Proximal Methods in Numerical Optimization

Lecture III – Proximal Gradient Methods

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

Povo, UniTN – March 4, 2026

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

Outline

Structured Optimization Problem

Problem (P)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \varphi(x) := f(x) + g(x)$$

Smooth part f

- $f : \mathbb{R}^n \to \mathbb{R}$ differentiable
- ∇f globally/locally Lipschitz/continuous
- possibly nonconvex

Nonsmooth part g

- $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ proper, lsc (unif. prox-bounded)
- prox-friendly: "simple", often convex
- $\operatorname{dom} g$ encodes constraints

Motivating Examples

Example 1: LASSO

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1$$

- f is $\mathcal{C}^\infty$, convex, $\nabla f(x) = A^\top(Ax-b)$
- g promotes sparsity
- soft-thresholding $\text{prox}_{\alpha g}$

Example 2: Constrained QP

$$\underset{x\in C}{\text{minimize}} \quad \frac{1}{2}x^\top Qx + q^\top x$$

- f is $\mathcal{C}^\infty$, quadratic
- $g = \delta_C$ (indicator of closed set C)
- $\text{prox}_{\alpha g} = \text{proj}_C$
- recovers projected gradient method

Further examples: group LASSO, nuclear norm minimization, sparse inverse covariance, neural-network training with ℓ_1 regularization, phase retrieval, matrix factorization...

Gradient Descent: A Quick Review

Problem: minimize$_x$ $f(x)$ with $f \in \mathcal{C}^1$.

Gradient Descent (GD)

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

Three views of the same step:

1. **Steepest descent:** follow direction $-\nabla f(x_k)$, scaled by α_k
2. **Linearization:** $x_{k+1} \in \arg\min_x \left[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \right]$ . . . (unbounded, TR)
3. **Quadratic model:**

$$Q(x; x_k, \alpha) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha} \|x - x_k\|^2$$

$$\alpha_k > 0 \quad \Longrightarrow \quad x_{k+1} \in \arg\min_x Q(x; x_k, \alpha_k) = x_k - \alpha_k \nabla f(x_k)$$

The quadratic model is the key to generalizing to structured problems $f + g$.

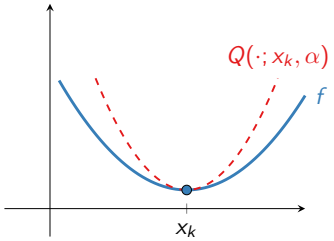
Proximal gradient methods

Quadratic Upper Models: Toward Majorization-Minimization

Suppose ∇f is (locally) L-Lipschitz, namely

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 = Q(y; x, 1/L)$$

The quadratic model $Q(\cdot; x, \alpha)$ is a (local) upper bound on f for all $\alpha \in (0, 1/L)$.



Minimizing the majorization model $Q(\cdot; x_k, 1/L) + g$:

$$\begin{aligned} x_{k+1} &\in \arg\min_x \{Q(x; x_k, 1/L) + g(x)\} \\ &= \arg\min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 + g(x) \right\} \\ &= \arg\min_x \left\{ g(x) + \frac{L}{2} \left\| x - x_k + \frac{\nabla f(x_k)}{L} \right\|^2 \right\} = \mathrm{prox}_{\alpha g}\left(x_k - \alpha \nabla f(x_k)\right) \end{aligned}$$

where $\alpha \equiv 1/L$. This is the **proximal gradient step** for $f + g$ at x_k with stepsize α.

Proximal Gradient Method (PGM)

Algorithm: basic PGM

1: Choose $x^0 \in \operatorname{dom} g$, stepsize $\alpha > 0$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad x_{k+1} \in \operatorname{prox}_{\alpha g}\big(x_k - \alpha \nabla f(x_k)\big)$
4: **end for**

Interpretation:

- Half-step: *gradient* on f
- Half-step: *proximal* on g
- Operator splitting: forward-backward, explicit-implicit

Fixed-point iteration.
Characterization for **convex** g:

$$x^\star \text{ stationary} \iff 0 \in \partial g(x^\star) + \nabla f(x^\star)$$
$$\iff x^\star = \operatorname{prox}_{\alpha g}(x^\star - \alpha \nabla f(x^\star))$$

