Metodi Numerici dell'Informatica

Regularization, smoothing and sparsity

Emanuele Rodolà
rodola@di.uniroma1.it

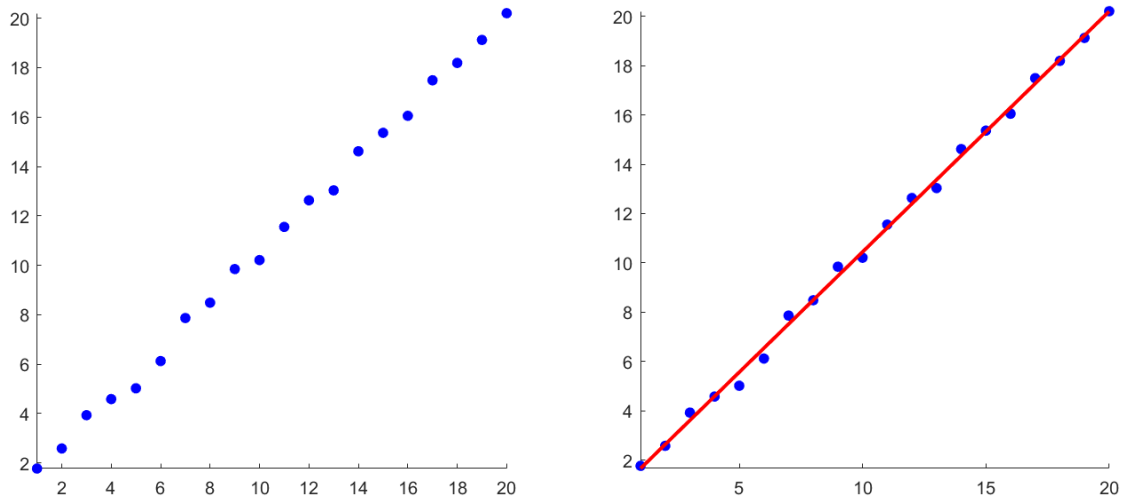
SAPIENZA
UNIVERSITÀ DI ROMA

2nd semester a.y. 2021/2022 · March 17, 2022

Motivation

Linear regression

We have seen fitting problems of the kind:



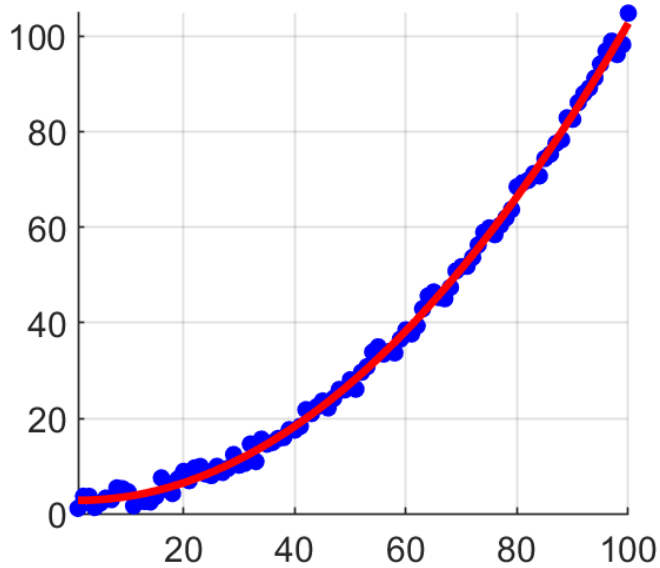
$$y_i = ax_i + b$$

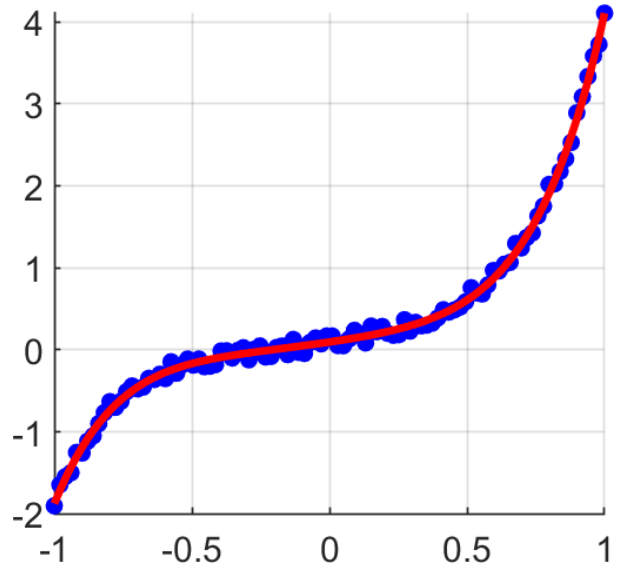
Which are formalized as the minimization problem:

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

Linear regression

In fact, we can address polynomial fitting as well:





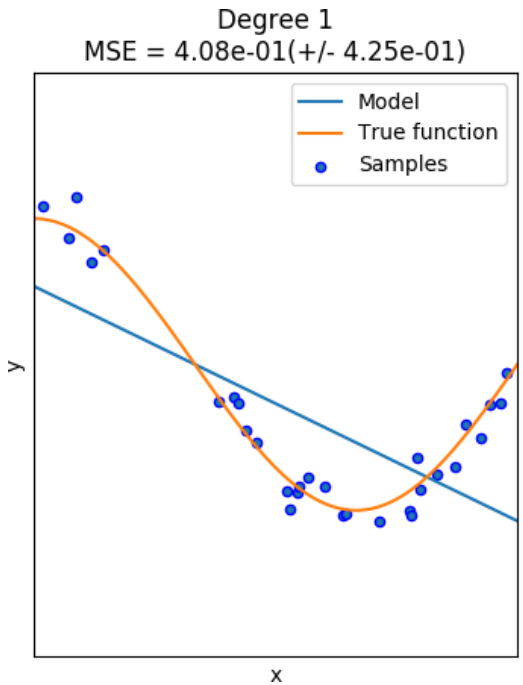
Because polynomials are still linear in the parameters:

