University of Minnesota
Driven to Discover℠

EE 5561: Image Processing and Applications

Lecture 11

Mehmet Akçakaya

Recap of Last Lecture

- General Inverse Problems
  - Least squares solution/maximum likelihood estimation in Gaussian noise
  - Regularization
  - Regularized least squares
  - Tikhonov regularization & variants
    - Energy-based regularizers

- Today:
  - More regularization based on sparsity

Sparsity

$$\arg\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{H}\mathbf{x}||_2^2 + \psi(\mathbf{x})$$

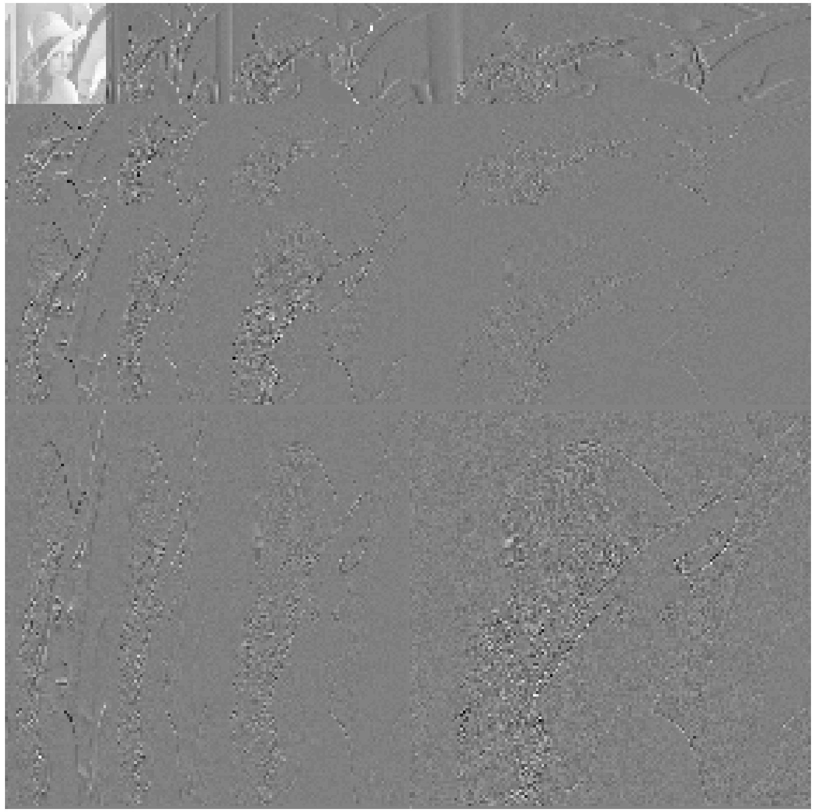
- Recall energy compaction for DCT/wavelets
- What happens when we throw away low-intensity transform coefficients?
  - Main idea behind compression
- This suggests a new regularization idea
  - Find **x** that “best fits” the data <u>and</u> has the least number of transform coefficients
- Let **W** be the transform, and **x** = **Wθ**
- We define the l_0 “norm” as

$$||\boldsymbol{\theta}||_0 = \text{number of non-zero coefficients of } \boldsymbol{\theta}$$

