Machine Learning

Regularization, smoothing and sparsity

Emanuele Rodolà
rodola@di.uniroma1.it

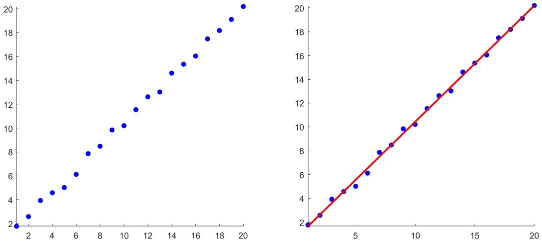
SAPIENZA
UNIVERSITÀ DI ROMA

2nd semester a.y. 2024/2025 · March 18, 2025

Motivation

Linear regression

We have seen fitting problems such as:



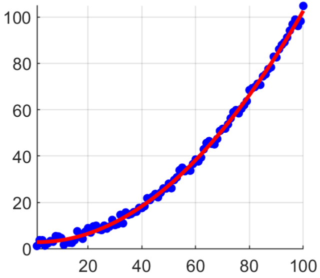
$$y_i = ax_i + b$$

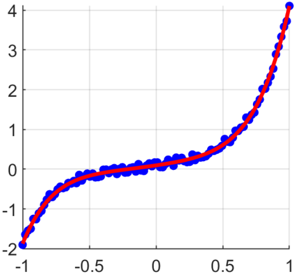
With the minimization problem:

$$\min_{a,b\in\mathbb{R}} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Linear regression

We did polynomial fitting as well:





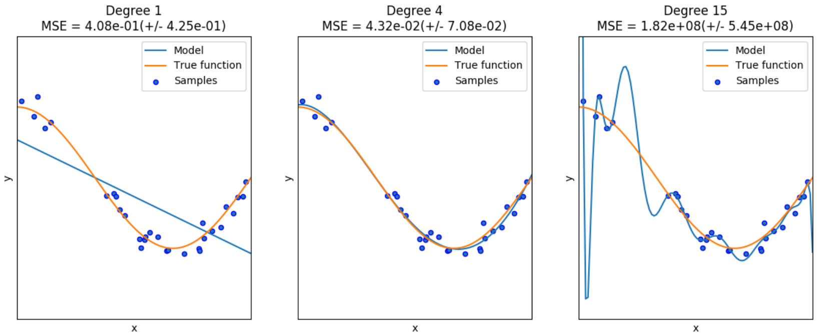
Because polynomials are still linear in the parameters:

$$y_i = b + \sum_{j=1}^{k} a_j x_i^j \quad \text{for all data points } i = 1, \ldots, n$$

There is a least-squares solution in closed form for any polynomial.

Quality of fitting

By the Stone-Weierstrass theorem, we can fit a polynomial in many cases:



Underfitting **Overfitting**

Linear regression: Matrix notation

In matrix notation, the MSE is simply:

$$\ell(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

Setting the gradient w.r.t. $\boldsymbol{\theta}$ to zero and solving for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Linear regression: Matrix notation

In matrix notation, the MSE is simply:

$$\ell(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

Setting the gradient w.r.t. $\boldsymbol{\theta}$ to zero and solving for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

In other words, $\boldsymbol{\theta}$ approximately satisfies:

$$\mathbf{X}\boldsymbol{\theta} \approx \mathbf{y}$$

where the residual error $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2$ is the smallest possible.

Normal equations

Consider the linear system:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

If an exact solution exists, then we can write

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Normal equations

Consider the linear system:

$$\mathbf{Ax} = \mathbf{b}$$

If an exact solution exists, then we can write

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

If an exact solution does not exist, we have the approximation problem:

$$\mathbf{Ax} \approx \mathbf{b}$$

Normal equations

Consider the linear system:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

If an exact solution exists, then we can write

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

If an exact solution does not exist, we have the approximation problem:

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}$$

which we rewrote using the normal equations:

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b}$$

Normal equations

Consider the linear system:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

If an exact solution exists, then we can write

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

If an exact solution does not exist, we have the approximation problem:

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}$$

which we rewrote using the normal equations:

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{b}$$

and the solution is:

$$\mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{b}$$

Normal equations

Consider the linear system:

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

If an exact solution exists, then we can write

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

If an exact solution does not exist, we have the approximation problem:

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}$$

which we rewrote using the normal equations:

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b}$$

and the solution is:

$$\mathbf{x} = \underbrace{(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top}_{\text{pseudo-inverse } \mathbf{A}^\dagger}\mathbf{b}$$

