Combining Newton and proximal-gradient for nonsmooth optimization

Gilles Bareilles

LJK, Univ. Grenoble Alpes

CANUM
13 - 17 juin 2022

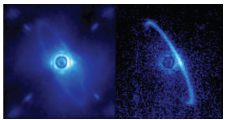
Context: linear inverse problems...

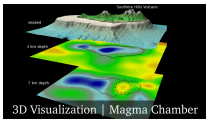
x_0 is the signal of interest, only accessible through measures y, via a (linear) model A and noise ξ:

$$y = Ax_0 + \xi$$

Examples

- Image processing diffraction of objective, low-res./damaged sensor
- Medical imaging computerized tomography, MRI, EEG
- Seismic imaging wave propagation
- Machine learning / Statistics regression





Context: linear inverse problems... are often ill posed

$$y = Ax_0 + \xi$$

Problem: not enough observations y to recover x_0 or y inconsistent.

Observation: we often know the solution to be sparse, piecewise constant, low-rank (matrix); we have ***prior knowledge*** on the ***structure*** of x_0.

How to leverage it?

Context: linear inverse problems... are often ill posed

$$y = Ax_0 + \xi$$

Problem: not enough observations y to recover x_0 or y inconsistent.

Observation: we often know the solution to be sparse, piecewise constant, low-rank (matrix); we have ***prior knowledge*** on the ***structure*** of x_0.

How to leverage it? Minimize the regularized empirical risk

$$x^\star \in \arg\min_x \underbrace{\sum_{i=1}^{n} \ell(A_i x, y_i)}_{\text{smooth empirical risk}} + \underbrace{\lambda r(x)}_{\text{nonsmooth regularizer}}$$

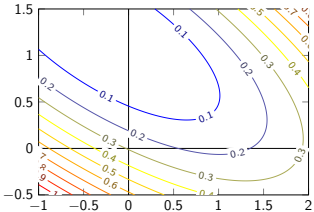
where

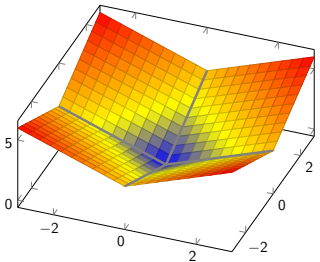
- ℓ is built from assumptions on noise ξ
- r is built to encode prior knowledge on structure of x_0

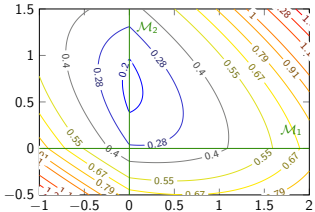
◇ Vaiter, S., Peyré, G., Fadili, J.: Low Complexity Regularization of Linear Inverse Problems (2015)

Nonsmooth regularizers and structure

▷ if x_0 has many zero entries, use $\|x\|_1 = \sum_{i=1}^n |x_i|$:



$\sum_{i=1}^n (A_i x - y_i)^2$ +  $\lambda\|x\|_1$ →



$\sum_{i=1}^n (A_i x - y_i)^2 + \lambda\|x\|_1$

LASSO

▷ if the matrix X_0 is low-rank, use $\|X\|_* = \sum_{i=1}^{\mathrm{rank}(X)} \sigma_i(X)$

Observation: there are smooth subspaces $\mathcal{M}$ such that r is ***smooth along*** and ***nonsmooth across***. These are ***structure manifolds***.

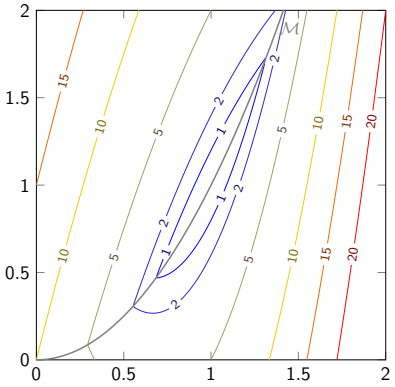
Composite problem, structure of minimizers

$$\text{Find } x^\star \in \arg\min_{x\in\mathbb{R}^n} F(x) \triangleq \underbrace{f(x)}_{\text{smooth}} + \underbrace{g(x)}_{\text{non smooth}}$$

Finding a minimizer of F nonsmooth amounts to:

- finding the right structure manifold $\mathcal{M}^\star \ni x^\star$
  e.g. the right sparsity pattern for ℓ_1
- minimizing F on that correct structure
  smooth problem on a smooth manifold

But, we never know when a manifold is the optimal one.



