ENGG5781 Matrix Analysis and Computations
Lecture 6: Least Squares Revisited

Wing-Kin (Ken) Ma

2022-23 First Term

Department of Electronic Engineering
The Chinese University of Hong Kong

Lecture 6: Least Squares Revisited

- Part I: regularization
- Part II: sparsity
  - ℓ_0 minimization
  - greedy pursuit, ℓ_1 minimization, and variations
  - majorization-minimization for ℓ_2–ℓ_1 minimization
  - dictionary learning
- Part III: LS with errors in $\mathbf{A}$
  - total LS
  - robust LS, and its equivalence to regularization

Part I: Regularization

Sensitivity to Noise

- Question: how sensitive is the LS solution when there is noise?

- Model:

$$\mathbf{y} = \mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\nu},$$

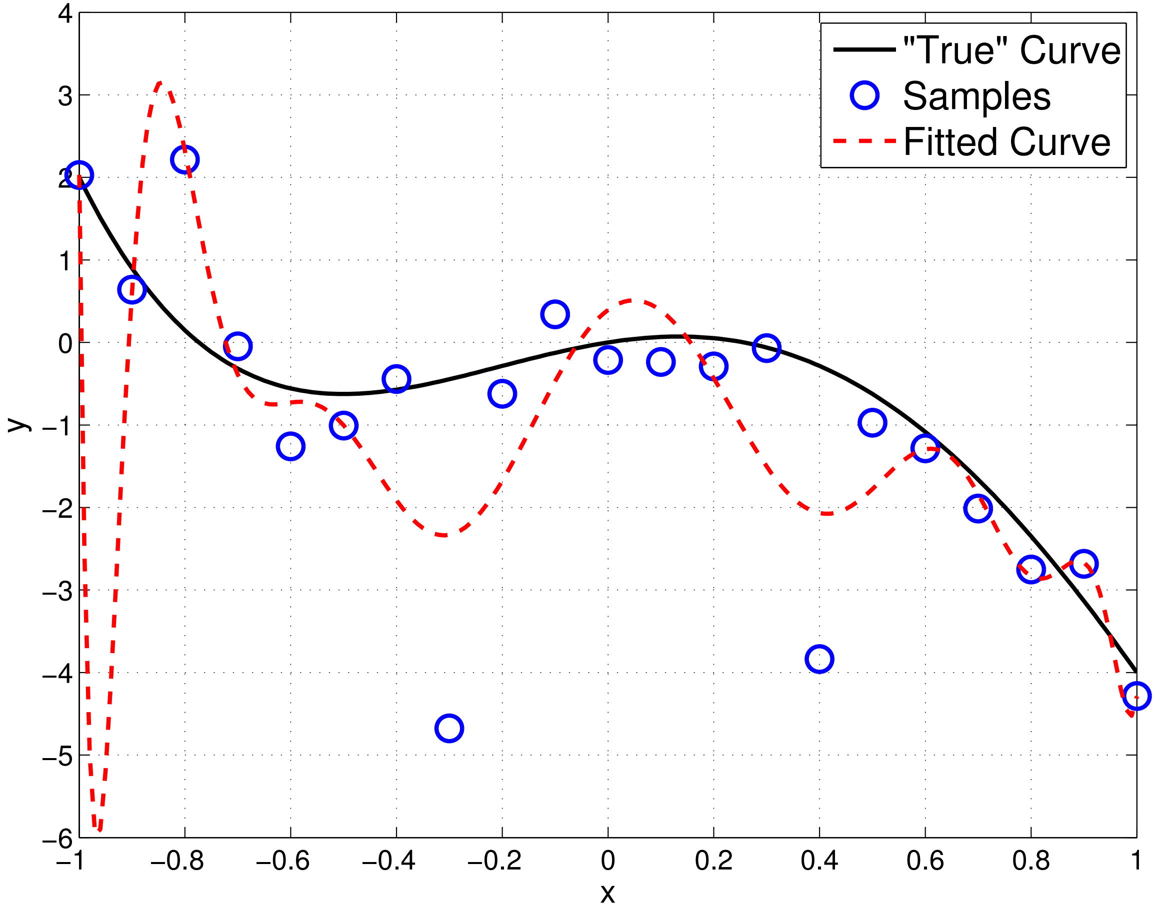
where $\bar{\mathbf{x}}$ is the true result; $\mathbf{A} \in \mathbb{R}^{m \times n}$ has full column rank; $\boldsymbol{\nu}$ is noise, modeled as a random vector with mean zero and covariance $\gamma^2 \mathbf{I}$.

- Mean square error (MSE) analysis: from $\mathbf{x}_{\sf LS} = \mathbf{A}^\dagger \mathbf{y} = \bar{\mathbf{x}} + \mathbf{A}^\dagger \boldsymbol{\nu}$ we get

$$\begin{aligned}
\mathrm{E}[\|\mathbf{x}_{\sf LS} - \bar{\mathbf{x}}\|_2^2] &= \mathrm{E}[\|\mathbf{A}^\dagger \boldsymbol{\nu}\|_2^2] = \mathrm{E}[\mathrm{tr}(\mathbf{A}^\dagger \boldsymbol{\nu}\boldsymbol{\nu}^T (\mathbf{A}^\dagger)^T)] = \mathrm{tr}(\mathbf{A}^\dagger \mathrm{E}[\boldsymbol{\nu}\boldsymbol{\nu}^T](\mathbf{A}^\dagger)^T] \\
&= \gamma^2 \mathrm{tr}(\mathbf{A}^\dagger (\mathbf{A}^\dagger)^T) = \gamma^2 \mathrm{tr}((\mathbf{A}^T\mathbf{A})^{-1}) \\
&= \gamma^2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2(\mathbf{A})}
\end{aligned}$$

- Observation: the MSE becomes very large if some $\sigma_i(\mathbf{A})$'s are close to zero.

Toy Demonstration: Curve Fitting



The same curve fitting example in Lecture 2. The "true" curve is the true $f(x)$ with model order $n = 4$. In practice, the model order may not be known and we may have to guess. The fitted curve above is done by LS with a guessed model order $n = 16$.

ℓ_2-Regularized LS

- Intuition: replace $\mathbf{x}_{\mathsf{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$ by

$$\mathbf{x}_{\mathsf{RLS}} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y},$$

  for some $\lambda > 0$, where the term $\lambda\mathbf{I}$ is added to improve the system conditioning, thereby attempting to reduce noise sensitivity

- how may we make sense out of such a modification?

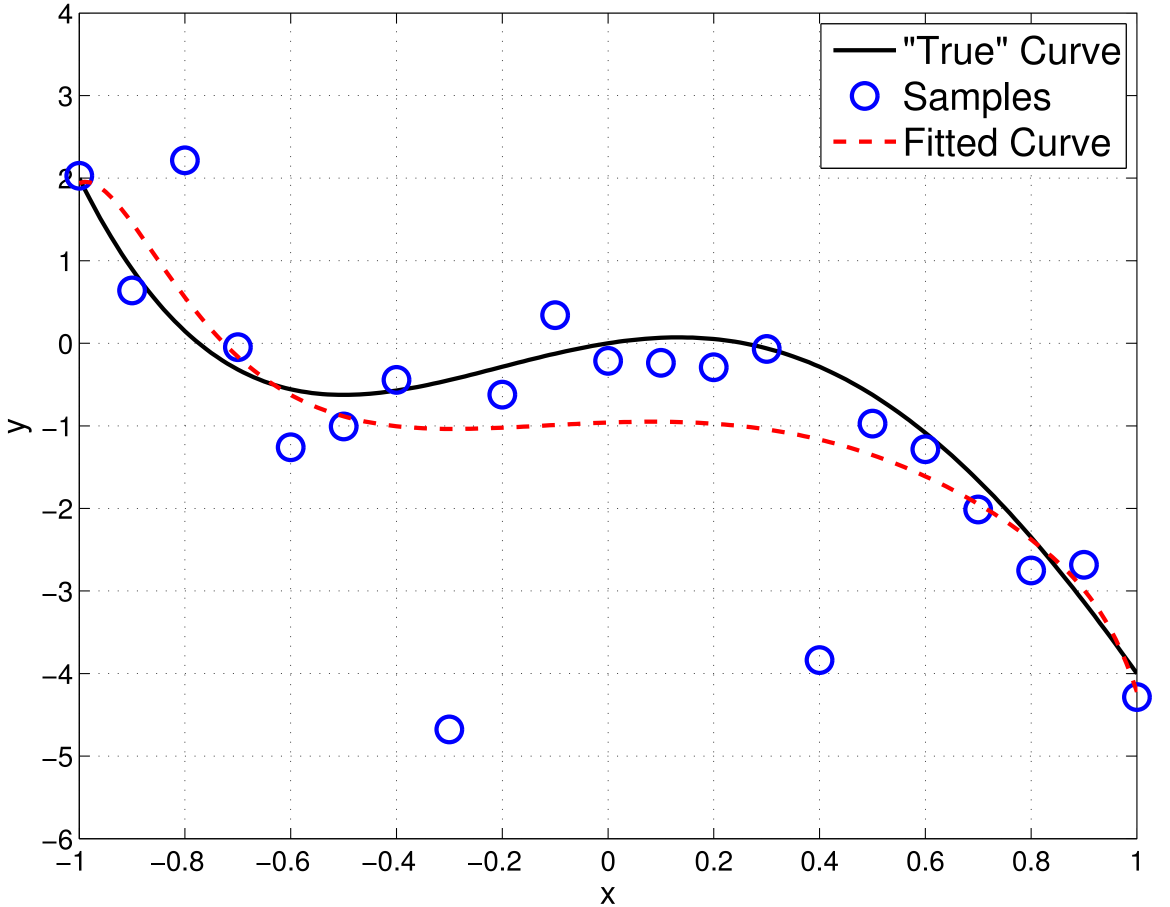
- ℓ_2-regularized LS: find an $\mathbf{x}$ that solves

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$$

  for some pre-determined $\lambda > 0$.

  - the solution is uniquely given by $\mathbf{x}_{\mathsf{RLS}} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y}$
  - the formulation says that we try to minimize both $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ and $\|\mathbf{x}\|_2^2$, and λ controls which one should be more emphasized in the minimization