Projected Gradient as a Special Case

Let $g = \delta_C$ (indicator of closed set $C \subseteq \mathbb{R}^n$):

$$\delta_C(x) = \begin{cases} 0 & x \in C \\ +\infty & x \notin C \end{cases}$$

Then:

$$\text{prox}_{\alpha\,\delta_C}(z) = \arg\min_{u \in C} \|u - z\|^2 = \text{proj}_C(z)$$

Projected Gradient Descent (PGD)

$$x_{k+1} \in \text{proj}_C\big(x_k - \alpha_k \nabla f(x_k)\big)$$

Examples: box constraints (ℓ_∞ ball), simplex projection, nuclear norm ball, PSD cone, complementarity constraints...

Block Coordinate Proximal Gradient

When $x = (x_1, \ldots, x_B)$ is partitioned into blocks:

$$\varphi(x) := f(x) + \sum_{b=1}^{B} g_b(x_b)$$

Block Coordinate Descent (BCD-PGM)

1: **for** $k = 0, 1, \ldots$ **do**
2: Pick block b_k
3: $(x_{b_k})_{k+1} = \mathrm{prox}_{\alpha g_{b_k}} \left((x_{b_k})_k - \alpha \nabla_{b_k} f(x_k) \right)$
4: **end for**

- Cheap per-iteration cost (partial gradient)
- Cyclic or random selection of blocks b_k for convergence to stationary points
- Suited for matrix factorization, dictionary learning, sparse NMF

Convergence theory

Global Lipschitz Gradient Assumption

Assumption (L)

∇f is L-Lipschitz continuous: there exists a finite L such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

- Descent lemma: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ for all x, y.

$$f(x_+) \leq f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{L}{2} \|x_+ - x\|^2$$

- Prox-grad step: $x_+ \in \mathrm{prox}_{\alpha g}(x - \alpha \nabla f(x)) = \arg\min_x \left\{ g(z) + \frac{1}{2\alpha} \|z - x\|^2 \right\}$.

$$g(x_+) + \frac{1}{2\alpha} \|x_+ - x + \alpha \nabla f(x)\|^2 \leq g(x) + \frac{\alpha}{2} \|\nabla f(x)\|^2$$

$$\implies \quad \varphi(x_+) \leq \varphi(x) + \frac{\alpha L - 1}{2\alpha} \|x_+ - x\|^2$$

Global Lipschitz Gradient Assumption II

Assumption (L)

∇f is L-Lipschitz continuous: there exists a finite L such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

$$\varphi(x_+) \leq \varphi(x) + \frac{\alpha L - 1}{2\alpha} \|x_+ - x\|^2$$

With $\alpha = \sigma/L$, $\sigma \in (0, 1)$, there is **sufficient decrease**:

$$x_{k+1} \in \mathrm{prox}_{\alpha g}(x_k - \alpha \nabla f(x_k)) \quad \Longrightarrow \quad \varphi(x_{k+1}) \leq \varphi(x_k) - \frac{1 - \sigma}{2\alpha} \|x_{k+1} - x_k\|^2$$

Issue: L often unknown or overly conservative $\Rightarrow$ use backtracking to find good α.

Gradient Mapping

Gradient mapping

Generalized gradient for $f + g$:

$$G_\alpha(x) := \frac{x - \operatorname{prox}_{\alpha g}(x - \alpha \nabla f(x))}{\alpha} = \frac{x - x_+}{\alpha}$$

Why it matters:

- Reduces to $\nabla f(x)$ when $g = 0$ (unconstrained, smooth).
- **Stationarity:** $G_\alpha(x^\star) = 0 \iff 0 \in \nabla f(x^\star) + \partial g(x^\star)$ for **convex** g.
- $\|G_\alpha(x)\|$ is a natural *measure of stationarity* at x.

$$\varphi(x_+) \leq \varphi(x) - \alpha \frac{1 - \sigma}{2} \|G_\alpha(x)\|^2 = \varphi(x) - \frac{1 - \sigma}{2\alpha} \|x_+ - x\|^2$$

PGM step: $x_{k+1} = x_k - \alpha G_\alpha(x_k)$.