Composite problem, structure of minimizers

$$\text{Find } x^\star \in \arg\min_{x\in\mathbb{R}^n} F(x) \triangleq \underbrace{f(x)}_{\text{smooth}} + \underbrace{g(x)}_{\text{non smooth}}$$

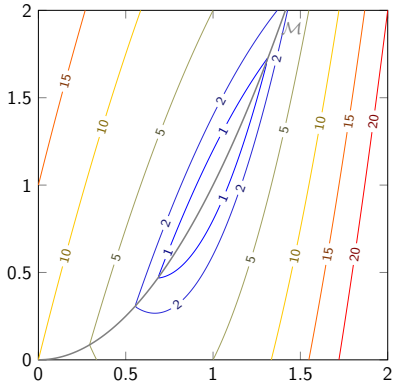
Finding a minimizer of F nonsmooth amounts to:

- finding the right structure manifold $\mathcal{M}^\star \ni x^\star$
  e.g. the right sparsity pattern for ℓ_1
- minimizing F on that correct structure
  smooth problem on a smooth manifold

But, we never know when a manifold is the optimal one.

Plan:

1. look at proximal gradient
2. look at Riemannian optimization
3. combine them



Proximal gradient: an optimization tool

▷ One proximal gradient step:

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

where

$$\mathbf{prox}_{\gamma g}(y) \triangleq \arg\min_u \left\{ g(u) + \frac{1}{2\gamma}\|u - y\|^2 \right\}.$$

▷ Forward-backward interpretation:

$$\begin{array}{lll} \text{explicit step on } f & & y = x - \gamma \nabla f(x) \\ \text{'implicit' step on } g & x_+ = \mathbf{prox}_{\gamma g}(y) \Leftrightarrow & x_+ = y - \gamma s \qquad \text{with } s \in \partial g(x_+) \end{array}$$

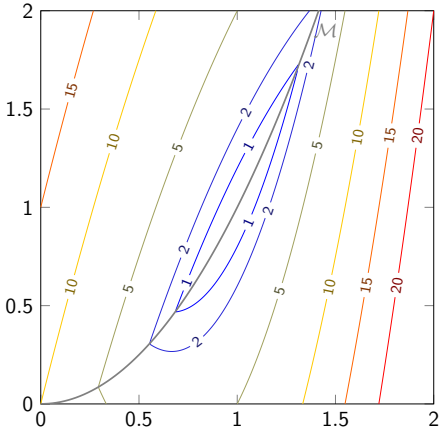
▷ If the step size γ is small enough $< 1/L$, where L is a Lipschitz constant for ∇f

✓ functional descent at every step hence converges to critical points

✗ "slow" convergence: $F(x_k) - F(x^\star) \leq O(1/k)$ in convex case

Proximal gradient: a structure detector

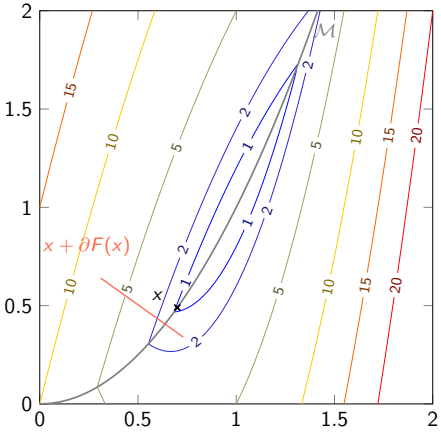
Near manifolds, the proximal gradient sends points to the manifold, smoothly:



$$g(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Proximal gradient: a structure detector

