6. Regularization functions

- multiobjective optimization problems
- optimal and Pareto optimal
- salarization and trade-off analysis
- regularized approximation
- Tikhonov and sparsity regularization
- smoothing
- maximum a posteriori estimation
- matrix regularizers

Multiobjective optimization

$$\begin{array}{ll} \text{minimize} & f_0(x) = (F_1(x), \ldots, F_k(x)) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

- $f_0 \colon \mathbf{R}^n \to \mathbf{R}^k$ is vector-valued objective; ith component $F_i \colon \mathbf{R}^n \to \mathbf{R}$
- also called *vector optimization* or *multicriteria optimization*

convex multiobjective problem if

- for all $x, y \in \mathbf{dom}\, f_0$ and $\theta \in [0, 1]$,

$$f_0(\theta x + (1-\theta)y) \preceq \theta f_0(x) + (1-\theta) f_0(y)$$

  i.e., F_i convex for all $i = 1, \ldots, k$

- f_i convex for $i = 1, \ldots, m$ and h_i affine for $i = 1, \ldots, p$

Interpretations

scalar optimization ($k = 1$):

- $f_0(x) \leq f_0(y)$: x no worse than y
- $f_0(x) \geq f_0(y)$: x no better than y
- $f_0(x) = f_0(y)$: x and y are equally good

find $x \in \mathbf{dom}\, f_0$ such that $f_0(x) \leq f_0(y)$ for all $y \in \mathbf{dom}\, f_0$

vector optimization:

- $f_0(x) \preceq f_0(y)$: x no worse than y in all objectives F_i
- $f_0(x) \succeq f_0(y)$: x no better than y in all objectives F_i
- $f_0(x) = f_0(y)$: x and y are equally good in all objectives F_i
- $f_0(x) \not\preceq f_0(y)$ and $f_0(x) \not\succeq f_0(y)$: x better than y in some objectives F_i and worse in others, *i.e.*, x and y are *incomparable*

'minimize' is ambiguous in the last case (which is, sadly, mostly the case)

Optimal value and points

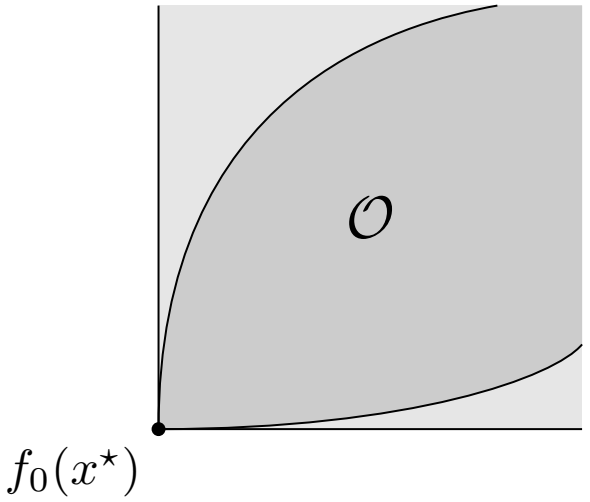
feasible point $x^\star$ is **optimal** if for *all* feasible x

$$f_0(x^\star) \preceq f_0(x) \quad \iff \quad F_i(x^\star) \leq F_i(x), \quad i = 1, \ldots, k$$

- $f_0(x^\star)$ is called **optimal value**
- implicit requirement: $f_0(x^\star)$ is comparable to $f_0(x)$ for all feasible x

set $\mathcal{O} \subseteq \mathbf{R}^k$ of **achievable objective values**:

$$\mathcal{O} = \left\{ f_0(x) \;\middle|\; \begin{array}{ll} x \in \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \cap \bigcap_{i=1}^{p} \mathbf{dom}\, h_i \\ f_i(x) \leq 0, & i = 1, \ldots, m \\ h_i(x) = 0, & i = 1, \ldots, p \end{array} \right\}$$



so $x^\star$ optimal $\iff$ $\mathcal{O} \subseteq f_0(x^\star) + \mathbf{R}^k_+$

most multiobjective problems do not have an optimal point

Pareto optimal values and points

feasible point $x^{\rm po}$ is **Pareto optimal** or *efficient* if for all feasible x

$$f_0(x) \preceq f_0(x^{\rm po}) \quad \Longrightarrow \quad f_0(x) = f_0(x^{\rm po})$$

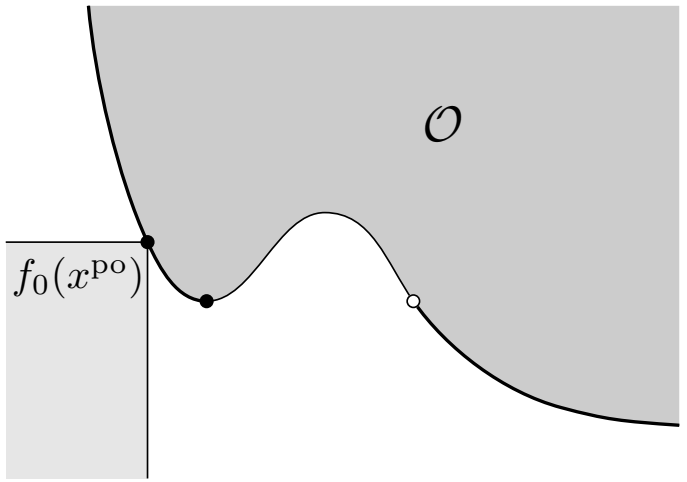
i.e., $x^{\rm po}$ is the best among all comparable feasible points

- $f_0(x^{\rm po})$ is called **Pareto optimal value**
- usually many Pareto optimal points
- feasible $x^{\rm po}$ Pareto optimal $\iff$

$$\mathcal{O} \cap (f_0(x^{\rm po}) - \mathbf{R}^k_+) = \{f_0(x^{\rm po})\}$$

- **Pareto front** is the set of values

$$\{f_0(x^{\rm po}) \mid x^{\rm po} \text{ Pareto optimal}\}$$



Scalarization

consider the (possibly nonconvex) scalar optimization problem

$$\begin{array}{ll} \text{minimize} & \lambda^T f_0(x) = \sum_{i=1}^k \lambda_i F_i(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

- variable $x \in \mathbf{R}^n$; vector-valued objective $f_0 \colon \mathbf{R}^n \to \mathbf{R}^k$
- parameter $\lambda \in \mathbf{R}^k_{++}$ is a given **weight vector**
- optimal point is Pareto optimal for the original multiobjective problem
  - to show this, suppose optimal x is not Pareto optimal, *i.e.*, for some y

$$f_0(y) \preceq f_0(x), \ f_0(y) \neq f_0(x) \implies \lambda^T (f_0(x) - f_0(y)) > 0$$

then $\lambda^T f_0(y) < \lambda^T f_0(x)$, contradicting optimality of x

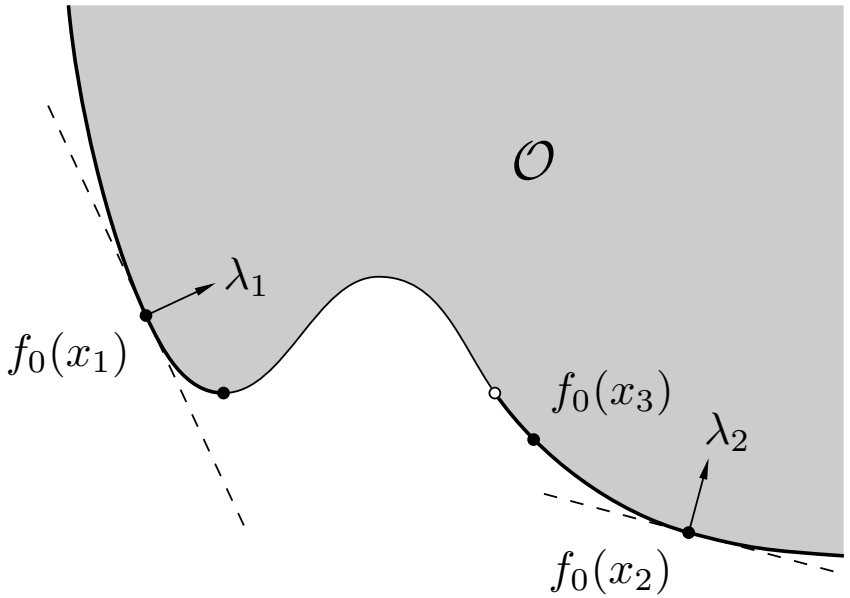
- varying $\lambda \succ 0$ gives (possibly) different Pareto optimal points
- but not all Pareto optimal points can be obtained this way

geometric interpretation:

$$x \text{ optimal} \iff \lambda^T(z - f_0(x)) \geq 0 \text{ for all } z \in \mathcal{O}$$