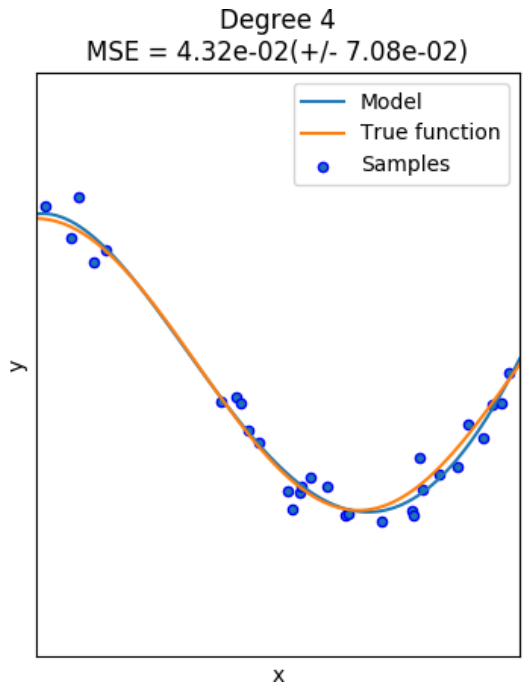
$$y_i = b + \sum_{j=1}^{k} a_j x_i^j \quad \text{for all data points } i = 1, \ldots, n$$

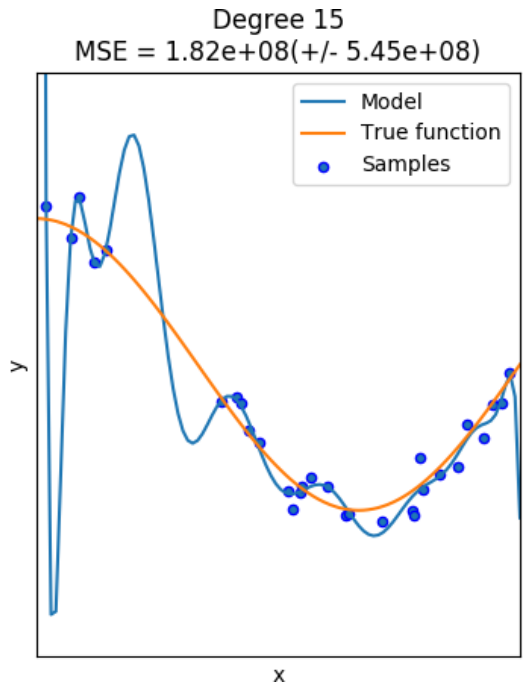
A least-squares solution is found in closed form for any polynomial.

Quality of fitting

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







Underfitting

Overfitting

Linear regression: Matrix notation

Using matrix notation, the MSE is simply written as:

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

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

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

Which gives us a closed form solution to linear regression.

Linear regression: Matrix notation

Using matrix notation, the MSE is simply written as:

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

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

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

Which gives us a closed form solution to linear regression.

In other words, $\boldsymbol{\theta}$ is an approximate solution that 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, as in all our examples so far, we consider 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, as in all our examples so far, we consider 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, as in all our examples so far, we consider 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, as in all our examples so far, we consider 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

We identify the following cases:

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

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

Types of linear systems

We identify the following cases:

- **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

We identify the following cases:

- **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)

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

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

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

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

Regularization

Under-determined case

In practice, $m < n$ means we do not have enough data points to solve unambiguously for the m parameters of our model.

We need to add some extra information to our problem.

Under-determined case

In practice, $m < n$ means we do not have enough data points to solve unambiguously for the m parameters of our model.

We need to add some extra information to our problem.

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

Under-determined case

In practice, $m < n$ means we do not have enough data points to solve unambiguously for the m parameters of our model.

We need to add some extra information to our problem.

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

These regularizers bring several benefits:

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

Under-determined case

In practice, $m < n$ means we do not have enough data points to solve unambiguously for the m parameters of our model.

We need to add some extra information to our problem.

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

These regularizers bring several benefits:

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

Under-determined case

In practice, $m < n$ means we do not have enough data points to solve unambiguously for the m parameters of our model.

We need to add some extra information to our problem.

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

These regularizers bring several benefits:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

In other words, to introduce Tikhonov regularization all we have to do is add α along the diagonal of $\mathbf{A}^\top\mathbf{A}$.

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

In other words, to introduce Tikhonov regularization all we have to do is add α along the diagonal of $\mathbf{A}^\top\mathbf{A}$.

Can be done for exact and over-determined problems as well!

Tikhonov regularization

Add a L_2 penalty to the least-squares energy function:

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

with some choice of $\alpha > 0$.

The penalty on the norm of $\mathbf{x}$ is asking for solutions with small entries.

To find a solution, take the gradient and equate to zero:

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

In other words, to introduce Tikhonov regularization all we have to do is add α along the diagonal of $\mathbf{A}^\top\mathbf{A}$.

Can be done for exact and over-determined problems as well!

Also known as ridge regression.

Example: Deblurring

Suppose you want to recover a sharp image from its blurry version:

(a) Sharp

(b) Blurry

Example: Deblurring

Suppose you want to recover a sharp image from its blurry version:

(a) Sharp (b) Blurry

This can be cast as a Tikhonov-regularized least-squares problem:

$$\min_{\mathbf{x}} \|\mathbf{x}_{\text{blurry}} - \mathbf{G}\mathbf{x}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

...provided that we know the blurring operator $\mathbf{G}$.