- PGM as GD with *gradient mapping*.
- Convergence requires $\left\|G_\alpha(x^k)\right\| \to 0$.

Sufficient Decrease

Lemma (Sufficient Decrease).

Under Assumption (L), with $\alpha \in (0, 1/L)$ and some $\sigma \in (0, 1)$:

$$\varphi(x_{k+1}) \leq \varphi(x_k) - \frac{\sigma\alpha}{2} \|G_\alpha(x_k)\|^2 \leq \varphi(x_k)$$

If $\varphi := f + g$ is bounded from below by $\varphi_\star$, then starting with $x_0 \in \operatorname{dom} \varphi$ and telescoping

$$\infty > \varphi(x_0) - \varphi_\star \geq \sum_{k=0}^{K} \varphi(x_k) - \varphi(x_{k+1}) \geq \frac{\sigma\alpha}{2} \sum_{k=0}^{K} \|G_\alpha(x_k)\|^2$$

meaning that $\|G_\alpha(x_k)\| \to 0$.

Backtracking Line Search

When L is unknown or large, use **adaptive stepsize**, typically based on **backtracking**:

Backtracking: one PGM iteration

Input: x_k, $\bar{\alpha} > 0$, $\beta \in (0,1)$
$\alpha_k \leftarrow \bar{\alpha}$
$x_{k+1} \leftarrow \mathrm{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))$
while $f(x_{k+1}) > Q(x_{k+1}; x_k, \alpha_k)$ **do**
    $\alpha_k \leftarrow \beta \alpha_k$
    $x_{k+1} \leftarrow \mathrm{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))$
end while

where $Q(y; x, \alpha) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\alpha} \|y - x\|^2$ is a quadratic model of f around x.

- Adaptive to local curvature
- Guarantees sufficient decrease at each iteration
- Terminates in finite steps (since $\alpha \leq 1/L$ satisfies the condition)
- ... even for locally Lipschitz ∇f, or less [DMT22, KM22]

Convergence: Convex Case

Additional Assumption (C):

f and g are convex; $\varphi_\star = \min \varphi > -\infty$.

Theorem (Sublinear Rate, $\mathcal{O}(1/k)$).

Under (L) + (C), with fixed $\alpha = 1/L$:

$$\varphi(x_k) - \varphi_\star \leq \frac{L \|x_0 - x_\star\|^2}{2k}$$

Key steps in proof:

- Sufficient decrease + convexity of $g \implies$ one-step progress inequality.
- Telescope the sufficient-decrease inequality over k steps.
- Use convexity of φ to bound $\varphi(x_k)$ by $\frac{1}{k}\sum_j \varphi(x_j)$.

Linear Convergence Under Strong Convexity

Assumption (SC):

f is μ-strongly convex ($\mu > 0$).

Theorem (Linear Rate).

Under (L) + (SC), with $\alpha = 1/L$:

$$\|x_k - x_\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x_\star\|^2$$

Condition number $\kappa = L/\mu$ governs the rate.

Derivation: The key contraction is

$$\|\mathcal{T}_\alpha(x) - \mathcal{T}_\alpha(y)\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|x - y\|^2$$

where $\mathcal{T}_\alpha(x) := \mathrm{prox}_{\alpha g}(x - \alpha \nabla f(x))$.

Convergence: Nonconvex Case

Without convexity, we cannot expect $\varphi(x_k) \to \varphi_\star$. Instead:

Theorem (Stationarity in $\mathcal{O}(K)$).

Under Assumption (L), backtracking PGM produces:

$$\min_{0 \le k \le K} \|G_\alpha(x_k)\|^2 \le \frac{2}{\alpha_{\min} K}(\varphi(x_0) - \varphi_\star)$$

Hence $G_\alpha(x_k) \to 0$ along a subsequence.