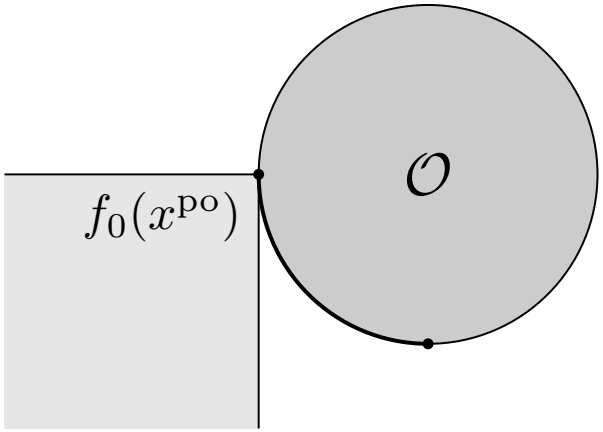
i.e., hyperplane $\{z \mid \lambda^T(z - f_0(x)) = 0\}$ supports $\mathcal{O}$ at $f_0(x)$

- $f_0(x_1)$, $f_0(x_2)$, $f_0(x_3)$ are Pareto optimal values
- $f_0(x_1)$ and $f_0(x_2)$ can be obtained by scalarization with $\lambda \succ 0$
- there is no supporting hyperplane of $\mathcal{O}$ at $f_0(x_3)$

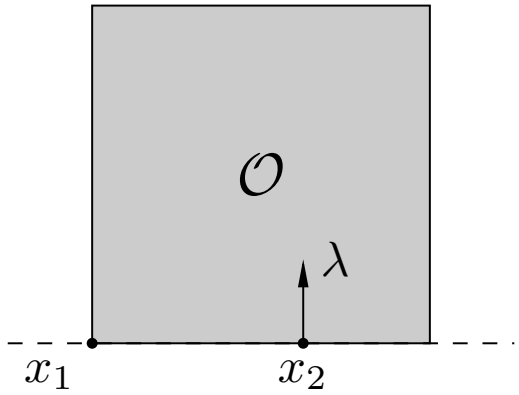


scalarization of convex problems:

- convex multiobjective problem has convex scalarization for any $\lambda \succeq 0$
- a partial converse holds: *all* Pareto optimal points can be obtained by scalarization with $\lambda \succeq 0$ $(\lambda \neq 0)$
- *not* all optimal points x from scalarization $\lambda \succeq 0$ are Pareto optimal



Pareto optimal value from $\lambda \succeq 0$
(but $\lambda \not\succeq 0$)



x_2 optimal from $\lambda \succeq 0$
but not Pareto optimal

- in practice: find all optimal x with $\lambda \succeq 0$, and then check Pareto optimality for $\lambda \not\succeq 0$

Trade-off analysis

consider biobjective problem with no optimal

$$\begin{array}{ll} \text{minimize} & (F_1(x), F_2(x)) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

- to achieve $F_1(x) - \delta$, how much worse do we have to accept in F_2?
- if we accept worsen $F_1(x) + \delta$, how much can we improve F_2?

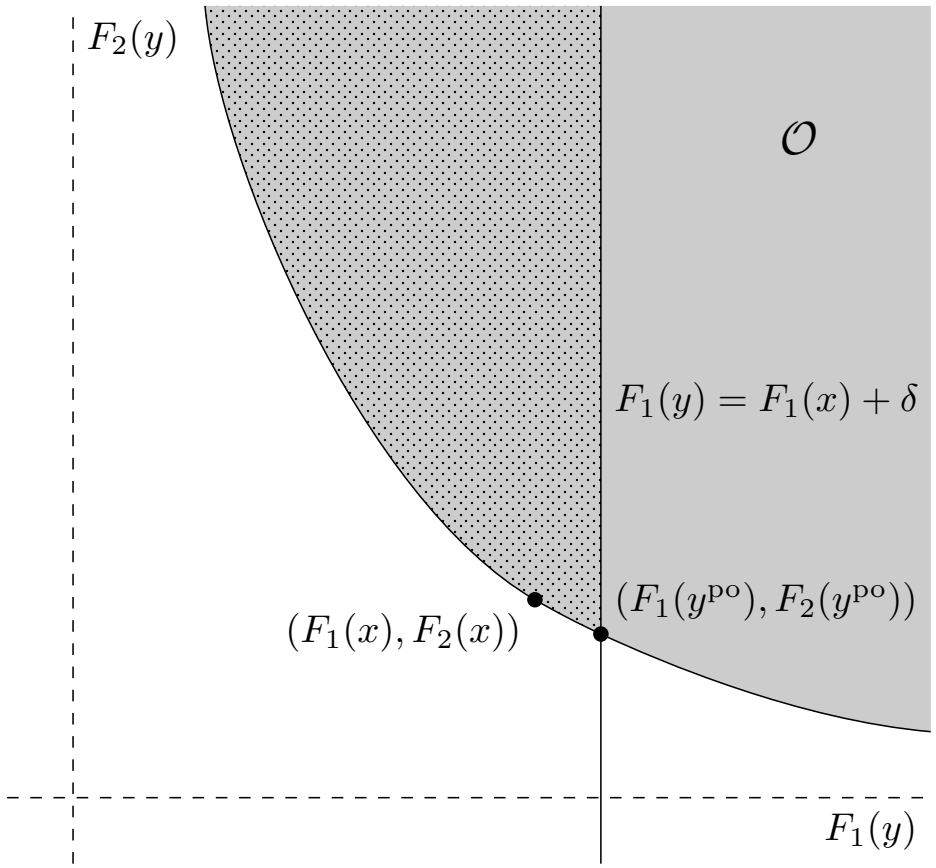
to answer these questions, we solve

$$\begin{array}{ll} \text{minimize} & F_2(y) \\ \text{subject to} & F_1(y) \leq F_1(x) + \delta \\ & f_i(y) \leq 0, \quad i = 1, \ldots, m \\ & h_i(y) = 0, \quad i = 1, \ldots, p \end{array}$$

- $y \in \mathbf{R}^n$ is the variable; $F_1(x), \delta \in \mathbf{R}$ are given

geometric interpretation: trade-off analysis problem equivalent to

$$\begin{array}{ll} \text{minimize} & F_2(y) \\ \text{subject to} & F_1(y) \leq F_1(x) + \delta \\ & (F_1(y), F_2(y)) \in \mathcal{O} \end{array}$$