Example: Deblurring

Suppose you want to recover a sharp image from its blurry version:

(a) Sharp (b) Blurry

This can be cast as a Tikhonov-regularized least-squares problem:

$$\min_{\mathbf{x}} \|\mathbf{x}_{\text{blurry}} - \mathbf{G}\mathbf{x}\|_2^2 + \alpha\|\mathbf{x}\|_2^2$$

...provided that we know the blurring operator $\mathbf{G}$.

Sparse problems

L_p norms

Let's take a second look at Tikhonov regularization:

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

L_p norms

Let's take a second look at Tikhonov regularization:

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

In general, one could consider different norms for the regularizer.

L_p norms

Let's take a second look at Tikhonov regularization:

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

In general, one could consider different norms for the regularizer.

First, generalize the L_2 norm to different power coefficients $p \geq 1$:

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left(|x_1 - y_1|^2 + |x_2 - y_2|^2\right)^{\frac{1}{2}}$$
$$\Downarrow$$
$$\|\mathbf{x} - \mathbf{y}\|_p = \left(|x_1 - y_1|^p + |x_2 - y_2|^p\right)^{\frac{1}{p}}$$

L_p norms

Let's take a second look at Tikhonov regularization:

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

In general, one could consider different norms for the regularizer.

First, generalize the L_2 norm to different power coefficients $p \geq 1$:

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left(|x_1 - y_1|^2 + |x_2 - y_2|^2\right)^{\frac{1}{2}}$$
$$\Downarrow$$
$$\|\mathbf{x} - \mathbf{y}\|_p = \left(|x_1 - y_1|^p + |x_2 - y_2|^p\right)^{\frac{1}{p}}$$

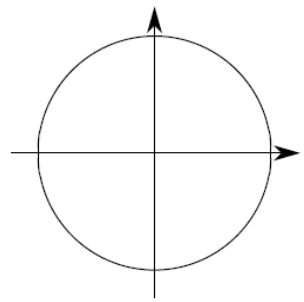
Then, generalize from $\mathbb{R}^2$ to $\mathbb{R}^k$:

$$\|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_{i=1}^{k} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

This definition gives us the L_p distance between vectors in $\mathbb{R}^k$.

L_p unit circles

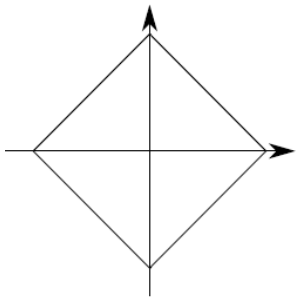
Let's have a look at unit circles using different L_p norms:



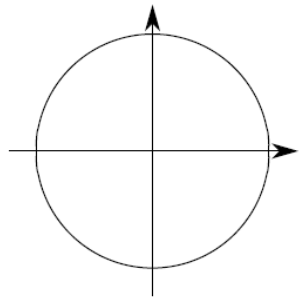
$\|\cdot\|_2$

L_p unit circles

Let's have a look at unit circles using different L_p norms:



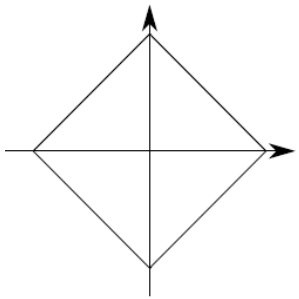
$\|\cdot\|_1$



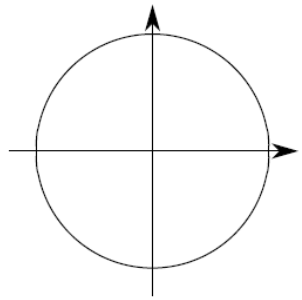
$\|\cdot\|_2$

L_p unit circles

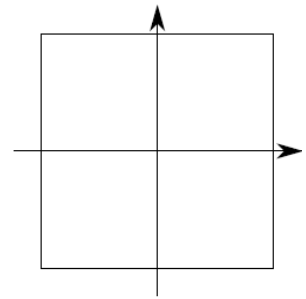
Let's have a look at unit circles using different L_p norms:



$\|\cdot\|_1$



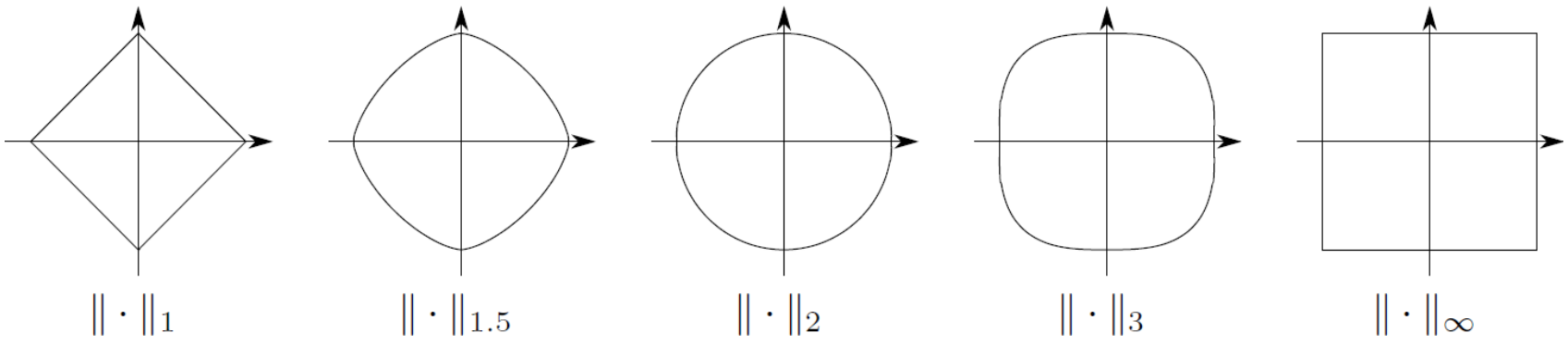
$\|\cdot\|_2$



$\|\cdot\|_\infty$

L_p unit circles

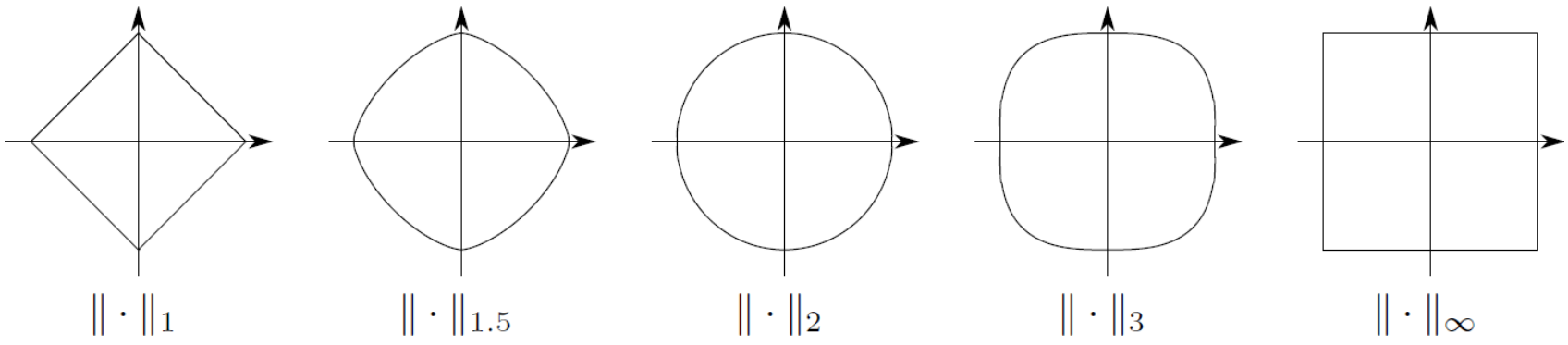
Let's have a look at unit circles using different L_p norms:



"Manhattan" **"max"**

L_p unit circles

Let's have a look at unit circles using different L_p norms:



"Manhattan" **"max"**

When used as regularizers, each norm $\|\cdot\|_p$ yields a different penalty.

Some special cases have convenient interpretations.

Interpretation as penalties

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

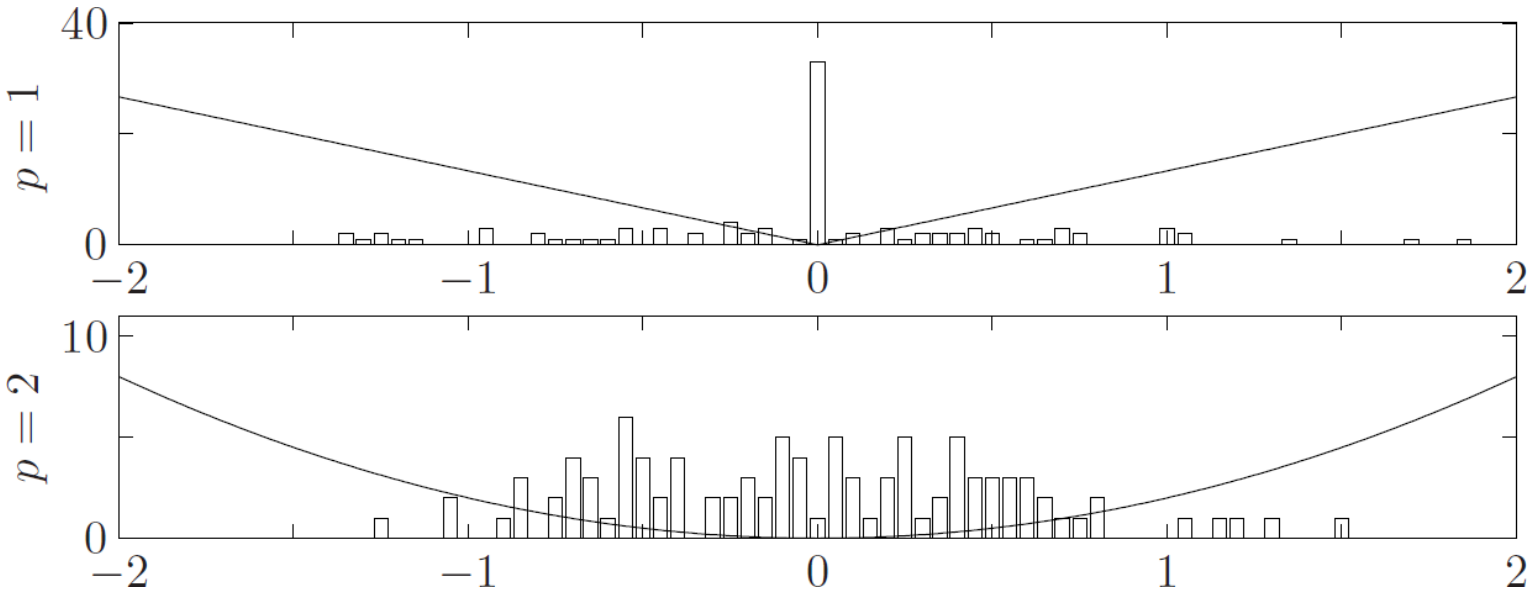
Depending on the choice of p, the values of $\mathbf{x}$ are penalized differently.

Interpretation as penalties

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

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

Roughly speaking, the shape of the penalty function is a measure of our dislike of a certain value x.