Toy Demonstration: Curve Fitting



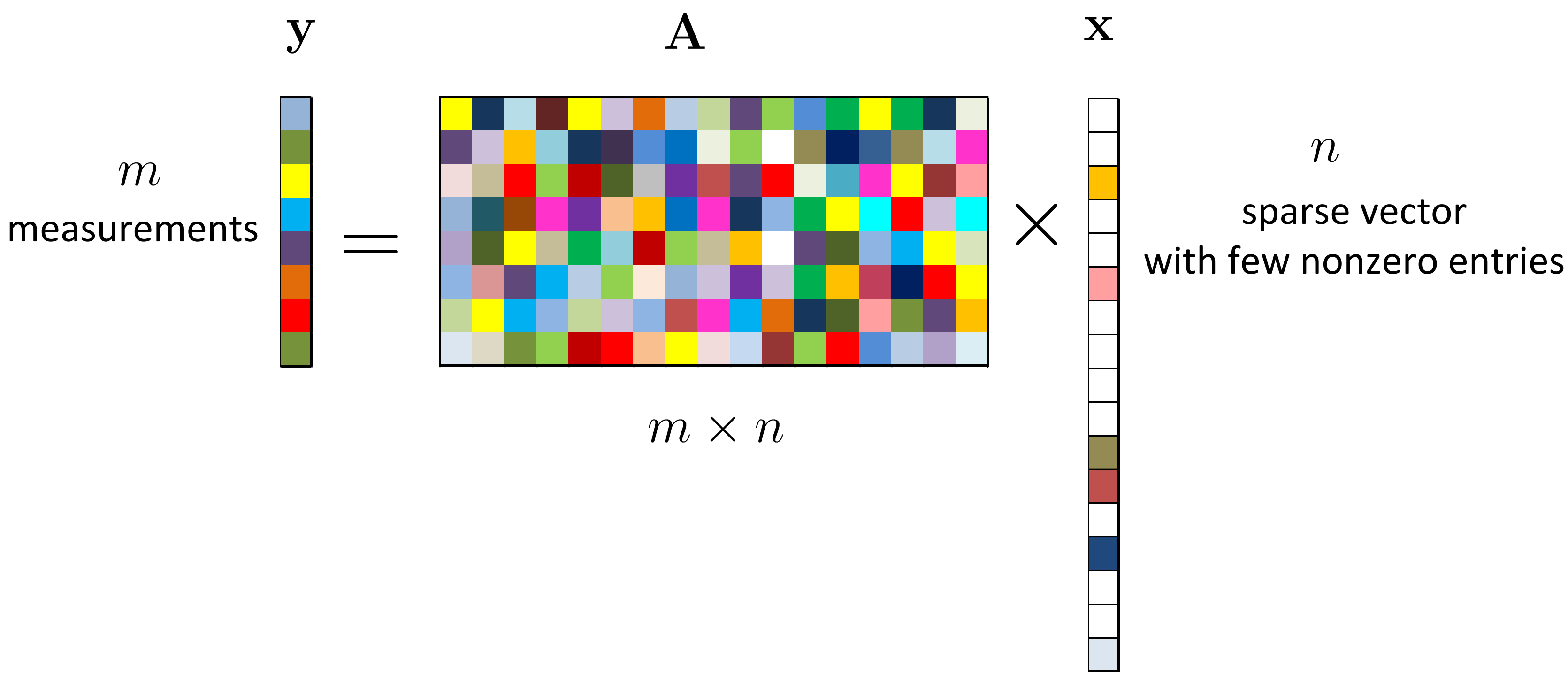
The fitted curve is done by ℓ_2-regularized LS with a guessed model order $n = 18$ and with $\lambda = 0.1$.

Part II: Sparsity

The Sparse Recovery Problem

Problem: given $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m < n$, find a sparsest $\mathbf{x} \in \mathbb{R}^n$ such that

$$\mathbf{y} = \mathbf{A}\mathbf{x}.$$



- by sparsest, we mean that $\mathbf{x}$ should have as many zero elements as possible.

A Sparsity Optimization Formulation

- let
$$\|\mathbf{x}\|_0 = \sum_{i=1}^{n} \mathbb{1}\{x_i \neq 0\}$$
denote the cardinality function
  - commonly called the "ℓ_0-norm", though it is not a norm.

- Minimum ℓ_0-norm formulation:
$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \ & \|\mathbf{x}\|_0 \\ \text{s.t.} \ & \mathbf{y} = \mathbf{A}\mathbf{x}. \end{aligned}$$

- Question: suppose that $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ is the vector we seek to recover. Can the min. ℓ_0-norm problem recover $\bar{\mathbf{x}}$ in an exact and unique fashion?
  - an answer lies in the notion of spark, which may be seen as a strong definition of rank

Spark

Spark: the spark of $\mathbf{A}$, denoted by $\mathrm{spark}(\mathbf{A})$, is the smallest number of linearly dependent columns of $\mathbf{A}$.

- let $\mathrm{spark}(\mathbf{A}) = k$. Then, k is the smallest number such that there exists a linearly dependent $\{\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_k}\}$ for some $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$[1].
- let $\mathrm{spark}(\mathbf{A}) = r + 1$. Then, $\{\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_r}\}$ is linearly independent for any $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}$
  - any collection of r columns of $\mathbf{A}$ is linearly independent, simply stated
- Comparison with rank: Let $\mathrm{rank}(\mathbf{A}) = r$ (not the same r above). Then, there exists a linearly independent $\{\mathbf{a}_{i_1}, \ldots, \mathbf{a}_{i_r}\}$ for some $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}$.
- Kruskal rank: this is an alternative definition of rank. The Kruskal rank of $\mathbf{A}$, denoted by $\mathrm{krank}(\mathbf{A})$, has its definition equivalent to $\mathrm{krank}(\mathbf{A}) = \mathrm{spark}(\mathbf{A}) - 1$.

[1] We leave it implicit that $i_k \neq i_j$ for any $k \neq j$.

Spark

- if any collection of m vectors in $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \subseteq \mathbb{R}^m$, with $n > m$, is linearly independent, then

$$\text{spark}(\mathbf{A}) = m + 1, \quad \text{rank}(\mathbf{A}) = m.$$

  - an example is Vandemonde matrices with distinct roots
  - some specifically designed bases also have this property

- but there also exist instances in which rank and spark are very different
  - let $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\} \in \mathbb{R}^m$ be linearly independent, and let $\mathbf{A} = [\ \mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_1\ ]$.
  - we have $\text{rank}(\mathbf{A}) = r$, but $\text{spark}(\mathbf{A}) = 2$

- to conclude, spark may be seen as a stronger definition of rank, and

$$\text{spark}(\mathbf{A}) - 1 \leq \text{rank}(\mathbf{A})$$

Perfect Recovery Guarantee of the Min. ℓ_0-Norm Problem

Theorem 6.1. Suppose that $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$. Then, $\bar{\mathbf{x}}$ is the unique solution to the minimum ℓ_0-norm problem if

$$\|\bar{\mathbf{x}}\|_0 < \frac{1}{2}\text{spark}(\mathbf{A}).$$

- Implication: if $\bar{\mathbf{x}}$ is sufficiently sparse, then the minimum ℓ_0-norm problem perfectly recovers $\bar{\mathbf{x}}$

- Proof sketch:

  1. let $\mathbf{x}^\star$ be a solution to the min. ℓ_0-norm problem. Let $\mathbf{e} = \bar{\mathbf{x}} - \mathbf{x}^\star$.
  2. $\mathbf{0} = \mathbf{A}\bar{\mathbf{x}} - \mathbf{A}\mathbf{x}^\star = \mathbf{A}\mathbf{e}$; $\|\mathbf{e}\|_0 \leq \|\bar{\mathbf{x}}\|_0 + \|\mathbf{x}^\star\|_0 \leq 2\|\bar{\mathbf{x}}\|_0$.
  3. suppose $\mathbf{e} \neq \mathbf{0}$. Then, $\mathbf{A}\mathbf{e} = \mathbf{0}, \|\mathbf{e}\|_0 \leq 2\|\bar{\mathbf{x}}\|_0 \implies \text{spark}(\mathbf{A}) \leq 2\|\bar{\mathbf{x}}\|_0$

Perfect Recovery Guarantee of the Min. ℓ_0-Norm Problem

- coherence: the coherence of $\mathbf{A}$ is defined as

$$\mu(\mathbf{A}) = \max_{j \neq k} \frac{|\mathbf{a}_j^T \mathbf{a}_k|}{\|\mathbf{a}_j\|_2 \|\mathbf{a}_k\|_2}.$$

  - measures how similar the columns of $\mathbf{A}$ are in the worst-case sense.

- a weak version of Theorem 6.1:

  Corollary 6.1. Suppose that $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$. Then, $\bar{\mathbf{x}}$ is the unique solution to the minimum ℓ_0-norm problem if

$$\|\bar{\mathbf{x}}\|_0 < \frac{1}{2}(1 + \mu(\mathbf{A})^{-1}).$$

  - Implication: perfect recovery may depend on how incoherent $\mathbf{A}$ is.
  - proof idea: show that $\mathrm{spark}(\mathbf{A}) \geq 1 + \mu(\mathbf{A})^{-1}$

On Solving the Minimum ℓ_0-Norm Problem

Question: How should we solve the minimum ℓ_0-norm problem

$$\begin{aligned} \min_{\mathbf{x}} \ & \|\mathbf{x}\|_0 \\ \text{s.t.} \ & \mathbf{y} = \mathbf{A}\mathbf{x}, \end{aligned}$$

or can it be efficiently solved?

- ℓ_0-norm minimization does not lead to a simple solution as in 2-norm min.
- the minimum ℓ_0-norm problem is NP-hard in general
  - what does that mean?
    - $*$ given any $\mathbf{y}, \mathbf{A}$, the problem is unlikely to be exactly solvable in polynomial time (i.e., in a complexity of $\mathcal{O}(n^p)$ for any $p > 0$)

Brute Force Search for the Minimum ℓ_0-Norm Problem

- notation: $\mathbf{A}_\mathcal{I}$ denotes a submatrix of $\mathbf{A}$ obtained by keeping the columns indicated by $\mathcal{I}$

- we may solve the ℓ_0-norm minimization problem via brute force search:

> **input:** $\mathbf{A}, \mathbf{y}$
> for all $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ do
>   if $\mathbf{y} = \mathbf{A}_\mathcal{I}\tilde{\mathbf{x}}$ has a solution for some $\tilde{\mathbf{x}} \in \mathbb{R}^{|\mathcal{I}|}$
>   record $(\tilde{\mathbf{x}}, \mathcal{I})$ as one of candidate solutions
> end
> **output:** a candidate solution $(\tilde{\mathbf{x}}, \mathcal{I})$ whose $|\mathcal{I}|$ is the smallest

- example: for $n = 3$, we test $\mathcal{I} = \{1\}, \mathcal{I} = \{2\}, \mathcal{I} = \{3\}, \mathcal{I} = \{1, 2\}, \mathcal{I} = \{2, 3\}, \mathcal{I} = \{1, 3\}, \mathcal{I} = \{1, 2, 3\}$

- manageable for very small n, too expensive even for moderate n

- how about a greedy search that searches less?

Greedy Pursuit

- consider a greedy search called the orthogonal matching pursuit (OMP)

Algorithm: OMP
input: $\mathbf{A}, \mathbf{y}$
set $\mathcal{I} = \emptyset$, $\hat{\mathbf{x}} = \mathbf{0}$
repeat

$$\mathbf{r} = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}$$

$$k = \arg \max_{j \in \{1,\ldots,n\}} |\mathbf{a}_j^T \mathbf{r}| / \|\mathbf{a}_j\|_2$$

$$\mathcal{I} := \mathcal{I} \cup \{k\}$$

$$\hat{\mathbf{x}} := \arg \min_{\mathbf{x} \in \mathbb{R}^n, \; x_i = 0 \; \forall i \notin \mathcal{I}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$$

until a stopping rule is satisfied, e.g., $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$ is sufficiently small
output: $\hat{\mathbf{x}}$

- note: there are many other greedy search strategies

Perfect Recovery Guarantee of Greedy Pursuit

- again, a key question is conditions under which OMP admits perfect recovery

- there are many such theoretical conditions, not only for OMP but also for other greedy algorithms

- one such result is as follows:

Theorem 6.2. Suppose that $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$. Then, OMP recovers $\bar{\mathbf{x}}$ if

$$\|\bar{\mathbf{x}}\|_0 < \frac{1}{2}(1 + \mu(\mathbf{A})^{-1}).$$

  – proof idea: show that OMP is guaranteed to pick a correct column at every stage.

