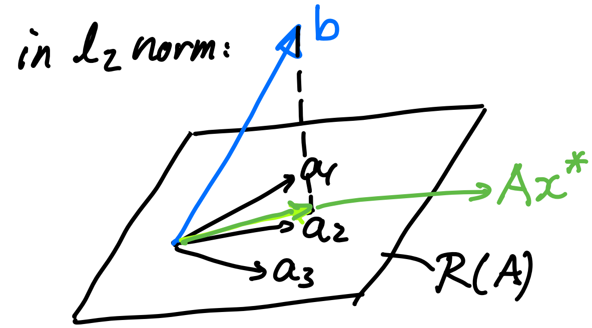
9. Approximation and data fitting

learning model given data

- norm approximation
- least-norm problems
- regularized approximation
- robust approximation

Norm approximation



$$\underset{x}{\text{minimize}} \quad \|Ax - b\|$$

($A \in \mathbf{R}^{m\times n}$ with $m \geq n$, $\|\cdot\|$ is a norm on $\mathbf{R}^m$)

interpretations of solution $x^\star = \operatorname{argmin}_x \|Ax - b\|$:

- **geometric**: $Ax^\star$ is point in $\mathcal{R}(A)$ closest to b
- **estimation**: linear measurement model

$$y = Ax + v$$

  y are measurements, x is unknown, v is measurement error

  given $y = b$, best guess of x is $x^\star$
- **optimal design**: x are design variables (input), Ax is result (output)

  $x^\star$ is design that best approximates desired result b

examples

- least-squares approximation ($\|\cdot\|_2$): solution satisfies

$$A^T A x = A^T b$$

($x^\star = (A^T A)^{-1} A^T b$ if $\mathbf{rank}\, A = n$)

- Chebyshev approximation ($\|\cdot\|_\infty$): can be solved as an LP

$\|Ax-b\|_\infty \le t$

$|a_i^T x - b_i| \le t$

$$\begin{array}{ll} \underset{t,x}{\text{minimize}} & t \\ \text{subject to} & -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1} \end{array}$$

- sum of absolute residuals approximation ($\|\cdot\|_1$): can be solved as an LP

$\|Ax-b\|_1 = \sum_{i=1}^{n} \underbrace{|a_i^T x - b_i|}$

$$\begin{array}{ll} \underset{y,x}{\text{minimize}} & \mathbf{1}^T y \\ \text{subject to} & -y \preceq Ax - b \preceq y \end{array}$$

$$\begin{array}{ll} \text{min} & \sum y_i \\ \text{s.t.} & |a_i^T x - b_i| \le y_i, \ \forall i \end{array}$$

Penalty function approximation

$$\begin{array}{ll} \text{minimize} & \phi(r_1) + \cdots + \phi(r_m) \\ \text{subject to} & r = Ax - b \end{array}$$

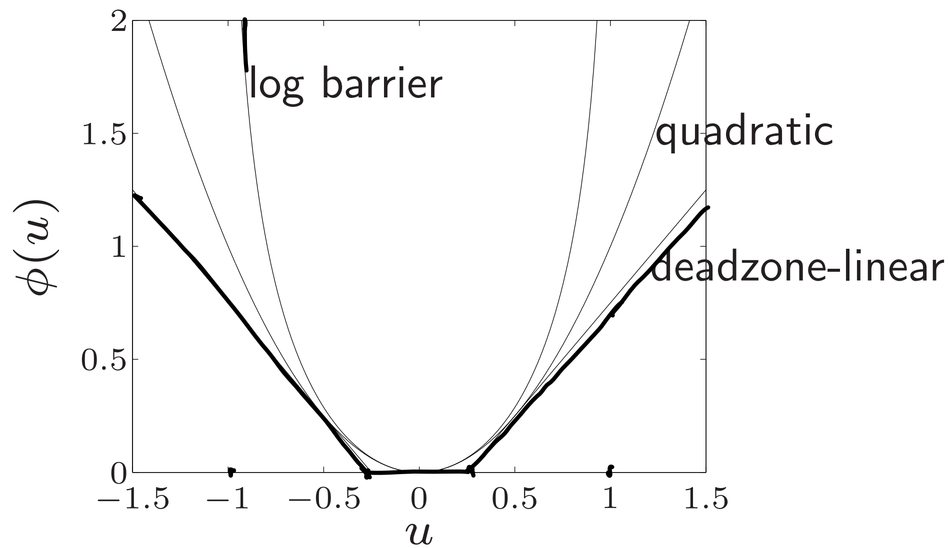
$x \in \mathbb{R}^n$, $r \in \mathbb{R}^m$

($A \in \mathbf{R}^{m \times n}$, $\phi : \mathbf{R} \to \mathbf{R}$ is a convex penalty function)

examples

- quadratic: $\phi(u) = u^2$
- deadzone-linear with width a:

$$\phi(u) = \max\{0, |u| - a\}$$



- log-barrier with limit a:

$$\phi(u) = \begin{cases} -a^2 \log(1 - (u/a)^2) & |u| < a \\ \infty & \text{otherwise} \end{cases}$$