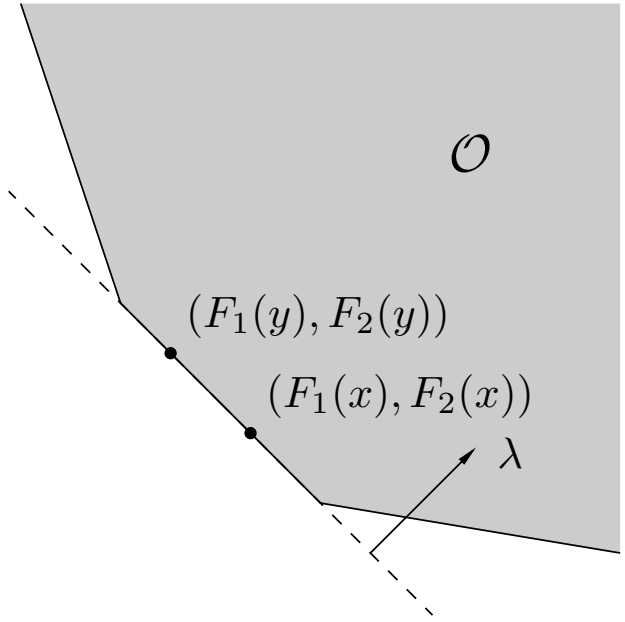


scalarization interpretation: for $\lambda^T f_0(x) = \sum_{i=1}^k \lambda_i F_i(x)$

- $\lambda_i > 0$ is the *importance*, *concern*, or *irritation* of objective F_i
- so adjusting $\lambda \succ 0$ we are trading-off between the objectives
- ratio λ_i/λ_j is the *relative importance* or *exchange rate* of F_i to F_j

example: Pareto optimal x, y with

$$\lambda_1 F_1(x) + \lambda_2 F_2(x) = \lambda_1 F_1(y) + \lambda_2 F_2(y)$$



suppose $\delta = F_1(x) - F_1(y) > 0$, then

$$F_2(y) = F_2(x) + (\lambda_1/\lambda_2)\delta$$

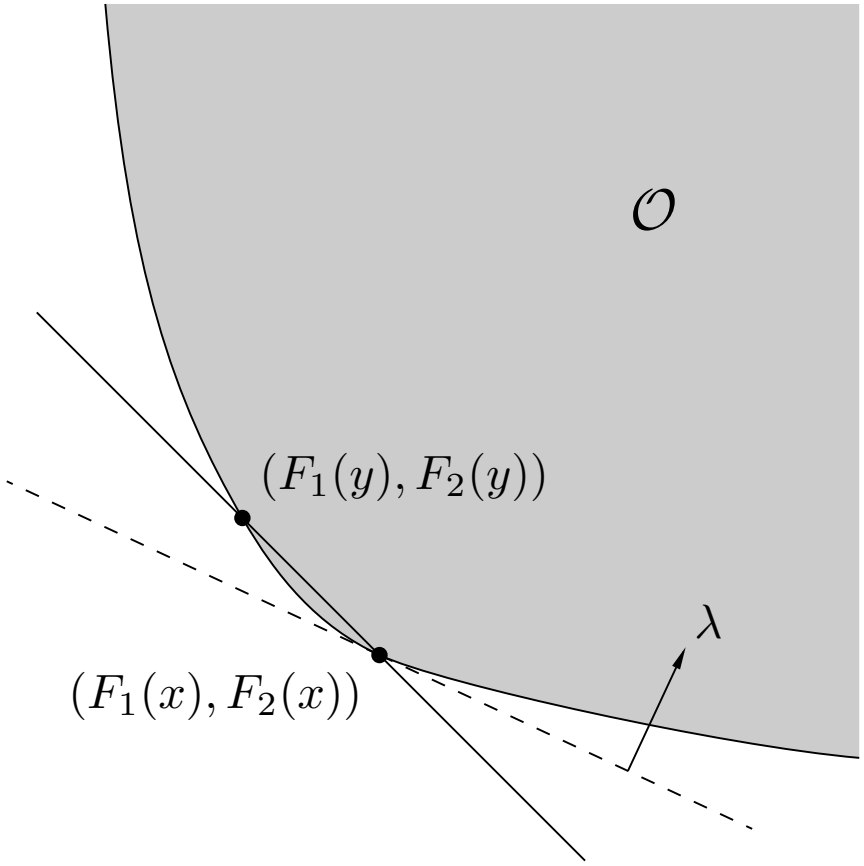
in other words,

$$\frac{F_2(y) - F_2(x)}{F_1(y) - F_1(x)} = -\frac{\lambda_1}{\lambda_2}$$

in the most general case, as $\delta = F_1(x) - F_1(y) \to 0$, we have

$$\lim_{\delta \to 0} \frac{F_2(y) - F_2(x)}{F_1(y) - F_1(x)} = -\frac{\lambda_1}{\lambda_2}$$

so λ gives the slope of the supporting hyperplane of $\mathcal{O}$ at $(F_1(x), F_2(x))$



Regularized approximation

$$\text{minimize} \quad (\phi_1(f(x) - b),\ \phi_2(x))$$

- $x \in \mathbf{R}^n$ is the variable; $f(x) - b \in \mathbf{R}^m$ is the residual
- $\phi_1 : \mathbf{R}^m \to \mathbf{R}$: residual penalty function (*primary objective*)
- $\phi_2 : \mathbf{R}^n \to \mathbf{R}$: regularization function (*secondary objective*)

relative scalarization:

$$\text{minimize} \quad \phi_1(f(x) - b) + \gamma\phi_2(x)$$

- $\gamma > 0$ is the *regularization parameter*
- varying γ gives different Pareto optimal points
- two extreme cases:
  - $\gamma \to 0$: ignore regularization, minimize residual
  - $\gamma \to \infty$: ignore residual, minimize regularization

Tikhonov regularization

$$\text{minimize} \quad (\phi_1(f(x) - b),\ \|x\|_2^2)$$

- $\|x\|_2^2$ called *Tikhonov regularization*; induce small solution x

example: Tikhonov regularization least squares

$$\text{minimize} \quad (\|Ax - b\|_2^2,\ \|x\|_2^2)$$

scalarize with $\lambda \succ 0$:

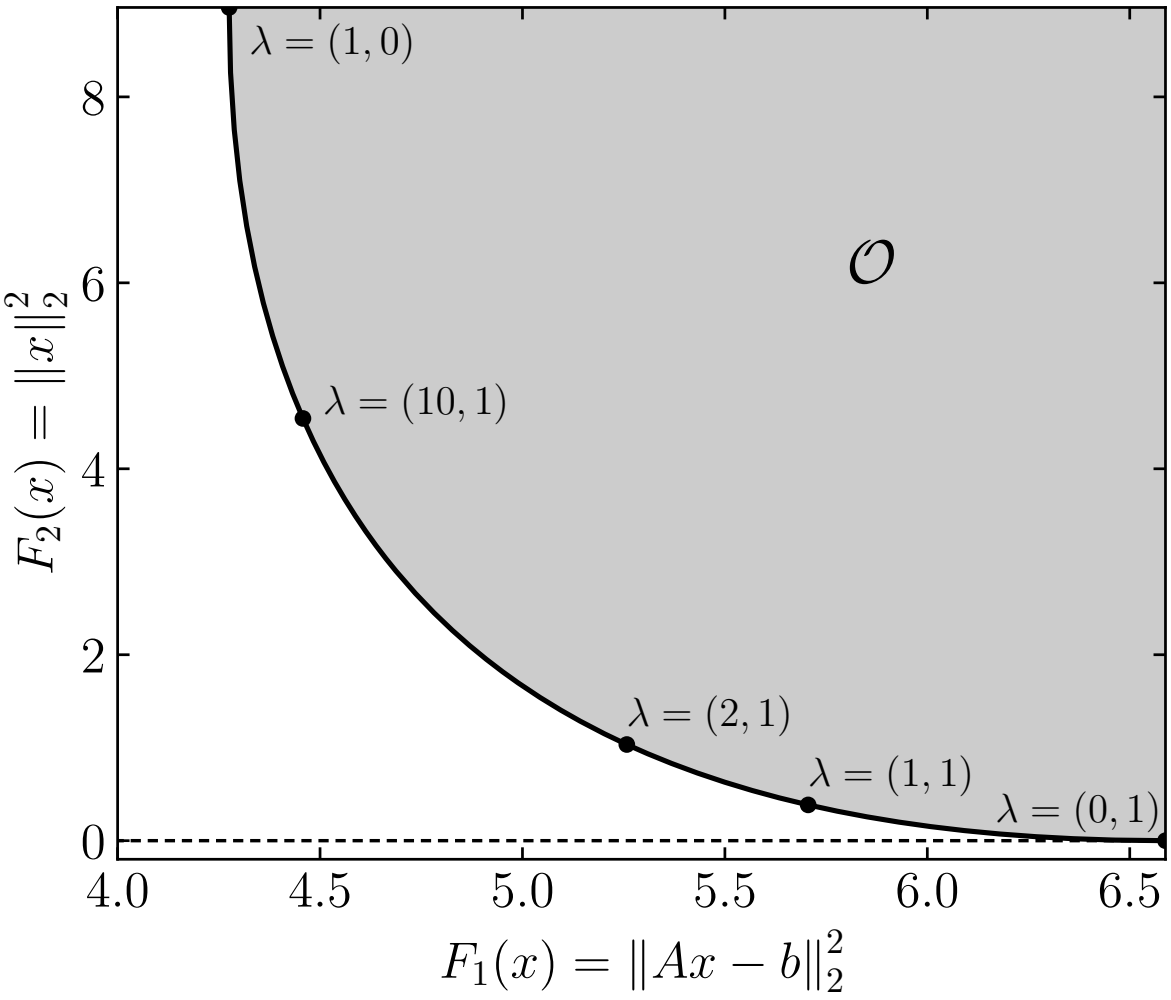
$$\lambda_1\|Ax - b\|_2^2 + \lambda_2\|x\|_2^2 = x^T(\lambda_1 A^T A + \lambda_2 I)x - 2\lambda_1 b^T Ax + \lambda_1 b^T b$$