Types of linear systems

- **Exact**: n linearly independent equations, $m = n$ parameters (matrix $\mathbf{A}$ is square)

$$\text{problem} : \mathbf{A}\mathbf{x} = \mathbf{b} \qquad \text{solution} : \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Types of linear systems

- **Exact**: n linearly independent equations, $m = n$ parameters (matrix $\mathbf{A}$ is square)

$$\text{problem} : \mathbf{Ax} = \mathbf{b} \qquad \text{solution} : \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

- **Over-determined**: n lin. ind. equations, $m < n$ parameters (matrix $\mathbf{A}$ is tall)

$$\text{problem} : \mathbf{Ax} \approx \mathbf{b} \qquad \text{solution} : \mathbf{x} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}$$

Types of linear systems

- **Exact**: n linearly independent equations, $m = n$ parameters (matrix $\mathbf{A}$ is square)

$$\text{problem} : \mathbf{A}\mathbf{x} = \mathbf{b} \qquad \text{solution} : \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

- **Over-determined**: n lin. ind. equations, $m < n$ parameters (matrix $\mathbf{A}$ is tall)

$$\text{problem} : \mathbf{A}\mathbf{x} \approx \mathbf{b} \qquad \text{solution} : \mathbf{x} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}$$

- **Under-determined**: n lin. ind. equations, $m > n$ parameters (matrix $\mathbf{A}$ is wide)

$$\text{problem} : \mathbf{A}\mathbf{x} \approx \mathbf{b} \qquad \text{solution} : \mathbf{???}$$

How to solve for $\mathbf{x}$ when we do not have enough data?

Regularization

General idea: Make additional assumptions, and write them as new terms in the optimization.

Under-determined case

General idea: Make additional assumptions, and write them as new terms in the optimization.

Main benefits:

- Impose a desired behavior of the solution (e.g. sparse, smooth)

Under-determined case

General idea: Make additional assumptions, and write them as new terms in the optimization.

Main benefits:

- Impose a desired behavior of the solution (e.g. sparse, smooth)
- Reduce the amount of necessary data

Under-determined case

General idea: Make additional assumptions, and write them as new terms in the optimization.

Main benefits:

- Impose a desired behavior of the solution (e.g. sparse, smooth)
- Reduce the amount of necessary data
- Make the optimization easier

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$\nabla_{\mathbf{x}} (\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}\|_2^2) = \mathbf{0}$$

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$\nabla_{\mathbf{x}}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\nabla_{\mathbf{x}}\|\mathbf{x}\|_2^2 = \mathbf{0}$$

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b} + 2\alpha\mathbf{x} = \mathbf{0}$$

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} + \alpha\mathbf{x} = \mathbf{0}$$

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$\mathbf{A}^\top \mathbf{A}\mathbf{x} + \alpha\mathbf{x} = \mathbf{A}^\top \mathbf{b}$$

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$(\mathbf{A}^\top\mathbf{A} + \alpha\mathbf{I})\mathbf{x} = \mathbf{A}^\top\mathbf{b}$$

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$(\mathbf{A}^\top\mathbf{A} + \alpha\mathbf{I})\mathbf{x} = \mathbf{A}^\top\mathbf{b}$$

In other words: just add α along the diagonal of $\mathbf{A}^\top\mathbf{A}$.

Tikhonov regularization

Add a L_2 penalty:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

with some choice of $\alpha > 0$. This penalizes large values in $\mathbf{x}$.

Let's find a solution:

$$(\mathbf{A}^\top\mathbf{A} + \alpha\mathbf{I})\mathbf{x} = \mathbf{A}^\top\mathbf{b}$$

In other words: just add α along the diagonal of $\mathbf{A}^\top\mathbf{A}$.

Also known as ridge regression.

Sparse problems

L_p norms

With Tikhonov regularization we had:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

L_p norms

With Tikhonov regularization we had:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_p^p$$

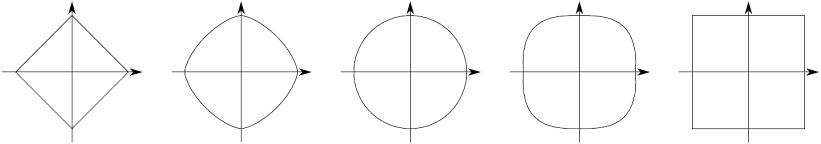
We can consider other norms for the regularizer!

L_p norms

With Tikhonov regularization we had:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_p^p$$

We can consider other norms for the regularizer!