Haar Wavelet

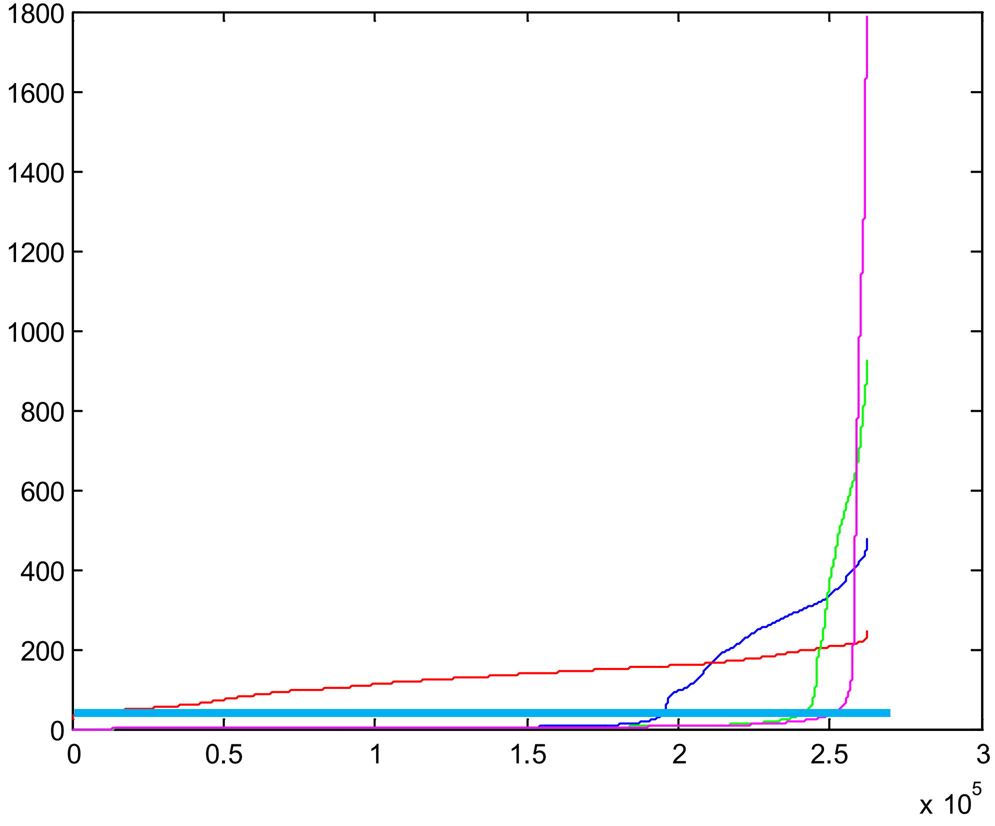
X

3-levels: $W_N X W_N^T$



Haar Wavelet

Energy "compaction":
sorted (abs value) image (red) vs. wavelet
1-level (blue), 2-level (green), 3-level(magenta) coefficients across the image



Wavelets

Original

Compression with Wavelet
(10-fold)

Compression with Wavelet
(20-fold)

Sparsity

- Goal: Find $\hat{\theta}$ that minimizes

$$\min_{\boldsymbol{\theta}} ||\mathbf{y} - \mathbf{HW}\boldsymbol{\theta}||_2^2 \qquad \text{subject to} \qquad ||\boldsymbol{\theta}||_0 \leq K$$

for some $K \in \mathbb{Z}^+$

$$\Updownarrow$$

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{\theta}||_0 \qquad \text{s. t.} \qquad ||\mathbf{y} - \mathbf{HW}\boldsymbol{\theta}||_2^2 \leq D$$

for some $D \in \mathbb{R}^+$

- $||\boldsymbol{\theta}||_0$ is hard to work with

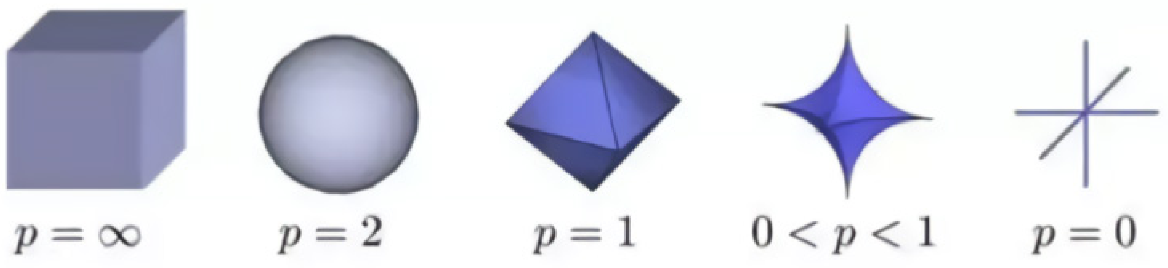
Sparsity

- The best convex approximation is $||\boldsymbol{\theta}||_1 = \sum_{i=1}^N |\theta_i|$
  - The reason why this is sparsity promoting has a lengthy history
  - We will try to give a high-level intuition
  - First consider the following two definitions