Proof sketch:

- Sufficient decrease: $\varphi(x_{k+1}) \le \varphi(x_k) - \frac{\alpha_k}{2} \|G_{\alpha_k}(x_k)\|^2$.
- Telescope: $\sum_{k=0}^{K-1} \frac{\alpha_k}{2} \|G_{\alpha_k}(x_k)\|^2 \le \varphi(x_0) - \varphi_\star < \infty$.
- $\alpha_k \ge \alpha_{\min} := \beta/L > 0$.

If ∇f not globally Lipschitz, $\alpha_{\min}$ depends on geometry around accumulation points...

Kurdyka–Łojasiewicz (KL) Framework

For finer convergence results in the nonconvex case, we use:

KL Property.

Function φ satisfies KL at $x_\star$ if there exists a desingularizing function ψ such that

$$\psi'(\varphi(x) - \varphi(x_\star))\,\|\partial\varphi(x)\| \geq 1$$

holds for all x near $x_\star$.

KL covers:

- Real-analytic, semi-algebraic functions (polynomials, ℓ_1-regularized, ...)
- Piecewise smooth functions
- Practically all functions encountered in ML/stats

Convergence with KL: whole sequence

Under KL property and sufficient decrease, PGM sequence **converges** to a critical point with rate determined by ψ (linear/sublinear depending on KL exponent).

Accelerated Gradient Methods: Nesterov's Idea

Can we do better than $\mathcal{O}(1/k)$ for convex problems?

Nesterov (1983): *Yes!* Using **momentum**.

Key idea — extrapolation: instead of applying prox-grad at x_k, apply it at a *momentum point / looking ahead*:

$$y_k = x_k + \tau_k(x_k - x_{k-1})$$

where $\{\tau_k\}$ is a carefully chosen sequence.

Lower bound (Nesterov, 1983)

No first-order method can achieve better than $\mathcal{O}(1/k^2)$ in general.

FISTA's revolution: Fast ISTA / PGM

Algorithm: FISTA — Beck & Teboulle [BT09]

$x_0 = y_1 \in \operatorname{dom} g$, $t_1 = 1$, $\alpha = 1/L$
for $k = 1, 2, \ldots$ **do**
  $x_k = \operatorname{prox}_{\alpha g}\big(y_k - \alpha \nabla f(y_k)\big)$
  $t_{k+1} = \frac{1}{2}\big(1 + \sqrt{1 + 4t_k^2}\big)$
  $y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$
end for

Theorem.

Under (L)+(C), FISTA achieves $\mathcal{O}(1/k^2)$:

$$\varphi(x_k) - \varphi_\star \leq \frac{2L\,\|x_0 - x_\star\|^2}{(k+1)^2}.$$

Momentum factor $\tau_k \equiv \frac{t_k - 1}{t_{k+1}} \to 1^-$ asymptotically, $y_k \approx 2x_k - x_{k-1}$.

- Momentum can hurt near stationary points $\rightsquigarrow$ restarting (empirically much faster, especially for ill-conditioned problems)
- Monotone variants to ensure $\varphi(x^k)$ decreasing (slightly slower in practice)

Acceleration in the Nonconvex Setting

Warning

Nesterov momentum does *not* guarantee acceleration for nonconvex φ.

iPiano [OCBP14]: inertial PGM for nonconvex f and convex g:

$$x_{k+1} = \operatorname{prox}_{\alpha_k g}\big(x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})\big)$$

Convergence under KL property + sufficient decrease condition on (α_k, β_k).

APGM-NC [LL15]: Nesterov acceleration with monitoring, for nonconvex f and nonconvex g. Achieves global convergence to stationary points in general, and complexity $\mathcal{O}(1/k^2)$ for convex problems.

Practical aspects

Termination Conditions

When do we stop? We need a computable, meaningful criterion.

Common criterion:

gradient mapping residual

$$\left\| G_\alpha(x^k) \right\| = \tfrac{1}{\alpha} \left\| x^k - x^{k+1} \right\| \le \varepsilon$$

Additional/complementary conditions:

- Iterate change: $\left\| x^k - x^{k-1} \right\| \le \varepsilon \max(1, \left\| x^k \right\|)$
- Objective change: $|\varphi(x^k) - \varphi(x^{k-1})| \le \varepsilon \max(1, |\varphi(x^k)|)$
- Dual residual (g convex): $\left\| x^k - \operatorname{prox}_g(x^k - \nabla f(x^k)) \right\| \le \varepsilon$
- Max iterations: $k \ge k_{\max}$

Practice: combine stationarity measure with iterate/objective change for robustness.