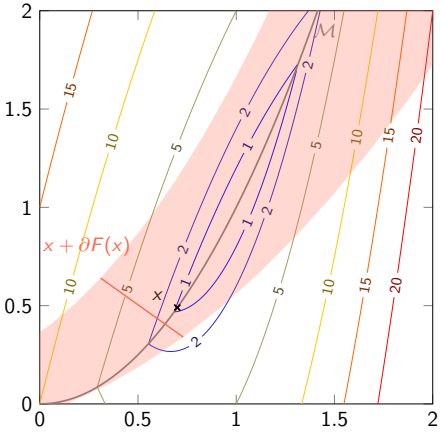
Near manifolds, the proximal gradient sends points to the manifold, smoothly:



$$g(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Proximal gradient: a structure detector

Near manifolds, the proximal gradient sends points to the manifold, smoothly:



$$g(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Proximal gradient: a structure detector

Near manifolds, the proximal gradient sends points to the manifold, smoothly:

Theorem (B., Iutzeler, Malick, '20)

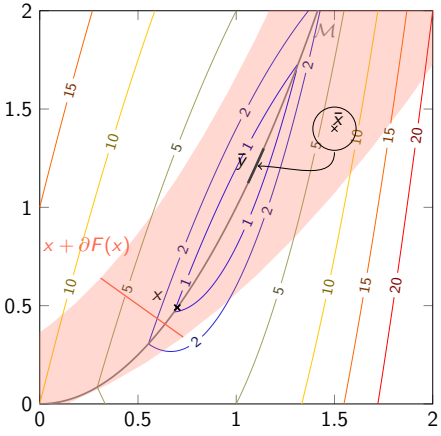
Take $\bar{x}$ *and* $\bar{y} = \mathbf{prox}_{\gamma g}(\bar{x} - \gamma \nabla f(\bar{x}))$ *such that*

- g *has structure* $\mathcal{M}$ *at* $\bar{y}$
- $\bar{x} - \gamma \nabla f(\bar{x}) - \bar{y} \in \mathrm{ri}\ \gamma \partial g(\bar{y})$

then, on a neighborhood $\mathcal{N}_{\bar{x}}$ *of* $\bar{x}$

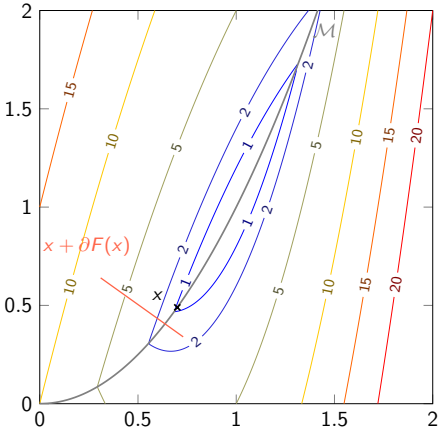
- *the prox-grad operator is* $\mathcal{M}$***-valued*** *and* $\mathcal{C}^1$

$\rightarrow$ How to move on $\mathcal{M}$?



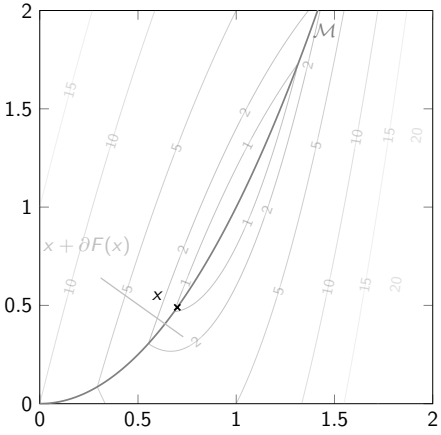
$$g(x) = 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Manifold optimization, in a nutshell



$$\arg\min_x 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

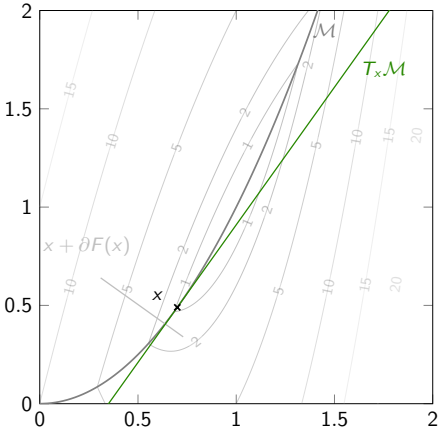
Manifold optimization, in a nutshell



$$\arg\min_x 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Manifold optimization, in a nutshell

