Convex Optimization with First Order Schemes

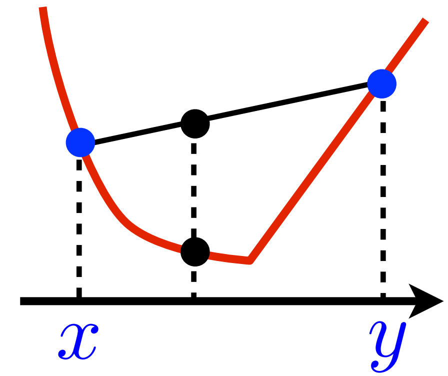
Gabriel Peyré

www.numerical-tours.com

Convex Optimization

Setting: $G : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$

$\mathcal{H}$: Hilbert space. Here: $\mathcal{H} = \mathbb{R}^N$.

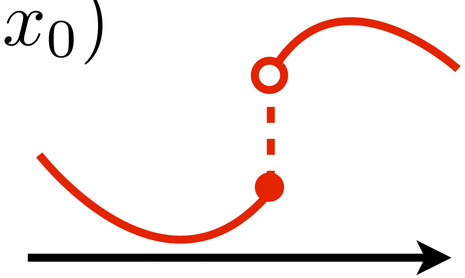


Problem: $\min_{x \in \mathcal{H}} G(x)$

Class of functions:

Convex: $G(tx + (1-t)y) \leqslant tG(x) + (1-t)G(y) \quad t \in [0,1]$

Lower semi-continuous: $\liminf_{x \to x_0} G(x) \geqslant G(x_0)$



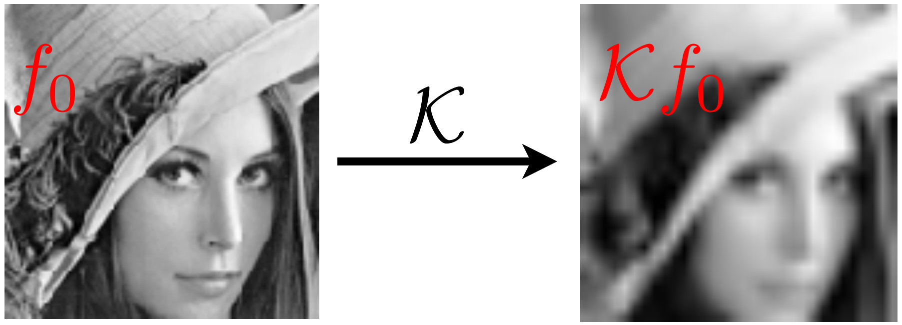
Proper: $\{x \in \mathcal{H} \setminus G(x) \neq +\infty\} \neq \emptyset$

Indicator: $$\iota_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if} \quad x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

($\mathcal{C}$ closed and convex)

Example: ℓ^1 Regularization

Inverse problem: measurements $y = \mathcal{K} f_0 + w$



$\mathcal{K} : \mathbb{R}^N \rightarrow \mathbb{R}^P, \quad P \leqslant N$

Model: $f_0 = \Psi x_0$ sparse in dictionary $\Psi \in \mathbb{R}^{N \times Q}, Q \geqslant N$.

$x \in \mathbb{R}^Q$ coefficients $\xrightarrow{\Psi}$ $f = \Psi x \in \mathbb{R}^N$ image $\xrightarrow{\mathcal{K}}$ $y = \mathcal{K} f \in \mathbb{R}^P$ observations

$\Phi = \mathcal{K} \circ \Psi \in \mathbb{R}^{P \times Q}$

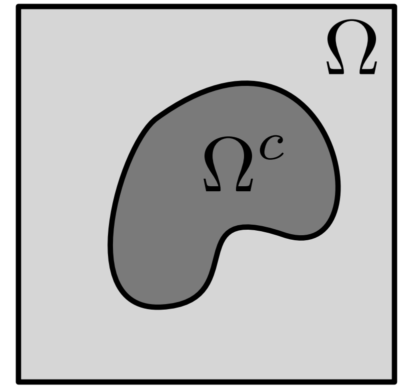
Sparse recovery: $f^\star = \Psi x^\star$ where $x^\star$ solves

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} \|y - \Phi x\|^2 + \lambda \|x\|_1$$

Fidelity — Regularization

Example: ℓ^1 Regularization

Inpainting: masking operator $\mathcal{K}$

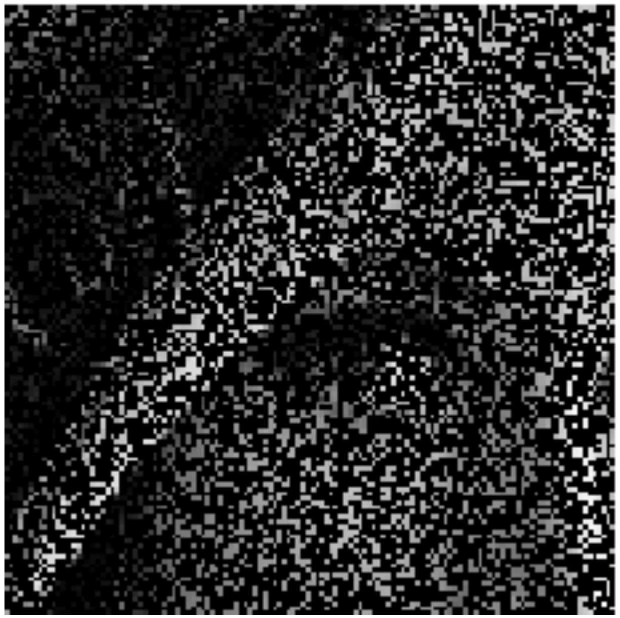


$$(\mathcal{K}f)_i = \begin{cases} f_i & \text{if} \quad i \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathcal{K} : \mathbb{R}^N \to \mathbb{R}^P \qquad P = |\Omega|$$

$\Psi \in \mathbb{R}^{N \times Q}$ translation invariant wavelet frame.