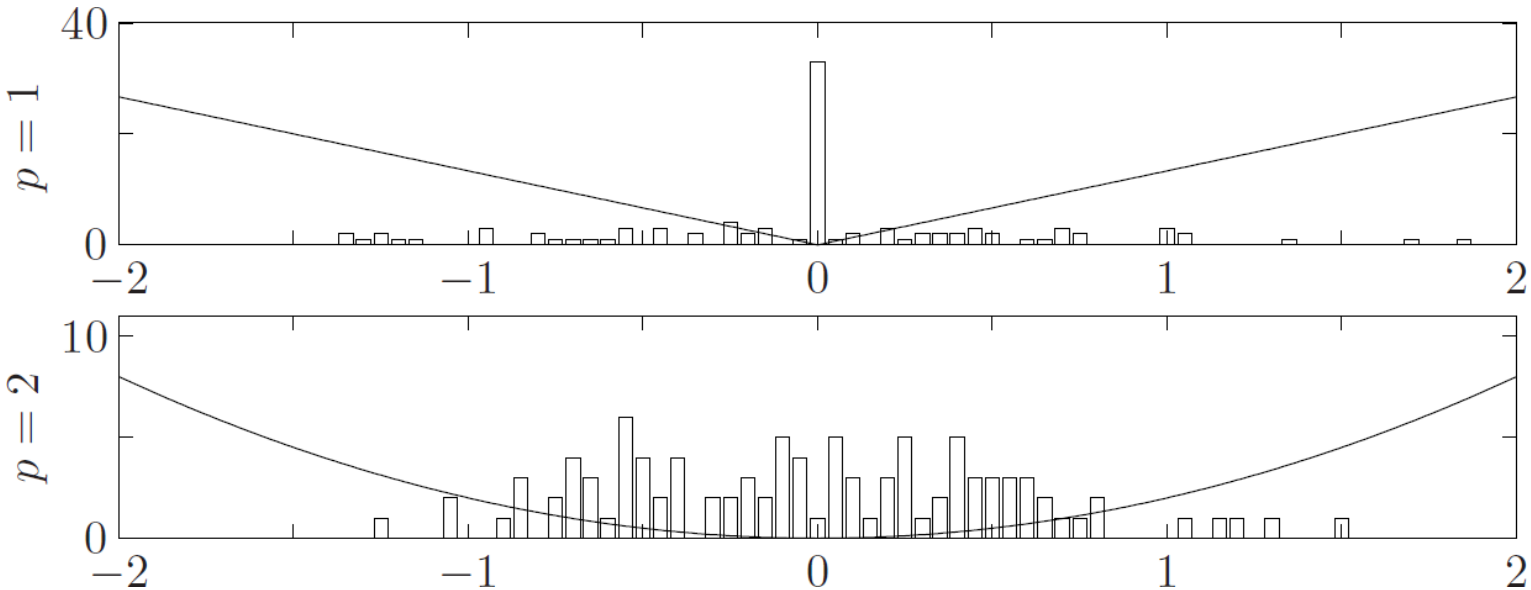


Interpretation as penalties

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

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

Roughly speaking, the shape of the penalty function is a measure of our dislike of a certain value x.



We see that the L_1 norm seems to favor sparse solutions.

Sparse solutions

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

For example, consider the problem

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

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 and the data term prevails.

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 and the data term prevails.
- For $\alpha \gg 0$, the solution $\mathbf{x}$ will contain lots of 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 and the data term prevails.
- For $\alpha \gg 0$, the solution $\mathbf{x}$ will contain lots of zeros.
- For in-between values of α, we are trading off between fidelity to the data and the number of non-zero elements of $\mathbf{x}$ (i.e. its sparsity).

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 and the data term prevails.
- For $\alpha \gg 0$, the solution $\mathbf{x}$ will contain lots of zeros.
- For in-between values of α, we are trading off between fidelity to the data and the number of non-zero elements of $\mathbf{x}$ (i.e. its sparsity).

Warning: This problem is not differentiable because of the L_1 norm! This means that we can not simply compute the gradient and equate to zero, in order to find a solution.

Sparse problems

Consider now the generic regularized 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 generic regularized 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.

A different form of sparsity arises when matrix $\mathbf{A}$ is sparse, that is, whenever $\mathbf{A}$ contains many zeros.

Sparse problems

Consider now the generic regularized 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.

A different form of sparsity arises when matrix $\mathbf{A}$ is sparse, that is, whenever $\mathbf{A}$ contains many zeros.

For example, $\mathbf{A}$ could be a tridiagonal matrix:

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

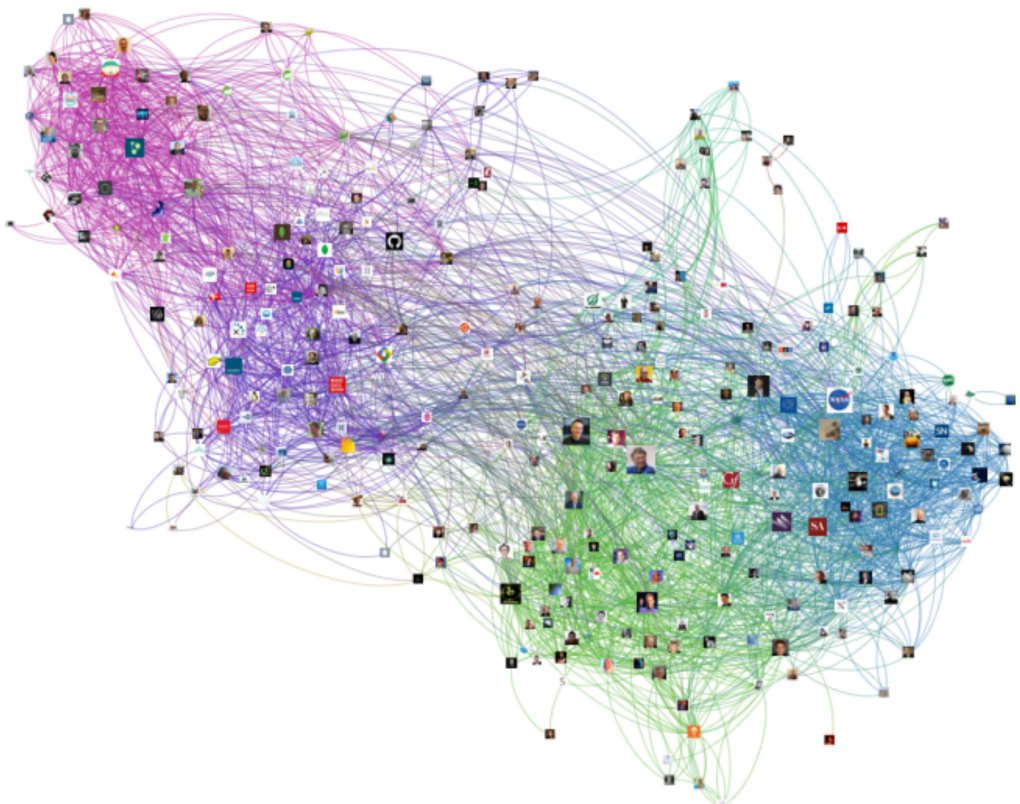
In the following, we will see some practical examples of sparse problems.