Elementary tools:

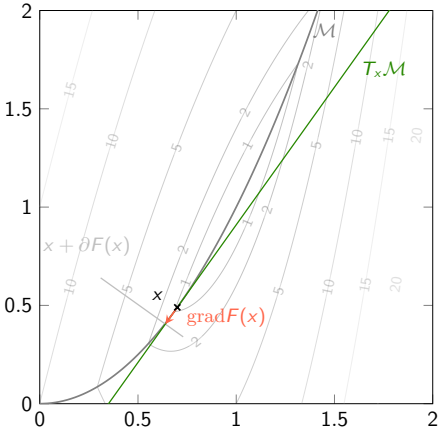


$$\arg\min_x 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Manifold optimization, in a nutshell

Elementary tools:

- Riemannian gradient
- Riemannian Hessian

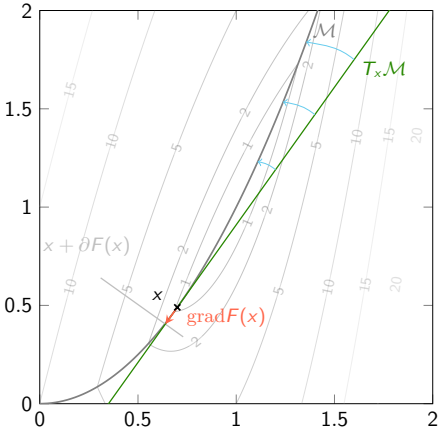


$$\arg\min_x 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Manifold optimization, in a nutshell

Elementary tools:

- Riemannian gradient
- Riemannian Hessian
- retraction: $T_x\mathcal{M} \to \mathcal{M}$



$$\arg\min_x 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Manifold optimization, in a nutshell

Elementary tools:

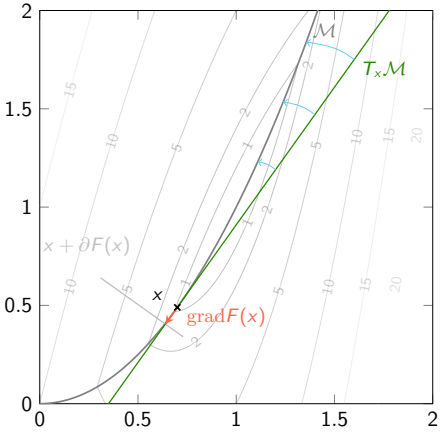
- Riemannian gradient
- Riemannian Hessian
- retraction: $T_x\mathcal{M} \to \mathcal{M}$

Typical Riemannian step "$\mathrm{ManUp}(x, \mathcal{M})$":

- find $d \in T_x\mathcal{M}$
- find a good step size $\alpha > 0$
- send αd to $\mathcal{M}$ with a retraction

$\to$ Many methods of smooth optim carry over manifold optimization.

◇ Boumal, N.: An introduction to optimization on smooth manifolds (2020).



$$\arg\min_x 10(1 - x_1)^2 + 5|x_1^2 - x_2|$$

Proposed algorithm

$$\textbf{prox}_{\gamma g}(y_{k-1} - \gamma \nabla f(y_{k-1})) \text{ gives } x_k \text{ and } \mathcal{M}_k \ni x_k$$
$$y_k = \mathrm{ManUp}(x_k, \mathcal{M}_k)$$

Theorem (B., Iutzeler, Malick, '20)

If $\gamma < 1/L$ and ManUp decreases function value, then any limit point $\bar{x}$ of (x_k) is a critical point: $0 \in \partial F(\bar{x})$.