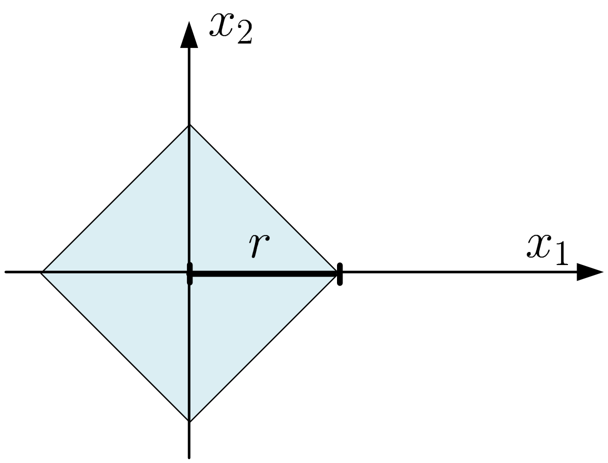
Convex Relexation

Another approximation approach is to replace $\|\mathbf{x}\|_0$ by a convex function:

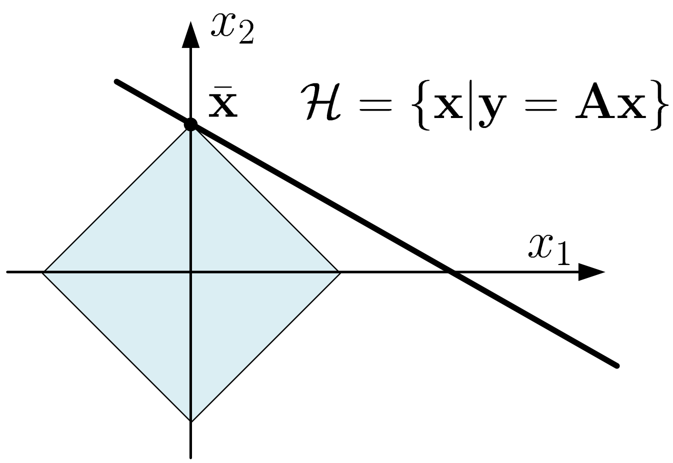
$$\begin{aligned} \min_{\mathbf{x}} \ & \|\mathbf{x}\|_1 \\ \text{s.t.} \ & \mathbf{y} = \mathbf{A}\mathbf{x}. \end{aligned}$$

- also known as basis pursuit in the literature
- convex, a linear program
- no closed-form solution (while the minimum 2-norm problem has)
- but the success of this minimum 1-norm problem, both in theory and practice, has motivated a large body of work on computationally efficient algorithms for it

Illustration of 1-Norm Geometry



(A)

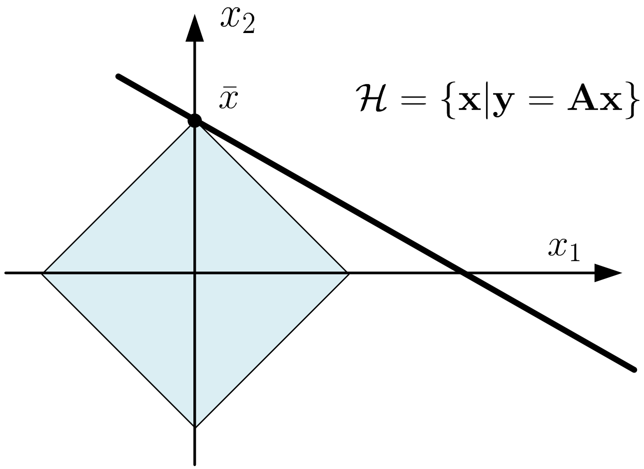


(B)

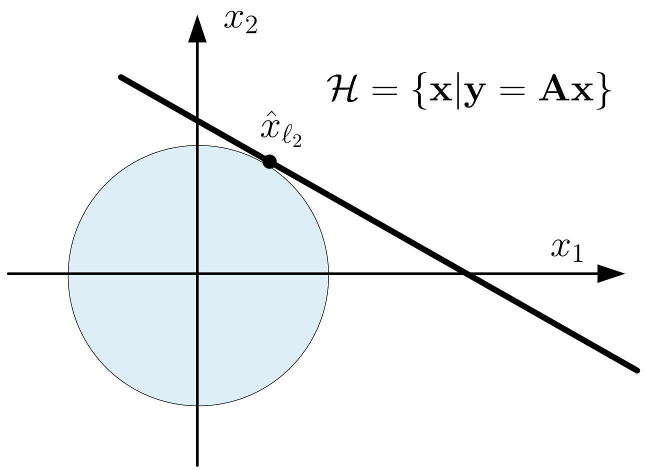
- Fig. A shows the 1-norm ball of radius r in $\mathbb{R}^2$. Note that the 1-norm ball ball is "pointy" along the axes.

- Fig. B shows the 1-norm recovery solution. The point $\bar{\mathbf{x}}$ is a "sparse" vector; the line $\mathcal{H}$ is the set of all $\mathbf{x}$ that satisfy $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Illustration of 1-Norm Geometry



(B)



(C)

- The 1-norm recovery problem is to pick out a point in $\mathcal{H}$ that has the minimum 1-norm. We can see that $\bar{\mathbf{x}}$ is such a point.

- Fig. C shows the geometry when 2-norm is used. We can see that the solution $\hat{\mathbf{x}}$ may not be sparse.

Perfect Recovery Guarantee of the Min. 1-Norm Problem

- again, researchers studied conditions under which the minimum 1-norm problem admits perfect recovery

- this has been an exciting topic, with many provable conditions such as the restricted isometry property (RIP), the nullspace property (NSP), ...
  - see the literature for details, and here is one: **[Yin'13]**

- a simple one is as follows:

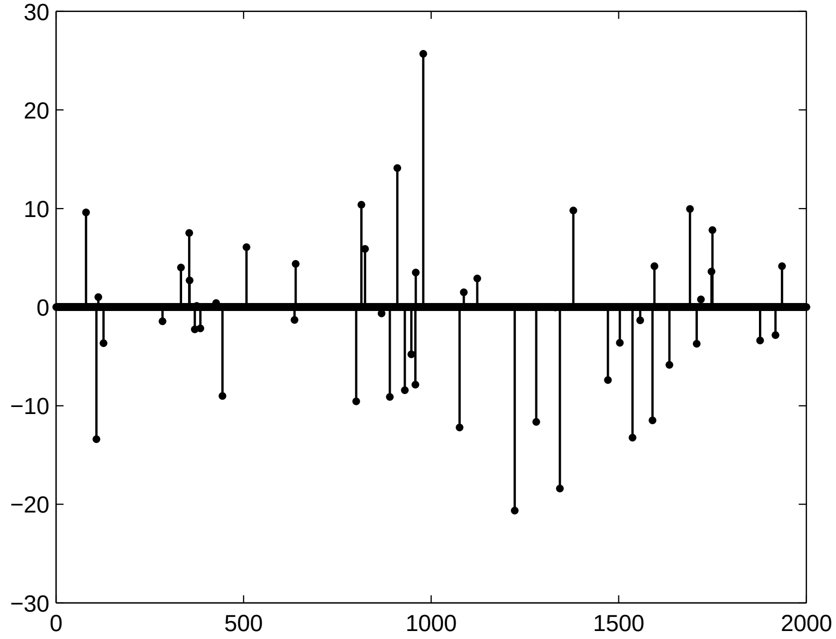
Theorem 6.3. Suppose that $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$. Then, $\bar{\mathbf{x}}$ is the unique solution to the minimum 1-norm problem if

$$\|\bar{\mathbf{x}}\|_0 < \frac{1}{2}(1 + \mu(\mathbf{A})^{-1}).$$

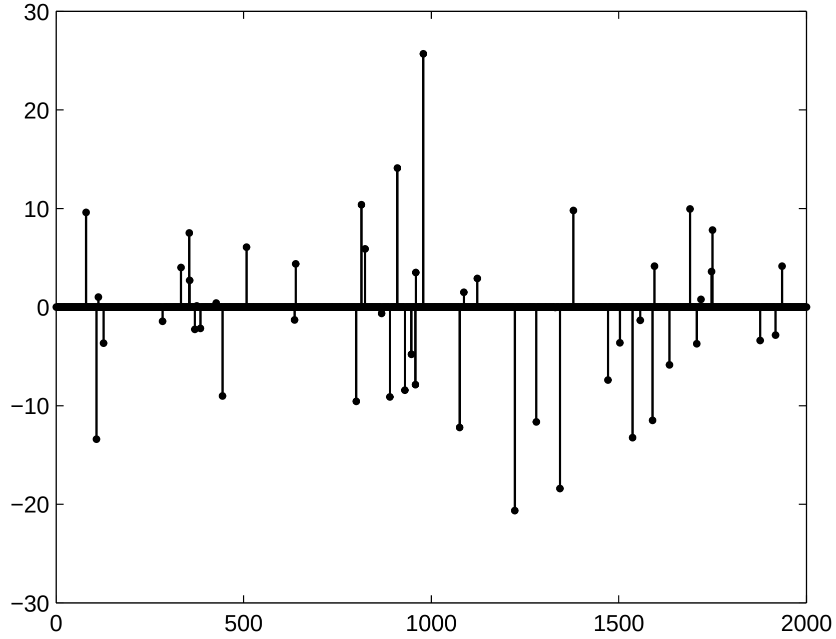
Toy Demonstration: Sparse Signal Reconstruction

- Sparse vector $\mathbf{x} \in \mathbb{R}^n$ with $n = 2000$ and $\|\mathbf{x}\|_0 = 50$.

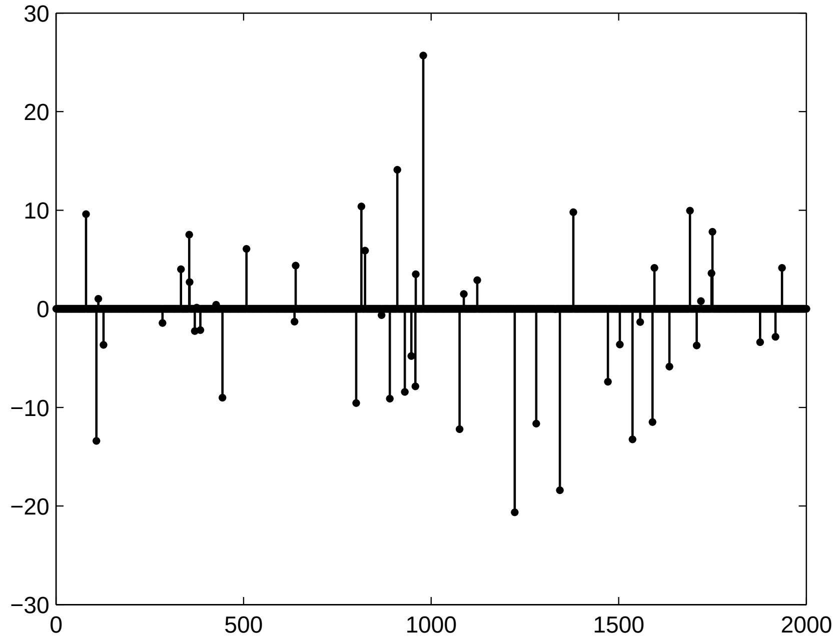
- $m = 400$ noise-free observations of $\mathbf{y} = \mathbf{A}\mathbf{x}$, a_{ij} is randomly generated.