Orignal $f_0 = \Psi x_0$

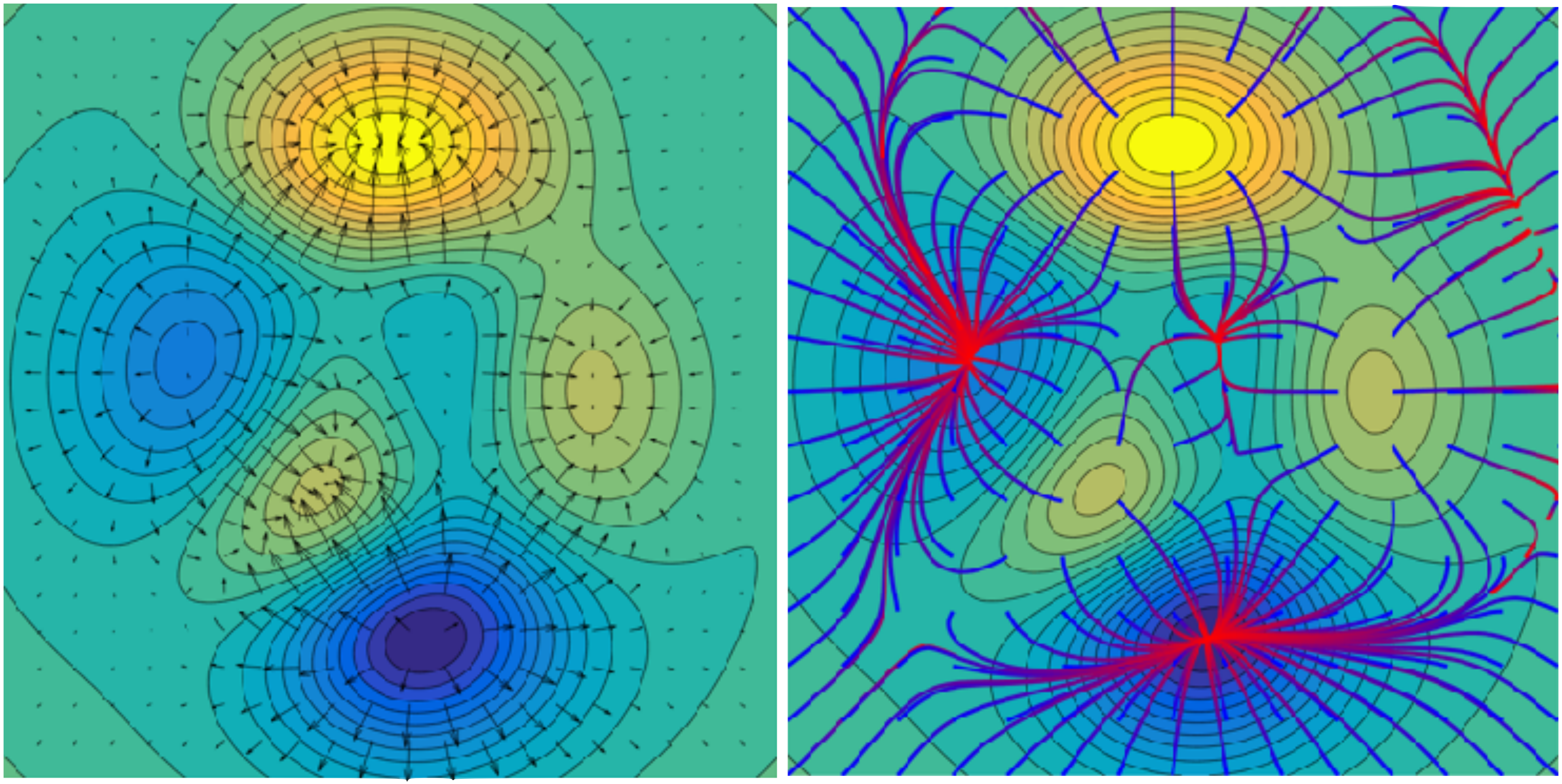


$y = \Phi x_0 + w$

Recovery $\Psi x^\star$

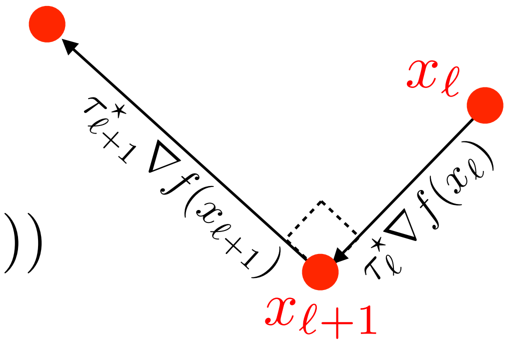
Overview

- Smooth optimization
- Subdifferential Calculus
- Proximal Calculus
- Forward Backward