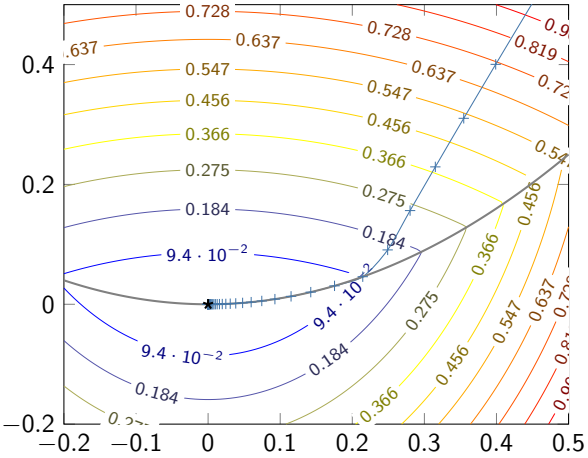
Take ManUp as a Riemannian Newton method and assume for a limit point $x^\star$ that

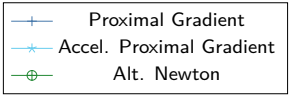
- *g has structure $\mathcal{M}^\star$ at $x^\star$*
- *$0 \in \mathrm{ri}\ \partial F(x^\star)$, $\mathrm{Hess}_{\mathcal{M}^\star} F(x^\star) \succ 0$ and $\mathrm{Hess}_{\mathcal{M}^\star}$ is locally Lipschitz around $x^\star$*

Then, after some finite time

- *$x_k \in \mathcal{M}^\star$*
- *x_k converges to $x^\star$ at a **quadratic rate**:* $\quad \|x_{k+1} - x^\star\| \leq C\|x_k - x^\star\|^2$

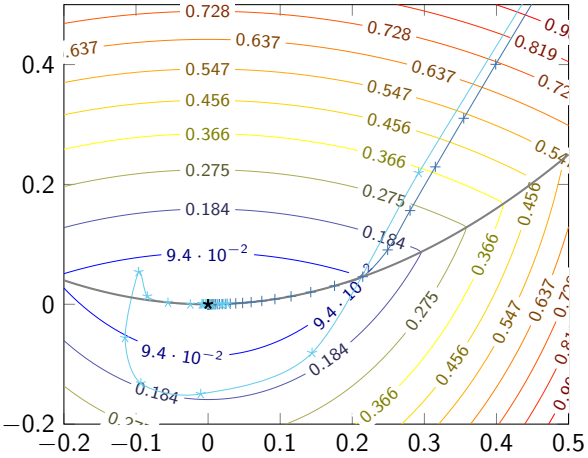
Illustrations

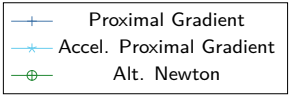




$$\min_{x \in \mathbb{R}^2} \underbrace{2x_1^2 + x_2^2}_{f(x)} + \underbrace{|x_1^2 - x_2|}_{g(x)}$$

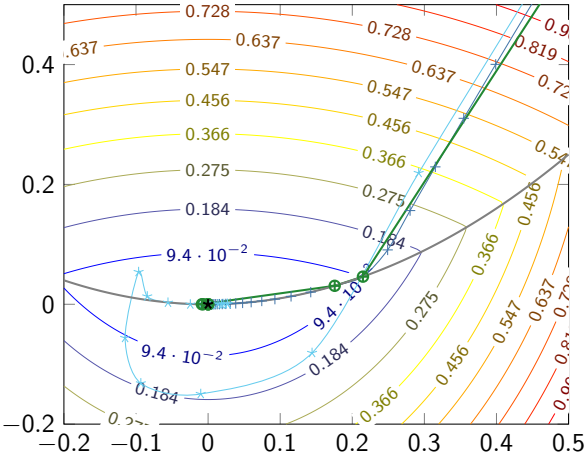
Illustrations

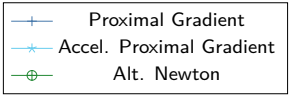




$$\min_{x \in \mathbb{R}^2} \underbrace{2x_1^2 + x_2^2}_{f(x)} + \underbrace{|x_1^2 - x_2|}_{g(x)}$$

