10. Regularization

- More on tradeoffs
- Regularization
- Effect of using different norms
- Example: hovercraft revisited

Laurent Lessard (www.laurentlessard.com)

Review of tradeoffs

Recap of tradeoffs:

- We want to make both $J_1(x)$ and $J_2(x)$ small subject to constraints.
- Choose a parameter $\lambda > 0$, solve

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & J_1(x) + \lambda J_2(x) \\ \text{subject to:} \quad & \text{constraints} \end{aligned}$$

- Each $\lambda > 0$ yields a solution $\hat{x}_\lambda$.
- Can visualize tradeoff by plotting $J_2(\hat{x}_\lambda)$ vs $J_1(\hat{x}_\lambda)$. This is called the **Pareto curve**.

Multi-objective tradeoff

- Similar procedure if we have more than two costs we'd like to make small, e.g. J_1, J_2, J_3
- Choose parameters $\lambda > 0$ and $\mu > 0$. Then solve:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & J_1(x) + \lambda J_2(x) + \mu J_3(x) \\ \text{subject to:} \quad & \text{constraints} \end{aligned}$$

- Each $\lambda > 0$ and $\mu > 0$ yields a solution $\hat{x}_{\lambda,\mu}$.
- Can visualize tradeoff by plotting $J_3(\hat{x}_{\lambda,\mu})$ vs $J_2(\hat{x}_{\lambda,\mu})$ vs $J_1(\hat{x}_{\lambda,\mu})$ on a 3D plot. You then obtain a **Pareto surface**.

Minimum-norm as a regularization

- When $Ax = b$ is underdetermined (A is wide), we can resolve ambiguity by adding a cost function, e.g. **min-norm** LS:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x\|^2 \\ \text{subject to:} \quad & Ax = b \end{aligned}$$

- Alternative approach: express it as a tradeoff!

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda\|x\|^2$$

  Tradeoffs of this type are called **regularization** and λ is called the *regularization parameter* or *regularization weight*

- If we let $\lambda \to \infty$, we just obtain $\hat{x} = 0$

- If we let $\lambda \to 0$, we obtain the minimum-norm solution!

Proof of minimum-norm equivalence

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda\|x\|^2$$

Equivalent to the least squares problem:

$$\underset{x}{\text{minimize}} \quad \left\| \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2$$

Solution is found via pseudoinverse (for tall matrix)

$$\begin{aligned} \hat{x} &= \left(\begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \right)^{-1} \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} b \\ 0 \end{bmatrix} \\ &= (A^{\mathsf{T}} A + \lambda I)^{-1} A^{\mathsf{T}} b \end{aligned}$$

Proof of minimum-norm equivalence

Solution of 2-norm regularization is:

$$\hat{x} = (A^\mathsf{T} A + \lambda I)^{-1} A^\mathsf{T} b$$

- Can't simply set $\lambda \to 0$ because A is **wide**, and therefore $A^\mathsf{T} A$ will not be invertible.
- Use the fact that: $A^\mathsf{T} A A^\mathsf{T} + \lambda A^\mathsf{T}$ can be factored two ways:

$$(A^\mathsf{T} A + \lambda I) A^\mathsf{T} = A^\mathsf{T} A A^\mathsf{T} + \lambda A^\mathsf{T} = A^\mathsf{T} (A A^\mathsf{T} + \lambda I)$$

$$(A^\mathsf{T} A + \lambda I) A^\mathsf{T} = A^\mathsf{T} (A A^\mathsf{T} + \lambda I)$$

$$A^\mathsf{T} (A A^\mathsf{T} + \lambda I)^{-1} = (A^\mathsf{T} A + \lambda I)^{-1} A^\mathsf{T}$$

Proof of minimum-norm equivalence

Solution of 2-norm regularization is:

$$\hat{x} = (A^\mathsf{T} A + \lambda I)^{-1} A^\mathsf{T} b$$

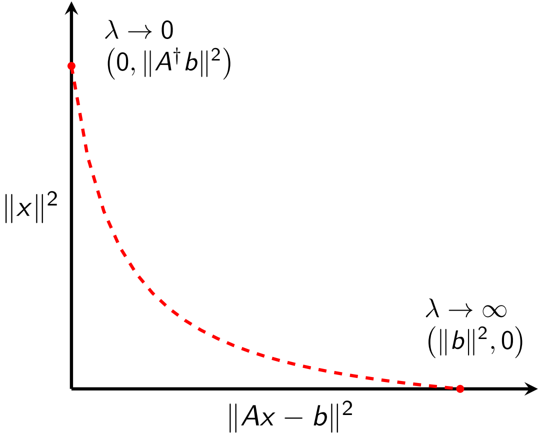
Also equal to:

$$\hat{x} = A^\mathsf{T} (A A^\mathsf{T} + \lambda I)^{-1} b$$

- Since AA^T is invertible, we can take the limit $\lambda \to 0$ by just setting $\lambda = 0$.
- In the limit: $\hat{x} = A^\mathsf{T}(AA^\mathsf{T})^{-1} b$. This is the exact solution to the minimum-norm least squares problem we found before!

