Regularization

CPSC 406 – Computational Optimization

Regularized least squares

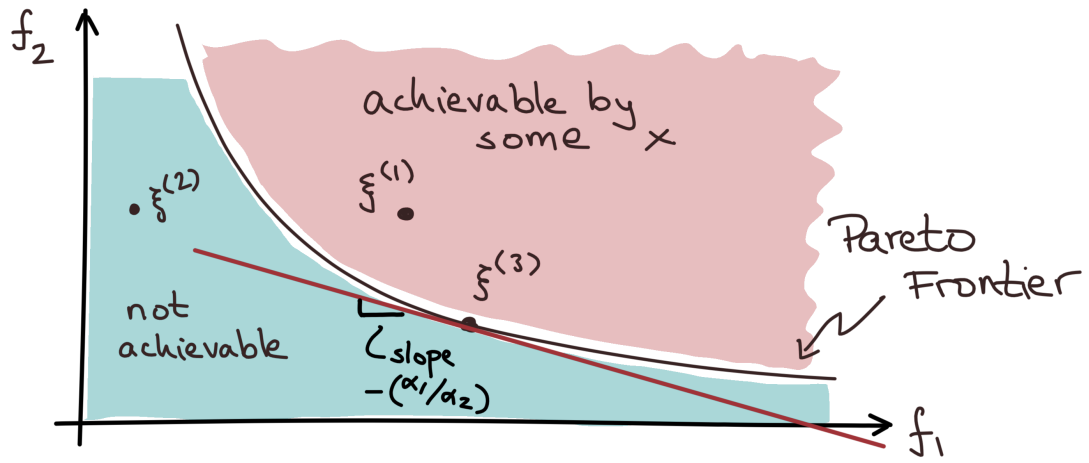
- competing objectives
- Tikhonov regularization (ridge regression)
- least-norm solutions

Multi-objective optimization

Many problems need to balance competing objectives, eg,

- choose x to minimize $f_1(x)$ **or** $f_2(x)$
- can make f_1 **or** f_2 small, but not both, eg

$$x^* = \underset{x}{\operatorname{argmin}}\, f_1(x) \quad \Longrightarrow \quad f_1(x^*) \leq f_2(x^*)$$



The Pareto frontier

$$\xi^{(i)} := \{f_1(x^{(i)}),\ f_2(x^{(i)})\}$$

- $\xi^{(1)}$ is not efficient
- $\xi^{(3)}$ dominates $\xi^{(1)}$
- $\xi^{(2)}$ is infeasible

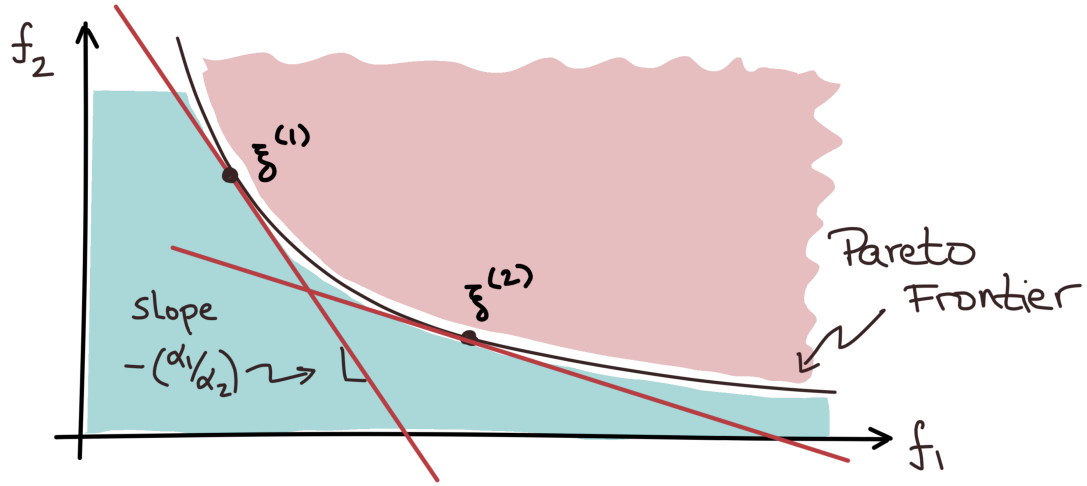
Weighted-sum objective

Common approach is to weight the sum of objectives:

$$\min_x \ \alpha_1 f_1(x) + \alpha_2 f_2(x)$$

- penalty parameters $\alpha_1, \alpha_2 > 0$
- ratio $\frac{\alpha_1}{\alpha_2}$ determines relative objective weights

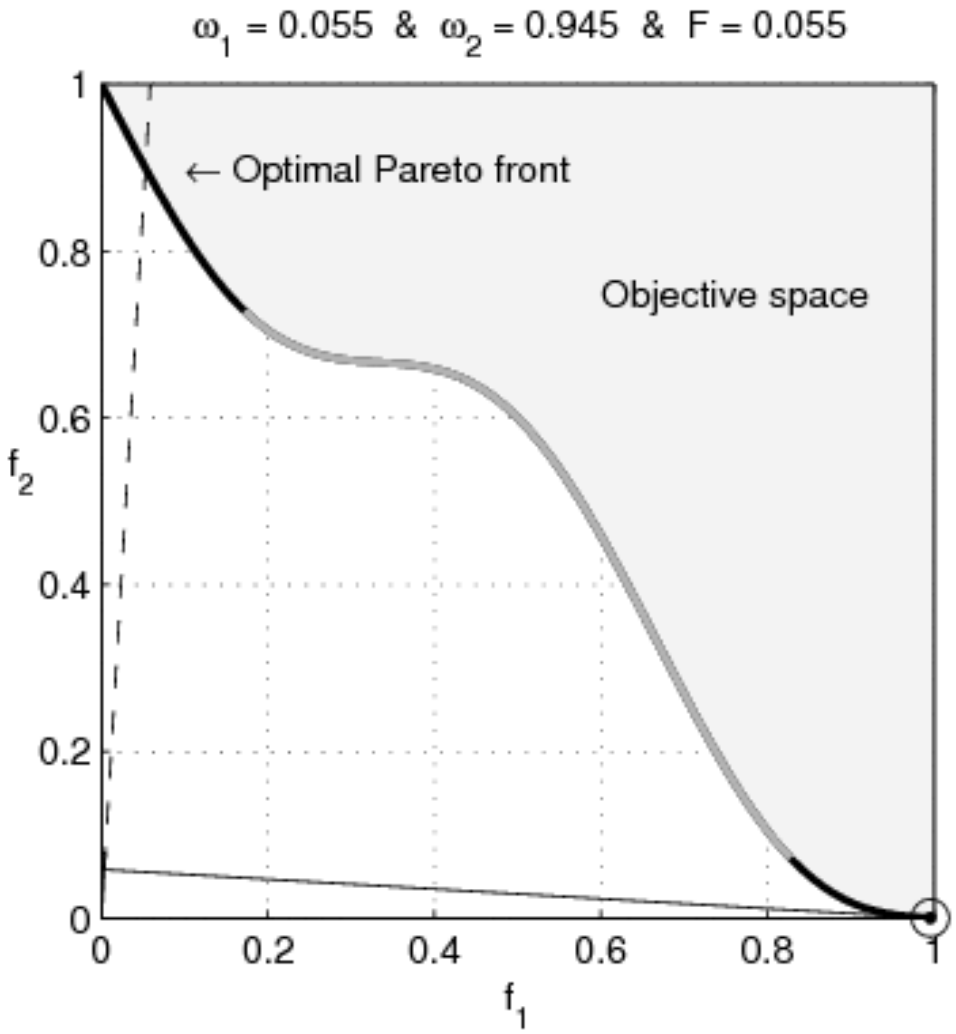
$$\operatorname*{argmin}_x \{ \alpha_1 f_1(x) + \alpha_2 f_2(x)\} = \operatorname*{argmin}_x \{ f_1(x) + \frac{\alpha_2}{\alpha_1} f_2(x)\}$$