Termination Conditions (to be precise)

When do we stop? We need a computable **measure of optimality/stationarity**.

From the PGM update we have

$$0 \in \partial g(x_{k+1}) + \frac{x_{k+1} - x_k}{\alpha_k} + \nabla f(x_k)$$

$$\iff \quad \nabla f(x_{k+1}) - \nabla f(x_k) - \frac{x_{k+1} - x_k}{\alpha_k} \in \partial g(x_{k+1}) + \nabla f(x_{k+1}) = \partial \varphi(x_{k+1})$$

Stationarity criterion:

$$\operatorname{dist}(0, \partial \varphi(x_{k+1})) \overset{\text{PGM}}{\leq} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) - \frac{x_{k+1} - x_k}{\alpha_k} \right\| \leq \varepsilon$$

Relaxing the Global Lipschitz Assumption

Global Lipschitz is often too strong:

- Unknown L, or L is very large (loose bound)
- Barrier / Nonconvex f: gradient may not be globally Lipschitz

Relaxations:

1. **Local Lipschitz + backtracking:** backtracking finds local L_k, guaranteed to terminate.
2. **Local descent lemma:** only require the backtracking condition (no global L).
3. **Proximal smoothness:** f prox-regular; weaker than convexity.
4. **Relative smoothness** [BBT17]: replace $\frac{1}{2}\|y-x\|^2$ by $D_h(y,x)$ (Bregman divergence); leads to *mirror descent / Bregman proximal gradient*.

Backtracking regularization

PGM with backtracking:

- prox-grad step: compute $x_+ \in \operatorname{prox}_{\alpha g}(x - \alpha \nabla f(x))$
- sufficient decrease: check if $\varphi(x_+) \leq \varphi(x_k) - \frac{\sigma}{\alpha} \|x_+ - x\|^2$
- backtracking: update α or break

For **nonconvex** f, backtracking is the standard practical approach. It is an adaptive (quadratic) regularization, not a linesearch procedure.

Even for **convex** f and g, adaptive schemes without checking descent [MM20, MM24, LTSP25]

Nonmonotone Descent

Observation: Forcing $\varphi(x_{k+1}) \leq \varphi(x_k)$ every step can be overly restrictive

Key idea: replace *monotone* sufficient decrease condition with

$$\varphi(x_{k+1}) \leq \Phi_k - \frac{\sigma}{\alpha_k} \|x_{k+1} - x_k\|^2$$

where $\Phi_k \geq \varphi(x_k)$ and $\{\Phi_k\}$ ensures convergence.

Max-type — Grippo, Lampariello & Lucidi [GLL86]

Given $M \in \mathbb{N}$ and $m(k) := \min\{M, k\}$,

$$\Phi_k := \max_{0 \leq j \leq m(k)} \varphi(x_{k-j})$$

Mean-type — Zhang & Hager [ZH04], [DM23]

Given $\eta \in (0, 1]$,

$$\Phi_k := (1-\eta)\Phi_{k-1} + \eta\varphi(x_k), \quad \Phi_0 := \varphi(x_0)$$

Benefits:

- Accepts larger steps; avoids zigzagging
- Bounded nonmonotonicity; specializes to monotone
- Convergence guaranteed, possibly under additional assumptions (max...) [KM22, DM23]

Spectral (Barzilai–Borwein) Stepsizes

Idea: estimate the local Lipschitz constant from secant information.