$a=1$

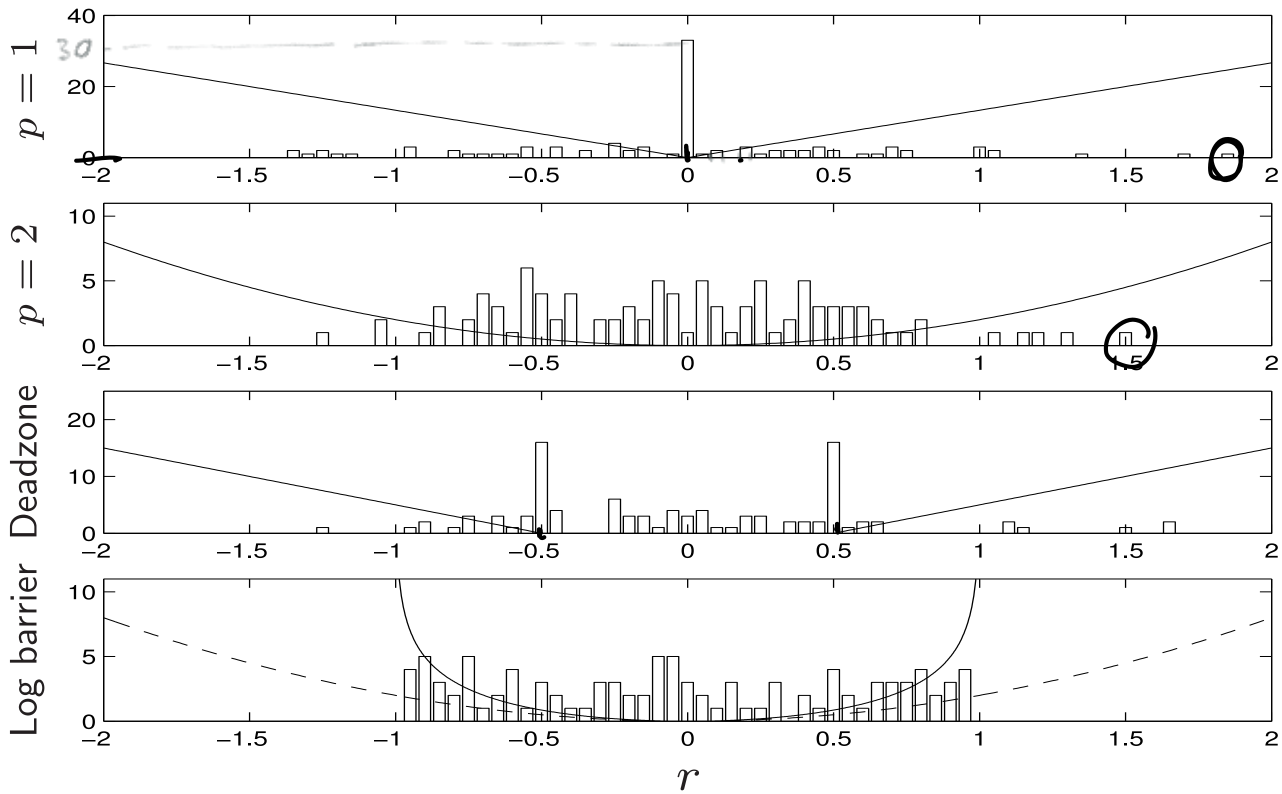
$r \in \mathbb{R}^{100}$, $x \in \mathbb{R}^{30}$ $\quad A \in \mathbb{R}^{100 \times 30}$

example ($m = 100$, $n = 30$): histogram of residuals for penalties

$$\phi(u) = |u|, \quad \phi(u) = u^2, \quad \phi(u) = \max\{0, |u| - a\}, \quad \phi(u) = -\log(1 - u^2)$$

l_1-norm

l_2-norm



sparse residuals allow for a few outliers (large residue)

→ no outlier allowed

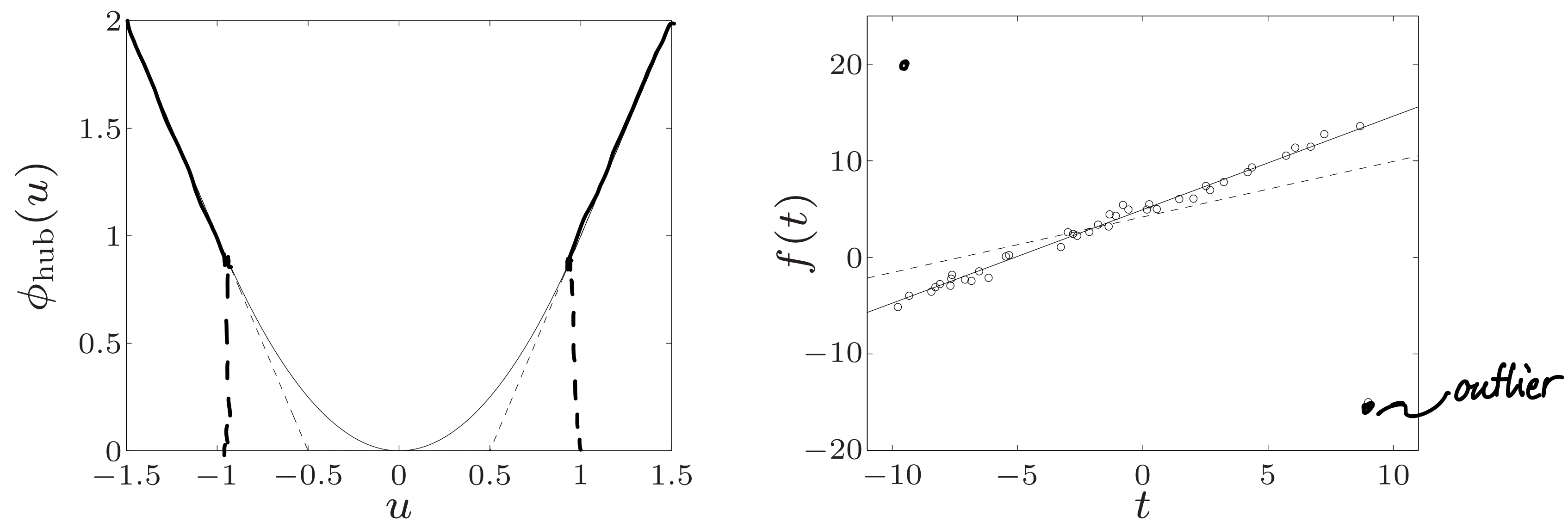
shape of penalty function has large effect on distribution of residuals

combines l_1 & l_2 norms:
behaves as l_1 norm for larger
residuals (away from zero)