ℓ_p norm: $||\boldsymbol{\theta}||_p = (\sum_{i=1}^N |\theta_i|^p)^{\frac{1}{p}}$ This is actually a norm for $p \geq 1$

ℓ_p ball: $\{\boldsymbol{\theta} : ||\boldsymbol{\theta}||_p = 1 \}$

  - Now let's see how these l_p balls look for different values of p



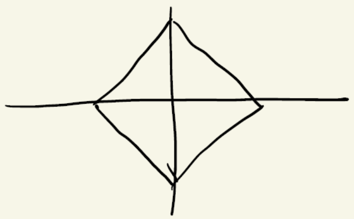
https://blog.mlreview.com/5b0e4be3938a

Sparsity

- In particular

ℓ_1-ball $\{\theta : \|\theta\|_1 = 1\}$

$|\theta_1| + |\theta_2| = 1$



"diamond"

– Now let's go back to our problem, but in the noiseless scenario, i.e.

$$\mathbf{y} = \mathbf{H}\mathbf{W}\boldsymbol{\theta} \triangleq \mathbf{A}\boldsymbol{\theta}$$

- As we discussed earlier, this is an interesting problem when **A** is either ill-conditioned or under-determined
- We will consider the latter case, i.e.

A is $\begin{array}{cc} & N \\ M & [\qquad] \end{array}$ $M < N$

Sparsity

- Suppose we solve

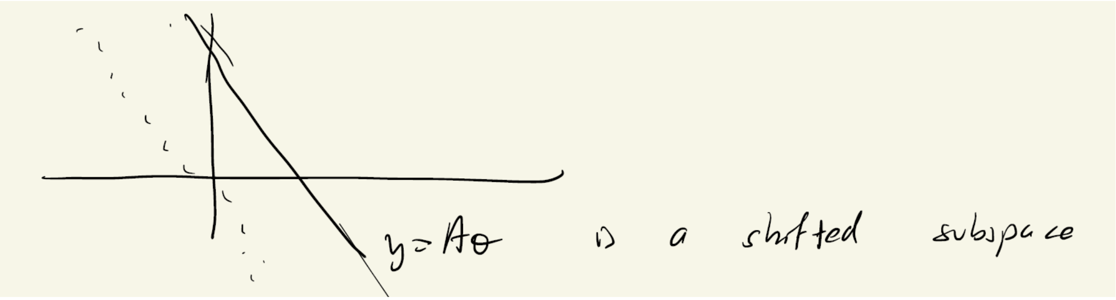
$$\min_{\boldsymbol{\theta}} ||\boldsymbol{\theta}||_p \qquad \text{s. t.} \qquad \mathbf{y} = \mathbf{A}\boldsymbol{\theta} \qquad\qquad \text{for } p \in \{0, 1, 2\}$$

- First look at all solutions to $\mathbf{y} = \mathbf{A}\boldsymbol{\theta}$

- How does this work?
  - Find any solution $\boldsymbol{\theta}'$ such that $\mathbf{y} = \mathbf{A}\boldsymbol{\theta}'$

    null-space of **A**
  - Then all the solutions lie in a shifted subspace given by $\boldsymbol{\theta}' + N(\mathbf{A})$
  - Why? Take any $\mathbf{w} \in N(\mathbf{A})$, then $\mathbf{A}\mathbf{w} = \mathbf{0}$
  - $(\mathbf{w} + \boldsymbol{\theta}')$ is another solution since $\mathbf{A}(\mathbf{w} + \boldsymbol{\theta}') = \mathbf{A}\mathbf{w} + \mathbf{A}\boldsymbol{\theta}' = \mathbf{0} + \mathbf{y} = \mathbf{y}$