Example: Graphs

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

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

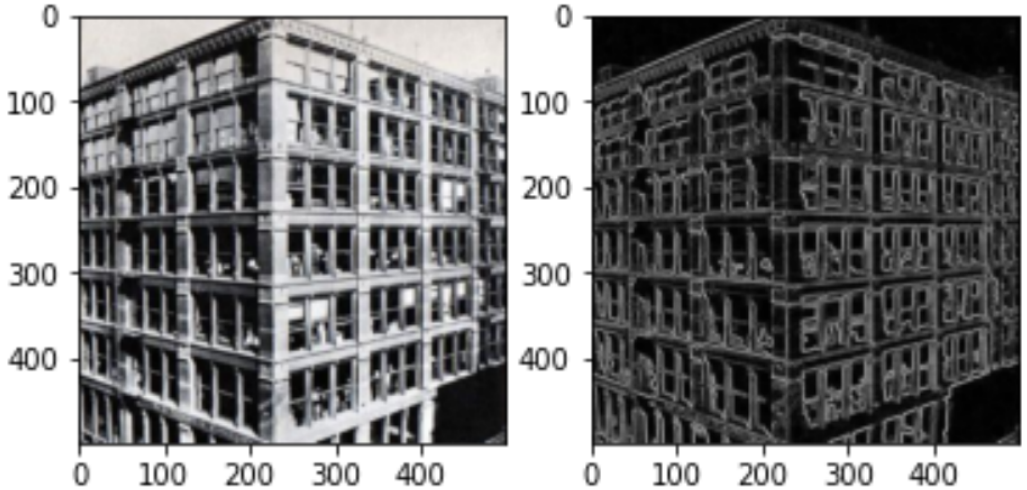
where $a_{ij} = 1$ if vertex v_i is connected to v_j by an edge.



Smoothing

Derivatives as a measure of smoothness

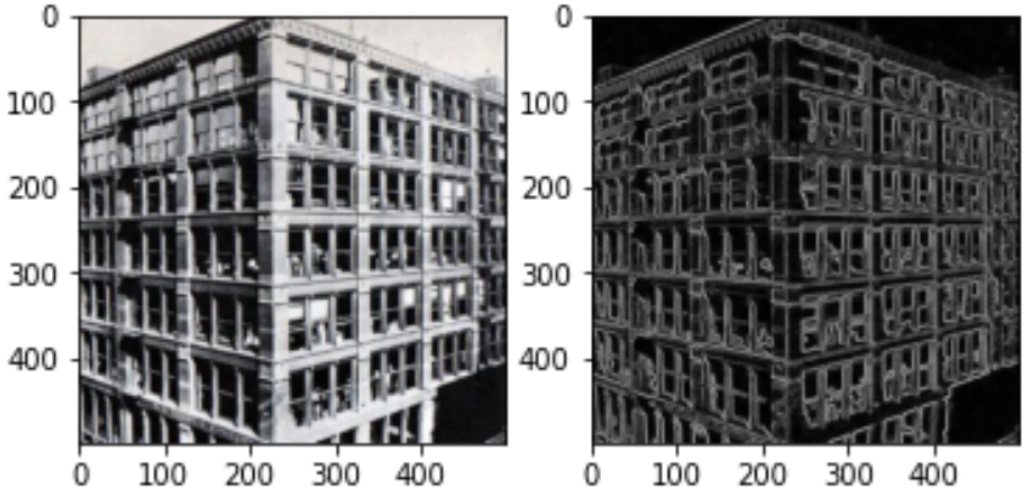
Recall the following exercise from the past lab session:



For a given image $\mathbf{x}$, the norm of the gradient $\|\nabla \mathbf{x}\|_2$ seems to capture its edges. Strong gradients correspond to sharp images.

Derivatives as a measure of smoothness

Recall the following exercise from the past lab session:



For a given image $\mathbf{x}$, the norm of the gradient $\|\nabla \mathbf{x}\|_2$ seems to capture its edges. Strong gradients correspond to sharp images.

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

Quadratic smoothing

We consider regularization terms of the form $\|\mathbf{D}\mathbf{x}\|$ in place of $\|\mathbf{x}\|$, where $\mathbf{D}$ is some differentiation operator.

Quadratic smoothing

We consider regularization terms of the form $\|\mathbf{Dx}\|$ in place of $\|\mathbf{x}\|$, where $\mathbf{D}$ is some differentiation operator.

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

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

Quadratic smoothing

We consider regularization terms of the form $\|\mathbf{Dx}\|$ in place of $\|\mathbf{x}\|$, where $\mathbf{D}$ is some differentiation operator.

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

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

Quadratic smoothing

We consider regularization terms of the form $\|\mathbf{Dx}\|$ in place of $\|\mathbf{x}\|$, where $\mathbf{D}$ is some differentiation operator.

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

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

For example, assume $\mathbf{x} \in \mathbf{R}^n$ represents a function sampled at n points.