Huber penalty function (with parameter M)

$$\phi_{\mathrm{hub}}(u) = \begin{cases} u^2 & |u| \leq M \\ M(2|u| - M) & |u| > M \end{cases}$$

linear growth for large u makes approximation less sensitive to outliers



- left: Huber penalty for $M = 1$
- right: affine function $f(t) = \alpha + \beta t$ fitted to 42 points t_i, y_i (circles) using quadratic (dashed) and Huber (solid) penalty

$\phi(u) = u^2$

- Robust statistics (robust to "outliers")

Announcements

- HW8 (last hw) due Wed by midnight, solutions will be posted 24 hrs after (or sooner if all HWs are submitted)
- Final exam: 10-hours take-home, start @ 9am
  either 3/9 or 3/10, see Canvas announcement.
- Final exams + solutions from previous years posted (Files/exams/)
  BUT this year's final is a different style (and includes some CVX/PY problems as well)
- Course evaluations are open!
  Please fill: depts use these in instructor/TA evaluations & allocating resources.
  I always read all feedback & much appreciate it — thanks!
- Office hours: as usual today (on Zoom), tomorrow, and Fri instead of TA session.

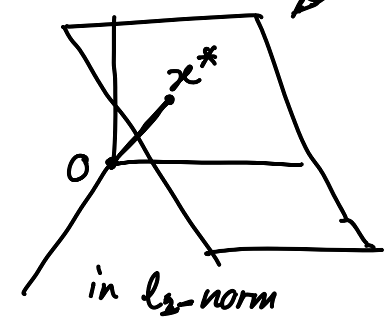
Least-norm problems

wide matrix

$m \times n$

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\| \\ \text{subject to} & Ax = b \end{array}$$

$\{x \mid Ax = b\}$ is an affine set



($A \in \mathbf{R}^{m \times n}$ with $m \leq n$, $\|\cdot\|$ is a norm on $\mathbf{R}^n$)

interpretations of solution $x^\star = \operatorname{argmin}_{Ax=b} \|x\|$:

- **geometric:** $x^\star$ is point in affine set $\{x \mid Ax = b\}$ with minimum distance to 0

- **estimation:** $b = Ax$ are (perfect) measurements of x; $x^\star$ is smallest ('most plausible') estimate consistent with measurements

- **design:** x are design variables (inputs); b are required results (outputs) desired

  $x^\star$ is smallest ('most efficient') design that satisfies requirements

$x(t)$ $\sum_t \|x(t)\|^2 \sim$ signal energy

$A_{m\times n}$ $m<n$

examples

- least-squares solution of linear equations ($\|\cdot\|_2$):

  can be solved via optimality conditions

$$2x + A^T\nu = 0, \qquad Ax = b$$

$\begin{cases} \min_x & \|x\|_2^2 \\ \text{s.t.} & Ax=b \leftarrow \nu \end{cases}$

$L(x,\nu) = \|x\|_2^2 + \nu^T(Ax-b)$

$\nabla_x L(x,\nu)\big|_{x=x^*} = 0$

$x^* = A^+ b$

- minimum sum of absolute values ($\|\cdot\|_1$): can be solved as an LP

$$\begin{array}{ll} \underset{x,y}{\text{minimize}} & \mathbf{1}^T y \\ \text{subject to} & -y \preceq x \preceq y, \quad Ax = b \end{array}$$

$\begin{cases} \min_x & \sum|x_i| \\ \text{s.t.} & Ax=b \end{cases}$

$|x_i| \le y_i$

  tends to produce sparse solution $x^\star$

extension: least-penalty problem

$$\begin{array}{ll} \text{minimize} & \phi(x_1) + \cdots + \phi(x_n) \\ \text{subject to} & Ax = b \end{array}$$

$\phi : \mathbf{R} \to \mathbf{R}$ is convex penalty function

training a NNet:
overparameterized ⇒ many perfect fits to data
but additional properties of learned x are desired
→ design appropriate loss functions
(generalization)

Regularized approximation

$$\left[\text{minimize}_{x} \ (\text{w.r.t. } \mathbf{R}^2_+) \quad (\overbrace{\underbrace{\|Ax-b\|}_{f_1(x)}}^{\text{fitting error}}, \overbrace{\underbrace{\|x\|}_{f_2(x)}}^{\text{"size"}})\right.$$

$A \in \mathbf{R}^{m\times n}$, norms on $\mathbf{R}^m$ and $\mathbf{R}^n$ can be different

interpretation: find good approximation $Ax \approx b$ with small x

- **estimation:** linear measurement model $y = Ax + v$, with prior knowledge that $\|x\|$ is small

$y = Ax$

- **optimal design**: small x is cheaper or more efficient, or the linear model $y = Ax$ is only valid for small x
- **robust approximation:** good approximation $Ax \approx b$ with small x is less sensitive to errors in A than good approximation with large x