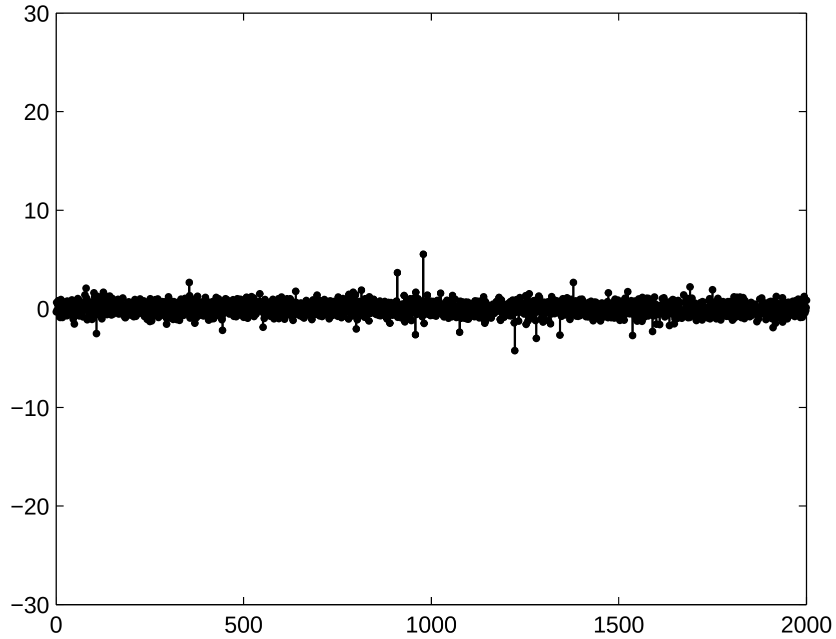
(a) Sparse source signal



(b) Recovery by 1-norm minimization



(c) Sparse source signal



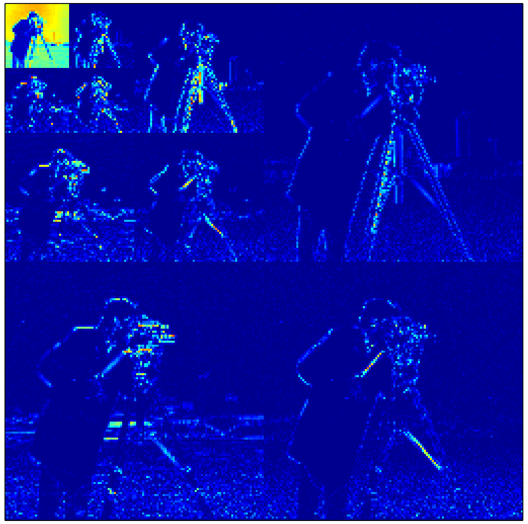
(d) Recovery by 2-norm minimization

Application: Compressive sensing (CS)

- Consider a signal $\tilde{\mathbf{x}} \in \mathbb{R}^n$ that has a sparse representation $\mathbf{x} \in \mathbb{R}^n$ in the domain of $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ (e.g. DCT or wavelet), i.e.,

$$\tilde{\mathbf{x}} = \boldsymbol{\Psi}\mathbf{x},$$

where $\mathbf{x}$ is sparse.



Left: the original image $\tilde{\mathbf{x}}$. Right: the corresponding coefficient $\mathbf{x}$ in the wavelet domain, which is sparse. Source: **[Romberg-Wakin'07]**

Application: CS

- To acquire $\mathbf{x}$, we use a sensing matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ to observe $\mathbf{x}$

$$\mathbf{y} = \mathbf{\Phi}\tilde{\mathbf{x}} = \mathbf{\Phi}\mathbf{\Psi}\mathbf{x}.$$

  Here, we have $m \ll n$, i.e., much few observations than the no. of unknowns

- Such a $\mathbf{y}$ will be good for compression, transmission and storage.

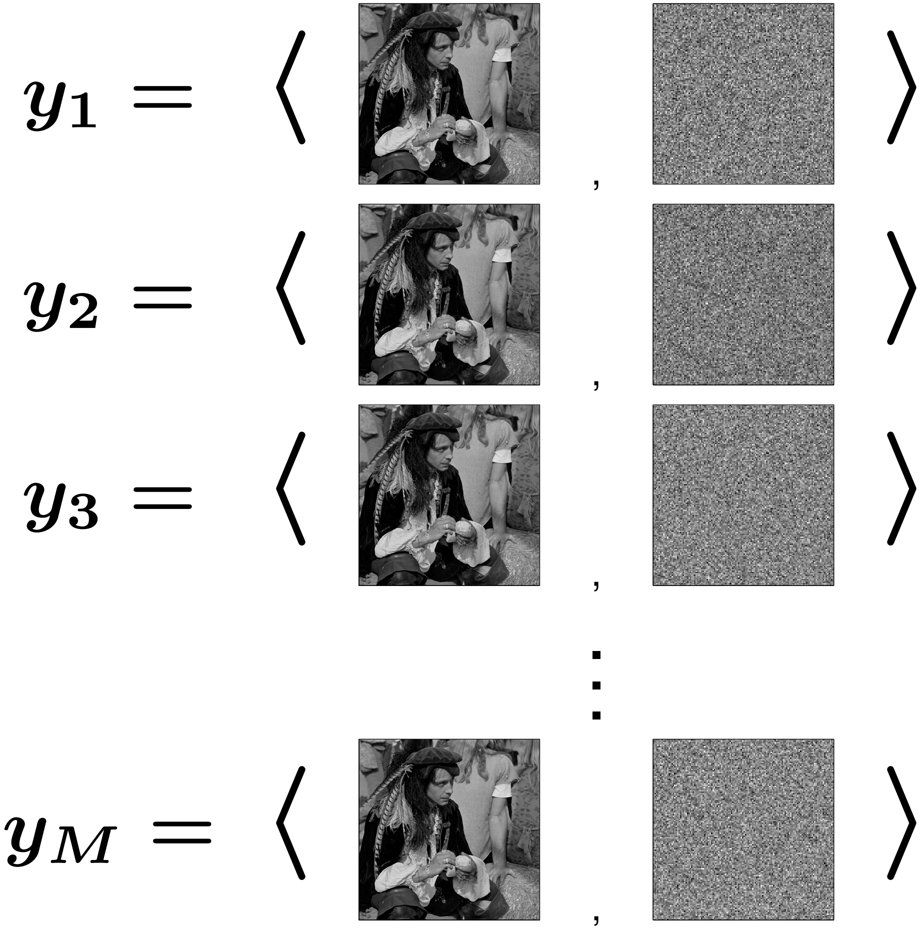
- $\tilde{\mathbf{x}}$ is recovered by recovering $\mathbf{x}$:

$$\begin{aligned} &\min \ \|\mathbf{x}\|_0 \\ &\ \text{s.t.}\ \mathbf{y} = \mathbf{A}\mathbf{x}, \end{aligned}$$

  where $\mathbf{A} = \mathbf{\Phi}\mathbf{\Psi}$

- how to choose $\mathbf{\Phi}$? CS research suggests that i.i.d. random $\mathbf{\Phi}$ will work well!

Application: CS



(a) measurements via i.i.d. random $\mathbf{\Phi}$

original (25k wavelets)

(b) original image

perfect recovery

(c) ℓ_1 recovery

Source: **[Romberg-Wakin'07]**

Variations

- when $\mathbf{y}$ is contaminated by noise, or when $\mathbf{y} = \mathbf{A}\mathbf{x}$ does not exactly hold, some variants of the previous min. 1-norm formulation may be considered:
  - basis pursuit denoising: given $\epsilon > 0$, solve

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \leq \epsilon$$

  - ℓ_1-regularized LS: given $\lambda > 0$, solve

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

  - Lasso: given $\tau > 0$, solve

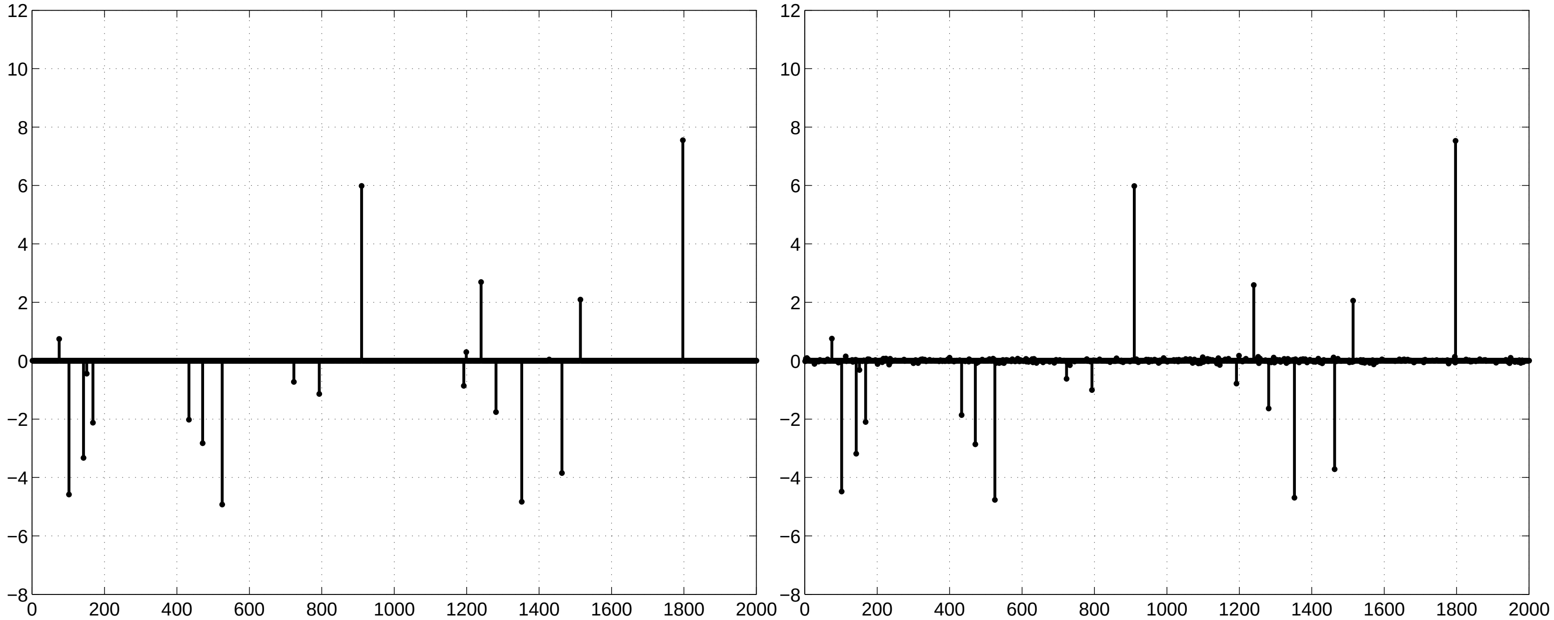
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_1 \leq \tau$$

- when outliers exist in $\mathbf{y}$ (i.e., some elements of $\mathbf{y}$ are badly corrupted), we also want the residual $\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{x}$ to be sparse; so,

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 + \lambda \|\mathbf{x}\|_1.$$

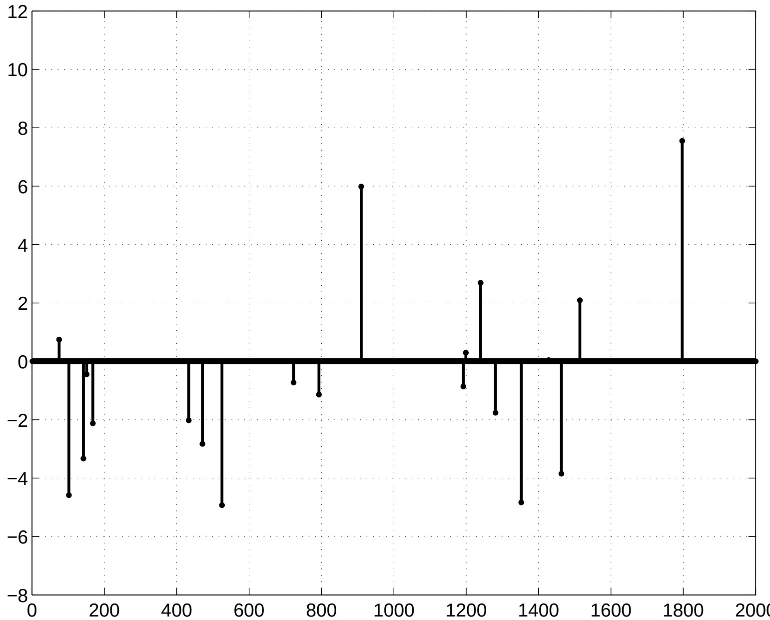
Toy Demonstration: Noisy Sparse Signal Reconstruction

- Sparse signal $\mathbf{x} \in \mathbb{R}^n$ with $n = 2000$ and $\|\mathbf{x}\|_0 = 20$.
- $m = 400$ noisy observations of $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\nu}$, both a_{ij} and ν_i are randomly generated.
- 1-norm regularized LS $\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$ is used. $\lambda = 0.1$.