Let $s_k := x_k - x_{k-1}$, $y_k := \nabla f(x_k) - \nabla f(x_{k-1})$. Mimic a diagonal quasi-Newton step:

BB Stepsizes — Barzilai & Borwein [BB88]

$$\alpha_k^{\mathrm{BB1}} = \frac{\langle s^k, s^k \rangle}{\langle s^k, y^k \rangle} \quad \text{(short step)}, \qquad \alpha_k^{\mathrm{BB2}} = \frac{\langle s^k, y^k \rangle}{\langle y^k, y^k \rangle} \quad \text{(long step)}$$

- Approximate the inverse of the average curvature
- In practice, must pair with **nonmonotone** globalization
- Dramatic **speedup** in practice (close to quasi-Newton)

Spectral Proximal Gradient Method

Algorithm: SPG [BMR00, JKMW23] / SpaRSA [WNF09]

1: x_0, $\alpha_0^{\mathrm{BB}} > 0$, $0 < \alpha_{\min} < \alpha_{\max}$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Starting from α_k^{BB}, by backtracking find α_k that delivers suff. decrease
4: $x_{k+1} \in \mathrm{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))$
5: Compute $s_{k+1} := x_{k+1} - x_k$, $y_{k+1} := \nabla f(x_{k+1}) - \nabla f(x_k)$
6: $\alpha_{k+1}^{\mathrm{BB}} \leftarrow \mathrm{clip}(\alpha^{\mathrm{BB}}(s_{k+1}, y_{k+1}), \alpha_{\min}, \alpha_{\max})$
7: **end for**

In practice, often faster than FISTA

Toward (quasi-)Newton methods

Limitations of First-Order Proximal Gradient

First-order methods: only use ∇f (gradient information).

Issues:

- Rate $\mathcal{O}(1/k)$ (or $\mathcal{O}(1/k^2)$ with momentum) — slow for ill-conditioned problems
- Strong convexity gives linear rate $\mathcal{O}((1-\mu/L)^k)$ but $\mu/L \ll 1$ in practice

Newton-type methods: use curvature (Hessian or approximation).

- Replace $\frac{1}{2\alpha}\|x - x_k\|^2$ by $\frac{1}{2}(x-x_k)^\top H_k(x-x_k)$ in the model of f
- *Variable metric prox-gradient:*

$$x_{k+1} \in \operatorname{prox}_g^{H_k}(x_k - H_k^{-1}\nabla f(x_k))$$

- Challenge: $\operatorname{prox}^{H_k}$ may not be easy to compute for arbitrary H_k

Proximal Newton-type methods for convex structured minimization [PB13, LSS14]

Variable Metric Proximal Operator

Definition.

For SPD matrix H, the H-proximal operator is:

$$\operatorname{prox}_g^H(z) := \arg\min_u \left\{ g(u) + \tfrac{1}{2}(u-z)^\top H(u-z) \right\}$$

Key issue: for general H, the above may have no closed form even when prox_g does.

Exceptions (tractable prox_g^H):

- $g = \lambda \|\cdot\|_1$ and *diagonal* $H \implies$ scaled soft-threshold
- $g = \delta_C$ and diagonal $H \implies$ scaled projection
- g separable and H diagonal $\implies$ coordinate-wise prox

Full (quasi-)Newton requires more sophisticated approaches to be practical.

PANOC — Stella et al. [STSP17, TSP18]

Idea: combine fast directions with safe PGM steps.

Look at the **inclusion** problem

$$\text{find } x \text{ such that } 0 \in \mathcal{T}_\alpha(x) := \frac{1}{\alpha}[x - \text{prox}_{\alpha g}(x - \alpha \nabla f(x))].$$

We seek a zero of the (set-valued) fixed-point residual $\mathcal{T}_\alpha$.

Newton-type scheme for this root-finding problem:

$$x_{k+1} = x_k - H_k \mathcal{T}_\alpha(x_k)$$

where H_k capture the curvature of $\mathcal{T}_\alpha$ around x_k.

$\rightsquigarrow$ H_k from **quasi-Newton** formulas

PANOC — Globalization

But, what about globalization?

Key observation.

The Moreau envelope is continuous.

PANOC algorithm:

1. Compute $x_{k+\frac{1}{2}} = \mathrm{prox}_{\alpha g}(x_k - \alpha \nabla f(x_k))$ (PGM step as anchor)
2. Build quasi-Newton direction on $\varphi_\alpha^{\mathrm{FB}}$ using L-BFGS
3. Linesearch to find x_{k+1} satisfying $\varphi_\alpha^{\mathrm{FB}}(x_{k+1}) \leq \varphi_\alpha^{\mathrm{FB}}(x_k) - \frac{\sigma}{\alpha}\|x_{k+1} - x_k\|^2$

Forward-Backward Envelope (FBE)

Definition [PB13].