$\underset{\sim}{A}x$

errors in A get multiplied by x_i's

Scalarized problem

$$\left[\ \underset{x}{\text{minimize}} \quad \|Ax - b\| + \gamma\|x\|\right.$$

- solution for $\gamma > 0$ traces out optimal trade-off curve
- other common method: minimize $\|Ax - b\|^2 + \delta\|x\|^2$ with $\delta > 0$

Tikhonov regularization : l_2-norm (ridge-regression)

$$\left[\ \underset{x}{\text{minimize}} \quad \|Ax - b\|_2^2 + \delta\|x\|_2^2\right.$$

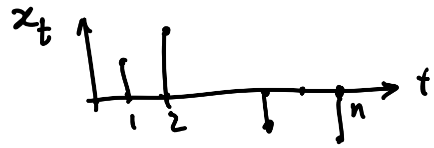
can be solved as a least-squares problem

$$\left[\ \underset{x}{\text{minimize}} \quad \left\| \underbrace{\begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix}}_{\tilde{A}} x - \underbrace{\begin{bmatrix} b \\ 0 \end{bmatrix}}_{\tilde{b}} \right\|_2^2\right. \qquad x^* = \tilde{A}^+ \tilde{b}$$

solution $x^\star = (A^TA + \delta I)^{-1}A^Tb$

Signal reconstruction

$$\text{minimize (w.r.t. } \mathbf{R}^2_+) \quad (\|\hat{x} - x_{\text{cor}}\|_2, \phi(\hat{x}))$$



- $x \in \mathbf{R}^n$ is unknown signal
- $x_{\text{cor}} = x + v$ is (known) corrupted version of x, with additive noise v
- variable $\hat{x}$ (reconstructed signal) is estimate of x
- $\phi : \mathbf{R}^n \to \mathbf{R}$ is regularization function or smoothing objective

examples: quadratic smoothing, total variation smoothing:

$$\hat{x} = \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix}$$

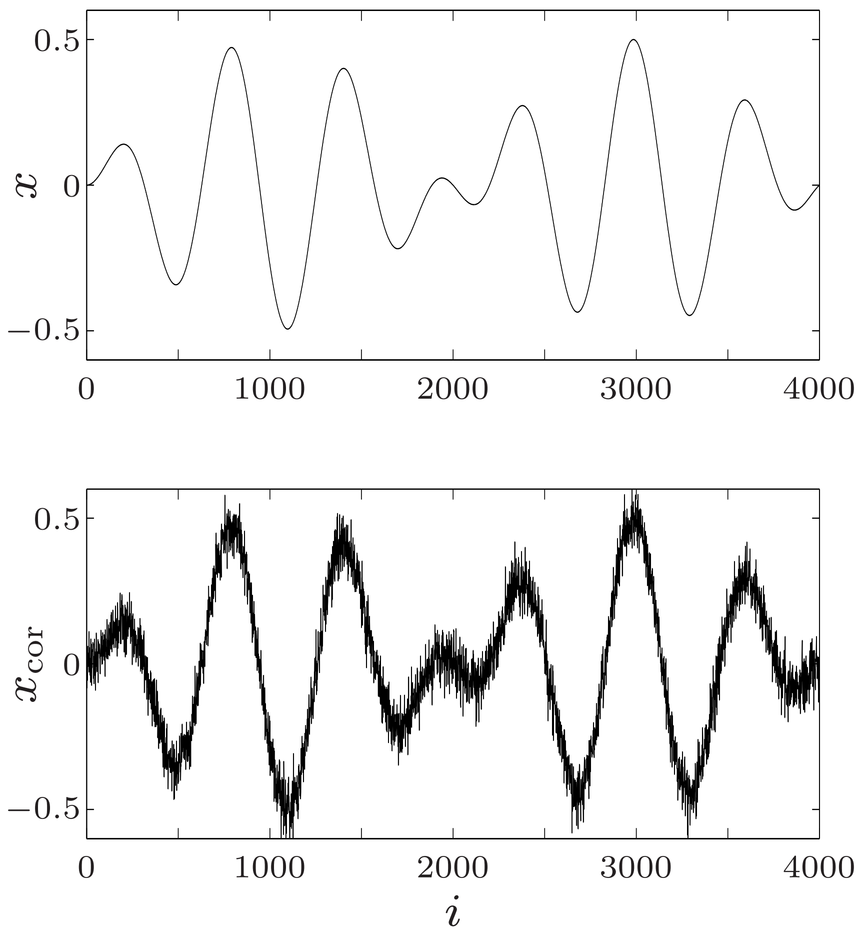
$$\phi_{\text{quad}}(\hat{x}) = \sum_{i=1}^{n-1} (\hat{x}_{i+1} - \hat{x}_i)^2, \qquad \phi_{\text{tv}}(\hat{x}) = \sum_{i=1}^{n-1} |\hat{x}_{i+1} - \hat{x}_i|$$

