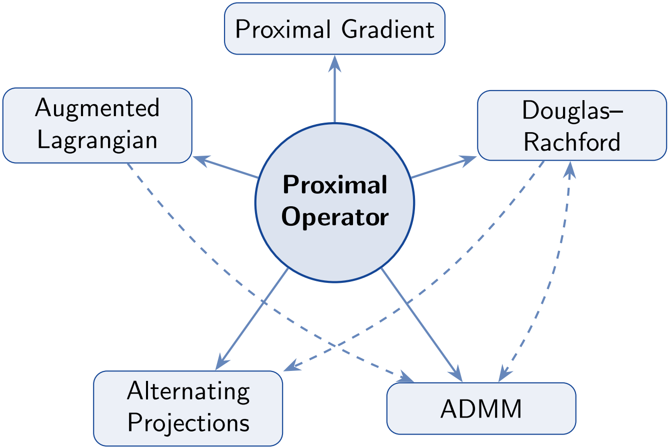
Outputs:

- Tight worst-case bound
- Worst-case function as a certificate
- **PEPit** [GMG$^+$24]: open-source python package

Proximal Methods in Numerical Optimization

Alberto DE MARCHI

University of the Bundeswehr Munich   alberto.demarchi@unibw.de   aldma.github.io



References I

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein.
Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers.
now, 2011.

Minh N. Dao, Mareike Dressler, Hongzhi Liao, and Vera Roshchina.
Douglas–Rachford is the best projection method, 2023.

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.

Yoel Drori and Marc Teboulle.
Performance of first-order methods for smooth convex minimization: a novel approach.
Mathematical Programming, 145(1):451–482, 2014.

Jonathan Eckstein and Dimitri P. Bertsekas.
On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators.
Mathematical Programming, 55(1):293–318, 1992.

References II

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.

Baptiste Goujaud, Céline Moucer, François Glineur, Julien M. Hendrickx, Adrien B. Taylor, and Aymeric Dieuleveut.
PEPit: computer-assisted worst-case analyses of first-order optimization methods in python.
Mathematical Programming Computation, 16(3):337–367, 2024.

Magnus R. Hestenes.
Multiplier and gradient methods.
Journal of Optimization Theory and Applications, 4(5):303–320, 11 1969.

Michael J. D. Powell.
A method for nonlinear constraints in minimization problems, pages 283–298.
Academic Press, 1969.

References III

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur.
Smooth strongly convex interpolation and exact worst-case performance of first-order methods.
Mathematical Programming, 161(1):307–345, 2017.