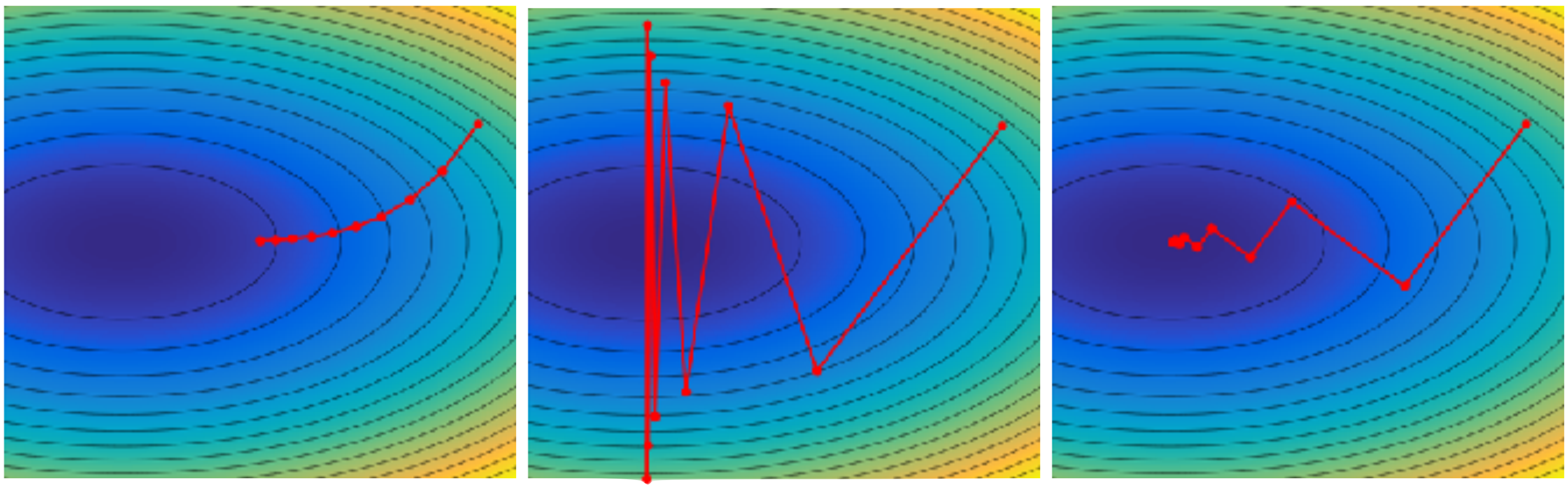


$$x_{\ell+1} = x_\ell - \tau_\ell \nabla f(x_\ell)$$

$$\tau_\ell^\star = \underset{\tau}{\operatorname{argmin}}\ f(x_\ell - \tau \nabla f(x_\ell))$$

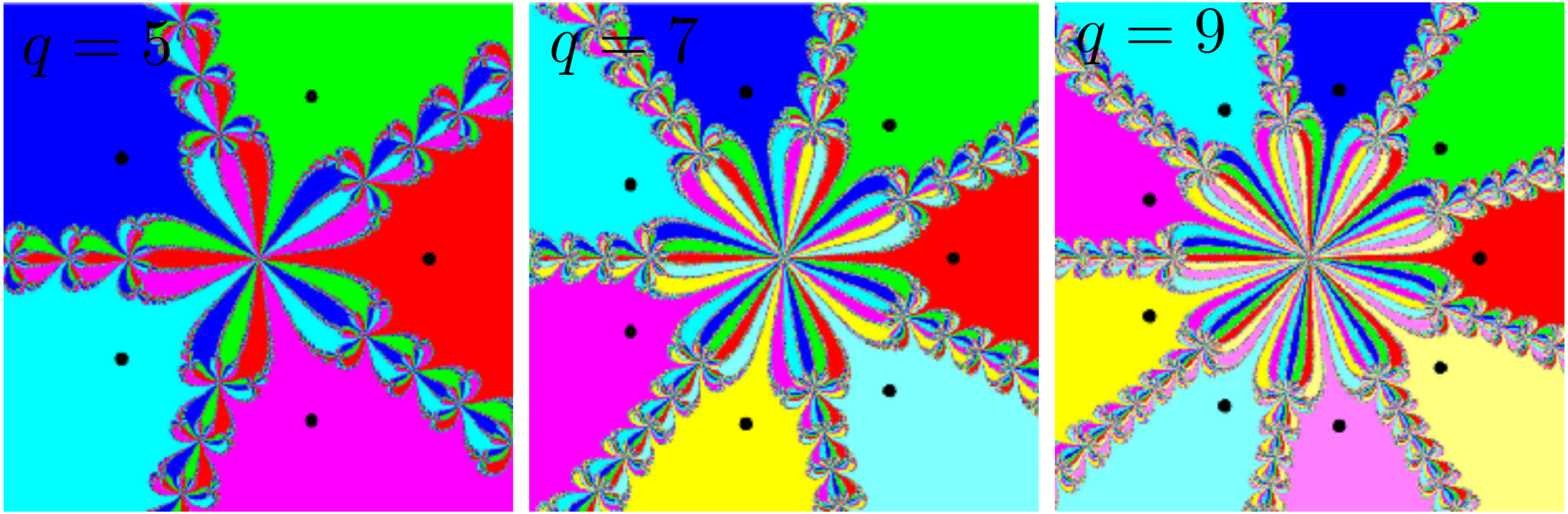


$$\nabla f(x_\ell) \perp \nabla f(x_{\ell+1})$$



Small τ_ℓ | Large τ_ℓ | Optimal $\tau_\ell = \tau_\ell^\star$

Newton method: $z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)}$



Attraction bassins for $f(z) = z^q - 1$

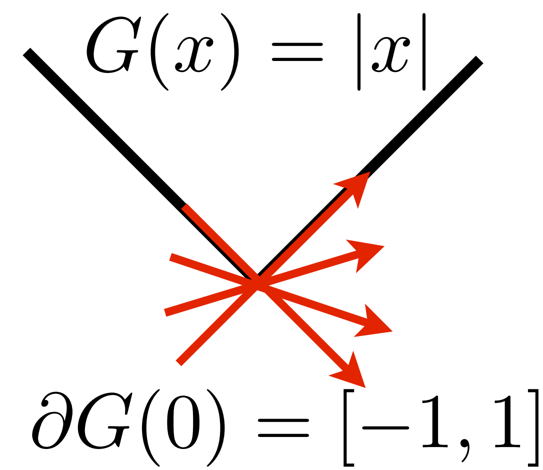
Overview

- Smooth optimization
- **Subdifferential Calculus**
- Proximal Calculus
- Forward Backward

Sub-differential

Sub-differential:

$$\partial G(x) = \{u \in \mathcal{H} \setminus \forall\, z,\ G(z) \geqslant G(x) + \langle u,\ z - x\rangle\}$$

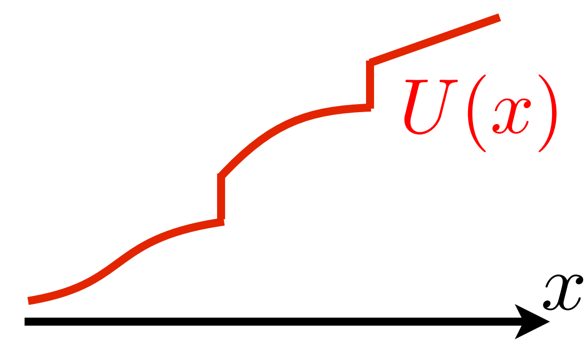


Smooth functions:

If F is C^1, $\partial F(x) = \{\nabla F(x)\}$

First-order conditions:

$$x^\star \in \underset{x \in \mathcal{H}}{\operatorname{argmin}}\ G(x) \iff 0 \in \partial G(x^\star)$$



Monotone operator: $\quad U(x) = \partial G(x)$

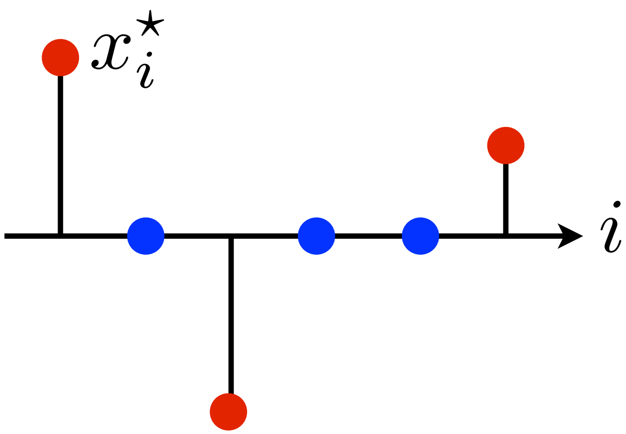
$$\forall\,(u, v) \in U(x) \times U(y),\quad \langle y - x,\ v - u\rangle \geqslant 0$$

Example: ℓ^1 Regularization

$$x^\star \in \underset{x \in \mathbb{R}^Q}{\text{argmin}}\ G(x) = \frac{1}{2}\|y - \Phi x\|^2 + \lambda \|x\|_1$$

$$\partial G(x) = \Phi^*(\Phi x - y) + \lambda \partial \| \cdot \|_1(x)$$

$$\partial \| \cdot \|_1(x)_i = \begin{cases} \text{sign}(x_i) & \text{if} \quad x_i \neq 0, \\ [-1, 1] & \text{if} \quad x_i = 0. \end{cases}$$



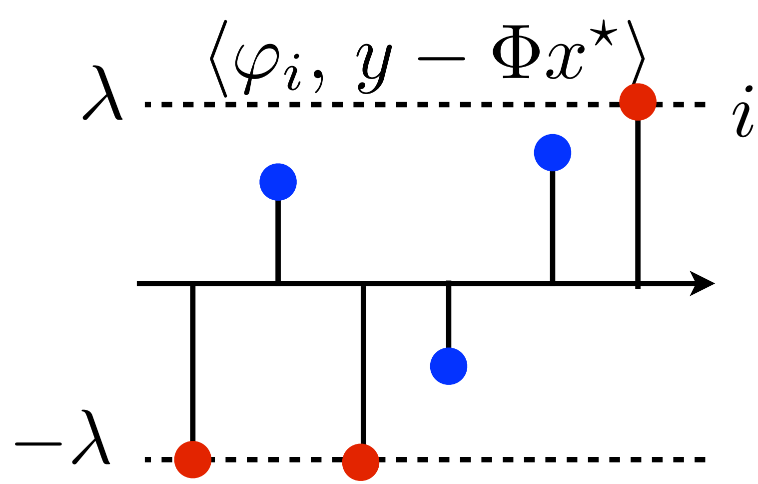
Support of the solution: ●

$$I = \{ i \in \{0, \ldots, N-1\} \setminus x_i^\star \neq 0 \}$$

First-order conditions:

$$\exists s \in \mathbb{R}^N, \quad \Phi^*(\Phi x^\star - y) + \lambda s = 0$$

$$\begin{cases} s_I = \text{sign}(x_I), \\ \|s_{I^c}\|_\infty \leqslant 1. \end{cases}$$



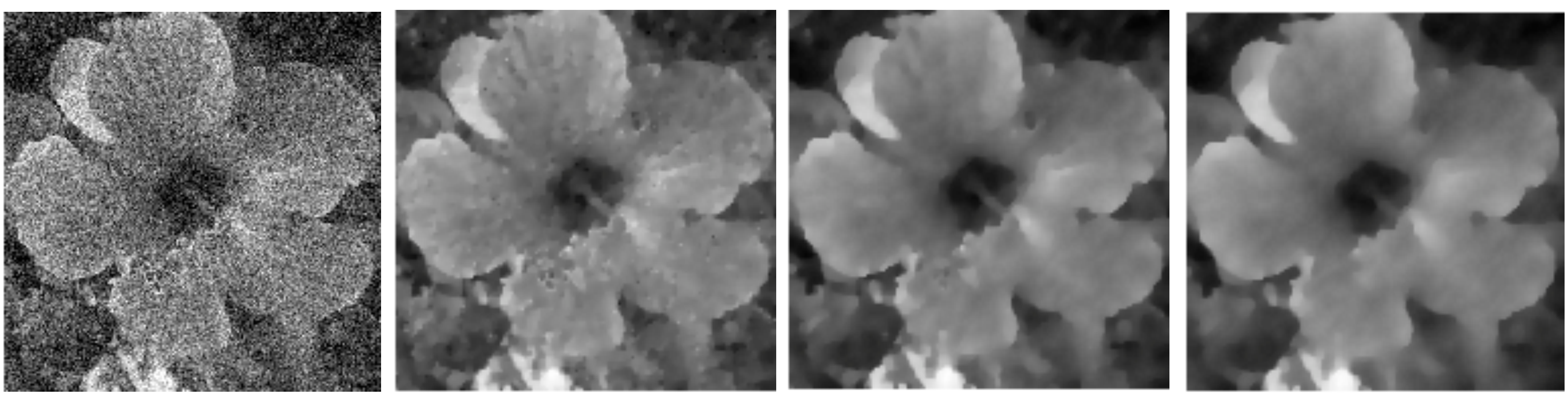
Example: Total Variation Denoising

Important: the optimization variable is f.