The *forward-backward envelope* of $\varphi := f + g$ is:

$$\varphi_\alpha^{\mathrm{FB}}(x) := \min_u \left\{ f(x) + \langle \nabla f(x), u - x \rangle + \tfrac{1}{2\alpha} \|u - x\|^2 + g(u) \right\}$$

Properties (for suff. small $\alpha > 0$):

- $\varphi_\alpha^{\mathrm{FB}}(x) \leq \varphi(x)$ for all x
- $\varphi_\alpha^{\mathrm{FB}}(x) = \varphi(x) \iff x$ is a fixed point / critical / stationary
- for convex g, gradient mapping $G_\alpha(x) = \alpha \nabla \varphi_\alpha^{\mathrm{FB}}(x)$

PANOC minimizes $\varphi_\alpha^{\mathrm{FB}}$ by combining fast smooth quasi-Newton extrapolation steps with safe prox-grad updates.

PANOC: Properties and Convergence [STSP17, TSP18]

Structure:

- One prox evaluation per iteration
- One gradient evaluation per iteration
- L-BFGS memory update (low cost)
- Falls back to PGM if quasi-Newton step rejected

Convergence:

- $\left\|G_\alpha(x^k)\right\| \to 0$ (nonconvex)
- Superlinear rate near non-degenerate solutions
- Formally: R-superlinear under second-order sufficient conditions

PANOC$^+$: extension to locally Lipschitz ∇f, also with inexact prox evaluations [DMT22]

Applications: embedded optimal control, MPC, robotics. Subsolver in `OpEn` [SFP20], `alpaqa` [PSP22], `Bazinga` [DMJKM23].

Regularized Proximal Newton-type Method

Since [PB13, LSS14], form the **second-order structured model** at iterate x_k:

$$m_k(x) \coloneqq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}(x - x_k)^\top B_k (x - x_k) + g(x)$$

where $B_k \approx \nabla^2 f(x_k)$ (exact Hessian or approximation) with regularization.

Proximal Newton-type subproblem

Solve

$$x_{k+1} \in \arg\min_x m_k(x)$$

Newton-type subproblem hardly available in closed-form, but a structured minimization: call PGM to solve it!

Ideas for practical proximal Newton-type methods: RPQN

RPQN [BFO19, Lec22]

Limited-memory quasi-Newton B_k with compact representation
$\implies$ small (memory M) system of equations to evaluate the scaled prox operator.

Nocedal's "Updating quasi-Newton matrices with limited storage" (1980)
Becker, Fadili & Ochs's "On quasi-Newton forward-backward splitting: Proximal calculus and convergence" (2019)
Kanzow & Lechner's "Efficient regularized proximal quasi-Newton methods for large-scale nonconvex composite optimization problems" (2024)

Ideas for practical proximal Newton-type methods: R2N

R2N [ABO22, DHO26]

- Solve the Newton-type proximal subproblem inexactly,
- ... obtaining at least the improvement of a prox-grad step (Cauchy decrease).
- Check if x_{k+1} delivers sufficient decrease for φ.
- If needed, recompute with an increased quadratic regularization.

Diouane, Habiboullah & Orban's "A proximal modified quasi-Newton method for nonsmooth regularized optimization" (2026)

Comparing the Methods: A Unified View

All methods minimize a *model* $m_k(x) \approx f(x) + g(x)$:

Method	**Model** $m_k(x)$	**Per-iter cost**
PGM	$\nabla f^\top d + \frac{1}{2\alpha}\|d\|^2 + g$	1 grad, 1 prox
PGM + BB	same, $\alpha_k = \alpha^{\text{BB}}$, nonmono	1 grad, 1 prox
FISTA	same + momentum	1 grad, 1 prox
PANOC	envelope $\varphi_\alpha^{\text{FB}}$ + L-BFGS	1 grad, 1 prox, QN update
R2N	$\nabla f^\top d + \frac{1}{2} d^\top B_k d + g$	inexact inner PGM solve
RPQN	same with compact represent.	1 grad, 1 inner root-finder

Trade-off: richer model $\implies$ fewer outer iterations but more expensive subproblem.
The *right* method depends on the problem structure, cost of g, Hessian availability.