is a convex quadratic in x, so the optimal point is given by

$$x = (A^T A + \gamma I)^{-1} A^T b, \quad \gamma = \lambda_2/\lambda_1$$

- removes rank requirement on A for least squares
- sometimes called *ridge regression*

- numerical results for $A \in \mathbf{R}^{100\times 10}$:



Sparsity regularization

$$\text{minimize} \quad (\phi_1(f(x) - b),\ \mathbf{card}\, x)$$

- $\mathbf{card}\, x = |\{i \mid x_i \neq 0,\ i = 1, \ldots, n\}|$: number of nonzero components
- induces sparse solution x, *i.e.*, many components of x are zero
- to find the Pareto front, solve for each $\mathbf{card}\, x = 0, 1, \ldots, n$:
  - if n is small, can enumerate all possible sparsity patterns of x
  - very difficult for large n

ℓ_1-norm regularization:

$$\text{minimize} \quad (\phi_1(f(x) - b),\ \|x\|_1)$$

- very useful convex approximation of $\mathbf{card}\, x$

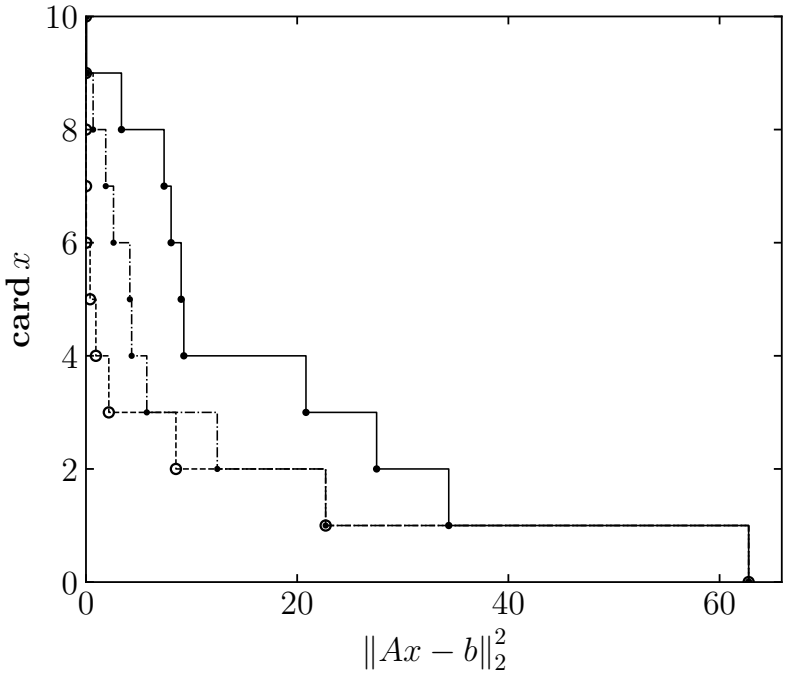
example: ℓ_1-norm regularized least squares, scalarized with $\gamma > 0$:

$$\text{minimize} \quad \|Ax - b\|_2^2 + \gamma \|x\|_1$$

- larger γ gives sparser solution x
- solution with $\mathbf{card}\, x = k$ from smallest γ approximates Pareto optimal
- **polishing**: fix sparsity pattern of x according to ℓ_1-norm heuristic solution and minimize $\|Ax - b\|_2^2$

numerical example for $A \in \mathbf{R}^{10\times 20}$:

- dashes: true Pareto optimal
- solid line: ℓ_1-norm approximation
- dashdots: ℓ_1-norm approximation and polishing



Smoothing

find estimation $\hat{x}$ of true signal $x \in \mathbf{R}^n$ from corrupted observation

$$x_{\rm obs} = x + v, \quad v \in \mathbf{R}^n \text{ is noise}$$

assumption: x is smooth, *i.e.*, $x_i \approx x_{i+1}$ for most $i = 1, \ldots, n-1$

regularized approximation formulation:

$$\text{minimize} \quad (\|\hat{x} - x_{\rm obs}\|_2^2, \ \phi(\hat{x}))$$

- **quadratic smoothing**: $\phi(\hat{x}) = \sum_{i=1}^{n-1} (\hat{x}_{i+1} - \hat{x}_i)^2$

  assumes x does not change too much from i to $i+1$

- **total variation smoothing**: $\phi(\hat{x}) = \sum_{i=1}^{n-1} |\hat{x}_{i+1} - \hat{x}_i|$

  assumes x is piecewise constant, *i.e.*, only a few changes for all i

scalarized smoothing problem with $\gamma > 0$:

$$\text{minimize} \quad \|\hat{x} - x_{\text{obs}}\|_2^2 + \gamma \|D\hat{x}\|$$

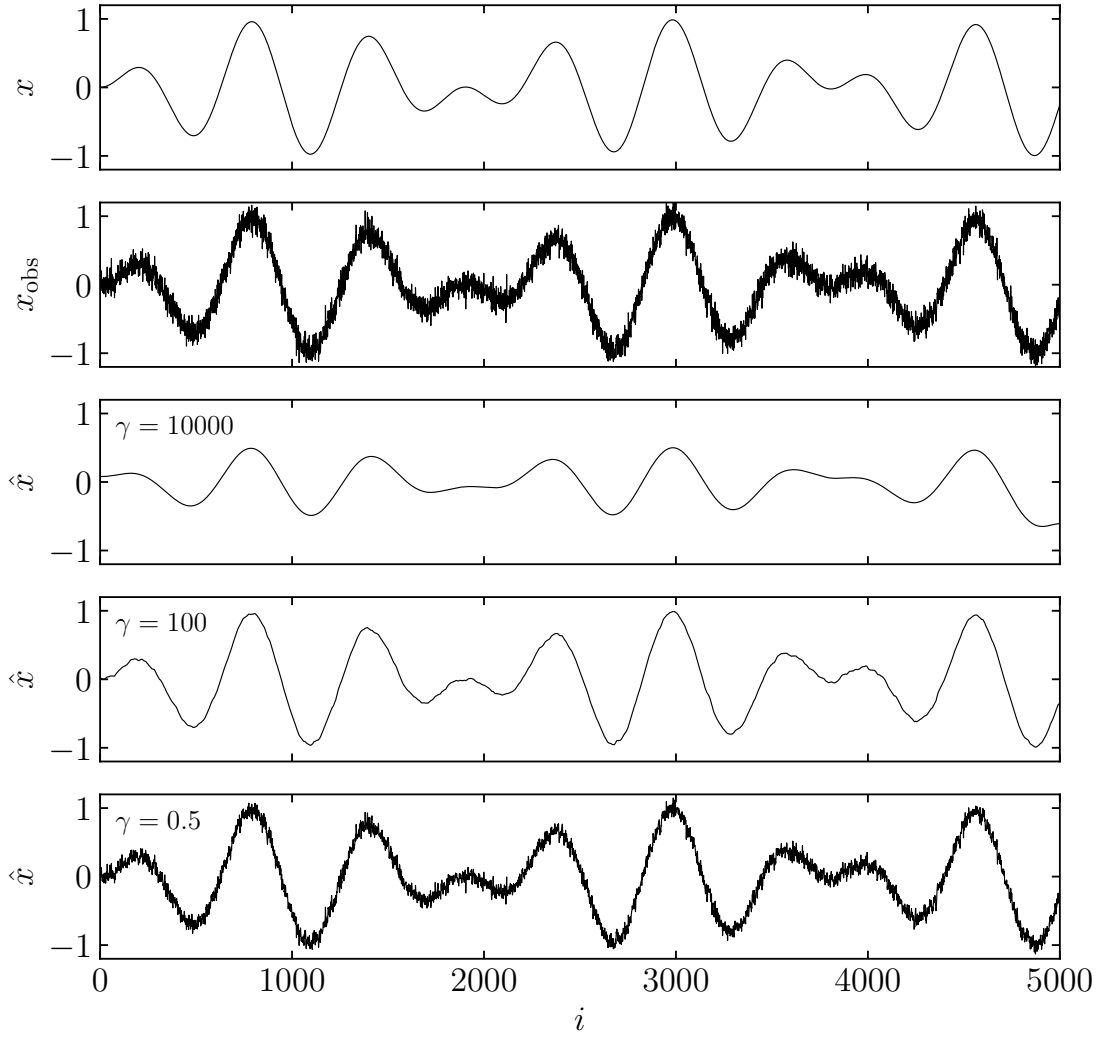
where $D \in \mathbf{R}^{(n-1)\times n}$ is the *first-order difference matrix*, given by

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

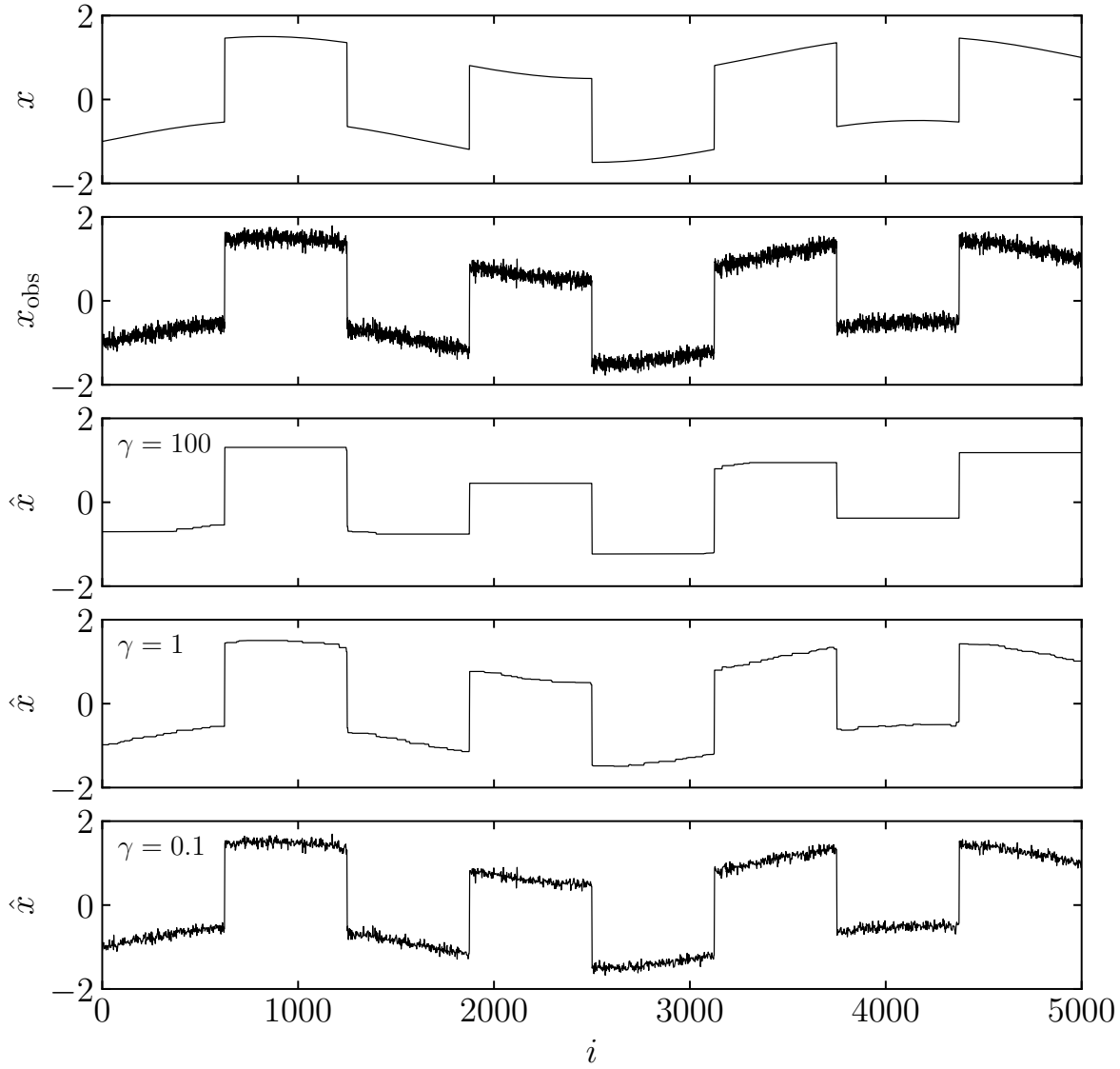
- $\|\cdot\|_1$ gives total variation smoothing
- $\|\cdot\|_2$ gives quadratic smoothing

examples:

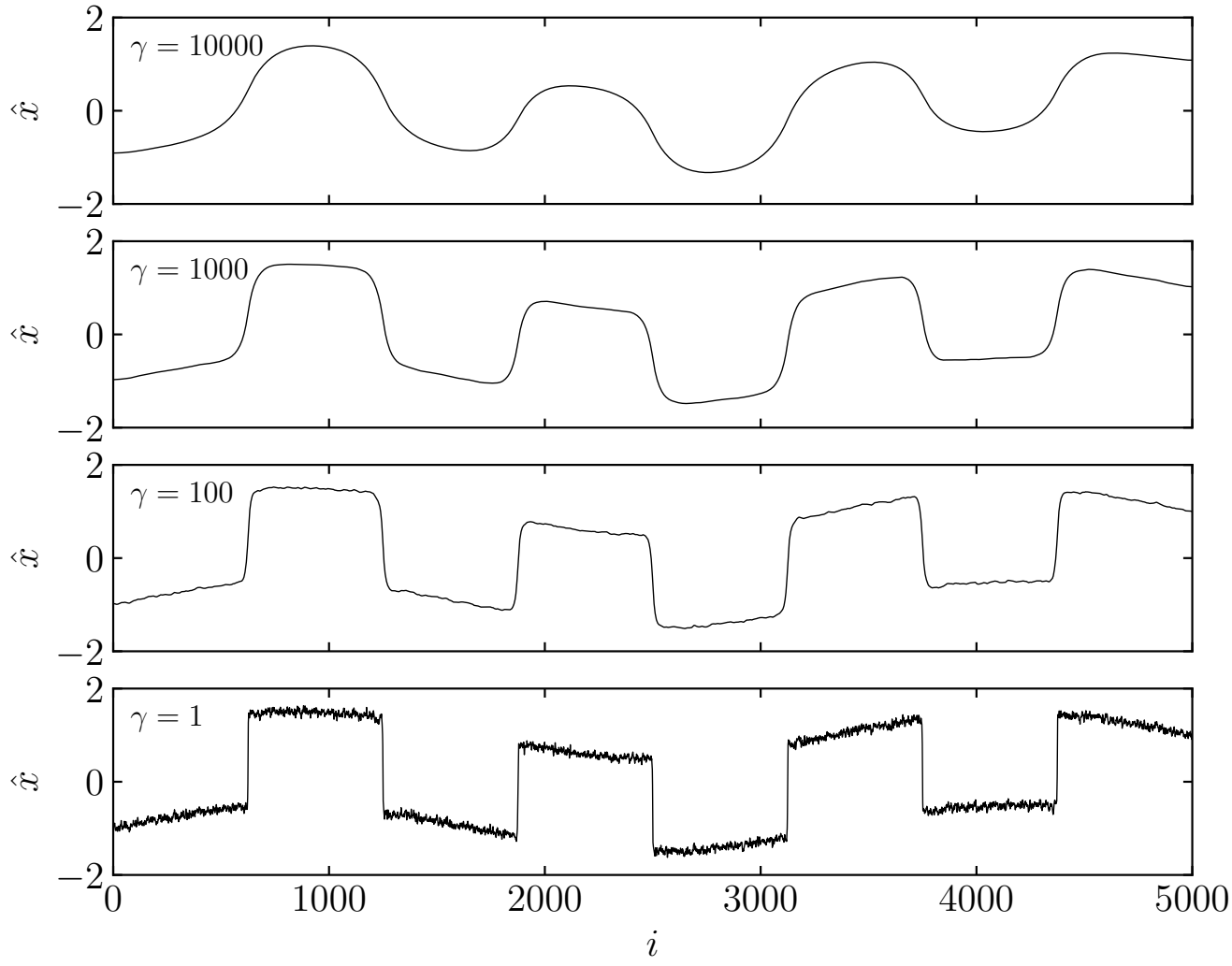
- quadratic smoothing for $x \in \mathbf{R}^{5000}$:



- total variation smoothing for $x \in \mathbf{R}^{5000}$:



- previous example under quadratic smoothing:



Maximum a posteriori (MAP) estimation

$$\text{maximize} \quad \log p_{y|x}(x, y) + \log p_x(x)$$

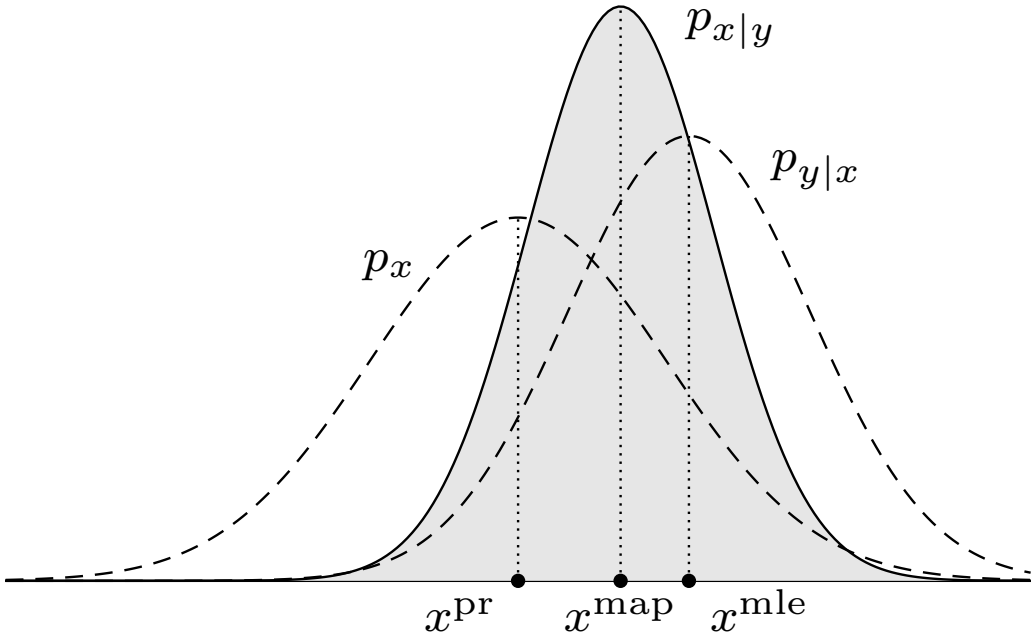
- variable $x \in \mathbf{R}^n$; observation $y \in \mathbf{R}^m$ is given
- $p_{y|x}$: $\mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}_+$: likelihood of y given x
- p_x: $\mathbf{R}^n \to \mathbf{R}_+$: prior probability of x
- according to Bayes' theorem: *posterior probability* of x given y is

$$p_{x|y}(x, y) = \frac{p_{y|x}(x, y)p_x(x)}{p_y(y)}$$

find x that maximizes posterior probability $p_{x|y}(x, y)$, *i.e.*, the most likely x given observation y and prior knowledge of x

(contrast with MLE, which only maximizes likelihood $p_{y|x}(x, y)$)

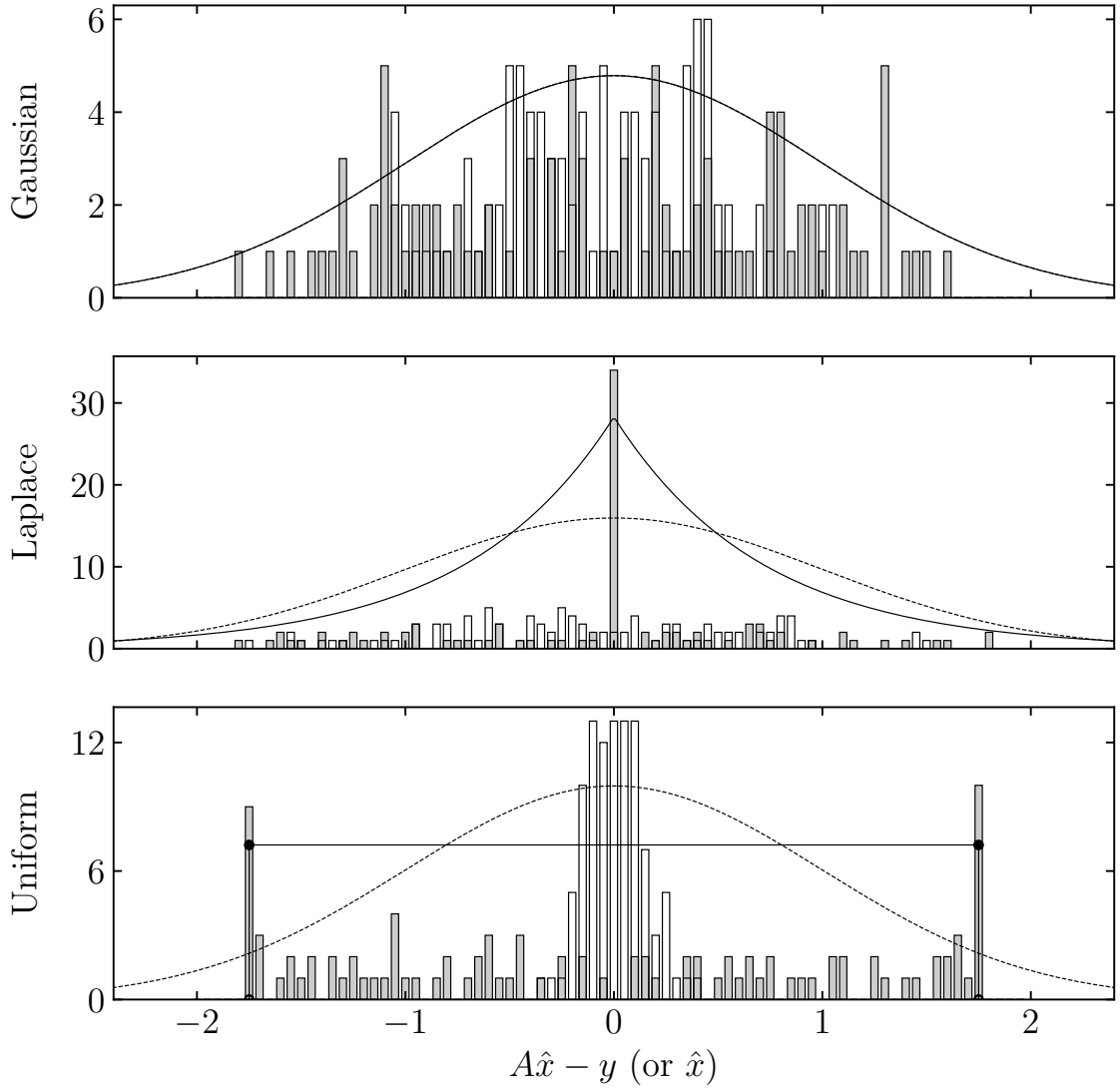
Bayesian interpretation:



regularization interpretation:

- MAP estimation is MLE with some kind of regularization function
- any regularized approximation can be interpreted as an MAP estimation by properly choosing the prior and likelihood densities

example: linear model $y = Ax + v$ ($A \in \mathbf{R}^{100\times 100}$) with IID Gaussian noise v, under Gaussian, Laplace, and uniform priors for x



(residual $A\hat{x} - y$ shown white; estimation $\hat{x}$ shown shaded)

Trade-off in MAP estimation

for linear model $y = Ax + v$ with IID noise $v_i \sim \mathcal{N}(0, 1)$, log-likelihood is

$$\log p_{y|x}(x, y) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\|Ax - y\|_2^2$$

suppose IID Gaussian prior $x \sim \mathcal{N}(0, \sigma_x^2 I)$, then

$$\log p_x(x) = -\frac{n}{2}\log(2\pi\sigma_x^2) - \frac{1}{2\sigma_x^2}\|x\|_2^2$$

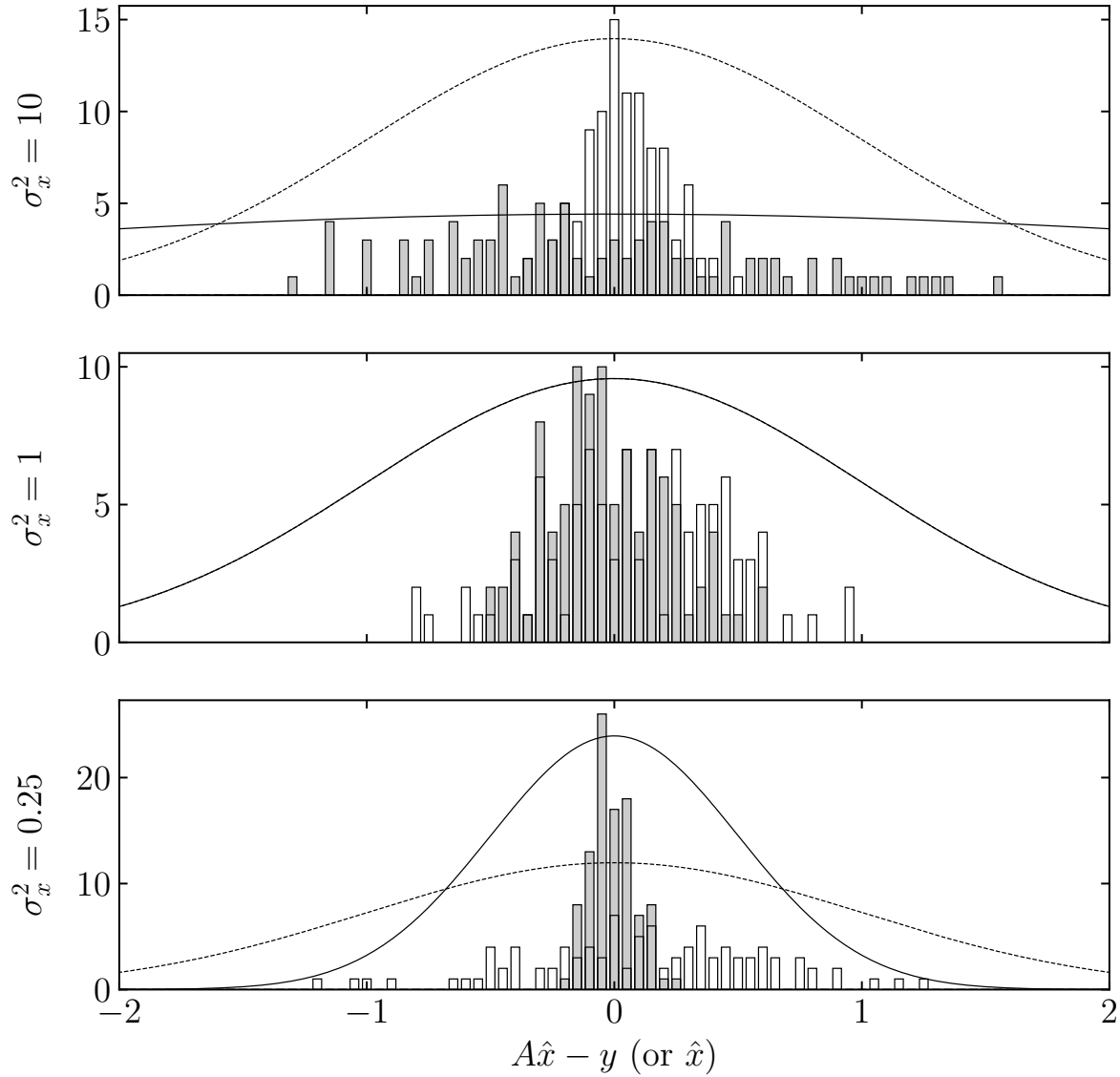
so MAP estimation solves

$$\text{minimize} \quad \|Ax - y\|_2^2 + (1/\sigma_x^2)\|x\|_2^2$$

- variable x, data y, and (hyper)parameter σ_x^2
- σ_x^2 controls the trade-off between likelihood and prior

trade-off parameter in MAP estimation usually implicit in density functions

example for $A \in \mathbf{R}^{100\times 100}$: (estimation $\hat{x}$ shown shaded)



larger σ_x^2 means larger variance in prior, so less regularization effect

Matrix regularizers

$$\text{minimize} \quad f_0(X) + \gamma\phi(X)$$

- variable $X \in \mathbf{R}^{m\times n}$; primary objective $f_0 \colon \mathbf{R}^{m\times n} \to \mathbf{R}$
- regularization function $\phi \colon \mathbf{R}^{m\times n} \to \mathbf{R}$, weighted by $\gamma > 0$

componentwise regularization:

- Tikhonov: $\phi(X) = \|X\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n X_{ij}^2$
- sparsity: $\phi(X) = \|X\|_{\text{sav}} = \sum_{i=1}^m \sum_{j=1}^n |X_{ij}|$

suppose $X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$, **columnwise sparsity** regularizer is

$$\phi(X) = \sum_{i=1}^n \|x_i\|_2$$

(note that there is no square on the ℓ_2-norm)

example: group lasso approximately solves the linear equations

$$\mathbf{tr}(A_i^T X) = b_i, \quad i = 1, \ldots, k$$

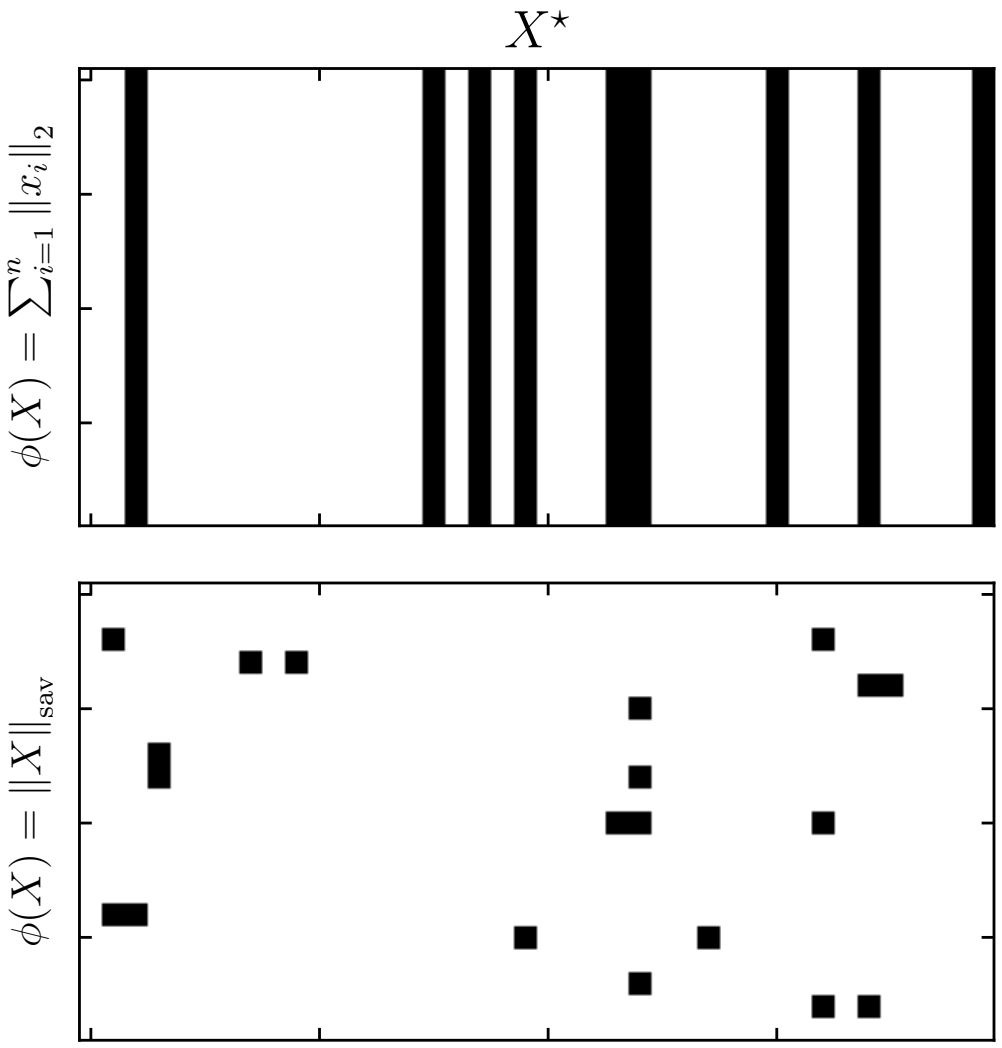
with columnwise sparsity regularization:

$$\text{minimize} \quad \sum_{i=1}^k \left(\sum_{j=1}^n a_{ij}^T x_j - b_i \right)^2 + \gamma \sum_{i=1}^n \|x_i\|_2$$

- variable $X \in \mathbf{R}^{m \times n}$; data $A_i \in \mathbf{R}^{m \times n}$, $b_i \in \mathbf{R}$
- $a_{ij} \in \mathbf{R}^m$ is the jth column of A_i (called *feature group*)
- $x_i \in \mathbf{R}^m$ is the ith column of X

- sparsity regularization encourages some x_i to be zero, $i.e.$, some feature groups are not used

numerical results for $m = 20$, $n = 40$, $k = 20$:
(compared with componentwise sparsity regularizer; zeros shown black)



Rank regularization

$$\text{minimize} \quad f_0(X) + \gamma \, \mathbf{rank}\, X$$

- variable $X \in \mathbf{R}^{m \times n}$; primary objective $f_0 \colon \mathbf{R}^{m \times n} \to \mathbf{R}$
- $\mathbf{rank}\, X$ counts the rank (number of nonzero singular values) of X, weighted by $\gamma > 0$
- generalizes cardinality sparsity regularization to matrices
- nonconvex so very hard to solve

usually solved via heuristic:

$$\text{minimize} \quad f_0(X) + \gamma \|X\|_*$$

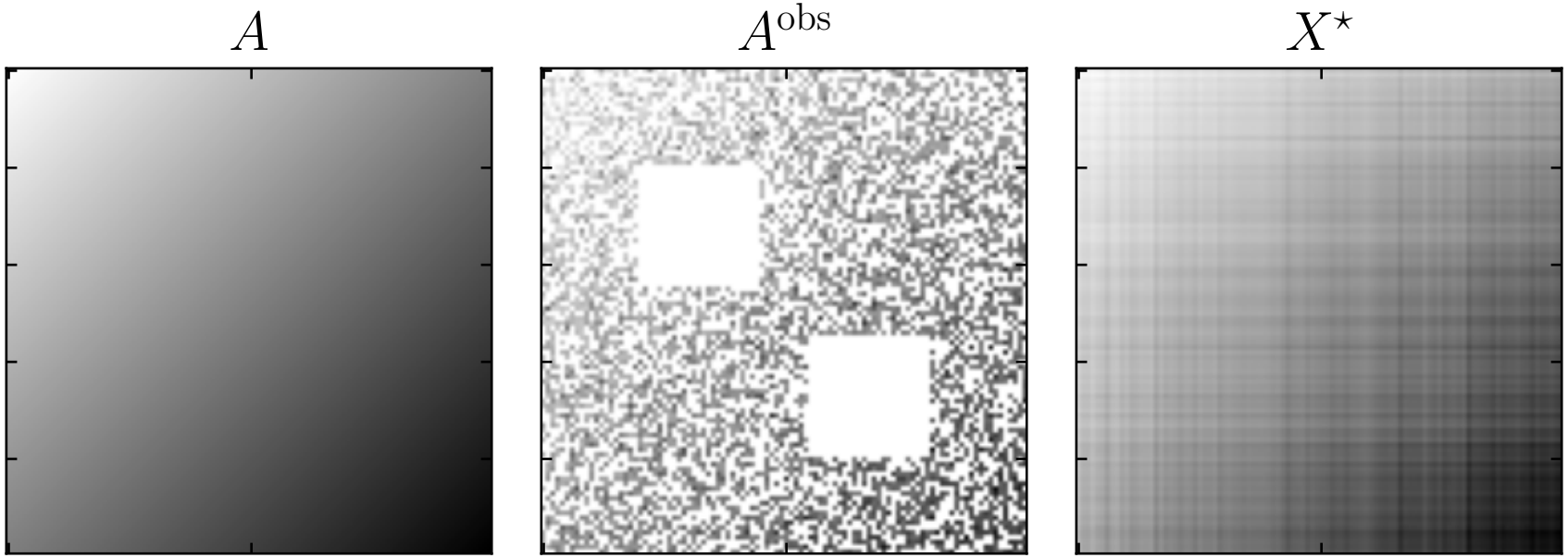
- $\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X)$ is the nuclear norm of X
- replace the number of nonzero singular values with their sum

example: low-rank matrix completion for $A \in \mathbf{R}^{m\times n}$

$$\text{minimize} \quad \sum_{(i,j)\in\mathcal{I}} \left(X_{ij} - A^{\text{obs}}_{ij}\right)^2 + \gamma\|X\|_*$$

- variable $X \in \mathbf{R}^{m\times n}$; observation A^{obs}_{ij} of A for $(i,j) \in \mathcal{I}$ is given

results for $A \in \mathbf{R}^{100\times 100}$ with rank 2; observations are corrupted with Gaussian noise and 60% of entries are missing:



singular values of A (circles bottom), A^{obs} (circles top), and $X^\star$ (dots):