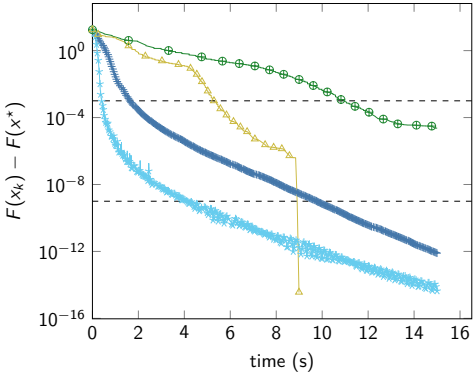
Illustrations

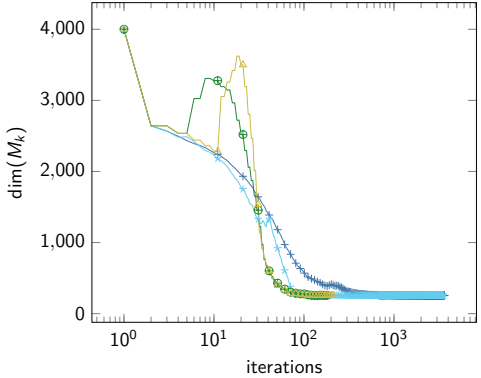


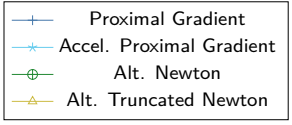


$$\min_{x \in \mathbb{R}^2} \underbrace{2x_1^2 + x_2^2}_{f(x)} + \underbrace{|x_1^2 - x_2|}_{g(x)}$$

Illustrations: logistic regression



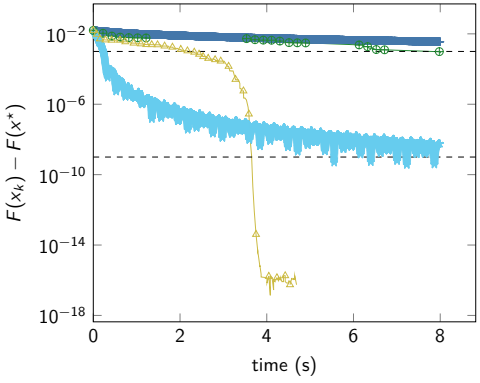


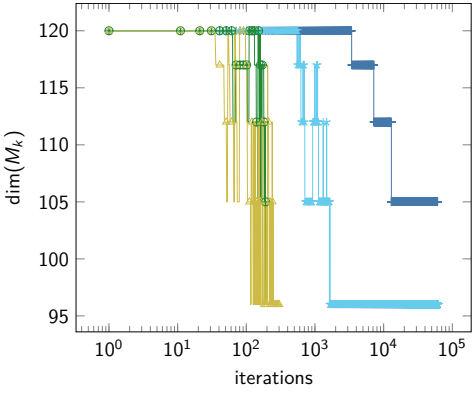


$$\min_{x \in \mathbb{R}^{4000}} \frac{1}{4000} \sum_{i=1}^{4000} \log(y_i \sigma(\langle A_i, x \rangle)) + 10^{-2} \|x\|_1,$$

The optimal manifold has dimension 249.

Illustrations: tracenorm regression





$$\min_{X \in \mathbb{R}^{10\times 12}} \frac{1}{2}\sum_{i=1}^{60} \left(\langle A_i, X\rangle - y_i\right)^2 + 10^{-2}\|X\|_*,$$

The optimal manifold is $\mathrm{rank}(X^\star) = 6$.

Take-away messages

- Proximal methods identify structure in composite additive problems
- We proposed to leverage this structure with Newton-like Riemannian steps

B. & Iutzeler & Malick: *Newton acceleration on manifolds identified by proximal-gradient methods*, Mathematical Programming, 2nd revision.

→ `https://arxiv.org/abs/2012.12936`
→ `https://github.com/GillesBareilles/NewtonRiemannAccel-ProxGrad`

Work in progress and perspectives

- Smarter manifold updates
- Explore composite structures beyond addition $f + g$, *e.g.* $f + g + h(L\cdot)$ or composition $g \circ c$
- Certify that a structure is optimal

Thank you!