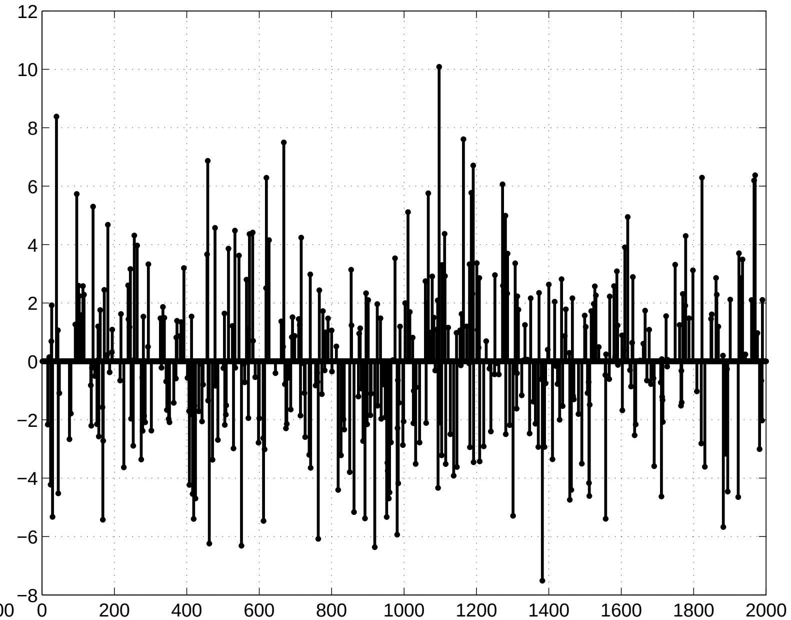


(a) Sparse source signal

(b) 1-norm regularized LS estimate

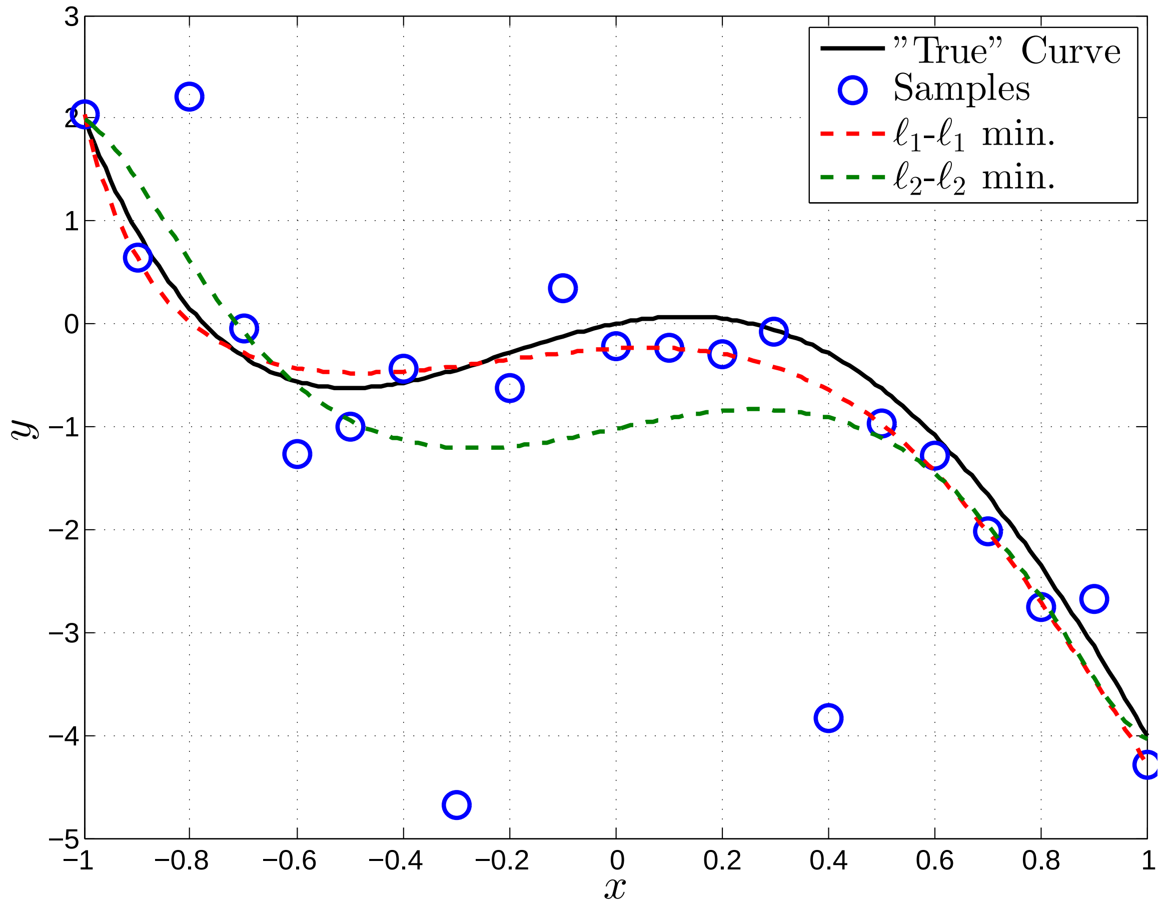


(c) Sparse source signal



(d) LS estimate

Toy Demonstration: Curve Fitting



The same curve fitting problem in Lecture 2. The guessed model order is $n = 18$.

ℓ_2-ℓ_2 min.: $\min \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2$

ℓ_1-ℓ_1 min.: $\min \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 + \lambda \|\mathbf{x}\|_1$

Total Variation (TV) Denoising

- Scenario:
  - estimate $\mathbf{x} \in \mathbb{R}^n$ from a noisy measurement $\mathbf{x}_{\rm cor} = \mathbf{x} + \boldsymbol{\nu}$.
  - $\mathbf{x}$ is known to be piecewise linear, i.e., for most i we have

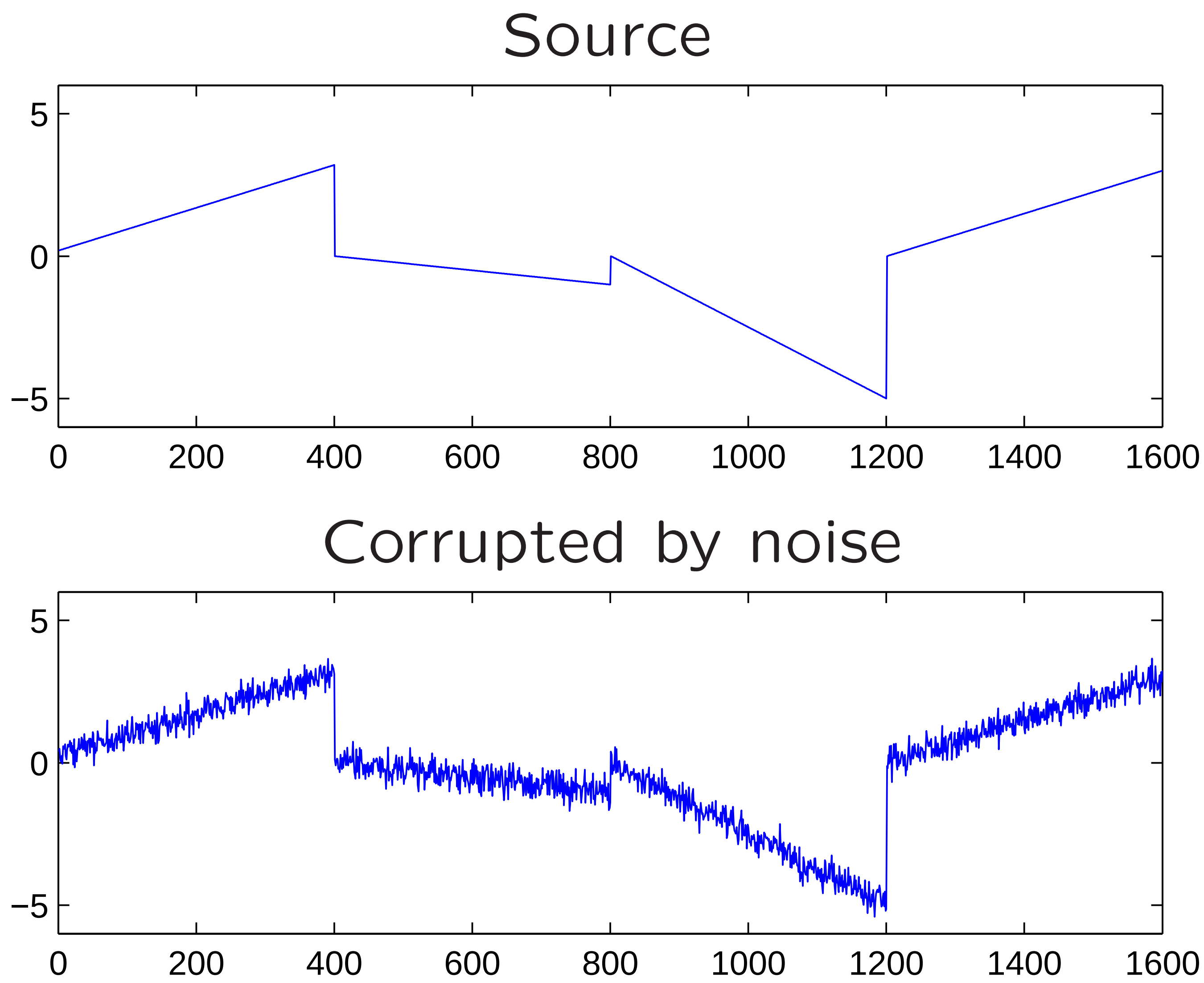
$$x_i - x_{i-1} = x_{i+1} - x_i \Longleftrightarrow -x_{i+1} + 2x_i - x_{i+1} = 0.$$

  - equivalently, $\mathbf{D}\mathbf{x}$ is sparse, where

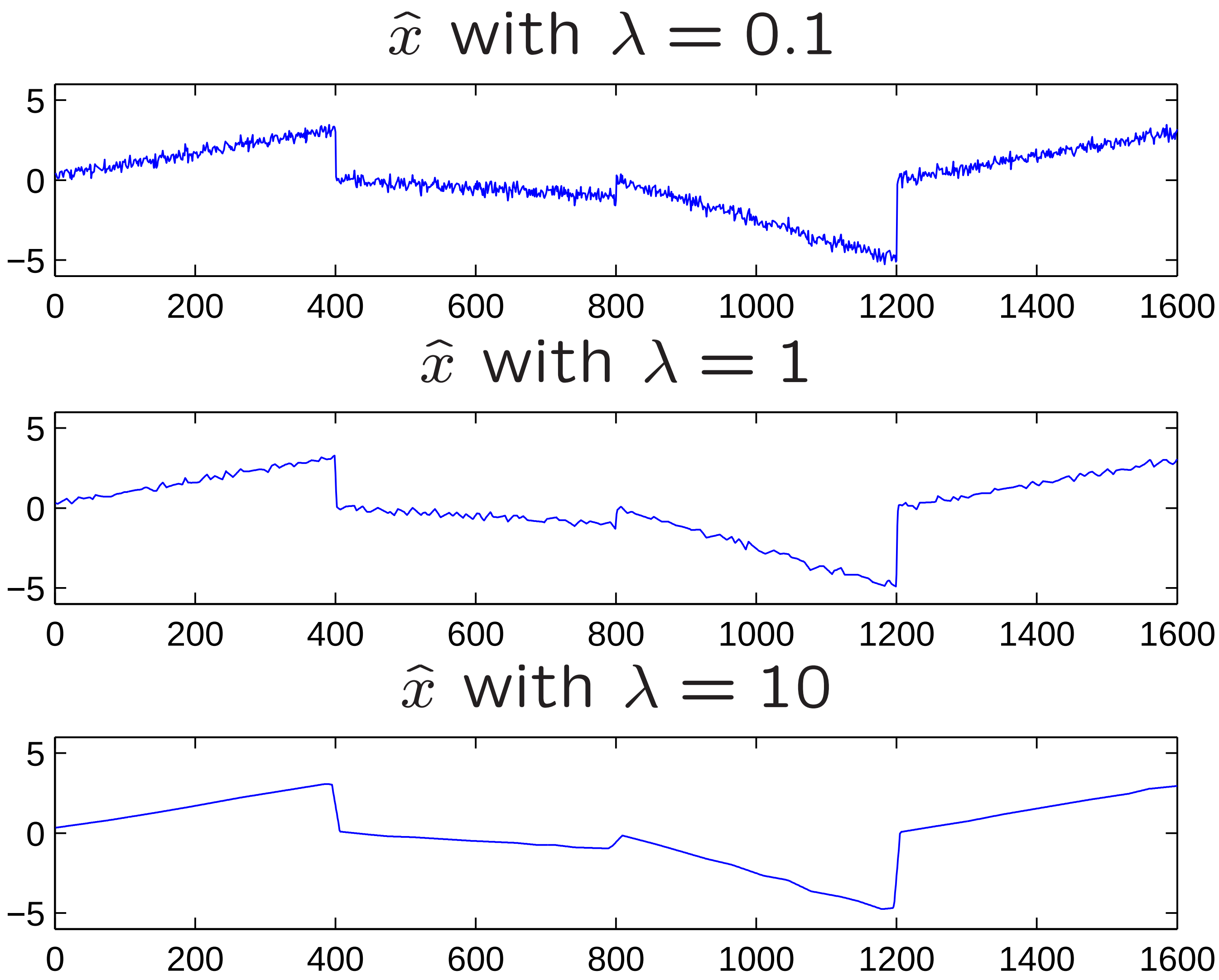
$$\mathbf{D} = \begin{bmatrix} -1 & 2 & 1 & 0 & \dots \\ 0 & -1 & 2 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & -1 & 2 & 1 \end{bmatrix}.$$