- Pictorially

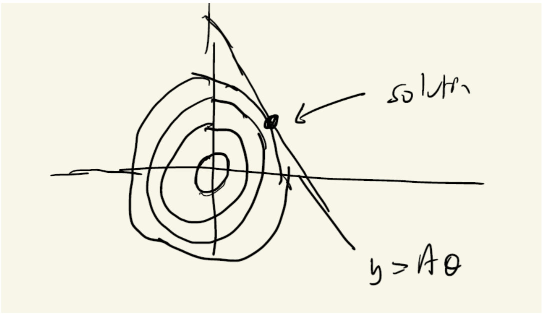


Sparsity

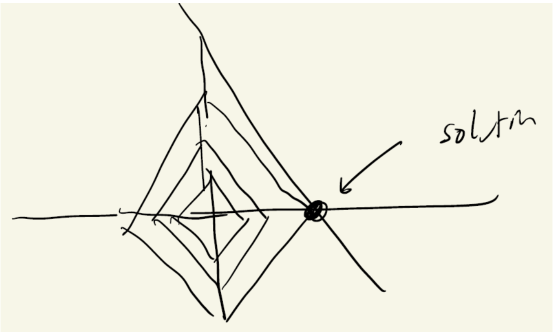
- For p = 2, the solution to

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{\theta}||_2 \qquad \text{s. t.} \qquad \mathbf{y} = \mathbf{A}\boldsymbol{\theta}$$

is the intersection of a scaled l_2 ball (i.e. sphere) and this shifted subspace



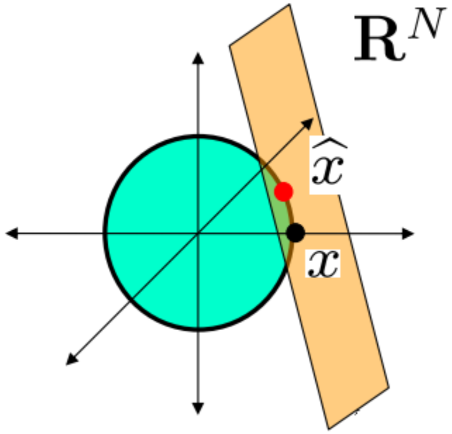
- For p = 1, it's the intersection of a "diamond" and this shifted subspace

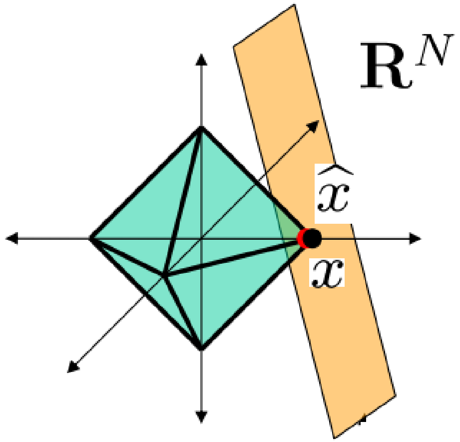


Note this ends up on the axes
→ same as the l_0 solution

Sparsity

- In high dimensions (M, N large), with appropriate choice of **A**, the null-space is oriented randomly
- Pictorially





- Same idea here: l_1 solution overlaps with the l_0 solution
- More dramatically at higher dimensions

Romberg & Wakin, “Compressed Sensing: A Tutorial”, 2007

Sparsity

- Easier to perform sparsity regularization in an unconstrained manner (i.e. in Lagrangian form)

$$\arg\min_{\boldsymbol{\theta}} ||\mathbf{y} - \mathbf{H}\mathbf{W}\boldsymbol{\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_1$$

- Questions:
  - How to choose **W**?
  - How to solve this objective function?
  - Better ways than the l_1 norm of transform-domain coefficients for using sparsity?

Solving Sparsity-Regularized LS

- First let's consider denoising again, i.e. **H = I**, or

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \qquad \mathbf{x} = \mathbf{W}\boldsymbol{\theta}, \qquad \boldsymbol{\theta} \text{ “sparse”}$$

$$\text{i.e. } \underbrace{\mathbf{y}}_{\text{image domain}} = \mathbf{W}\underbrace{\boldsymbol{\theta}}_{\text{transform domain}} + \mathbf{n}$$

- For now assume, **W** is orthogonal (e.g. wavelet or DCT), then

$$\mathbf{W}^T\mathbf{y} = \boldsymbol{\theta} + \mathbf{W}^T\mathbf{n}$$

- Recall if $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad \mathbf{W}^T\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{W}^T\mathbf{W}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$
- Thus without loss of generality, let

$$\mathbf{y}' = \mathbf{W}^T\mathbf{y}, \quad \mathbf{n}' = \mathbf{W}^T\mathbf{n}$$