In practice, choose solver depending on most/least expensive oracles at hand.
Some comparisons in [DHO26].

Numerical examples

Sparse linear regression with ProximalAlgorithms.jl / RegularizedOptimization.jl

$$\underset{x}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|Ax - b\|^2}_{=:f(x)} + \lambda R(x)$$

$$\nabla f(x) = A^\top (Ax - b)$$
$$\nabla f(x) - \nabla f(y) = A^\top A(x - y)$$

[link to tutorial: PG]

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.


Jonathan Bartzilai and Jonathan M. Borwein.
Two-point step size gradient methods.
IMA Journal of Numerical Analysis, 8(1):141–148, 1988.


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.


Stephen Becker, Jalal Fadili, and Peter Ochs.
On quasi-Newton forward-backward splitting: Proximal calculus and convergence.
SIAM Journal on Optimization, 29:2445–2481, 2019.


Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan.
Nonmonotone spectral projected gradient methods on convex sets.
SIAM Journal on Optimization, 10(4):1196–1211, 2000.


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.

References II


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.


L. Grippo, F. Lampariello, and S. Lucidi.
A nonmonotone line search technique for Newton's method.
SIAM Journal on Numerical Analysis, 23(4):707–716, 1986.


Xiaoxi Jia, Christian Kanzow, Patrick Mehlitz, and Gerd Wachsmuth.
An augmented Lagrangian method for optimization problems with structured geometric constraints.
Mathematical Programming, 199(1):1365–1415, 2023.

References III


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.


Theresa Lechner.
Proximal Methods for Nonconvex Composite Optimization Problems.
PhD thesis, 2022.


Huan Li and Zhouchen Lin.
Accelerated proximal gradient methods for nonconvex programming.
In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.


Jason D. Lee, Yuekai Sun, and Michael A. Saunders.
Proximal Newton-type methods for minimizing composite functions.
SIAM Journal on Optimization, 24(3):1420–1443, 2014.


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.

References IV


Yura Malitsky and Konstantin Mishchenko.
Adaptive gradient descent without descent.
In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6702–6712. PMLR, 13–18 Jul 2020.


Yura Malitsky and Konstantin Mishchenko.
Adaptive proximal gradient method for convex optimization, 2024.


Jorge Nocedal.
Updating quasi-Newton matrices with limited storage.
Mathematics of Computation, 35(151):773–782, 1980.


Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock.
iPiano: Inertial proximal algorithm for nonconvex optimization.
SIAM Journal on Imaging Sciences, 7(2):1388–1419, 2014.


Panagiotis Patrinos and Alberto Bemporad.
Proximal Newton methods for convex composite optimization.
In *52nd IEEE Conference on Decision and Control*, pages 2358–2363, 2013.

References V


Pieter Pas, Mathijs Schuurmans, and Panagiotis Patrinos.
Alpaqa: A matrix-free solver for nonlinear MPC and large-scale nonconvex optimization.
In *2022 European Control Conference (ECC)*, pages 417–422, 2022.


Pantelis Sopasakis, Emil Fresk, and Panagiotis Patrinos.
OpEn: Code generation for embedded nonconvex optimization.
IFAC-PapersOnLine, 53(2):6548–6554, 2020.
21st IFAC World Congress.


Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos.
A simple and efficient algorithm for nonlinear model predictive control.
In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, 2017.


Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos.
Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms.
SIAM Journal on Optimization, 28(3):2274–2303, 2018.


Stephen J. Wright, Robert D. Nowak, and Mario A. T. Figueiredo.
Sparse reconstruction by separable approximation.
IEEE Transactions on Signal Processing, 57(7):2479–2493, 2009.

References VI


Hongchao Zhang and William W. Hager.
A nonmonotone line search technique and its application to unconstrained optimization.
SIAM Journal on Optimization, 14(4):1043–1056, 2004.