$$f^\star \in \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \; \frac{1}{2}\|y - f\|^2 + \lambda J(f)$$

Finite difference gradient: $\nabla : \mathbb{R}^N \to \mathbb{R}^{N \times 2}$ $\qquad (\nabla f)_i \in \mathbb{R}^2$

Discrete TV norm: $J(f) = \sum_i \|(\nabla f)_i\|$



$\lambda = 0$ (noisy) $\longrightarrow \lambda$

Example: Total Variation Denoising

$$f^\star \in \underset{f \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\|y - f\|^2 + \lambda J(f)$$

$$J(f) = G(\nabla f) \qquad G(u) = \sum_i \|u_i\|$$

Composition by linear maps: $\partial(J \circ A) = A^* \circ (\partial J) \circ A$

$$\partial J(f) = -\mathrm{div}\,(\partial G(\nabla f))$$

$$\partial G(u)_i = \begin{cases} \frac{u_i}{\|u_i\|} & \text{if} \quad u_i \neq 0, \\ \{\eta \in \mathbb{R}^2 \setminus \|\eta\| \leqslant 1\} & \text{if} \quad u_i = 0. \end{cases}$$

First-order conditions: $\exists\, v \in \mathbb{R}^{N \times 2},\ f^\star = y + \lambda \mathrm{div}(v)$

$$\begin{cases} \forall\, i \in I,\ v_i = \frac{\nabla f_i^\star}{\|\nabla f_i^\star\|}, \\ \forall\, i \in I^c,\ \|v_i\| \leqslant 1 \end{cases} \qquad I = \{i \setminus (\nabla f^\star)_i \neq 0\}$$

Overview

- Smooth optimization
- Subdifferential Calculus
- **Proximal Calculus**
- Forward Backward

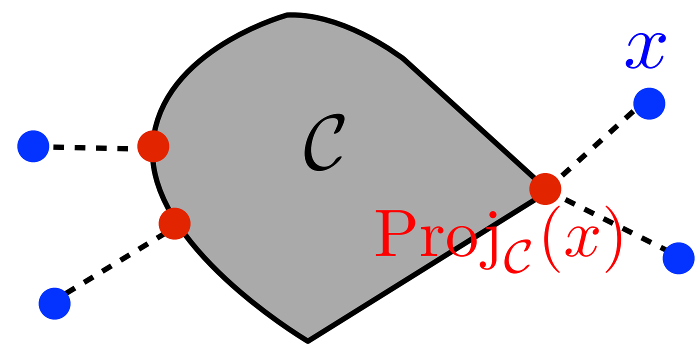
Proximal Operators

Proximal operator of G:

$$\mathrm{Prox}_{\gamma G}(x) = \underset{z}{\mathrm{argmin}} \; \frac{1}{2}\|x - z\|^2 + \gamma G(z)$$

Indicators: $G(x) = \iota_{\mathcal{C}}(x)$

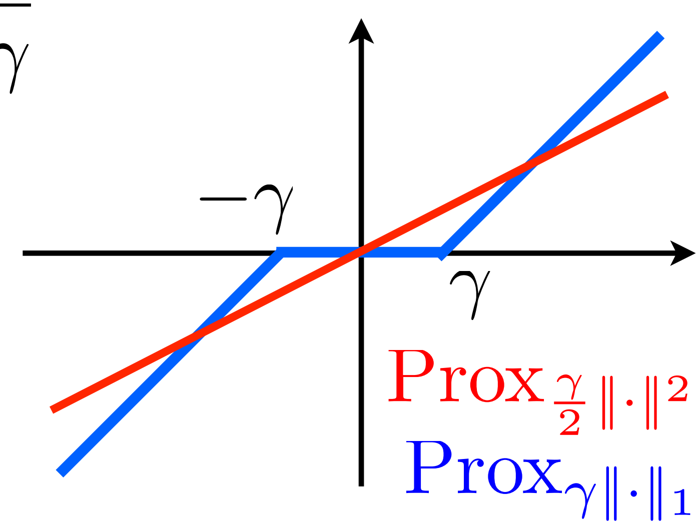
$$\mathrm{Prox}_{\gamma G}(x) = \mathrm{Proj}_{\mathcal{C}}(x) = \underset{z \in \mathcal{C}}{\mathrm{argmin}} \; \|x - z\|$$



ℓ^2 norm squared: $\mathrm{Prox}_{\frac{\gamma}{2}\|\cdot\|^2}(x) = \dfrac{x}{1+\gamma}$

ℓ^1 norm: $G(x) = \|x\|_1 = \sum_i |x_i|$

$$\mathrm{Prox}_{\gamma G}(x)_i = \max\left(0, 1 - \frac{\gamma}{|x_i|}\right) x_i$$



Proximal Calculus

Separability: $G(x) = G_1(x_1) + \ldots + G_n(x_n)$

$$\mathrm{Prox}_G(x) = (\mathrm{Prox}_{G_1}(x_1), \ldots, \mathrm{Prox}_{G_n}(x_n))$$

Quadratic functionals: $G(x) = \frac{1}{2}\|\Phi x - y\|^2$

$$\mathrm{Prox}_{\gamma G} = (\mathrm{Id} + \gamma \Phi^* \Phi)^{-1} \Phi^*$$

$$= \Phi^* (\mathrm{Id} + \gamma \Phi \Phi^*)^{-1}$$

Composition by tight frame: $A \circ A^* = \mathrm{Id}$

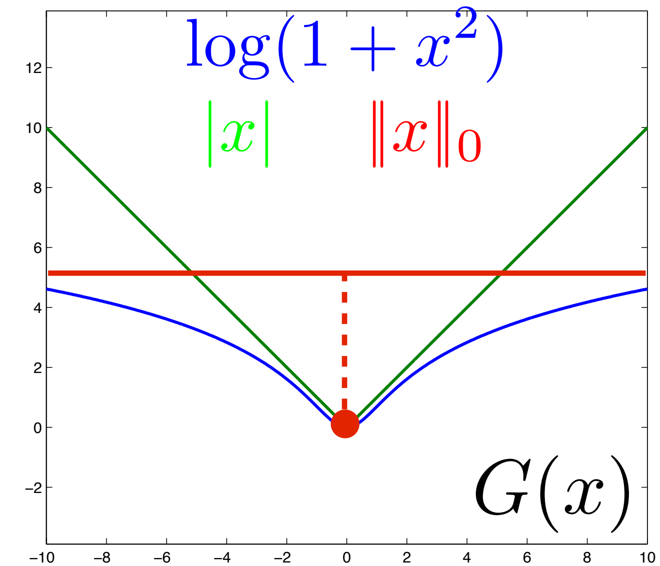
$$\mathrm{Prox}_{G \circ A}(x) = A^* \circ \mathrm{Prox}_G \circ A + \mathrm{Id} - A^* \circ A$$

