

Lecture 23: Proximal Maps and Shrinkage

Charles A. Bouman

Purdue University

Monday, October 19, 2026

Outline

- 1 Definition of the proximal map $H(y)$
- 2 The cart-and-spring picture
- 3 Denoising interpretation: $H(y)$ is a MAP estimate
- 4 Same structure as the augmented Lagrangian dual iteration
- 5 Well-definedness, and repeated application of H
- 6 The L_1 proximal map and the shrinkage function $S_\lambda(y_i)$
- 7 The general problem $Y = AX + W$ with a Laplacian prior
- 8 ICD updates become shrinkage steps; the pseudocode
- 9 Where this breaks: $\|Bx\|_1$, and what comes next

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

Today's Big Ideas

- The proximal map is a simple but powerful operation. It takes a point and returns the point that minimizes a convex function plus a penalty for moving away from where you started.
- The proximal map turns up in two places we have already been. It is the MAP estimate for denoising, and it has the same structure as the augmented Lagrangian dual update.
- For the L_1 prior the proximal map is shrinkage. Small components are set to zero and larger ones are pulled toward zero.
- Shrinkage also solves the general inverse problem one pixel at a time. Each ICD update with a Laplacian prior is a shrinkage step, and the result is the global minimum.

The guarantee is lost as soon as the L_1 norm is applied to a matrix times the image, which is what total variation does. Fixing that is the job of the next two lectures.

Definition of the proximal map

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

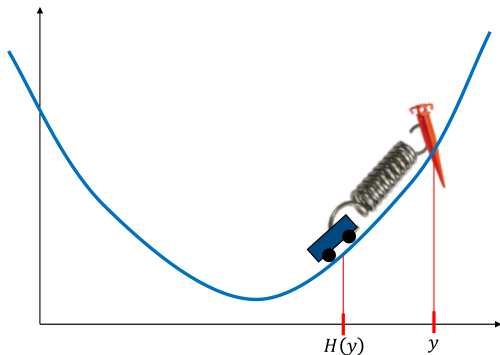
$$H(y) = \arg \min_{x \in \Omega} \left\{ \frac{1}{2} \|y - x\|^2 + u(x) \right\} . \quad (9.31)$$

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, in which case $u(x)$ may be any proper closed convex function $f : \mathbb{R}^N \rightarrow \mathbb{R} \cup \{\infty\}$.
- The input y and the output $H(y)$ are both points in $\mathbb{R}^N$: the proximal map is a map from $\mathbb{R}^N$ to Ω .

This is a minor detour in Chapter 9, but the result is used again in Sections 9.5 and in Chapter 10.

Graphical illustration: the stake and the spring



The input, y , can be thought of as a stake in the ground holding one end of a spring. The other end of the spring, with unit spring constant, holds a cart on wheels that would otherwise roll down the hill of the function being minimized. The cart moves toward the bottom of $u(x)$, but it is restricted by the spring, so instead it reaches an equilibrium point where the force of the hill balances the force of the spring.

A caveat about the picture

The book's footnote

This physical analogy is not exact because in this case the spring extends diagonally rather than horizontally, as the proximal map definition requires. But the emotional intuition gained by looking at the figure is correct.

- In the definition, the penalty is $\frac{1}{2}||y - x||^2$, a function of the difference between the two positions y and x only.
- Two forces trade off: the term $u(x)$ pulls x toward the minimum of u , and the term $\frac{1}{2}||y - x||^2$ pulls x toward y .

Interpretation: the proximal map is a MAP denoiser

Consider the problem in which $Y = X + W$, where W is white Gaussian noise with unit variance and our objective is to recover X from Y .

- **Denoising**, the recovery of X from Y in this simplest of inverse problems, has been the focus of an enormous amount of attention over the years.
- Perhaps this is not only because denoising is of practical importance but also because it presents a pure theoretical problem in image modeling.

The MAP cost function for this denoising problem is given by

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

where $u(x) = -\log p(x)$ is the convex negative log prior distribution. From this, we see that the MAP estimate, $\hat{x} = H(y)$, is given by the proximal mapping of (9.31).

Same structure as the augmented Lagrangian dual update

Recall the dual-space iteration of the augmented Lagrangian from last lecture:

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

It is also interesting to note that by substituting in $u(x) = -g(x)$, the proximal mapping has the same structure as the update equation for the augmented Lagrangian iteration in (9.26).

- Maximizing g is minimizing $-g$; the quadratic term is the same penalty on moving away from the previous point.

Well defined, and what repeated application does

- Notice that $f(x; y) = \frac{1}{2}||y - x||^2 + u(x)$ is a **strictly convex** function on Ω . Therefore, the global minimum must be unique, and $H(y)$ must always be well defined for $y \in \mathbb{R}^N$.

In problem 7 at the end of the chapter, we show that iterative application of a proximal mapping converges to the global minimum of the function $u(x)$. So 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 $u(x)$, or, more formally,

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

In retrospect, this is the same reason that the augmented Lagrangian iteration of (9.26) is guaranteed to converge to the solution of the dual optimization problem.

Special case: the L_1 prior

We can gain insight into the value of the proximal mapping by considering some illustrative special cases. One very interesting case occurs when we use the L_1 prior of Section 6.3 as the regularizing prior.

