Lecture 23: Proximal Maps and Shrinkage

Charles A. Bouman

Purdue University

Monday, October 19, 2026

Outline

1. The proximal map
   - Definition and picture
   - The ADMM updates and the dual update are proximal maps
2. The L_1 proximal map and the shrinkage function
3. ICD with a Laplacian prior: each update is a shrinkage step
4. Where ICD fails: the regularizer $||Bx||_1$

Reading: Section 9.4, *Foundations of Computational Imaging*.

Today's Big Ideas

1. The proximal map minimizes a convex function plus distance from y.
2. ADMM steps, the clip, the denoiser, and dual steps: proximal maps.
3. For L_1, the proximal map shrinks: small values go to zero.
4. Laplacian prior: ICD steps shrink and reach the global minimum.
5. For total variation, ICD does not in general reach the global minimum.

Definition of the proximal map

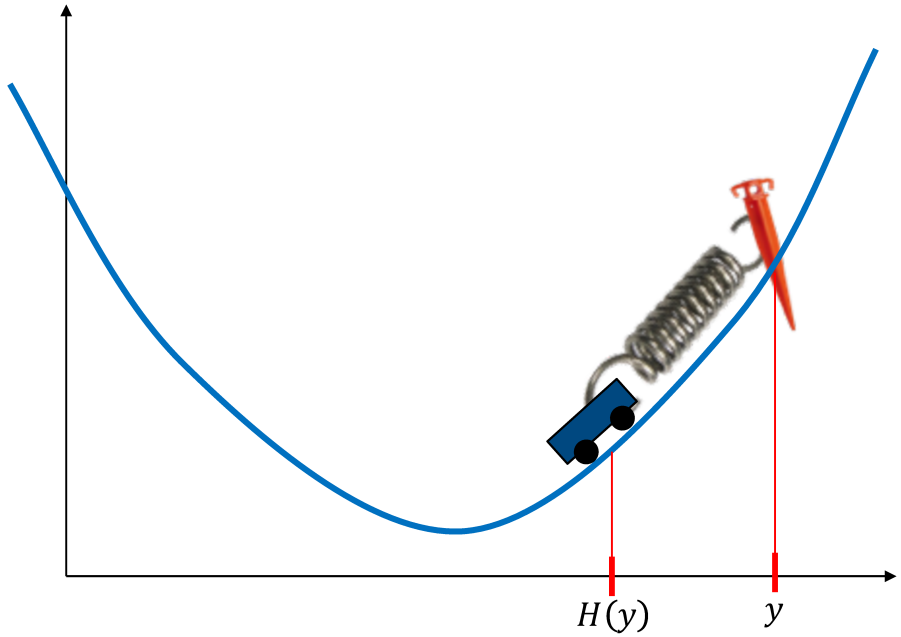
The **proximal map**, $H(y)$, of a function $u(x)$ is defined as

$$H(y) = \underset{x \in \Omega}{\arg\min} \left\{ \frac{1}{2} ||y - x||^2 + u(x) \right\} \ . \tag{9.31}$$

Here $u(x)$ is a convex function on the closed convex nonempty set $\Omega \subset \mathbb{R}^N$.

- More generally, u may be an extended convex function. Then u may be any proper closed convex function $u : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$.
- The input y is a point in $\mathbb{R}^N$, and the output $H(y)$ is a point in Ω.

The stake and the spring



- The input y is a stake holding a spring with unit spring constant.
- The other end holds a cart that would roll down the hill $u(x)$.
- The cart stops where the two forces balance. That point is $H(y)$.
- Not exact (the spring should be horizontal), but the intuition is right.

The ADMM updates are proximal maps

Scaling: for $a > 0$,

$$\arg\min_{x \in \Omega} \left\{ \frac{a}{2} ||y - x||^2 + u(x) \right\}$$

is the proximal map of u/a, evaluated at y.

The split-variable ADMM updates have this form.

- $\hat{x} \leftarrow \arg\min_x \{ f(x) + \frac{a}{2} ||x - (\hat{v} - u)||^2 \}$ is the proximal map of f/a at $\hat{v} - u$.
- $\hat{v} \leftarrow \arg\min_v \{ g(v) + \frac{a}{2} ||(\hat{x} + u) - v||^2 \}$ is the proximal map of g/a at $\hat{x} + u$.

For positivity, $g = \tilde{g}$ and the proximal map is the clip $\hat{v} = \text{clip}\{\hat{x} + u, \mathbb{R}^{+N}\}$.

The proximal map is a MAP denoiser

Let $Y = X + W$, where W is white Gaussian noise with unit variance.

- **Denoising** recovers X from Y: the simplest inverse problem.