- We will solve $\quad \mathbf{y}' = \boldsymbol{\theta} + \mathbf{n}'$

Denoising with Sparsity

- Here everything is in transform domain
- For ease of notation, we will just use $\mathbf{y} = \boldsymbol{\theta} + \mathbf{n}$
- Hence we are solving the following objective function (for denoising)

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{2}||\mathbf{y} - \boldsymbol{\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_1$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_{k=1}^{N}|y_k - \theta_k|^2 + \lambda\sum_{k=1}^{N}|\theta_k|$$

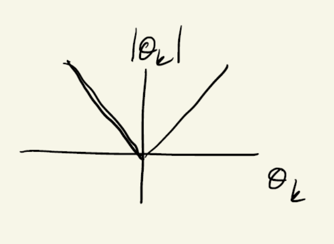
$$= \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{N}\left(\frac{1}{2}|y_k - \theta_k|^2 + \lambda|\theta_k|\right)$$

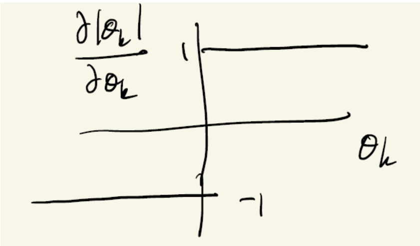
  - Hence we can minimize this separately for each *k*

Denoising by Thresholding

– For each k, we are solving

$$\min_{\theta_k} \frac{1}{2}|y_k - \theta_k|^2 + \lambda|\theta_k|$$

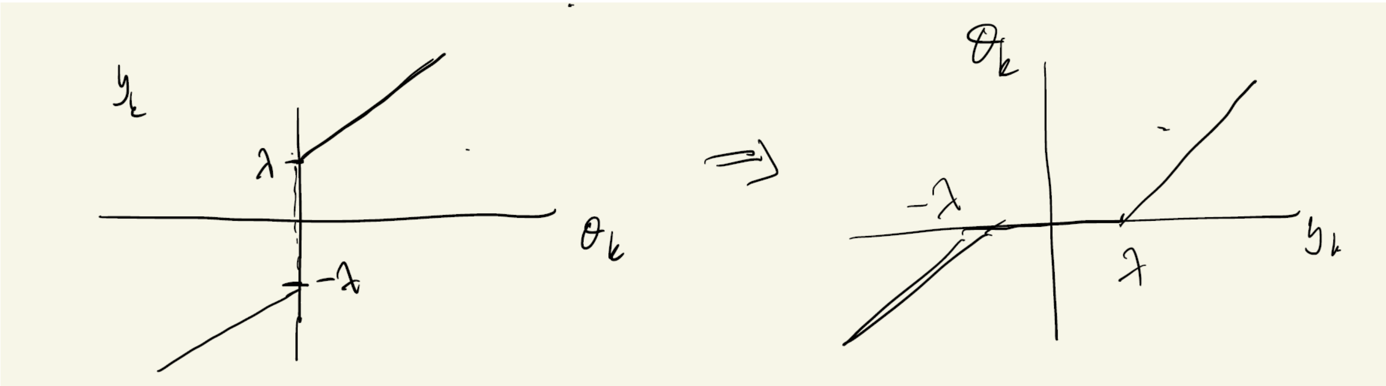




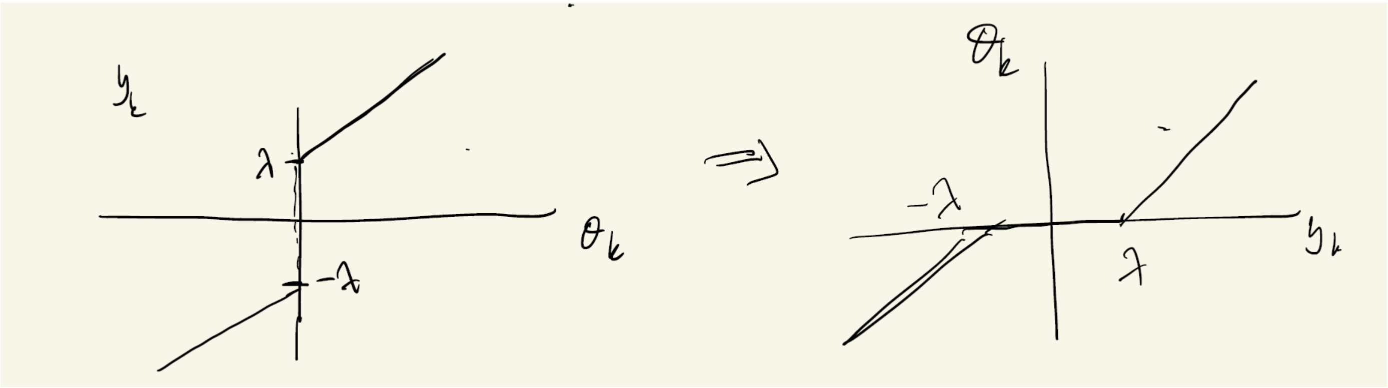
- Take derivative with respect to θ_k and set to 0