In this case, $u(x) = \lambda \|x\|_1$, where $\lambda \geq 0$ is a scalar parameter, and the proximal mapping is given by

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

Since $\|x\|_1$ is a convex function, we know that the function being minimized is a strictly convex function of x and therefore has a unique global minimum.

We can calculate the exact minimum by first noticing that the function being minimized is separable, so that it can be written as the sum of independent terms:

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

For each component y_i , the solution can therefore be computed by independently minimizing with respect to each component x_i .

The scalar solution: three cases

Minimizing with respect to x_i results in the solution

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

- This result can be easily verified by considering the three cases (problem 8 at the end of the chapter).

The shrinkage function

The **shrinkage function** is the name given to the expression of (9.34), which occurs so commonly that it deserves one. With a little bit of manipulation, we may express the shrinkage function in a more compact form given by

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

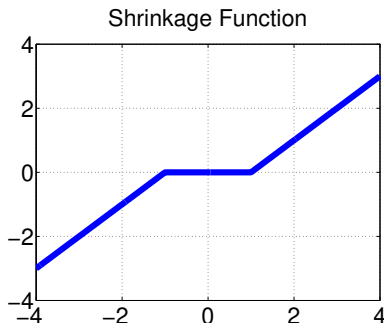
where $S_\lambda(y_i)$ is defined as the shrinkage function applied to the scalar y_i .

Using this more compact notation, we can also express the solution to the proximal mapping of (9.33) as

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

where $S_\lambda(y)$ is now interpreted as applying the shrinkage function to each component of the vector y .

The shrinkage function, $\lambda = 1$



- It acts like a **soft threshold** that sets all values to zero for $|y_i| \leq \lambda$ but remains continuous. Other values are “shrunk” toward zero, with positive values being moved down and negative values being moved up.
- Intuitively for the denoising case, if the value of $|y_i|$ is below λ , then it seems likely that the component contains mostly **noise**, so the value is set to zero.
- However, if the value of $|y_i|$ is above λ , then it seems likely that the component contains mostly **signal**, and it is retained, but with attenuation.

The general problem $Y = AX + W$

In fact, this framework extends to the more general case when our inverse problem is based on the model $Y = AX + W$, where A is an $M \times N$ matrix and $W \sim N(\mathbf{0}, \Lambda^{-1})$.

If we assume that the components of X are i.i.d. with a Laplacian distribution, then the prior density is given by

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

and the corresponding MAP optimization problem is then given by

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

- This MAP estimate has a very similar form to the proximal map of (9.36). So perhaps it is not surprising that the ICD update equation for solving this problem will incorporate the shrinkage function.

Rewriting the cost for a single pixel

In order to see this, we may use (7.13) to rewrite the MAP cost function as a function of the single pixel in the form

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

where x_{old} is the previous value of x_s and θ_1 and θ_2 are given in (7.14) and (7.15) as

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

- The quadratic data term is the same as in Chapter 7; only the prior term $\lambda|x_s|$ is new.

The ICD update is a shrinkage step

Using this simplified expression, it is possible to write the ICD update as

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

Notice that this update now has the form of the shrinkage operator given in (9.34), so the ICD update can be written as

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

- The unregularized quadratic minimum is $x_{old} - \theta_1/\theta_2$; the L_1 prior then shrinks it 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

In this case the ICD algorithm provably converges to the global minimum even though the MAP cost function is not continuously differentiable and therefore does not meet the conditions of the ICD convergence theorem of Chapter 7.

- Intuitively, the ICD algorithm converges in this case because the updated directions of the ICD algorithm are perpendicular to the discontinuous manifolds of the L_1 norm that occur when $x_i = 0$ for any i .
- The convergence result is due to Alliney and Ruzinsky, IEEE Trans. on Signal Processing, 1994, as cited in the text.

Where this breaks: $\|Bx\|_1$

However, in the more general problem with the form

$$\hat{x} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|y - Ax\|_A^2 + \|Bx\|_1 \right\}, \quad (9.38)$$

where B is an $N \times N$ matrix, the ICD updates **do not in general converge to the global minimum**.

- This means that the ICD updates cannot be used to guarantee a solution to important problems such as the total variation regularization problem of Section 6.3.
- Fortunately, in Section 9.5.1, we will introduce a very clever method for constrained optimization that can be used to solve problems with the form of (9.38).

Summary

- Proximal map: $H(y) = \arg \min_{x \in \Omega} \{\frac{1}{2} \|y - x\|^2 + u(x)\}$; the cost is strictly convex, so $H(y)$ is always well defined.
- $H(y)$ is the MAP estimate for denoising, $Y = X + W$ with unit noise variance and $u(x) = -\log p(x)$.
- With $u(x) = -g(x)$ it has the same structure as the augmented Lagrangian dual update; repeated application of H converges to $\arg \min_x u(x)$.
- For $u(x) = \lambda \|x\|_1$ the map separates and equals shrinkage, $S_\lambda(y_i) = \text{sign}(y_i) \max\{|y_i| - \lambda, 0\}$: components below λ are zeroed, larger ones retained with attenuation.
- For $Y = AX + W$ 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.
- For $\|Bx\|_1$, which includes total variation, ICD does not in general converge to the global minimum.

Reading: Section 9.4. Next lecture: Section 9.5, Variable Splitting and the ADMM Algorithm.