- TV denoising: estimate $\mathbf{x}$ by solving

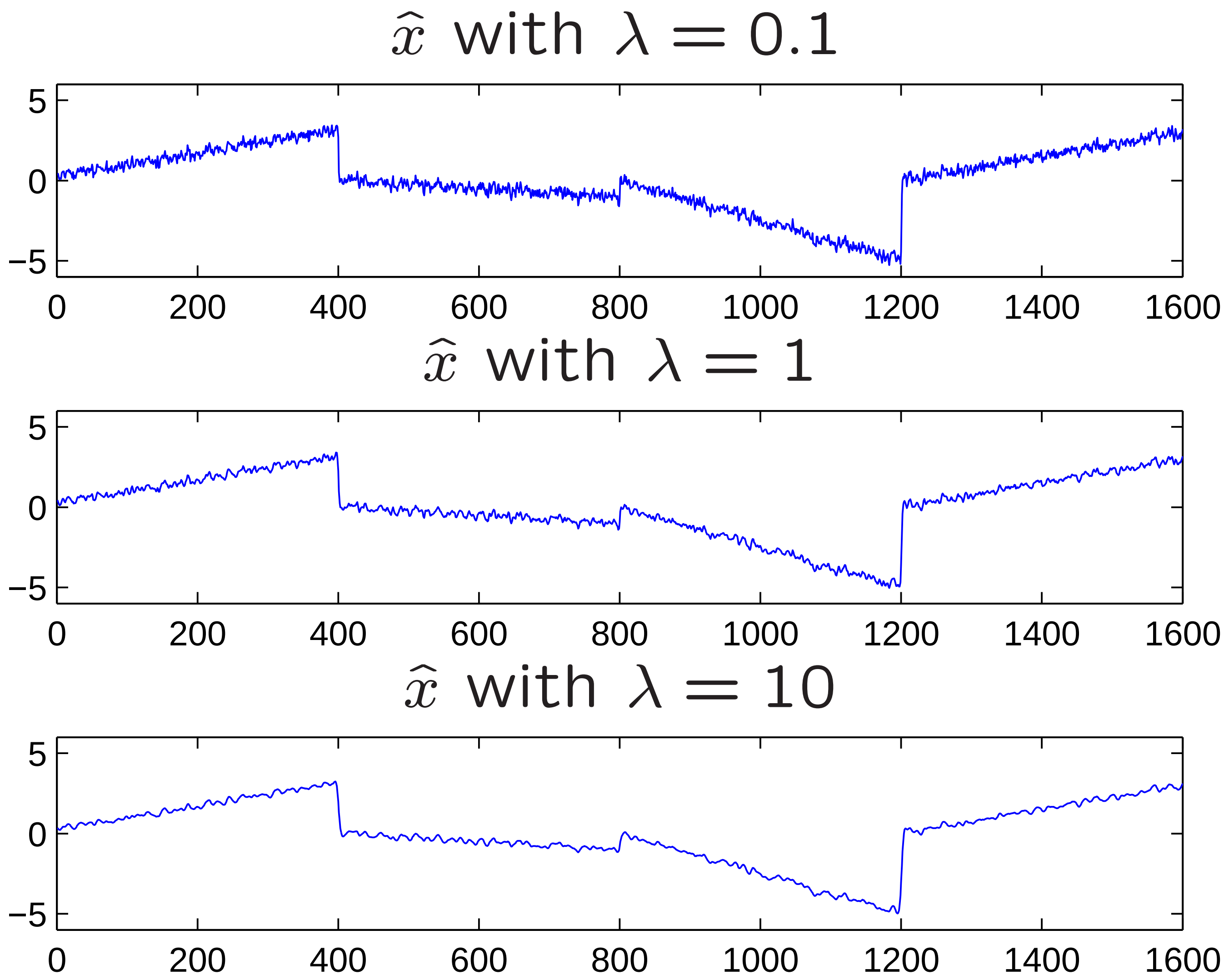
$$\min_{\mathbf{x}} \|\mathbf{x}_{\rm cor} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{D}\mathbf{x}\|_1$$



Original $\mathbf{x}$ and corrupted $\mathbf{x}_{\rm cor}$



TV denoised signals for various λ's.

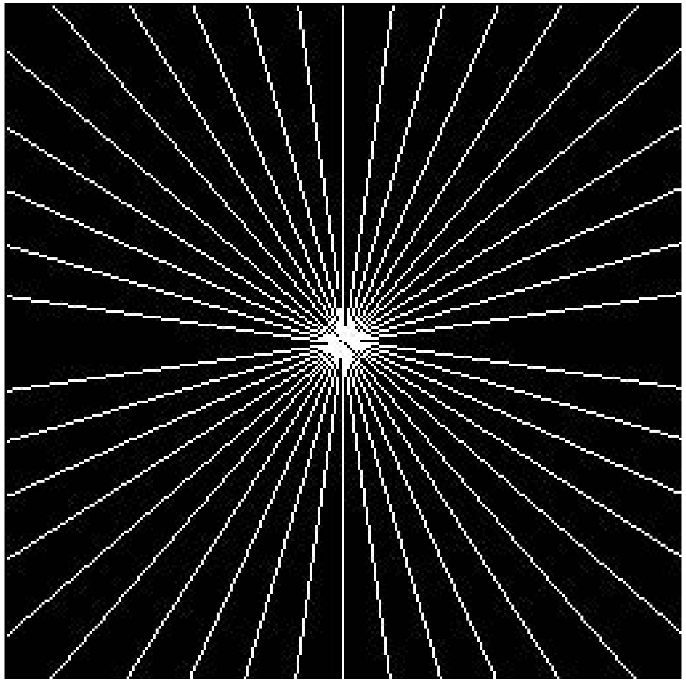


TV denoised signals via ℓ_2 regularization and for various λ's.

Application: Magnetic Resonance Imaging (MRI)

Problem: MRI image reconstruction.

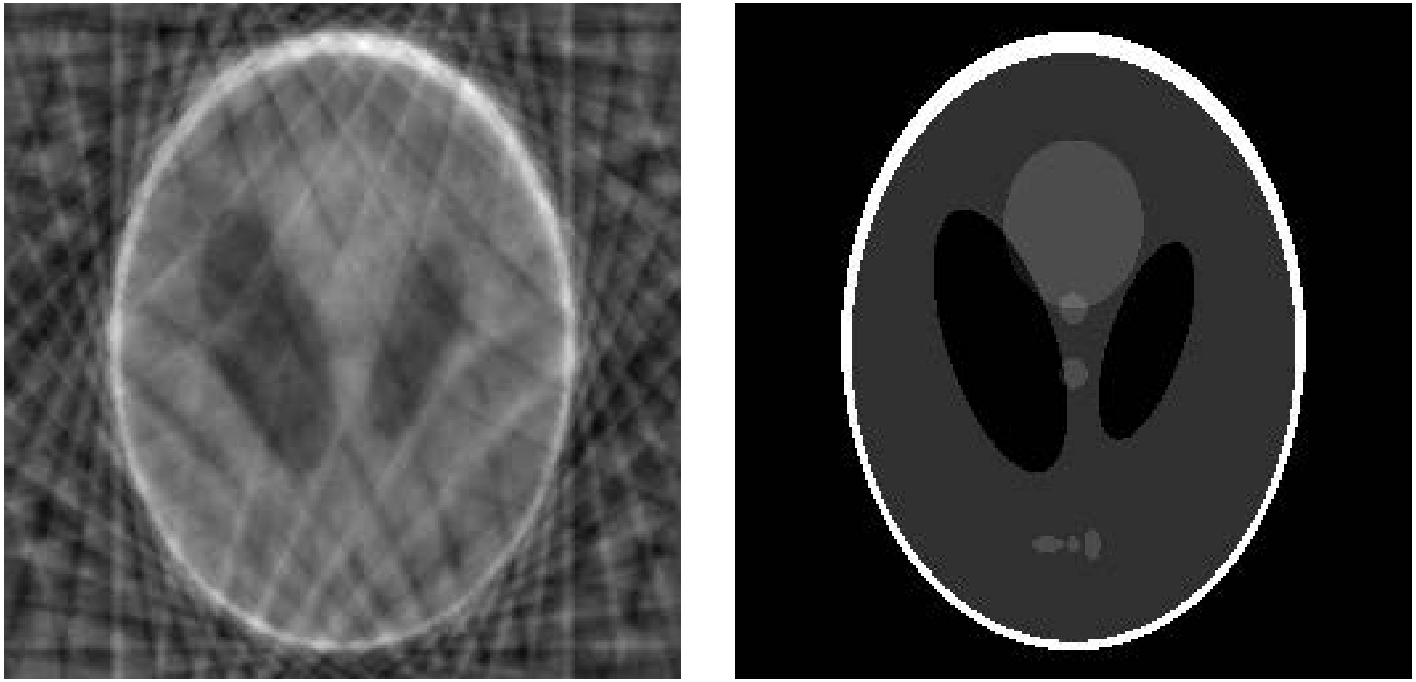
(a)



(b)

Fig. a shows the original test image. Fig. b shows the sampling region in the frequency domain. Fourier coefficients are sampled along 22 approximately radial lines. Source: **[Candès-Romberg-Tao'06]**

Application: MRI



(c) (d)

Fig. c is the recovery by filling the unobserved Fourier coefficients to zero. Fig. d is the recovery by a TV minimization problem. Source: **[Candès-Romberg-Tao'06]**

Efficient Computations of the $\ell_2 - \ell_1$ Minimization Solution

- consider the $\ell_2 - \ell_1$ minimization problem

$$\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

- as mentioned, the problem is convex and there are many optimization algorithms custom-designed for it
  - some keywords for such algorithms: majorization-minimization (MM), ADMM, fast proximal gradient (or the so-called FISTA), Frank-Wolfe,...
- Aim: get some flavor of one particular algorithm, namely, MM, that is sufficiently "matrix" and is suitable for large-scale problems