$$= \|d\|_1$$

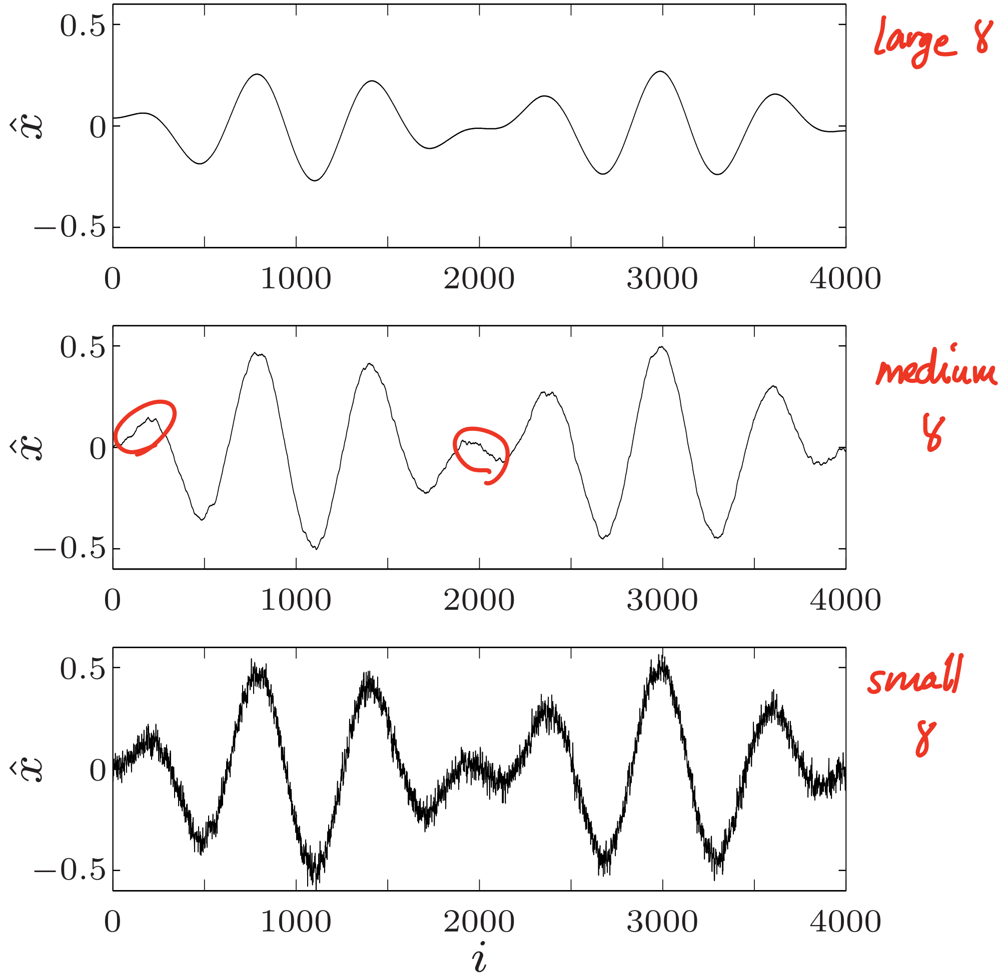
$$d = \begin{bmatrix} \hat{x}_2 - \hat{x}_1 \\ \hat{x}_3 - \hat{x}_2 \\ \vdots \\ \hat{x}_n - \hat{x}_{n-1} \end{bmatrix} \in \mathbb{R}^{n-1} \qquad \phi_{\text{quad}}(\hat{x}) = \|d\|_2^2$$

quadratic smoothing example

$\hat{x} = \operatorname{argmin}_x \; \underbrace{\|x - x_{cor}\|_2^2} + \gamma \, \underbrace{\phi_{quad}(x)}_{\text{smoothness}}$



original signal x and noisy signal x_{cor}

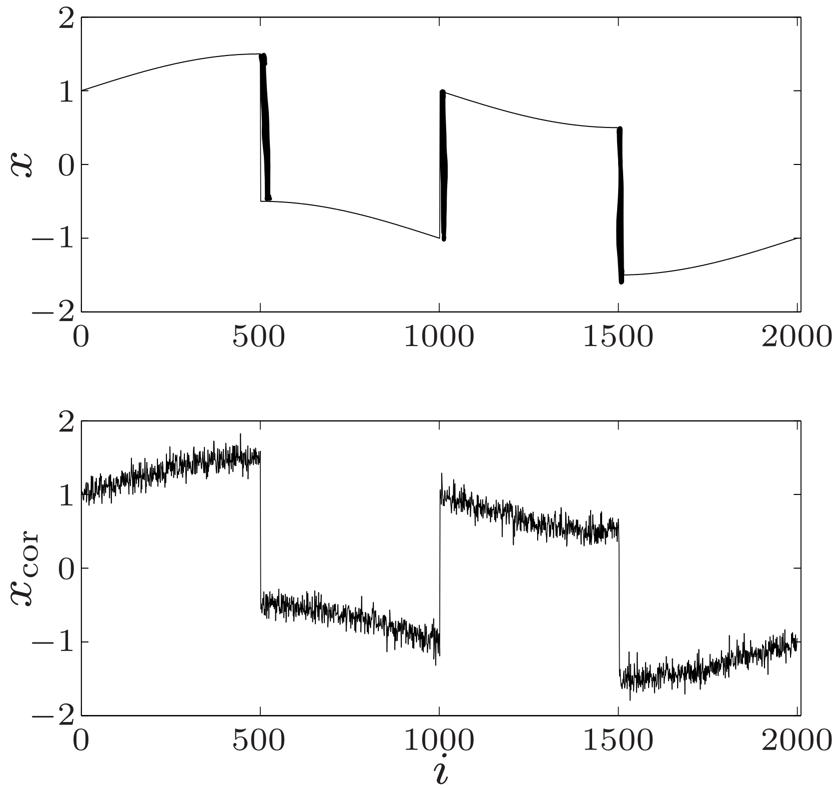


three solutions on trade-off curve $\|\hat{x} - x_{\mathrm{cor}}\|_2$ versus $\phi_{\mathrm{quad}}(\hat{x})$