$$(\theta_k - y_k) + \lambda\,\mathrm{sign}(\theta_k) = 0$$

$$y_k = \theta_k + \lambda\,\mathrm{sign}(\theta_k)$$



Denoising by Thresholding



$$\hat{\theta}_k = \begin{cases} y_k - \lambda & y_k > \lambda \\ 0 & |y_k| < \lambda \\ y_k + \lambda & y_k < -\lambda \end{cases} = \begin{cases} |y_k| - \lambda & y_k > \lambda \\ 0 & |y_k| < \lambda \\ -(|y_k| - \lambda) & y_k < -\lambda \end{cases}$$

$$\hat{\theta}_k = \max(|y_k| - \lambda, 0) \cdot \mathrm{sign}(y_k) \triangleq S_\lambda(y_k)$$

soft-thresholding function

Denoising by Thresholding

Original

Noisy

Soft-thresholding

Hard-thresholding

Solving Sparsity-Regularized LS

- How to choose λ?
  - For denoising there are certain techniques, e.g. Stein's unbiased risk estimate (SURE)
  - With more complicated **H** → usually empirical

- **Recap:** $$\arg\min_{\boldsymbol{\theta}} \frac{1}{2}||\mathbf{y} - \boldsymbol{\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_1$$
  - Solution is $$\hat{\theta}_k = \max(|y_k| - \lambda, 0) \cdot \mathrm{sign}(y_k) \triangleq S_\lambda(y_k)$$

Solving Sparsity-Regularized LS

- Now let's go back to

$$\arg\min_{\boldsymbol{\theta}} ||\mathbf{y} - \mathbf{A}\boldsymbol{\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_1 \qquad (\mathbf{A} = \mathbf{HW})$$

- This is of the form (more generally)

$$\arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \qquad \text{with } f(\cdot) \text{ convex}$$

- No closed form solution necessarily… Simplest solution is gradient descent
  - Start at $\boldsymbol{\theta}^{(0)}$, move in the negative direction of the gradient

$$\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)} - \eta_k \nabla f(\boldsymbol{\theta}^{(k-1)})$$

k: iteration number

  - Then

$$f(\boldsymbol{\theta}^{(0)}) > f(\boldsymbol{\theta}^{(1)}) > \ldots$$

For appropriate choices of η_k
(+ some details)

Gradient Descent

- Not easy in our case
  - Derivative of the l_1 norm is not continuous
- Let's reinterpret gradient descent
  - Consider the quadratic approximation

$$f(\mathbf{z}) = f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \mathbf{z} - \boldsymbol{\theta} \rangle + \frac{1}{2\eta} ||\mathbf{z} - \boldsymbol{\theta}||_2^2$$

replaces Taylor series

  - Then

$$\min_{\mathbf{z}} f(\mathbf{z}) = \min_{\mathbf{z}} \left\{ f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \mathbf{z} - \boldsymbol{\theta} \rangle + \frac{1}{2\eta} ||\mathbf{z} - \boldsymbol{\theta}||_2^2 \right\}$$

$$= \min_{\mathbf{z}} \left\{ 2\eta \langle \nabla f(\boldsymbol{\theta}), \mathbf{z} \rangle + ||\mathbf{z}||_2^2 - 2 \langle \boldsymbol{\theta}, \mathbf{z} \rangle \right\}$$