MM for $\ell_2 - \ell_1$ Minimization: LS as an Example

- to see the insight of MM, we start with the plain old LS

$$\min_{\mathbf{x}} \ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2.$$

- observe that for a given $\bar{\mathbf{x}}$, one has

$$\begin{aligned} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 &= \|\mathbf{y} - \mathbf{A}\bar{\mathbf{x}} - \mathbf{A}(\mathbf{x} - \bar{\mathbf{x}})\|_2^2 \\ &= \|\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}\|_2^2 - 2(\mathbf{x} - \bar{\mathbf{x}})^T\mathbf{A}^T(\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}) + \|\mathbf{A}(\mathbf{x} - \bar{\mathbf{x}})\|_2^2 \\ &\leq \|\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}\|_2^2 - 2(\mathbf{x} - \bar{\mathbf{x}})^T\mathbf{A}^T(\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}) + c\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \end{aligned}$$

  for any $\mathbf{x} \in \mathbb{R}^n$ and for any $c \geq \sigma_{\max}^2(\mathbf{A})$

MM for $\ell_2 - \ell_1$ Minimization: LS as an Example

- let $c \geq \sigma_{\max}^2(\mathbf{A})$, and let

$$g(\mathbf{x}, \bar{\mathbf{x}}) = \|\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}\|_2^2 - 2(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{A}^T (\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}) + c\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2$$

- we have

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \leq g(\mathbf{x}, \bar{\mathbf{x}}), \quad \text{for any } \mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^n$$
$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 = g(\mathbf{x}, \mathbf{x}), \quad \text{for any } \mathbf{x} \in \mathbb{R}^n$$

- also,

$$\arg \min_{\mathbf{x} \in \mathbb{R}^n} \; g(\mathbf{x}, \bar{\mathbf{x}}) = \tfrac{1}{c}\mathbf{A}^T(\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}) + \bar{\mathbf{x}}$$

- Idea: given an initial point $\mathbf{x}^{(0)}$, do

$$\mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \; g(\mathbf{x}, \mathbf{x}^{(k)}) = \tfrac{1}{c}\mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^{(k)}) + \mathbf{x}^{(k)}, \quad k = 1, 2, \ldots$$

  - note: not very interesting at this moment as the above iteration is the same as gradient descent with step size $1/c$

MM for $\ell_2 - \ell_1$ Minimization: General MM Principle

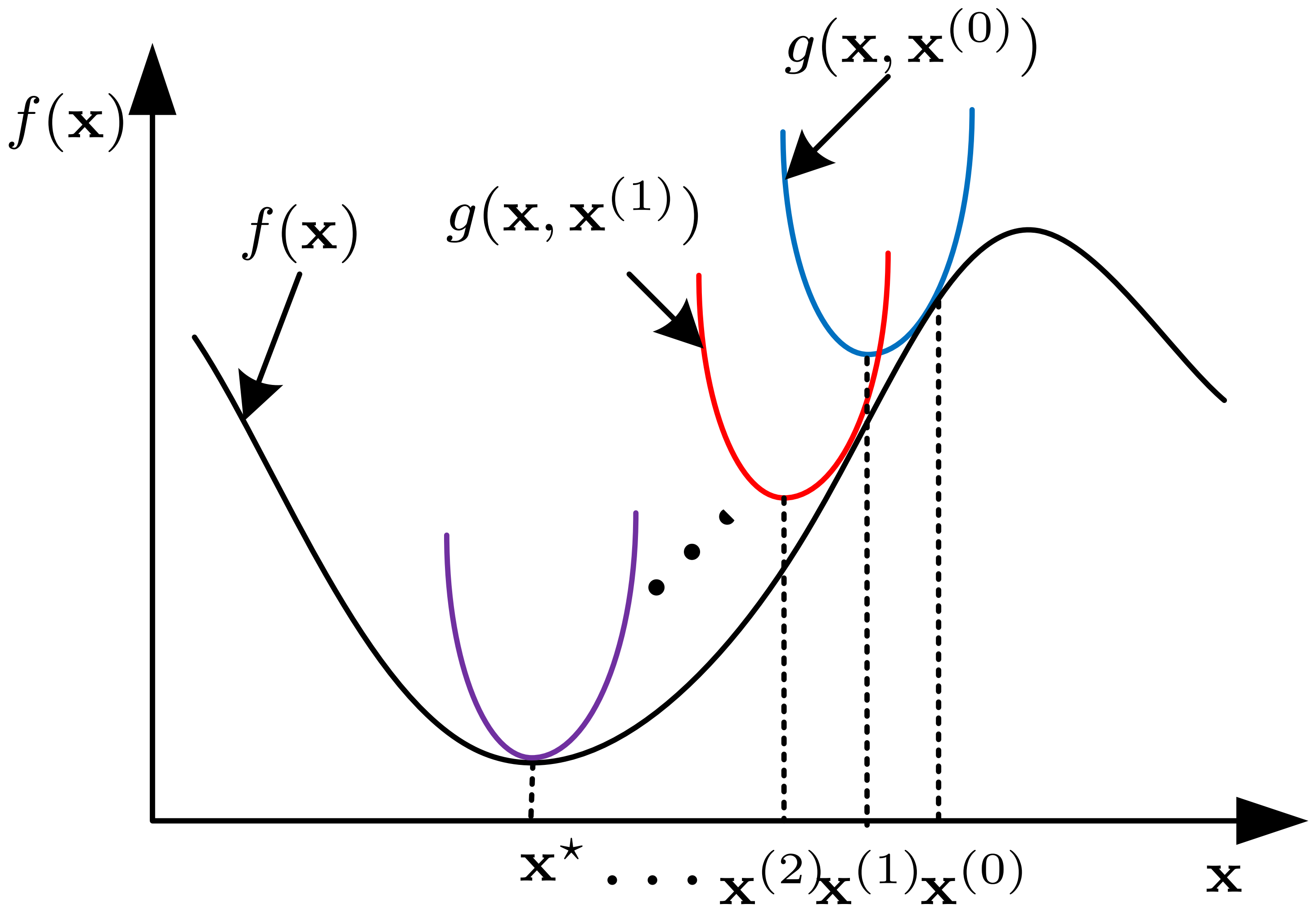
- the example shown above is an instance of MM
- general MM principle:
  - consider a general optimization problem

    $$\min_{\mathbf{x} \in \mathcal{C}} \; f(\mathbf{x})$$

    and suppose that f is hard to minimize directly
  - let $g(\mathbf{x}, \bar{\mathbf{x}})$ be a surrogate function that is easy to minimize and satisfies

    $$f(\mathbf{x}) \leq g(\mathbf{x}, \bar{\mathbf{x}}) \text{ for all } \mathbf{x}, \bar{\mathbf{x}}, \qquad f(\mathbf{x}) = g(\mathbf{x}, \mathbf{x}) \text{ for all } \mathbf{x}$$
  - MM algorithm: $\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x} \in \mathcal{C}} \; g(\mathbf{x}, \mathbf{x}^{(k)}), k = 1, 2, \ldots$
  - as a basic result, $f(\mathbf{x}^{(0)}) \geq f(\mathbf{x}^{(1)}) \geq f(\mathbf{x}^{(2)}) \ldots$
  - suppose that f is convex and $\mathcal{C}$ is convex. MM is guaranteed to converge to an optimal solution under some mild assumption **[Razaviyayn-Hong-Luo'13]**

MM for $\ell_2 - \ell_1$ Minimization: General MM Principle



MM for $\ell_2 - \ell_1$ Minimization

- now consider applying MM to the $\ell_2 - \ell_1$ minimization problem

$$\min_{\mathbf{x}} \ \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

- let $c \geq \sigma_{\max}^2(\mathbf{A})$, and let

$$g(\mathbf{x}, \bar{\mathbf{x}}) = \tfrac{1}{2}\left(\|\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}\|_2^2 - 2(\mathbf{x} - \bar{\mathbf{x}})^T\mathbf{A}^T(\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}) + c\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2\right) + \lambda\|\mathbf{x}\|_1$$

  - simply plug the same surrogate for $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ we saw previously

- it can be shown that

$$\mathbf{x}^{(k+1)} = \mathrm{soft}\left(\tfrac{1}{c}\mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x}^{(k)}) + \mathbf{x}^{(k)}, \lambda/c\right)$$

  where soft is called the soft-thresholding operator and is defined as follows: if $\mathbf{z} = \mathrm{soft}(\mathbf{x}, \delta)$ then $z_i = \mathrm{sign}(x_i)\max\{|x_i| - \delta, 0\}$

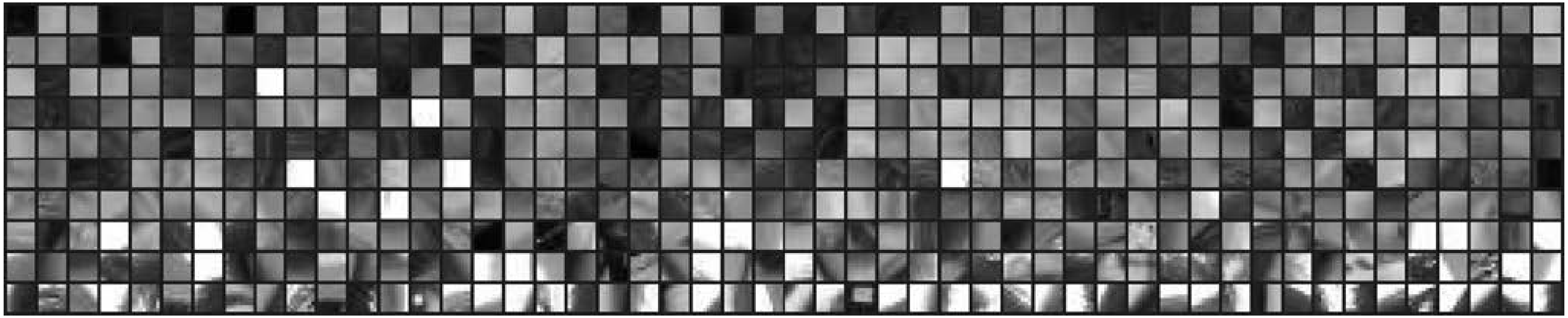
Dictionary Learning

- previously $\mathbf{A}$ is assumed to be given
- how about learning a fat $\mathbf{A}$ from data, as in matrix factorization?
- Dictionary learning (DL): given $\tau > 0$ and $\mathbf{Y} \in \mathbb{R}^{m \times n}$, solve

$$\begin{aligned} \min_{\mathbf{A} \in \mathbb{R}^{m \times k}, \mathbf{B} \in \mathbb{R}^{k \times n}} \ & \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{A}\mathbf{b}_i\|_2^2 \\ \text{s.t.} \ & \|\mathbf{b}_i\|_0 \leq \tau, \quad i = 1, \ldots, n \end{aligned}$$