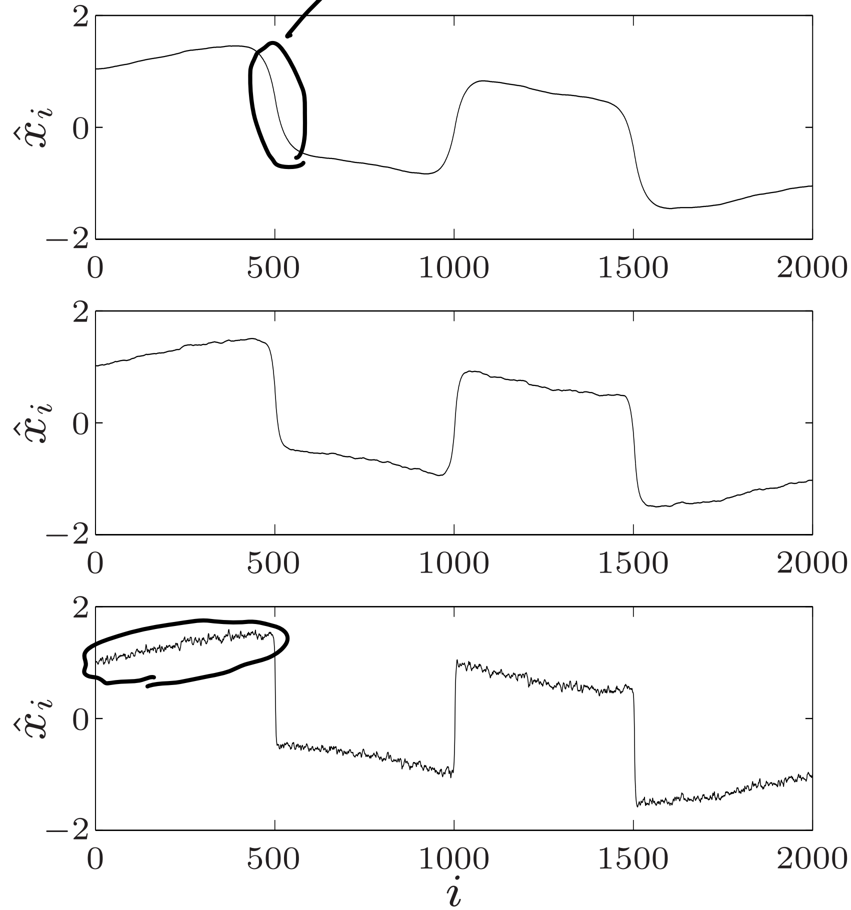
total variation reconstruction example

$\|x - x_{cor}\|_2^2 + 8\,\phi_{tv}(x)$

blurred edge

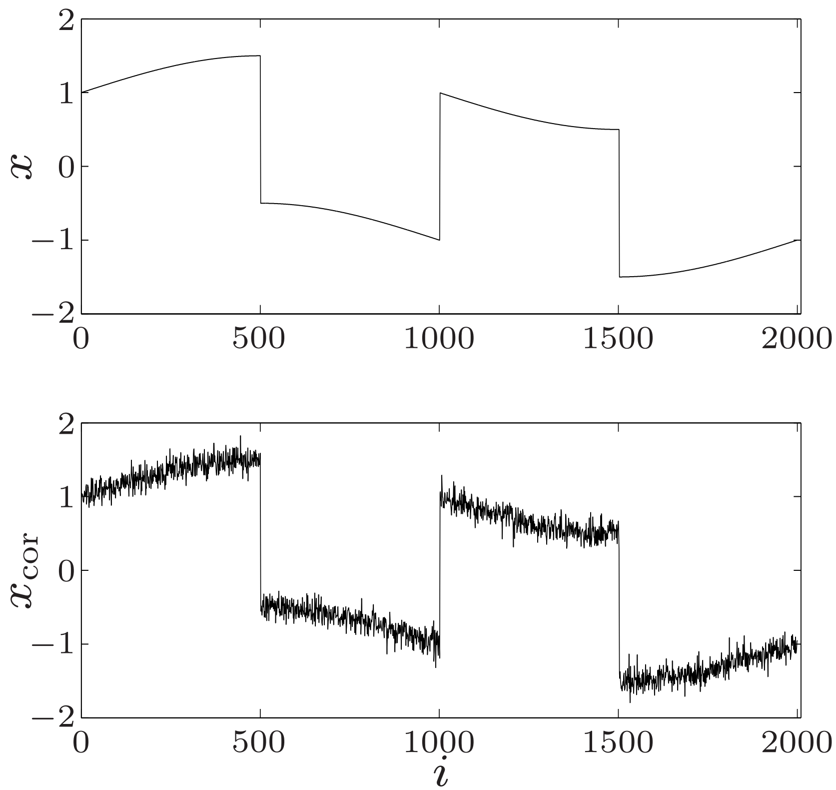


original signal x and noisy signal x_{cor}

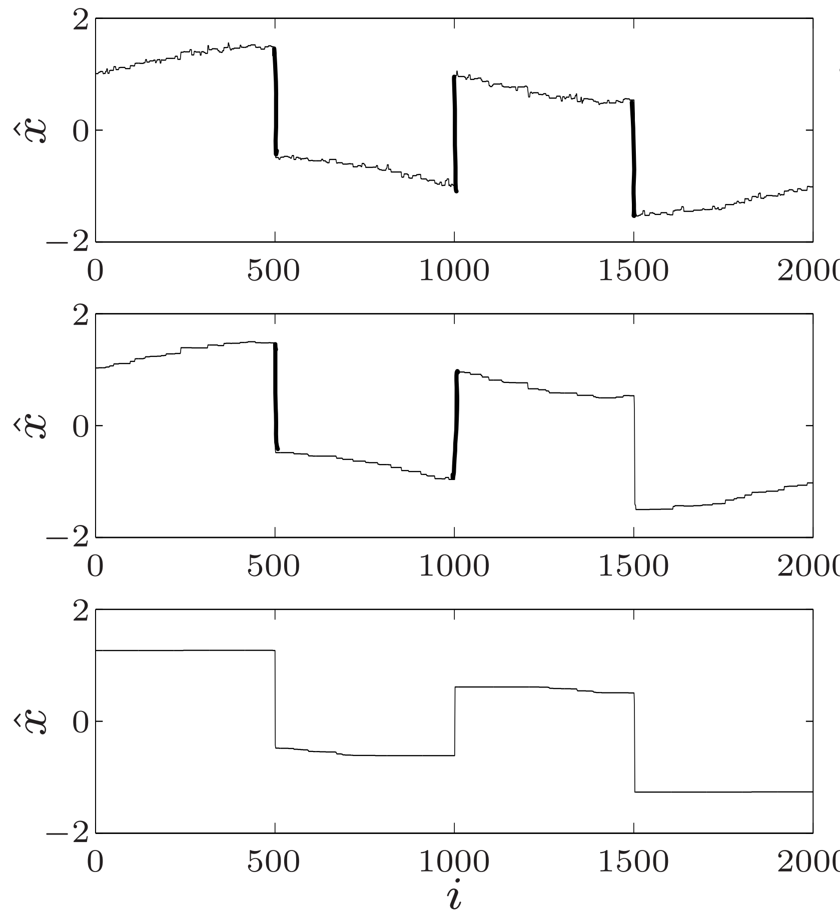


three solutions on trade-off curve $\|\hat{x} - x_{\mathrm{cor}}\|_2$ versus $\phi_{\mathrm{quad}}(\hat{x})$

quadratic smoothing smooths out noise **and** sharp transitions in signal



original signal x and noisy signal x_{cor}



small γ

large γ

three solutions on trade-off curve $\|\hat{x} - x_{\mathrm{cor}}\|_2$ versus $\phi_{\mathrm{tv}}(\hat{x})$

$$\phi_{tv}(\hat{x}) = \sum_{i=1}^{n-1} |\hat{x}_{i+1} - \hat{x}_i|$$

total variation smoothing preserves sharp transitions in signal

Robust approximation

minimize$_x$ $\|Ax - b\|$ with uncertain A

two approaches:

- **stochastic**: assume A is random, minimize$_x$ $\mathbf{E}_A \|Ax - b\|$