Ortho-basis A: $\mathrm{Prox}_{G \circ A} = A^* \circ \mathrm{Prox}_G \circ A$

Non-convex Proximal Operators

Proximal operator of G:

$$\mathrm{Prox}_{\gamma G}(x) = \underset{z}{\mathrm{argmin}}\ \frac{1}{2}\|x - z\|^2 + \gamma G(z)$$

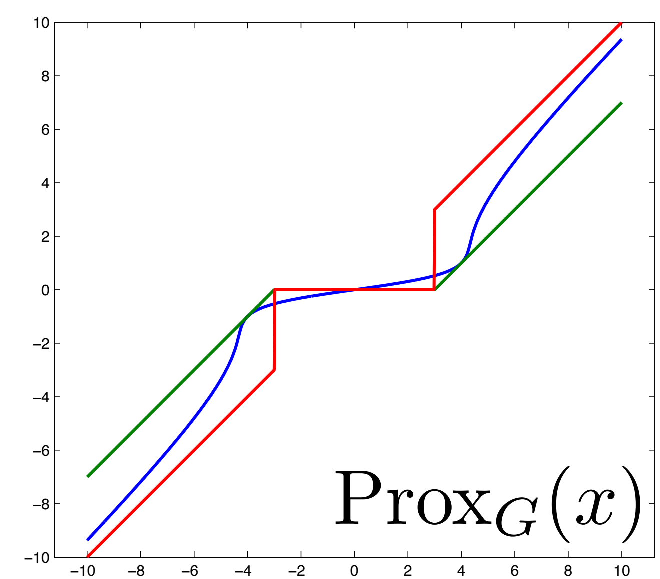


$$G(x) = \|x\|_1 = \sum_i |x_i|$$

$$\mathrm{Prox}_{\gamma G}(x)_i = \max\left(0, 1 - \frac{\gamma}{|x_i|}\right) x_i$$

$$G(x) = \|x\|_0 = |\,\{i \setminus x_i \neq 0\}\,|$$

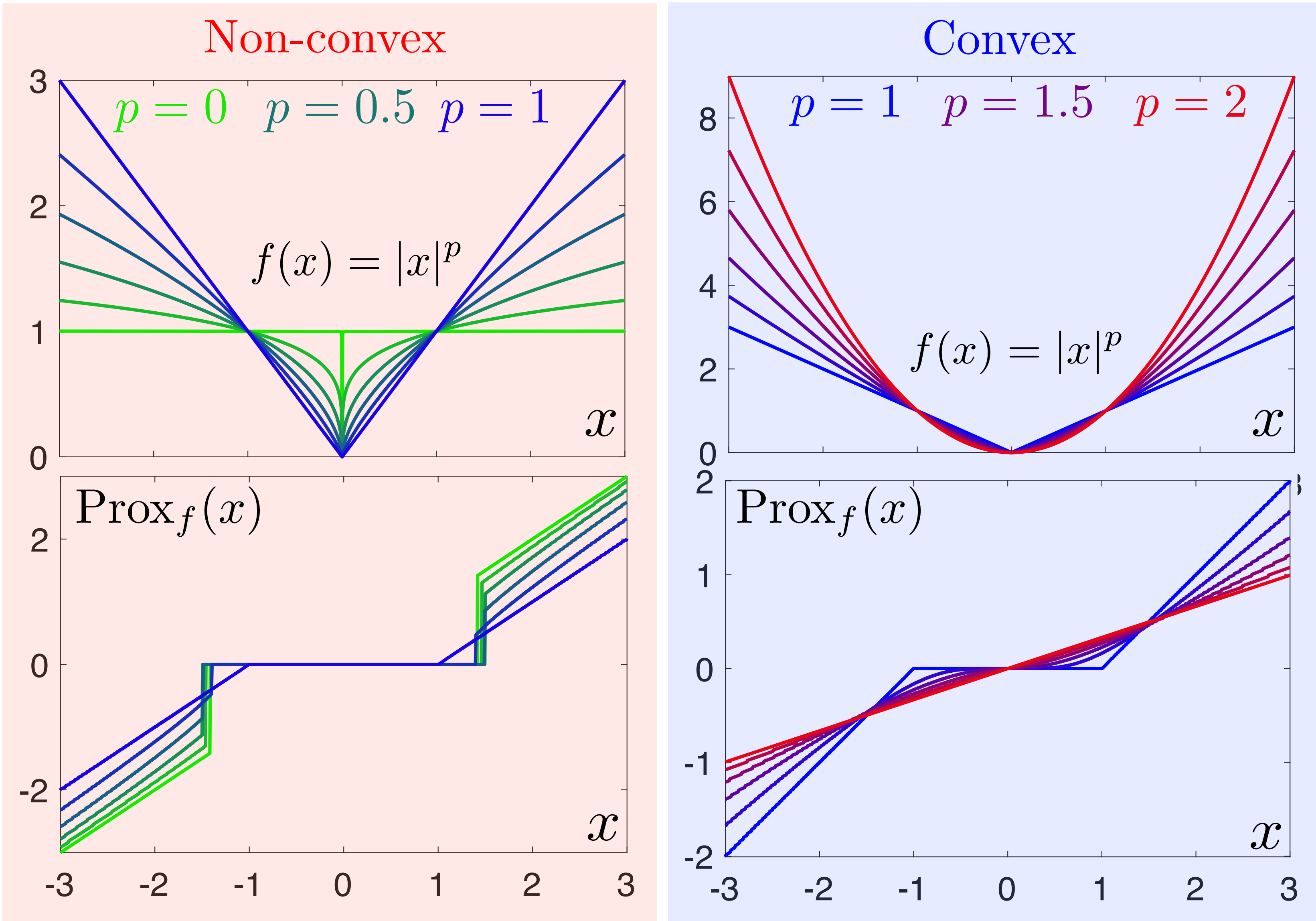
$$\mathrm{Prox}_{\gamma G}(x)_i = \begin{cases} x_i & \text{if} \quad |x_i| \geqslant \sqrt{2\gamma}, \\ 0 & \text{otherwise.} \end{cases}$$