Tradeoff visualization

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda \|x\|^2$$



Regularization

Regularization: Additional penalty term added to the cost function to encourage a solution with desirable properties.

Regularized least squares:

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda R(x)$$

- $R(x)$ is the regularizer (penalty function)
- λ is the regularization parameter
- The model has different names depending on $R(x)$.

Regularization

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda R(x)$$

1. If $R(x) = \|x\|^2 = x_1^2 + x_2^2 + \cdots + x_n^2$
   It is called: L_2 *regularization*, *Tikhonov regularization*, or *Ridge regression* depending on the application. It has the effect of **smoothing** the solution.

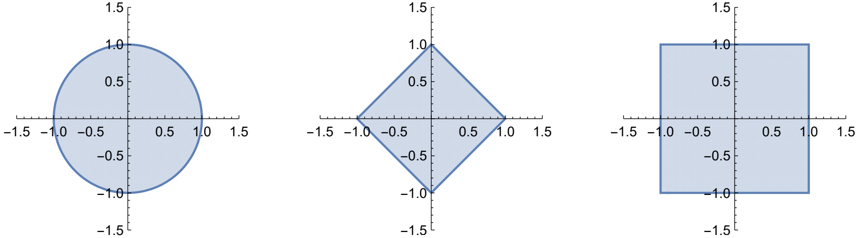
2. If $R(x) = \|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$
   It is called: L_1 *regularization* or *LASSO*. It has the effect of **sparsifying** the solution ($\hat{x}$ will have few nonzero entries).

3. $R(x) = \|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$
   It is called L_∞ *regularization* and it has the effect of **equalizing** the solution (makes most components equal).

Norm balls

For a norm $\|\cdot\|_p$, the **norm ball** of radius r is the set:

$$B_r = \{x \in \mathbb{R}^n \mid \|x\|_p \leq r\}$$



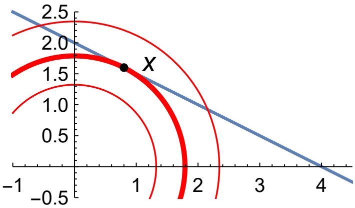
$\|x\|_2 \leq 1$	$\|x\|_1 \leq 1$	$\|x\|_\infty \leq 1$
$x^2 + y^2 \leq 1$	$\lvert x\rvert + \lvert y\rvert \leq 1$	$\max\{\lvert x\rvert, \lvert y\rvert\} \leq 1$

Simple example

Consider the minimum-norm problem for different norms:

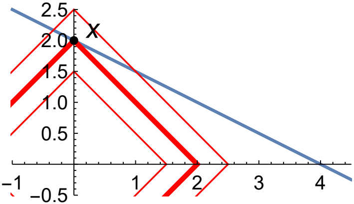
$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \|x\|_p \\ \text{subject to:} \quad & Ax = b \end{aligned}$$

- set of solutions to $Ax = b$ is an affine subspace
- solution is point belonging to smallest norm ball
- for $p = 2$, this occurs at the perpendicular distance

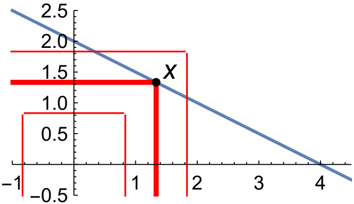


Simple example

- for $p = 1$, this occurs at one of the axes.
- sparsifying behavior



- for $p = \infty$, this occurs at equal values of coordinates
- equalizing behavior



Another simple example

Suppose we have data points $\{y_1, \dots, y_m\} \subset \mathbb{R}$, and we would like to find the best estimator for the data, according to different norms. Suppose data is sorted: $y_1 \le \cdots \le y_m$.

$$\underset{x}{\text{minimize}} \quad \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} x \\ \vdots \\ x \end{bmatrix} \right\|_p$$

- $p = 2$: $\hat{x} = \frac{1}{m}(y_1 + \cdots + y_m)$. This is the **mean** of the data.
- $p = 1$: $\hat{x} = y_{\lceil m/2 \rceil}$. This is the **median** of the data.
- $p = \infty$: $\hat{x} = \frac{1}{2}(y_1 + y_m)$. This is the **mid-range** of the data.