- **worst-case:** set $\mathcal{A}$ of possible values of A, minimize$_x$ $\sup_{A\in\mathcal{A}} \|Ax - b\|$

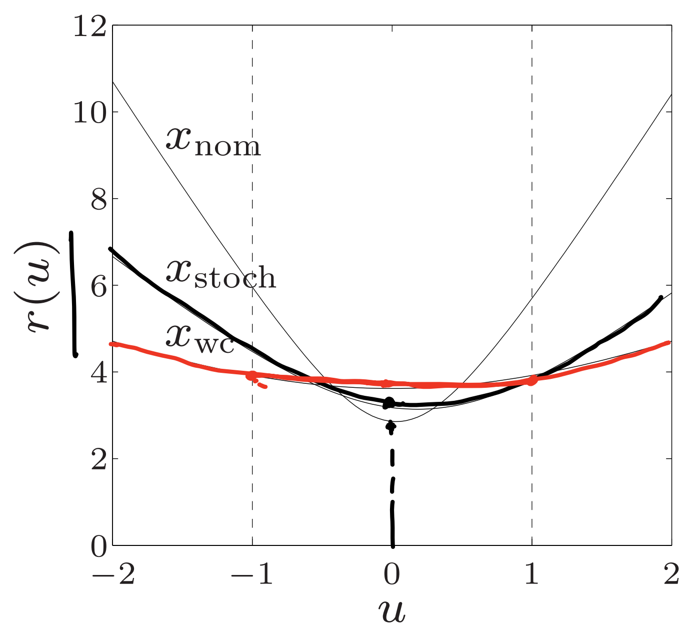
tractable only in special cases (certain norms $\|\cdot\|$, distributions, sets $\mathcal{A}$)

$u \in \mathbb{R}$

example: $A(u) = A_0 + uA_1$

- $x_{\rm nom}$ minimizes$_x$ $\|A_0 x - b\|_2^2$
- $x_{\rm stoch}$ minimizes $\mathbf{E}\,\|A(u)x - b\|_2^2$
  with u uniform on $[-1, 1]$
- $x_{\rm wc}$ minimizes$_x$ $\sup_{-1\le u\le 1} \|A(u)x - b\|_2^2$

figure shows $r(u) = \|A(u)x - b\|_2$



- $\mathbb{E}\|A(u)x-b\|_2^2 = \mathbb{E}_u \|(A_0+uA_1)x-b\|_2^2$

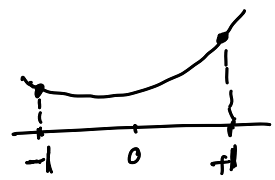
$$= \mathbb{E}_u \|A_0x-b+uA_1x\|_2^2$$

$$= \mathbb{E}_u\left(\|A_0x-b\|_2^2 + \|\underline{u}A_1x\|_2^2\right)$$

$$= \|A_0x-b\|_2^2 + \|A_1x\|_2^2 \underbrace{\mathbb{E}_u u^2}_{\text{variance of } U[-1,1]}$$

$$\rightarrow \left[\min_x \|A_0x-b\|_2^2 + (\mathbb{E}_u u^2)\|A_1x\|_2^2 \right. \rightsquigarrow \text{least squares}$$