Gradient Descent

$$\min_{\mathbf{z}} f(\mathbf{z}) = \min_{\mathbf{z}} \left\{ f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \mathbf{z} - \boldsymbol{\theta} \rangle + \frac{1}{2\eta} ||\mathbf{z} - \boldsymbol{\theta}||_2^2 \right\}$$

$$= \min_{\mathbf{z}} \left\{ ||\mathbf{z}||_2^2 - 2\langle \boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}), \mathbf{z} \rangle \right\}$$

$$= \min_{\mathbf{z}} \left\{ ||\mathbf{z}||_2^2 - 2\langle \boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}), \mathbf{z} \rangle + ||\boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta})||_2^2 \right\}$$

does not depend on **z**

$$= \min_{\mathbf{z}} \left|\left| \mathbf{z} - \left(\boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}) \right) \right|\right|_2^2$$

same as our gradient step

Proximal Gradient Descent

- Now we will use this to get at a method called the proximal gradient descent

$$\arg\min_{\boldsymbol{\theta}} ||\mathbf{y} - \mathbf{A}\boldsymbol{\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_1$$

is (also) of the form

$$\arg\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) + h(\boldsymbol{\theta}) \qquad \text{(both convex)}$$

- We do the quadratic approximation on g, but not on h

$$\min_{\boldsymbol{\theta}} \left\{ g(\mathbf{z}) + \langle \nabla g(\mathbf{z}), \boldsymbol{\theta} - \mathbf{z} \rangle + \frac{1}{2\eta}||\boldsymbol{\theta} - \mathbf{z}||_2^2 + h(\boldsymbol{\theta}) \right\}$$

$$= \min_{\boldsymbol{\theta}} \frac{1}{2\eta} \left|\left| \boldsymbol{\theta} - \big(\mathbf{z} - \eta \nabla g(\mathbf{z})\big) \right|\right|_2^2 + h(\boldsymbol{\theta})$$

Proximal Gradient Descent

- This objective function looks like denoising with $h(\cdot)$ as regularizer

$$\min_{\boldsymbol{\theta}} \frac{1}{2\eta} \left|\left| \boldsymbol{\theta} - \big(\mathbf{z} - \eta \nabla g(\mathbf{z})\big) \right|\right|_2^2 + h(\boldsymbol{\theta})$$

- Proximal operator is exactly this!
  - It's the solution to

$$\text{prox}_{h,\eta}(\mathbf{u}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\eta} ||\boldsymbol{\theta} - \mathbf{u}||_2^2 + h(\boldsymbol{\theta})$$

- Thus, we set iterations as

$$\boldsymbol{\theta}^{(t)} = \text{prox}_{h,\eta}\big(\boldsymbol{\theta}^{(t-1)} - \eta \nabla g(\boldsymbol{\theta}^{(t-1)})\big)$$

  - Also note we know how to solve the proximal operator for $h(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_1$ (soft-thresholding)

Proximal Gradient Descent

- Overall strategy can be described in two steps

$$\mathbf{z}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \eta \nabla g(\boldsymbol{\theta}^{(t-1)}) \longleftarrow \text{data fidelity}$$

$$\boldsymbol{\theta}^{(t)} = \mathrm{prox}_{h,\eta}(\mathbf{z}^{(t)}) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}||\boldsymbol{\theta} - \mathbf{z}^{(t)}||_2^2 + \eta h(\boldsymbol{\theta}) \longleftarrow \text{regularization/ proximal}$$

- For our problem

$$g(\boldsymbol{\theta}) = ||\mathbf{y} - \mathbf{A}\boldsymbol{\theta}||_2^2 \qquad \nabla g(\boldsymbol{\theta}) = -\mathbf{A}^T(\mathbf{y} - \mathbf{A}\boldsymbol{\theta})$$

$$h(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_1 \quad \Longrightarrow \quad \mathrm{prox}_{h,\eta}(\mathbf{z}) = S_\eta(\mathbf{z})$$

- Thus, overall

$$\mathbf{z}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \eta \mathbf{A}^T(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}^{(t-1)})$$

$$\boldsymbol{\theta}^{(t)} = S_\eta(\mathbf{z}^{(t)})$$