$\|\cdot\|_1$ $\|\cdot\|_{1.5}$ $\|\cdot\|_2$ $\|\cdot\|_3$ $\|\cdot\|_\infty$

"Manhattan" **"max"**

Some special cases have convenient interpretations.

Interpretation as penalties

Recall that $\|\mathbf{x}\|_p^p = |x_1|^p + |x_2|^p + \cdots + |x_n|^p$.

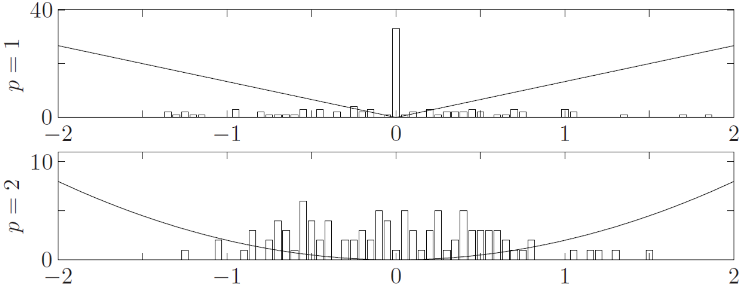
Depending on p, the values of $\mathbf{x}$ are penalized differently.

Interpretation as penalties

Recall that $\|\mathbf{x}\|_p^p = |x_1|^p + |x_2|^p + \cdots + |x_n|^p$.

Depending on p, the values of $\mathbf{x}$ are penalized differently.

Roughly: the shape of the penalty function is a measure of our dislike.

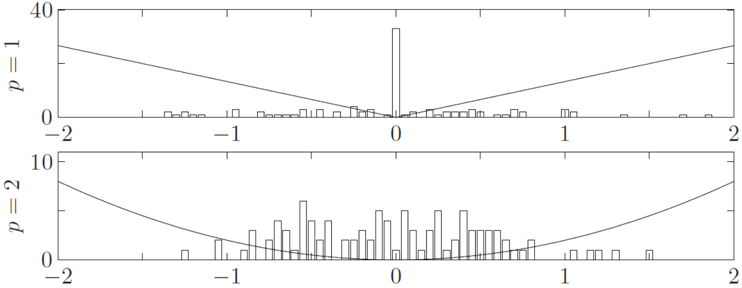


Interpretation as penalties

Recall that $\|\mathbf{x}\|_p^p = |x_1|^p + |x_2|^p + \cdots + |x_n|^p$.

Depending on p, the values of $\mathbf{x}$ are penalized differently.

Roughly: the shape of the penalty function is a measure of our dislike.



L_1 norm favors sparse solutions.

Sparse solutions

Regularization with the L_1 norm is a heuristic to find sparse solutions.

For example, consider the problem

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}\|_1$$

Observe that:

Sparse solutions

Regularization with the L_1 norm is a heuristic to find sparse solutions.

For example, consider the problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_1$$

Observe that:

- For $\alpha \approx 0$, this is basically least squares.

Sparse solutions

Regularization with the L_1 norm is a heuristic to find sparse solutions.

For example, consider the problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_1$$

Observe that:

- For $\alpha \approx 0$, this is basically least squares.
- For $\alpha \gg 0$, the solution $\mathbf{x}$ will contain many zeros.

Sparse solutions

Regularization with the L_1 norm is a heuristic to find sparse solutions.

For example, consider the problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}\|_1$$

Observe that:

- For $\alpha \approx 0$, this is basically least squares.
- For $\alpha \gg 0$, the solution $\mathbf{x}$ will contain many zeros.
- Otherwise: trade-off between data fidelity and sparsity of $\mathbf{x}$.

Sparse solutions

Regularization with the L_1 norm is a heuristic to find sparse solutions.

For example, consider the problem

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{x}\|_1$$

Observe that:

- For $\alpha \approx 0$, this is basically least squares.
- For $\alpha \gg 0$, the solution $\mathbf{x}$ will contain many zeros.
- Otherwise: trade-off between data fidelity and sparsity of $\mathbf{x}$.

Warning: This problem is not differentiable because of the L_1 norm!

Sparse problems

Consider now the general problem:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \alpha \rho(\mathbf{x})$$

for some $p \geq 1$, $\alpha \geq 0$, and regularization function ρ.

Sparse problems

Consider now the general problem:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \alpha\rho(\mathbf{x})$$

for some $p \geq 1$, $\alpha \geq 0$, and regularization function ρ.