Its derivative can be approximated as $\mathbf{\Delta x}$, where $\mathbf{\Delta}$ is:

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

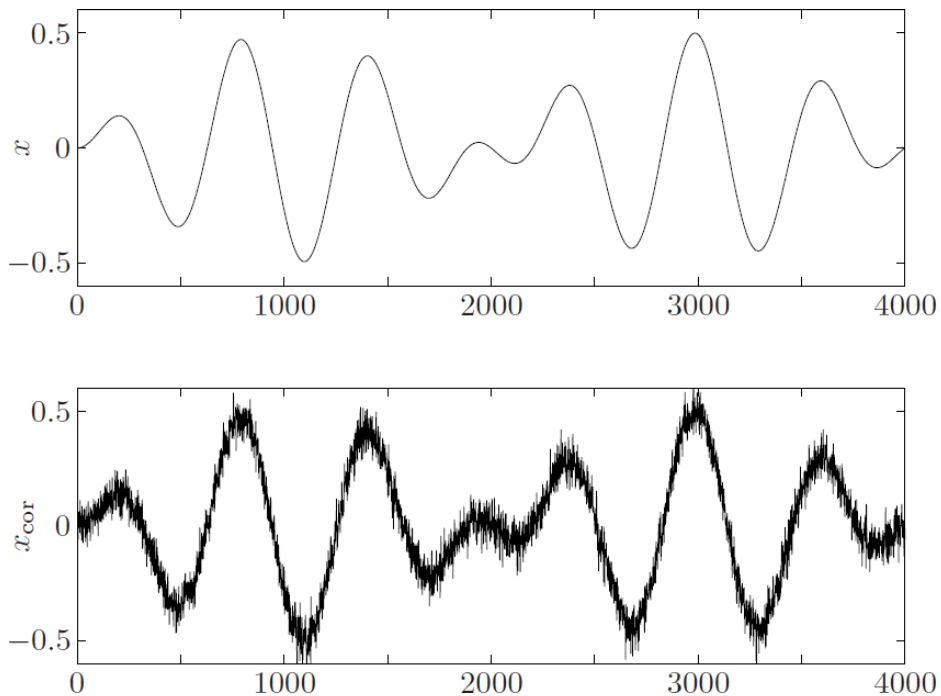
Example: Denoising

We are given a corrupted audio signal $\mathbf{x}_{\text{cor}}$, and want to denoise it.

We optimize the following quadratic smoothing problem:

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

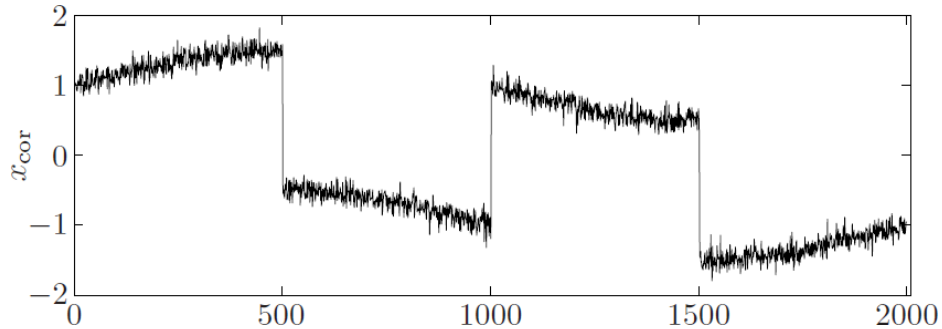
with $\mathbf{\Delta}$ defined as in the previous slide.



Total variation reconstruction

If the original signal is smooth, quadratic smoothing works well.

However, consider the noisy signal:

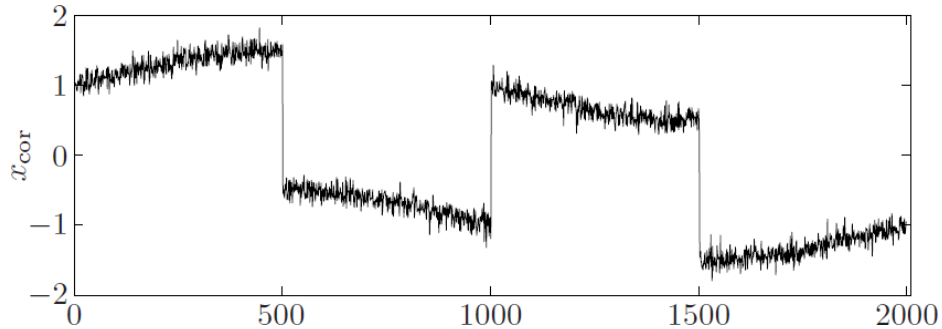


Quadratic smoothing will treat the jumps as noise, and attenuate them.

Total variation reconstruction

If the original signal is smooth, quadratic smoothing works well.

However, consider the noisy signal:



Quadratic smoothing will treat the jumps as noise, and attenuate them.

To preserve occasional big jumps, consider the smoothing function:

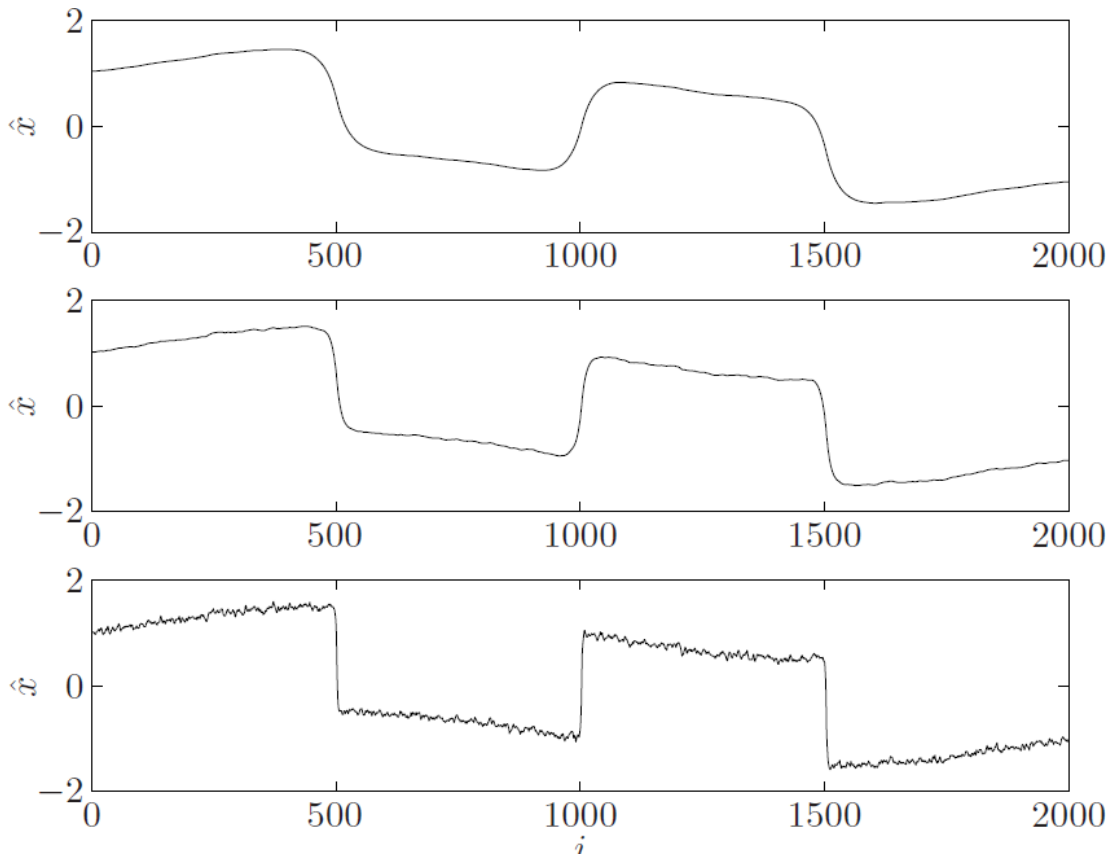
$$\|\mathbf{\Delta x}\|_1 = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

with the same $\mathbf{\Delta}$ used for quadratic smoothing.

Total variation reconstruction

Example of applying quadratic smoothing to a noisy signal with big discontinuities:

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



Total variation reconstruction

In other words, the problem:

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

is using a L_1 regularization on the derivatives of the signal.

This corresponds to seeking a solution $\mathbf{x}$ with sparse discontinuities.

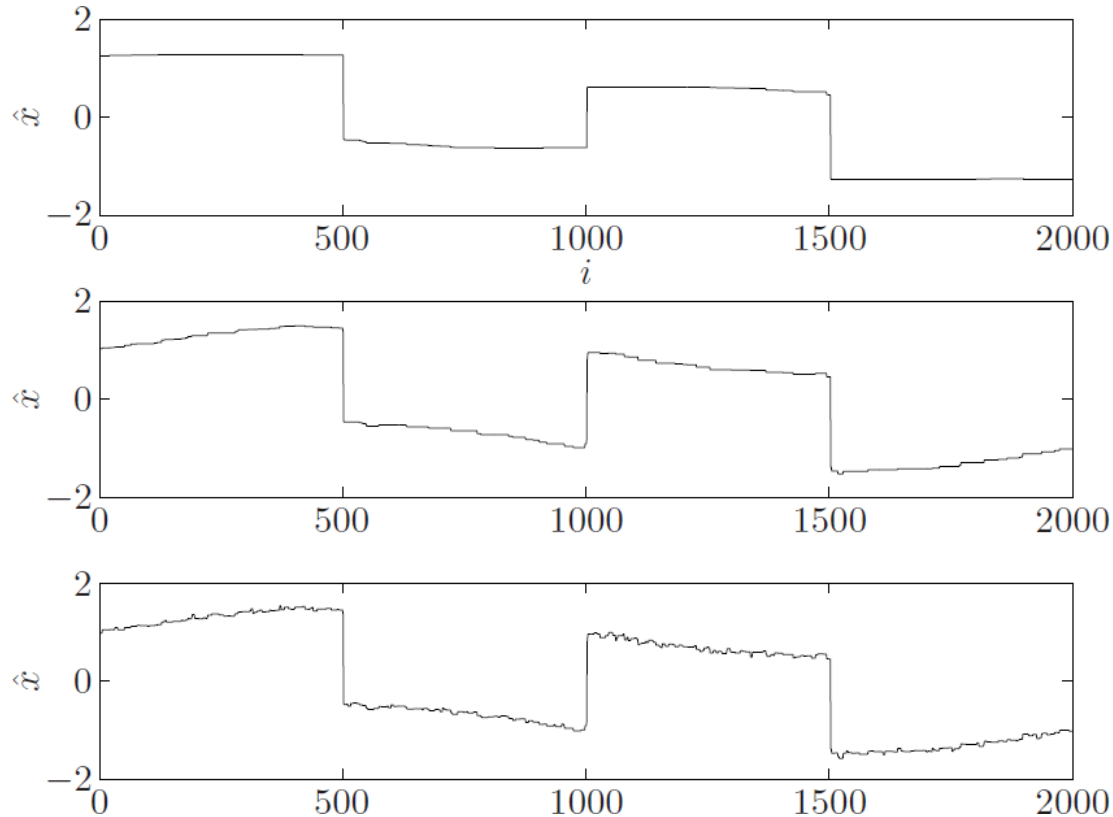
Total variation reconstruction

In other words, the problem:

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

is using a L_1 regularization on the derivatives of the signal.

This corresponds to seeking 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"