The MAP cost function for denoising is

$$f(x; y) = \frac{1}{2}||y - x||^2 + u(x) \ ,$$

where $u(x) = -\log p(x)$ is the convex negative log prior distribution.

So the MAP estimate is $\hat{x} = H(y)$, the proximal map of (9.31).

Same structure as the augmented Lagrangian dual update

The dual-space iteration of the augmented Lagrangian is

$$\lambda^{(k+1)} = \underset{\lambda \in \mathbb{R}^K}{\arg\max} \left\{ g(\lambda) - \frac{1}{2a} ||\lambda - \lambda^{(k)}||^2 \right\} . \tag{9.26}$$

Substitute $u = -g$. Then the proximal map has the same structure as the dual update (9.26).

- Maximizing g is the same as minimizing $-g$.
- The quadratic term penalizes moving away from the previous $\lambda^{(k)}$.

The proximal map is well defined

- The function $f(x; y) = \frac{1}{2}||y - x||^2 + u(x)$ is strictly convex on Ω.
- So its minimum is unique, and $H(y)$ is defined for every y.

Repeated application (proved in Problem 7)

The recursion

$$y^{(k+1)} = \arg\min_{x \in \Omega} \left\{ \frac{1}{2}||y^{(k)} - x||^2 + u(x) \right\}$$

converges to the minimum of $u(x)$:

$$\lim_{k \to \infty} y^{(k)} = \arg\min_{x} u(x) \ .$$

This is the same reason that the augmented Lagrangian iteration (9.26) converges to the solution of the dual problem.

Special case: the L_1 prior

Let $u(x) = \lambda||x||_1$, where $\lambda \geq 0$ is a scalar parameter. Then the proximal map is

$$H(y) = \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}||y - x||^2 + \lambda||x||_1 \right\} . \tag{9.33}$$

- The function is strictly convex, so its global minimum is unique.
- The function is separable:

$$f(x; y) = \sum_{i=1}^{N} \left\{ \frac{1}{2}(y_i - x_i)^2 + \lambda|x_i| \right\} .$$

- So each x_i is found by one independent scalar minimization.

The scalar solution: three cases

Minimizing with respect to x_i gives

$$\underset{x_i \in \mathbb{R}}{\arg\min} \left\{ \frac{1}{2}(y_i - x_i)^2 + \lambda |x_i| \right\} = \begin{cases} y_i - \lambda & \text{for } y_i > \lambda \ , \\ 0 & \text{for } |y_i| \leq \lambda \ , \\ y_i + \lambda & \text{for } y_i < -\lambda \end{cases} . \qquad (9.34)$$

- The result is verified by considering the three cases (Problem 8).

The shrinkage function

The **shrinkage function** is the expression of (9.34). In compact form it is

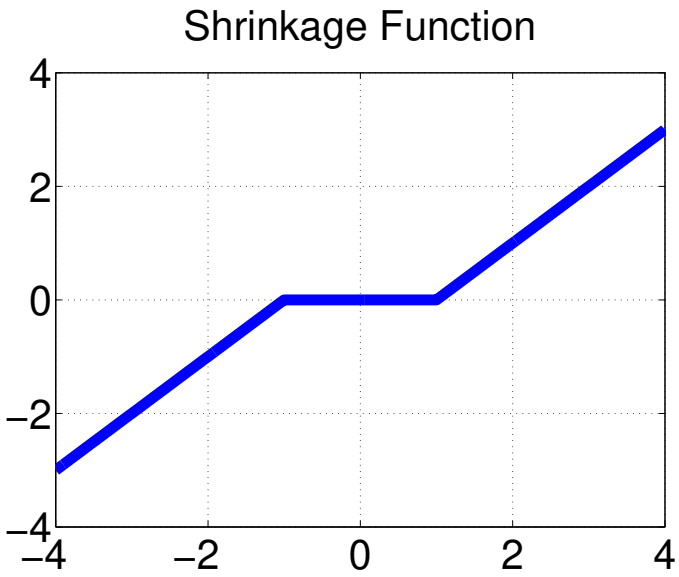
$$S_\lambda(y_i) \triangleq \text{sign}(y_i)\max\{|y_i| - \lambda, 0\} \ . \tag{9.35}$$

The solution of the L_1 proximal map (9.33) is then

$$\underset{x\in\mathbb{R}^N}{\arg\min}\left\{\frac{1}{2}||y-x||^2 + \lambda||x||_1\right\} = S_\lambda(y) \ , \tag{9.36}$$

where $S_\lambda(y)$ applies the shrinkage function to each component of the vector y.

The shrinkage function, $\lambda = 1$



- It is a continuous soft threshold. Values with $|y_i| \leq \lambda$ are set to zero.
- Other values shrink toward zero: positive ones down, negative ones up.
- For denoising, $|y_i| < \lambda$ is likely **noise**: set it to zero.
- For denoising, $|y_i| > \lambda$ is likely **signal**: keep it, but reduce it.