It could happen that matrix $\mathbf{A}$ is sparse.

Sparse problems

Consider now the general problem:

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + \alpha\rho(\mathbf{x})$$

for some $p \geq 1$, $\alpha \geq 0$, and regularization function ρ.

It could happen that matrix $\mathbf{A}$ is sparse.

For example, $\mathbf{A}$ is tridiagonal:

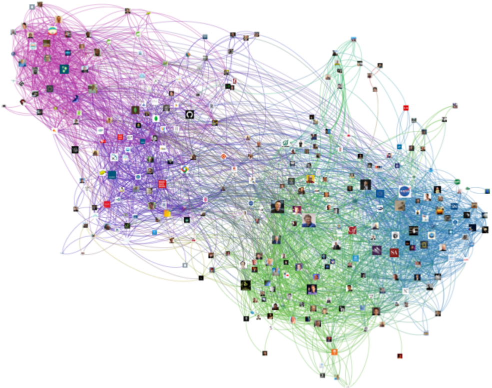
$$A = \begin{pmatrix} v_1 & w_1 & & & & \\ u_2 & v_2 & w_2 & & & \\ & u_3 & v_3 & w_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & u_{n-1} & v_{n-1} & w_{n-1} \\ & & & & u_n & v_n \end{pmatrix}$$

Example: Graphs

A graph with n nodes can be encoded as a $n \times n$ adjacency matrix:

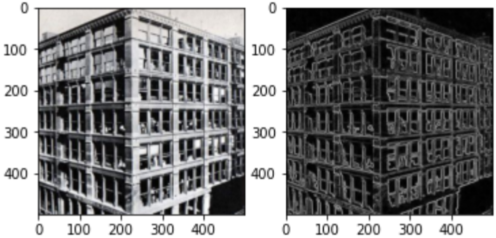
$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \end{pmatrix}$$

where $a_{ij} = 1$ if vertex v_i is connected to v_j by an edge.



Smoothing

Derivatives as a measure of smoothness

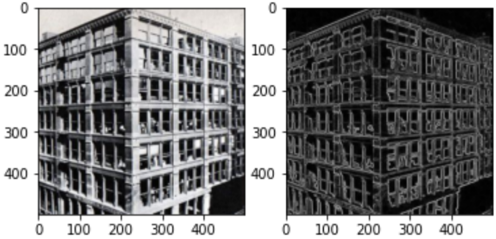


$\mathbf{x}$ $\qquad\qquad$ $\|\nabla \mathbf{x}\|_2$

The norm of the gradient captures the edges!

Sharp images have strong gradients.

Derivatives as a measure of smoothness



$\mathbf{x}$ $\|\nabla \mathbf{x}\|_2$

The norm of the gradient captures the edges!

Sharp images have strong gradients.

$\|\nabla \mathbf{x}\|_2$ as a penalty would promote smooth solutions.

Quadratic smoothing

More in general, consider $\|\mathbf{Dx}\|$ with $\mathbf{D}$ some differentiation operator.

Quadratic smoothing

More in general, consider $\|\mathbf{D}\mathbf{x}\|$ with $\mathbf{D}$ some differentiation operator.

$\|\mathbf{D}\mathbf{x}\|$ is a measure of the variation or smoothness of $\mathbf{x}$.

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \alpha\|\mathbf{D}\mathbf{x}\|_2^2$$

Quadratic smoothing

More in general, consider $\|\mathbf{D}\mathbf{x}\|$ with $\mathbf{D}$ some differentiation operator.

$\|\mathbf{D}\mathbf{x}\|$ is a measure of the variation or smoothness of $\mathbf{x}$.

$$\min_{\mathbf{x}} \underbrace{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2}_{\text{data term}} + \alpha \underbrace{\|\mathbf{D}\mathbf{x}\|_2^2}_{\text{smoothness}}$$

Quadratic smoothing

More in general, consider $\|\mathbf{Dx}\|$ with $\mathbf{D}$ some differentiation operator.

$\|\mathbf{Dx}\|$ is a measure of the variation or smoothness of $\mathbf{x}$.

$$\min_{\mathbf{x}} \underbrace{\|\mathbf{Ax}-\mathbf{b}\|_2^2}_{\text{data term}} + \alpha \underbrace{\|\mathbf{Dx}\|_2^2}_{\text{smoothness}}$$

Example: $\mathbf{x} \in \mathbb{R}^n$ represents a function sampled at n points.