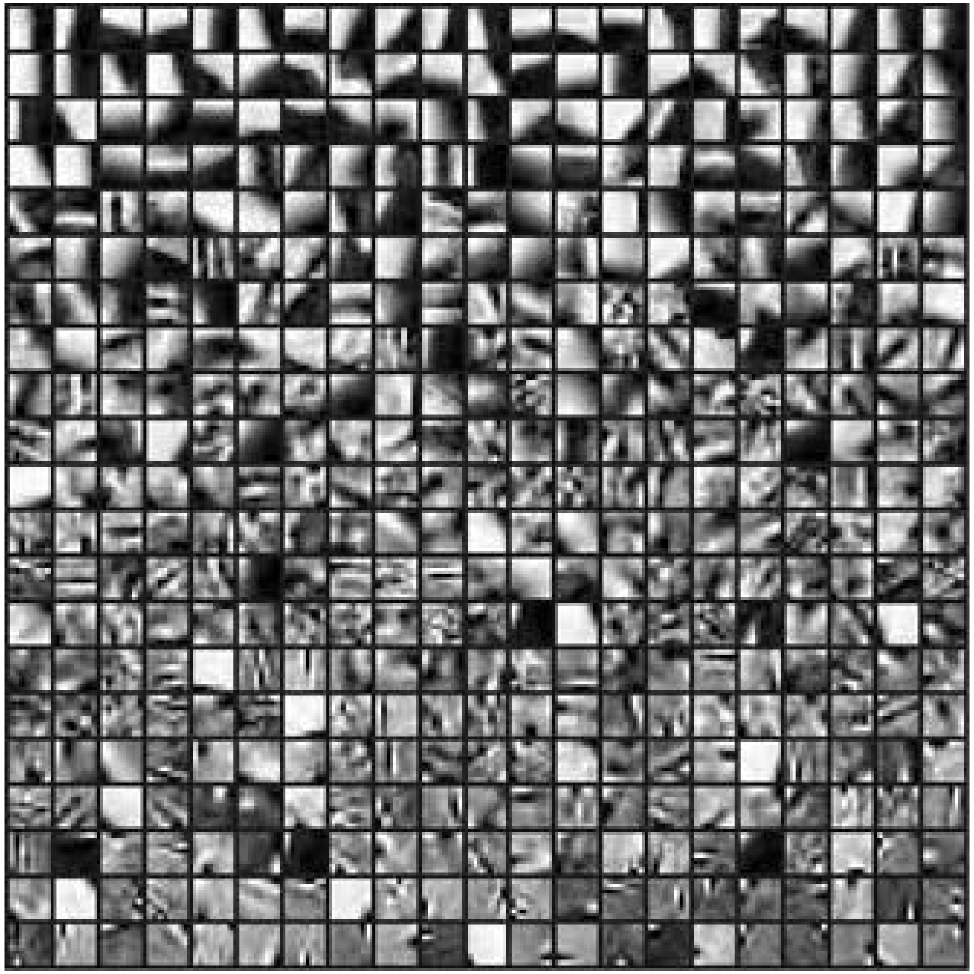
  - DL considers $k \geq m$, and $\mathbf{A}$ is called an overcomplete dictionary
  - DL is handled by alternating optimization—the same approach in matrix fac.

Dictionary Learning



A collection of 500 random image blocks. Source: **[Aharon-Elad-Bruckstein'06]**.

Dictionary Learning



The learned dictionary. Source: **[Aharon-Elad-Bruckstein'06]**.

Part III: LS with Errors in $\mathbf{A}$

LS with Errors in $\mathbf{A}$

- Scenario: errors exist in the system matrix $\mathbf{A}$
- Aim: mitigate the effects of the system matrix errors on the LS solution
- there are many ways to do so, and we look at two
- Total LS (TLS):

$$\min_{\mathbf{x}\in\mathbb{R}^n,\ \mathbf{\Delta}\in\mathbb{R}^{m\times n}} \|\mathbf{y}-(\mathbf{A}+\mathbf{\Delta})\mathbf{x}\|_2^2 + \|\mathbf{\Delta}\|_F^2$$

  – minimally perturb the system matrix for best fitting in the Euclidean sense

- Robust LS :

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{\Delta}\in\mathcal{U}} \|\mathbf{y}-(\mathbf{A}+\mathbf{\Delta})\mathbf{x}\|_2^2$$

  for some pre-determined uncertainty set $\mathcal{U}\subset\mathbb{R}^{m\times n}$

  – robustify the LS via a worst-case means

Total LS

$$\min_{\mathbf{x}\in\mathbb{R}^n,\ \boldsymbol{\Delta}\in\mathbb{R}^{m\times n}} \|\mathbf{y} - (\mathbf{A} + \boldsymbol{\Delta})\mathbf{x}\|_2^2 + \|\boldsymbol{\Delta}\|_F^2$$

- does not seem to have a closed-form solution at first sight
- turns out to have a closed-form solution under some mild assumptions
- assume $\mathbf{A}$ to be of full column rank with $m \geq n+1$
- let $\mathbf{C} = [\ \mathbf{A}\ \mathbf{y}\ ]$, and let $\mathbf{v}_{n+1}$ be the $(n+1)$th right singular value of $\mathbf{C}$. If

$$\mathrm{rank}(\mathbf{C}) = n+1, \qquad v_{n+1,n+1} \neq 0,$$

then

$$\mathbf{x}_{\mathsf{TLS}} = -\frac{1}{v_{n+1,n+1}} \begin{bmatrix} v_{1,n+1} \\ \vdots \\ v_{n,n+1} \end{bmatrix}$$

is a TLS solution

– see **[Golub-Van Loan'12]** for further discussion on issues like $v_{n+1,n+1} \neq 0$

Proof Sketch of the TLS Solution

- idea: turn the TLS problem to a low-rank matrix approximation problem
- by a change of variables

$$\mathbf{C} = [\ \mathbf{A}\ \mathbf{y}\ ] \in \mathbb{R}^{m\times(n+1)}, \qquad \mathbf{D} = [\ \boldsymbol{\Delta}\ (\mathbf{A}+\boldsymbol{\Delta})\mathbf{x}\ ] \in \mathbb{R}^{m\times(n+1)},$$

the TLS problem can be formulated as

$$\min_{\mathbf{x},\mathbf{D}} \|\mathbf{C}-\mathbf{D}\|_F^2 \qquad \text{s.t. } \mathbf{D}\begin{bmatrix}\mathbf{x}\\ -1\end{bmatrix} = \mathbf{0} \tag{$\dagger$}$$

- the constraint in ($\dagger$), together with $m \geq n+1$, implies $\mathrm{rank}(\mathbf{D}) \leq n$
- or, we can equivalently rewrite ($\dagger$) as

$$\min_{\mathbf{x},\mathbf{D}} \|\mathbf{C}-\mathbf{D}\|_F^2 \qquad \text{s.t. } \mathrm{rank}(\mathbf{D}) \leq n,\ \mathbf{D}\begin{bmatrix}\mathbf{x}\\ -1\end{bmatrix} = \mathbf{0}$$

Proof Sketch of the TLS Solution

- consider a *relaxation* of (†):

$$\min_{\mathbf{D}} \|\mathbf{C} - \mathbf{D}\|_F^2 \qquad \text{s.t. } \operatorname{rank}(\mathbf{D}) \le n, \tag{‡}$$

  where we drop the constraint $\mathbf{D}\begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}$

- let $\mathbf{D}^\star$ be a solution to (‡). If there exists an $\mathbf{x}$ such that $\mathbf{D}^\star \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}$, $\mathbf{D}^\star$ is also a solution to (†) and $\mathbf{x}$ is a TLS solution

- let $\mathbf{C} = \sum_{i=1}^{n+1} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ be the SVD

- by the Eckart-Young-Mirsky theorem, a solution to (‡) is $\mathbf{D}^\star = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.

- as a basic fact of SVD, we have $\mathbf{D}^\star \mathbf{v}_{n+1} = \mathbf{0}$.

- thus, if $v_{n+1,n+1} \neq 0$, we have the desired TLS solution

Robust LS

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{\Delta}\in\mathcal{U}} \|\mathbf{y} - (\mathbf{A} + \mathbf{\Delta})\mathbf{x}\|_2$$

- consider the case of $\mathcal{U} = \{\mathbf{\Delta} \in \mathbb{R}^{m\times n} \mid \|\mathbf{\Delta}\|_2 \leq \lambda\}$ for some $\lambda > 0$
- the robust LS problem can be shown to be equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_2$$

- Observations and Implications:
  - the equivalent form of the robust LS is very similar to (but not exactly the same as) the previous ℓ_2-regularized LS
  - robustification is equivalent to regularization
- it can be shown that the same equivalence holds if we replace the uncertainty set by $\mathcal{U} = \{\mathbf{\Delta} \in \mathbb{R}^{m\times n} \mid \|\mathbf{\Delta}\|_F \leq \lambda\}$

Proof Sketch of the Robust LS Equivalence Result

- by the definition of induced norms, we have

$$\|\mathbf{\Delta}\|_2 \leq \lambda \quad \Longleftrightarrow \quad \|\mathbf{\Delta x}\|_2 \leq \lambda \|\mathbf{x}\|_2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

- then, for any $\mathbf{x} \in \mathbb{R}^n$ and for any $\mathbf{\Delta} \in \mathcal{U}$,

$$\begin{aligned} \|\mathbf{y} - (\mathbf{A} + \mathbf{\Delta})\mathbf{x}\|_2 &\leq \|\mathbf{y} - \mathbf{Ax}\|_2 + \|\mathbf{\Delta x}\|_2 \\ &\leq \|\mathbf{y} - \mathbf{Ax}\|_2 + \lambda \|\mathbf{x}\|_2, \end{aligned} \qquad (*)$$

  and note that the 1st equality above holds if $\mathbf{y} - \mathbf{Ax} = -\alpha \mathbf{\Delta x}$ for some $\alpha \geq 0$, and the 2nd equality above holds if $\mathbf{x}$ is the 1st right singular vector of $\mathbf{\Delta}$

- consider the case of $\mathbf{x} \neq \mathbf{0}$, $\mathbf{y} - \mathbf{Ax} \neq \mathbf{0}$. It can be verified that

$$\mathbf{\Delta} = -\frac{\lambda}{\|\mathbf{y} - \mathbf{Ax}\|_2 \|\mathbf{x}\|_2} (\mathbf{y} - \mathbf{Ax}) \mathbf{x}^T$$

  attains the equalities in $(*)$ and lies in $\mathcal{U}$

- the other cases of $\mathbf{x}$ are handled in a similar fashion

More Robust LS Equivalences

- denote $\mathcal{U}_{q,p} = \{ \boldsymbol{\Delta} \in \mathbb{R}^{m \times n} \mid \|\boldsymbol{\Delta}\mathbf{x}\|_p \leq \lambda \|\mathbf{x}\|_q \ \forall \mathbf{x} \}$, where $p, q \geq 1$. We have

$$\min_{\mathbf{x} \in \mathbb{R}^n} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{q,p}} \|\mathbf{y} - (\mathbf{A} + \boldsymbol{\Delta})\mathbf{x}\|_p = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_p + \lambda \|\mathbf{x}\|_q$$

- proof: almost the same as the previous case

- some interesting special cases:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{2,1}} \|\mathbf{y} - (\mathbf{A} + \boldsymbol{\Delta})\mathbf{x}\|_2 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda \|\mathbf{x}\|_1$$

$$\min_{\mathbf{x} \in \mathbb{R}^n} \max_{\substack{\boldsymbol{\Delta} \in \mathbb{R}^{m \times n} \\ \|\boldsymbol{\delta}_i\|_1 \leq \lambda \ \forall i}} \|\mathbf{y} - (\mathbf{A} + \boldsymbol{\Delta})\mathbf{x}\|_1 = \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1 + \lambda \|\mathbf{x}\|_1$$

- Implication: ℓ_1 regularization may also be seen as an act of robustification

- suggested reading: **[Bertsimas-Copenhaver'17]**, including extension to PCA

References

[Yin'13], W. Yin, *Sparse Optimization Lecture: Sparse Recovery Guarantees*, 2013. Available online at http://www.math.ucla.edu/~wotaoyin/summer2013/slides/Lec03_SparseRecoveryGuarantees.pdf

[Romberg-Wakin'07] J. Romberg and M. Walkin, *Compressed Sensing: A tutorial*, in IEEE SSP Workshop, 2017. Available online at http://web.yonsei.ac.kr/nipi/lectureNote/Compressed%20Sensing%20by%20Romberg%20and%20Wakin.pdf

[Candès-Romberg-Tao'06] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[Aharon-Elad-Bruckstein'06] M. Aharon, M.I Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Image Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[Razaviyayn-Hong-Luo'13] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[Golub-Van Loan'12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd edition, JHU Press, 2012.

[Bertsimas-Copenhaver'17] D. Bertsimas and M. S. Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, 2017.