Julia demo: Data Norm.ipynb

Example: hovercraft revisited

One-dimensional version of the hovercraft problem:

- Start at $x_1 = 0$ with $v_1 = 0$ (at rest at position zero)
- Finish at $x_{50} = 100$ with $v_{50} = 0$ (at rest at position 100)
- Same simple dynamics as before:

$$\begin{aligned} x_{t+1} &= x_t + v_t \\ v_{t+1} &= v_t + u_t \end{aligned} \qquad \text{for: } t = 1, 2, \ldots, 49$$

- Decide thruster inputs $u_1, u_2, \ldots, u_{49}$.
- This time: minimize $\|u\|_p$

Example: hovercraft revisited

$$
\begin{aligned}
\underset{x_t, v_t, u_t}{\text{minimize}} \quad & \|u\|_p \\
\text{subject to:} \quad & x_{t+1} = x_t + v_t && \text{for } t = 1, \dots, 49 \\
& v_{t+1} = v_t + u_t && \text{for } t = 1, \dots, 49 \\
& x_1 = 0, \quad x_{50} = 100 \\
& v_1 = 0, \quad v_{50} = 0
\end{aligned}
$$

- This model has 150 variables, but very easy to understand.
- We can simplify the model considerably...

Model simplification

$$x_{t+1} = x_t + v_t \qquad \text{for: } t = 1, 2, \ldots, 49$$
$$v_{t+1} = v_t + u_t$$

$$\begin{aligned} v_{50} &= v_{49} + u_{49} \\ &= v_{48} + u_{48} + u_{49} \\ &= \ldots \\ &= v_1 + (u_1 + u_2 + \cdots + u_{49}) \end{aligned}$$

Model simplification

$$\begin{aligned} x_{t+1} &= x_t + v_t \\ v_{t+1} &= v_t + u_t \end{aligned} \qquad \text{for: } t = 1, 2, \dots, 49$$

$$\begin{aligned} x_{50} &= x_{49} + v_{49} \\ &= x_{48} + 2v_{48} + u_{48} \\ &= x_{47} + 3v_{47} + 2u_{47} + u_{48} \\ &= \dots \\ &= x_1 + 49v_1 + (48u_1 + 47u_2 + \cdots + 2u_{47} + u_{48}) \end{aligned}$$

Model simplification

$$x_{t+1} = x_t + v_t \qquad \text{for: } t = 1, 2, \dots, 49$$
$$v_{t+1} = v_t + u_t$$

Constraint can be rewritten as:

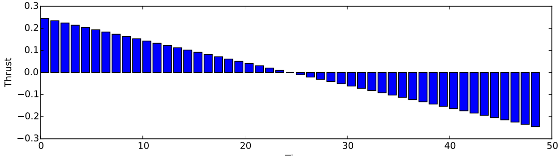
$$\begin{bmatrix} 48 & 47 & \dots & 2 & 1 & 0 \\ 1 & 1 & \dots & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{49} \end{bmatrix} = \begin{bmatrix} x_{50} - x_1 - 49 v_1 \\ v_{50} - v_1 \end{bmatrix}$$

so we don't need the intermediate variables x_t and v_t!

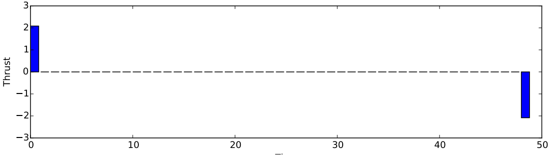
Julia demo: Hover 1D.ipynb

Results

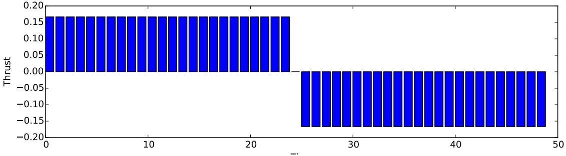
1. Minimizing $\|u\|_2^2$ (smooth)



2. Minimizing $\|u\|_1$ (sparse)

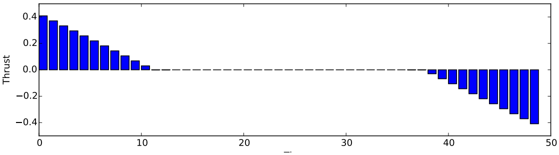


3. Minimizing $\|u\|_\infty$ (equalized)

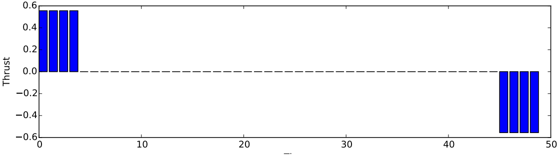


Tradeoff studies

1. Minimizing $\|u\|_2^2 + \lambda\|u\|_1$ (smooth and sparse)



2. Minimizing $\|u\|_\infty + \lambda\|u\|_1$ (equalized and sparse)



3. Minimizing $\|u\|_2^2 + \lambda\|u\|_\infty$ (equalized and smooth)