Its derivative is approximated as $\mathbf{\Delta x}$, with $\mathbf{\Delta}$:

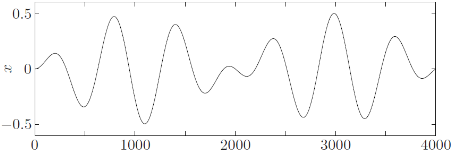
$$\begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

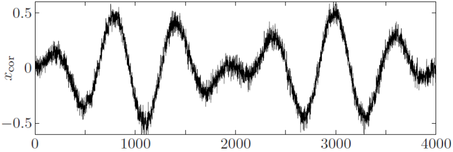
Example: Denoising

We are going to denoise a corrupted audio signal $\mathbf{x}_{\mathrm{cor}}$:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\mathrm{cor}}\|_2^2 + \alpha \|\mathbf{\Delta x}\|_2^2$$

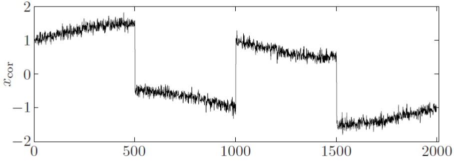
with $\mathbf{\Delta}$ defined as in the previous slide.





Total variation reconstruction

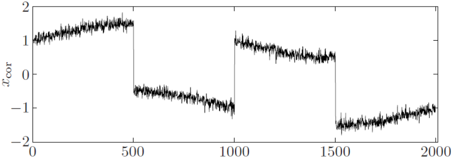
Now consider the noisy signal:



Smoothing will treat the jumps as noise, and attenuate them!

Total variation reconstruction

Now consider the noisy signal:



Smoothing will treat the jumps as noise, and attenuate them!

To preserve occasional big jumps, consider the penalty:

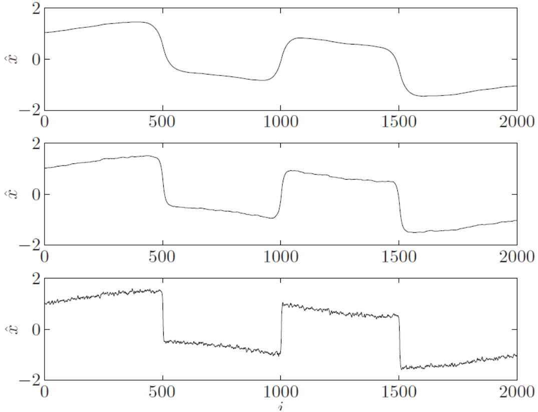
$$\|\boldsymbol{\Delta}\mathbf{x}\|_1 = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

with the same $\boldsymbol{\Delta}$ as before.

Total variation reconstruction

With quadratic smoothing, this is what we get:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\text{cor}}\|_2^2 + \alpha \|\boldsymbol{\Delta}\mathbf{x}\|_2^2$$



Total variation reconstruction

Instead, the problem:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\text{cor}}\|_2^2 + \alpha \|\mathbf{\Delta x}\|_1$$

uses L_1 regularization on the derivatives of the signal.

It seeks a solution $\mathbf{x}$ with sparse discontinuities.

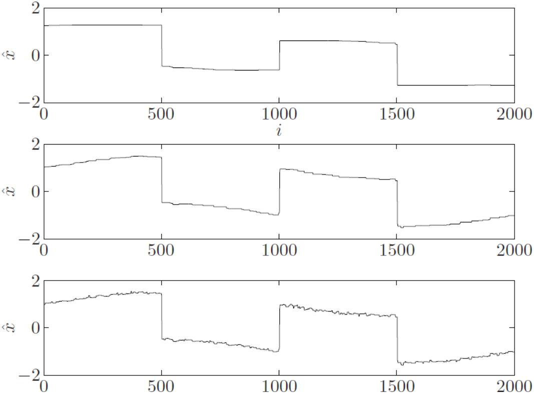
Total variation reconstruction

Instead, the problem:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\mathrm{cor}}\|_2^2 + \alpha \|\boldsymbol{\Delta}\mathbf{x}\|_1$$

uses L_1 regularization on the derivatives of the signal.

It seeks a solution $\mathbf{x}$ with sparse discontinuities.



Suggested reading

For least squares and Tikhonov regularization, read Sections 4.1.2 and 4.1.3 of the book:

J. Solomon, "Numerical Algorithms"

For more on regularization and smoothing, read Sections 6.3.2 and 6.3.3 of the book:

Boyd and Vandenberghe, "Convex Optimization"