Two efficient solutions on the Pareto Frontier

- $x^{(1)}$ solution for $\alpha_1 > \alpha_2$
- $x^{(2)}$ solution for $\alpha_2 > \alpha_1$

Nonconvex Pareto frontier



Tikhonov regularization

$$\min_x \ \tfrac{1}{2}\|Ax - b\|^2 + \tfrac{1}{2}\lambda\|Dx\|^2$$

- known as **ridge regression** in statistics and machine learning
- $\|Dx\|^2$ is the **regularization penalty** (often $D := I$)
- λ is the positive **regularization parameter**
- equivalent expression for objective

$$\|Ax - b\|^2 + \lambda\|Dx\|^2 = \left\| \begin{bmatrix} A \\ \sqrt{\lambda}D \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2$$

- Normal equations (unique solution if D full rank)

$$(A^\intercal A + \lambda D^\intercal D)x = A^\intercal b$$

Question

Given tall $m \times n$ matrix A with SVD $A = U\Sigma V^T$, where Σ is the diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. Let

$$M = \begin{bmatrix} A \\ \sqrt{\lambda}\, I \end{bmatrix}$$

For $i = 1, \ldots, n$, are the singular values of M are

a. σ_i

b. $\sigma_i^2 + \lambda$

c. $\sqrt{\sigma_i^2 + \lambda}$

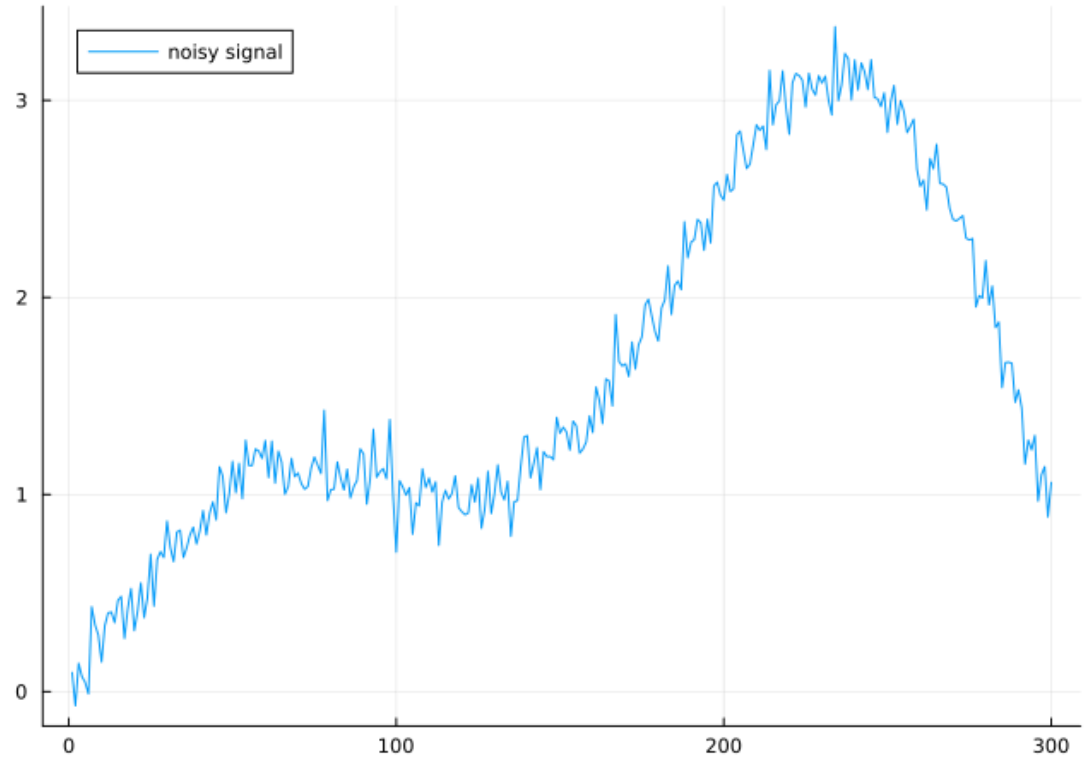
d. $\sigma_i + \sqrt{\lambda}$

Hint: Recall that the singular values of M are the square roots of the eigenvalues of $M^T M$.

Example — signal denoising

observe noisy measurements y of a signal

$$b = \hat{x} + \eta \quad \text{where} \quad \begin{cases} x^\natural \in \mathbb{R}^n \text{ is true signal} \\ \eta \in \mathbb{R}^n \text{ is noise} \end{cases}$$

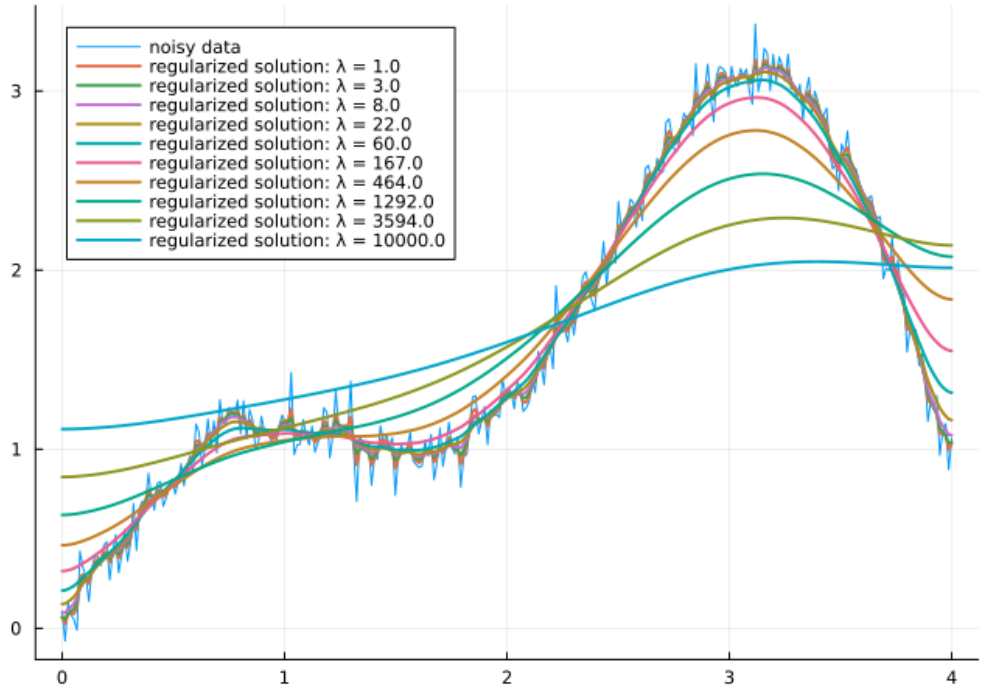


Example – least-squares formulation

- Naive least squares fits noise perfectly: b solves $\min_x \|x - b\|^2$
- assumption: $x^\natural$ is **smooth** $\implies$ balance fidelity to data against smoothness

$$\min_x \underbrace{\|x - b\|^2}_{f_1(x)} + \underbrace{\lambda \sum_{i=1}^{n-1} \|x_i - x_{i+1}\|^2}_{f_2(x)}$$

- $f_2(x)$ is penalizes **jumps** in signal



Example — matrix notation

- Define the $(n-1) \times n$ finite difference matrix

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \quad \Longrightarrow \quad \sum_{i=1}^{n-1} \|x_i - x_{i+1}\|^2 = \|Dx\|^2$$