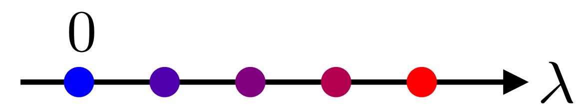
$$G(x) = \sum_i \log(1 + |x_i|^2)$$

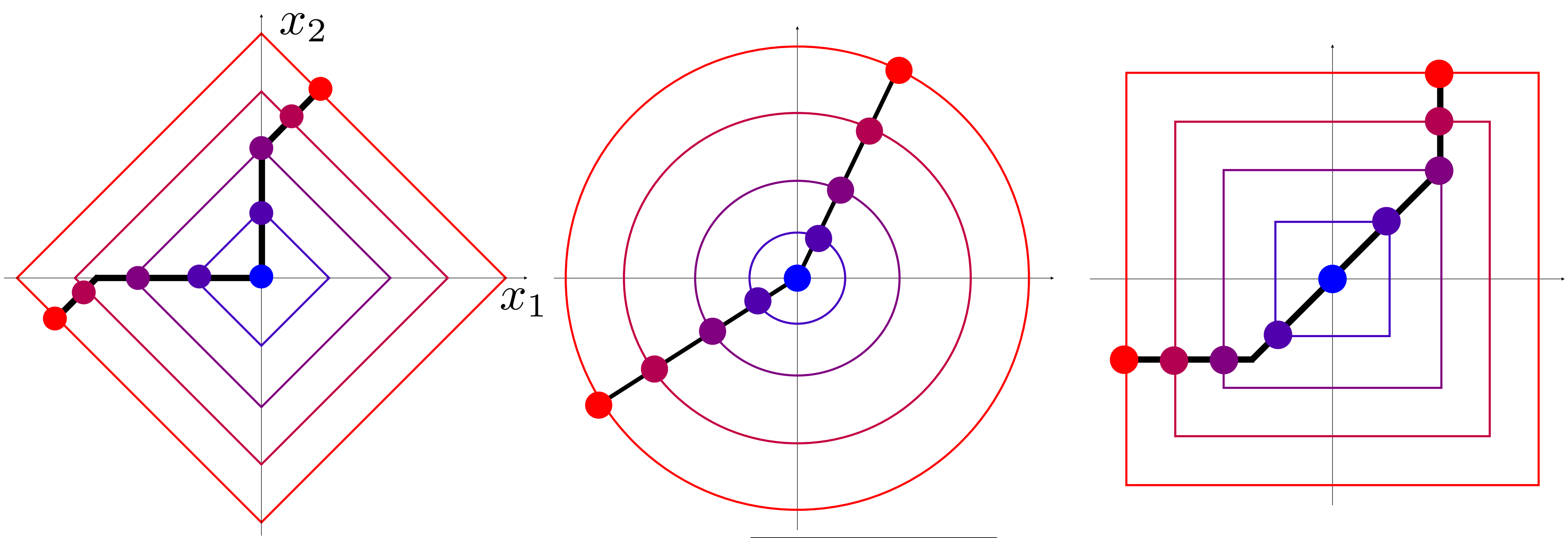
$\longrightarrow$ 3rd order polynomial root.

$$\mathrm{Prox}_f(x) = \underset{x'}{\mathrm{argmin}} \, \frac{1}{2}\|x - x'\|^2 + f(x')$$



$$\mathrm{Prox}_{\lambda f}(x) = \underset{x'}{\mathrm{argmin}}\ \frac{1}{2}\|x - x'\|^2 + \lambda f(x')$$





$f(x) = |x_1| + |x_2|$ $\quad$ $f(x) = \sqrt{|x_1|^2 + |x_2|^2}$ $\quad$ $f(x) = \max(|x_1|, |x_2|)$

Prox and Subdifferential

Resolvant of ∂G:

$$z = \mathrm{Prox}_{\gamma G}(x) \iff 0 \in z - x + \gamma \partial G(z)$$

$$\iff x \in (\mathrm{Id} + \gamma \partial G)(z) \iff z = (\mathrm{Id} + \gamma \partial G)^{-1}(x)$$

Inverse of a set-valued mapping:

$$\text{where } x \in U(y) \iff y \in U^{-1}(x)$$

$$\mathrm{Prox}_{\gamma G} = (\mathrm{Id} + \gamma \partial G)^{-1} \quad \text{is a single-valued mapping}$$

Fix point: $x^\star \in \underset{x}{\mathrm{argmin}}\, G(x)$

$$\iff 0 \in \partial G(x^\star) \iff x^\star \in (\mathrm{Id} + \gamma \partial G)(x^\star)$$

$$\iff x^\star = (\mathrm{Id} + \gamma \partial G)^{-1}(x^\star) = \mathrm{Prox}_{\gamma G}(x^\star)$$

Gradient and Proximal Descents

Gradient descent: $x^{(\ell+1)} = x^{(\ell)} - \gamma_\ell \nabla G(x^{(\ell)})$ **[explicit]**

G is C^1 and ∇G is L-Lipschitz

Theorem: If $0 < \gamma_\ell < 2/L$, $x^{(\ell)} \rightarrow x^\star$ a solution.