The general problem $Y = AX + W$

Let $Y = AX + W$, where A is an $M \times N$ matrix and $W \sim N(\mathbf{0}, \Lambda^{-1})$.

Let the components of X be i.i.d. with a Laplacian distribution. Then the prior density is

$$p_\lambda(x) = (\lambda/2)^N \exp\{-\lambda||x||_1\} \ ,$$

and the MAP optimization problem is

$$\hat{x} = \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}||y - Ax||_\Lambda^2 + \lambda||x||_1 \right\} \ . \tag{9.37}$$

- This problem has a form similar to the proximal map (9.36).
- So we expect the ICD update to contain the shrinkage function.

The MAP cost as a function of one pixel

As a function of the single pixel x_s, the MAP cost is

$$f_s(x_s) = \theta_1(x_s - x_{old}) + \frac{1}{2}\theta_2(x_s - x_{old})^2 + \lambda|x_s| + \text{const} \ ,$$

where x_{old} is the previous value of x_s, and

$$\begin{aligned} \theta_1 &= -(y - Ax)^t \Lambda A_{*,s} \ , \\ \theta_2 &= A_{*,s}^t \Lambda A_{*,s} \ . \end{aligned}$$

- The terms in θ_1 and θ_2 come from the quadratic data term.
- The term $\lambda|x_s|$ comes from the Laplacian prior.

The ICD update is a shrinkage step

Completing the square, the ICD update is

$$x_s \leftarrow \arg\min_{x_s} \left\{ \frac{1}{2} \left(x_s - \left[x_{old} - \frac{\theta_1}{\theta_2} \right] \right)^2 + \frac{\lambda}{\theta_2} |x_s| \right\} .$$

This update has the form of (9.34), so

$$x_s \leftarrow S_{\lambda/\theta_2} \left(x_{old} - \frac{\theta_1}{\theta_2} \right) .$$

- The point $x_{old} - \theta_1/\theta_2$ minimizes the quadratic data term alone.
- The L_1 prior then shrinks that point by λ/θ_2.

ICD algorithm using L_1 prior

ICD Algorithm Using L_1 Prior:

Initialize $e \leftarrow y - Ax$

For K iterations $\{$

  For each pixel $s \in S$ $\{$

$$x_{old} \leftarrow x_s$$

$$\theta_1 \leftarrow -e^t \Lambda A_{*,s}$$

$$\theta_2 \leftarrow A_{*,s}^t \Lambda A_{*,s}$$

$$x_s \leftarrow S_{\lambda/\theta_2}\left(x_{old} - \frac{\theta_1}{\theta_2}\right)$$

$$e \leftarrow e - A_{*,s}(x_s - x_{old}) \ .$$

  $\}$

$\}$

The algorithm is guaranteed to converge to the global minimum of the convex cost function.

Why ICD converges here

Convergence (Alliney and Ruzinsky, 1994)

ICD with the L_1 prior converges to the global minimum of (9.37).

- Not smooth, so the smooth-cost ICD theorem does not apply.
- ICD still converges: its steps are perpendicular to the surfaces $x_i = 0$.

Where ICD fails: $||Bx||_1$

Consider the more general problem

$$\hat{x} = \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}||y - Ax||_\Lambda^2 + ||Bx||_1 \right\} \ , \tag{9.38}$$

where B is an $N \times N$ matrix.

- ICD updates **do not in general reach the global minimum**.
- So ICD cannot guarantee a solution for total variation.

Summary

1. The proximal map returns the point that minimizes u plus a penalty for moving away from y: $H(y) = \arg\min_{x\in\Omega}\{\frac{1}{2}||y-x||^2 + u(x)\}$.
2. The ADMM updates are the proximal maps of f/a and g/a, the clip is the proximal map of $\tilde{g}$, the MAP denoiser is the proximal map of $-\log p(x)$, and the dual update $\lambda^{(k+1)} = \arg\max_\lambda\{g(\lambda) - \frac{1}{2a}||\lambda - \lambda^{(k)}||^2\}$ is the proximal map of $-g/a$.
3. For the L_1 prior it is the shrinkage function: $S_\lambda(y_i) = \text{sign}(y_i)\max\{|y_i| - \lambda, 0\}$.
4. With a Laplacian prior, each ICD update is a shrinkage step, $x_s \leftarrow S_{\lambda/\theta_2}(x_{old} - \theta_1/\theta_2)$, and ICD converges to the global minimum.
5. For $\frac{1}{2}||y - Ax||^2_\Lambda + ||Bx||_1$, which includes total variation, ICD does not in general converge to the global minimum.

Next lecture: replace Bx with a new variable v and add the constraint $Bx = v$.