- least-squares objective

$$\|x - b\|^2 + \lambda \|Dx\|^2 = \left\| \begin{bmatrix} I \\ \sqrt{\gamma} D \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2$$

- Normal equations

$$(I + \gamma D^\mathsf{T} D) x = b$$

Sparse matrices

D is sparse: only $2n - 2$ nonzero entries (2 per row)

```
using SparseArrays: spdiagm
finiteDiff(n) = spdiagm(0 => ones(n), +1 => -ones(n-1))[1:n-1,:]
finiteDiff(4)
```

```
3×4 SparseArrays.SparseMatrixCSC{Float64, Int64} with 6 stored entries:
 1.0  -1.0    ·     ·
  ·    1.0  -1.0    ·
  ·     ·    1.0  -1.0
```

Equivalent formulations

$$\min_x \{ f_1(x) + \lambda f_2(x) \}$$
$$\min_x \{ f_1(x) \mid f_2(x) \leq \tau \}$$
$$\min_x \{ f_2(x) \mid f_1(x) \leq \sigma \}$$

▶ Code

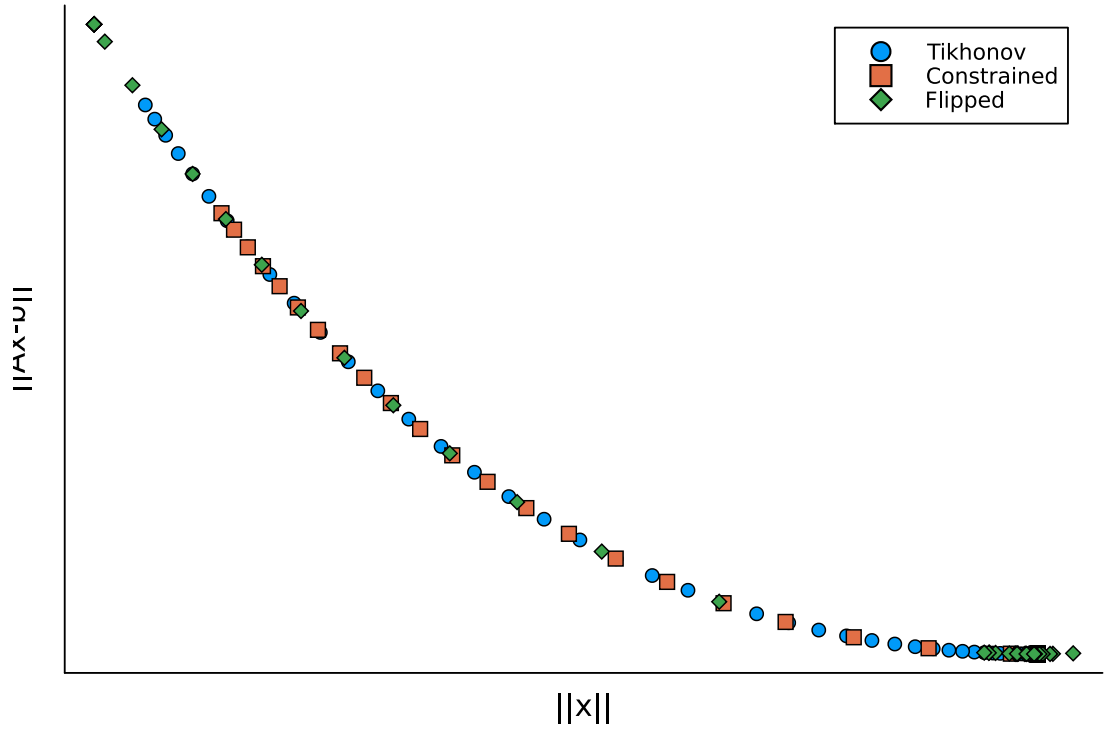


Figure 1