Sub-gradient descent: $x^{(\ell+1)} = x^{(\ell)} - \gamma_\ell v^{(\ell)}, \quad v^{(\ell)} \in \partial G(x^{(\ell)})$

Theorem: If $\gamma_\ell \sim 1/\ell$, $x^{(\ell)} \rightarrow x^\star$ a solution.

$\longrightarrow$ Problem: slow.

Proximal-point algorithm: $x^{(\ell+1)} = \mathrm{Prox}_{\gamma_\ell G}(x^{(\ell)})$ **[implicit]**

Theorem: If $\gamma_\ell \geqslant c > 0$, $x^{(\ell)} \rightarrow x^\star$ a solution.

$\longrightarrow$ $\mathrm{Prox}_{\gamma G}$ hard to compute.

Overview

- Smooth optimization
- Subdifferential Calculus
- Proximal Calculus
- **Forward Backward**

Proximal Splitting Methods

Solve $\min_{x \in \mathcal{H}} E(x)$

Problem: $\mathrm{Prox}_{\gamma E}$ is not available.

Splitting: $E(x) = F(x) + \sum_i G_i(x)$

Smooth — Simple

Iterative algorithms using: $\nabla F(x)$ and $\mathrm{Prox}_{\gamma G_i}(x)$

Forward-Backward: $\xrightarrow{\text{solves}}$ $F + G$

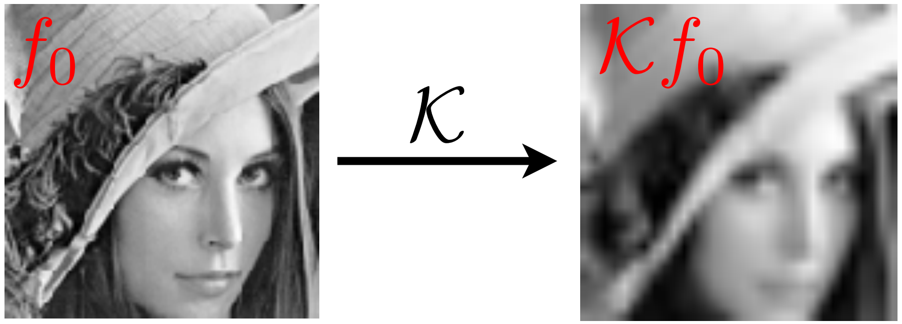
Douglas-Rachford: $\longrightarrow$ $\sum G_i$

Primal-Dual: $\longrightarrow$ $\sum G_i \circ A$

Generalized FB: $\longrightarrow$ $F + \sum G_i$

Smooth + Simple Splitting

Inverse problem: measurements $y = \mathcal{K} f_0 + w$



$\mathcal{K} : \mathbb{R}^N \to \mathbb{R}^P, \quad P \leqslant N$

Model: $f_0 = \Psi x_0$ sparse in dictionary Ψ.

Sparse recovery: $f^\star = \Psi x^\star$ where $x^\star$ solves

$$\min_{x \in \mathbb{R}^N} \underbrace{F(x)}_{\text{Smooth}} + \underbrace{G(x)}_{\text{Simple}}$$

Data fidelity: $F(x) = \frac{1}{2}\|y - \Phi x\|^2 \qquad \Phi = \mathcal{K} \circ \Psi$

Regularization: $G(x) = \|x\|_1 = \sum_i |x_i|$

Forward-Backward

Fix point equation:

$$x^\star \in \underset{x}{\text{argmin}}\ F(x) + G(x) \iff 0 \in \nabla F(x^\star) + \partial G(x^\star)$$

$$\iff (x^\star - \gamma \nabla F(x^\star)) \in x^\star + \gamma \partial G(x^\star)$$

$$\iff x^\star = \text{Prox}_{\gamma G}(x^\star - \gamma \nabla F(x^\star))$$

Forward-backward:

$$x^{(\ell+1)} = \text{Prox}_{\gamma G}\left(x^{(\ell)} - \gamma \nabla F(x^{(\ell)})\right)$$

Projected gradient descent: $G = \iota_{\mathcal{C}}$

Theorem: Let ∇F be L-Lipschitz.

If $\gamma < 2/L$, $x^{(\ell)} \rightarrow x^\star$ a solution of $(\star)$

Example: L1 Regularization

$$\min_x \frac{1}{2}\|\Phi x - y\|^2 + \lambda\|x\|_1 \quad \Longleftrightarrow \quad \min_x F(x) + G(x)$$

$$F(x) = \frac{1}{2}\|\Phi x - y\|^2$$

$$\nabla F(x) = \Phi^*(\Phi x - y) \qquad L = \|\Phi^*\Phi\|$$

$$G(x) = \lambda\|x\|_1$$

$$\mathrm{Prox}_{\gamma G}(x)_i = \max\left(0, 1 - \frac{\gamma\lambda}{|x_i|}\right) x_i$$

Forward-backward $\Longleftrightarrow$ Iterative soft thresholding

Convergence Speed

$$\min_x E(x) = F(x) + G(x)$$

∇F is L-Lipschitz.

G is simple.

Theorem: If $L > 0$, FB iterates $x^{(\ell)}$ satisfies

$$E(x^{(\ell)}) - E(x^\star) = O(1/\ell)$$

C degrades with $L \to 0$.

Multi-steps Accelerations

Beck-Teboule accelerated FB: $t^{(0)} = 1$

$$x^{(\ell+1)} = \mathrm{Prox}_{1/L}\left(y^{(\ell)} - \frac{1}{L}\nabla F(y^{(\ell)}) \right)$$

$$t^{(\ell+1)} = \frac{1+\sqrt{1+4(t^{(\ell)})^2}}{2}$$

$$y^{(\ell+1)} = x^{(\ell+1)} + \frac{t^{(\ell)}-1}{t^{(\ell+1)}}(x^{(\ell+1)} - x^{(\ell)})$$

(see also Nesterov method)

Theorem: If $L > 0$, $E(x^{(\ell)}) - E(x^\star) = O(1/\ell^2)$

Complexity theory: optimal in a worse-case sense.