- convex quadratic fct of u $\leftarrow \|(A_0+uA_1)x-b\|_2^2$



$$\min_x \underbrace{\sup_{-1\le u\le 1} \|(A_0+uA_1)x-b\|_2^2}$$

$$= \max\left\{\underbrace{\|(A_0+A_1)x-b\|_2^2}_{\text{convex in } x}, \underbrace{\|(A_0-A_1)x-b\|_2^2}_{\text{convex in } x}\right\}$$

stochastic robust LS with $A = \bar{A} + U$, U random, $\mathbf{E}\,U = 0$, $\mathbf{E}\,U^TU = P$

$$\text{minimize} \quad \mathbf{E}\,\|(\bar{A} + U)x - b\|_2^2$$

- explicit expression for objective:

$$\begin{aligned}
\mathbf{E}\,\|Ax - b\|_2^2 &= \mathbf{E}\,\|\bar{A}x - b + Ux\|_2^2 \\
&= \|\bar{A}x - b\|_2^2 + \mathbf{E}_U\, x^T U^T U x \\
&= \|\bar{A}x - b\|_2^2 + x^T P x
\end{aligned}$$

- hence, robust LS problem is equivalent to LS problem

$$\underset{x}{\text{minimize}} \quad \|\bar{A}x - b\|_2^2 + \|P^{1/2}x\|_2^2$$

- for $P = \delta I$, get Tikhonov regularized problem

$$\text{minimize} \quad \|\bar{A}x - b\|_2^2 + \delta\|x\|_2^2$$

skipped

worst-case robust LS with $\mathcal{A} = \{\bar{A} + u_1 A_1 + \cdots + u_p A_p \mid \|u\|_2 \leq 1\}$

$$\text{minimize} \quad \sup_{A \in \mathcal{A}} \|Ax - b\|_2^2 = \sup_{\|u\|_2 \leq 1} \|P(x)u + q(x)\|_2^2$$

where $P(x) = \begin{bmatrix} A_1 x & A_2 x & \cdots & A_p x \end{bmatrix}$, $q(x) = \bar{A}x - b$

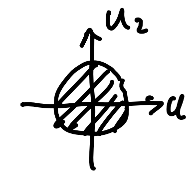
- from page 7-32, strong duality holds between the following problems

$$\begin{array}{ll} \text{maximize} & \|Pu + q\|_2^2 \\ \text{subject to} & \|u\|_2^2 \leq 1 \end{array} \qquad \begin{array}{ll} \text{minimize} & t + \lambda \\ \text{subject to} & \begin{bmatrix} I & P & q \\ P^T & \lambda I & 0 \\ q^T & 0 & t \end{bmatrix} \succeq 0 \end{array}$$

- hence, robust LS problem is equivalent to SDP

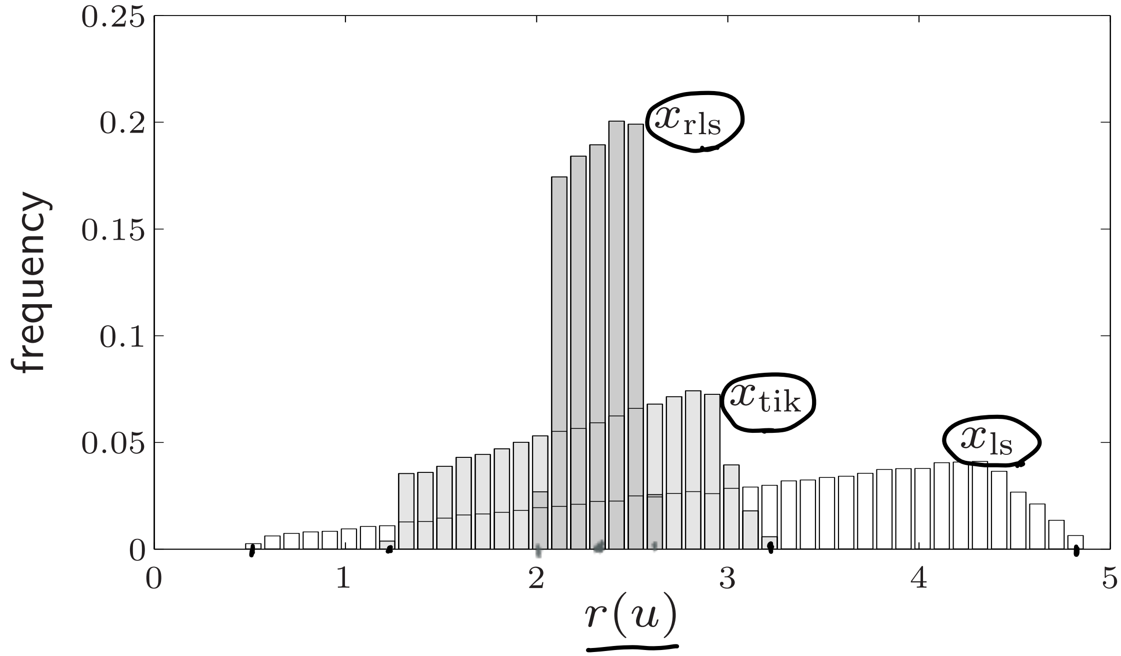
$$\begin{array}{ll} \text{minimize} & t + \lambda \\ \text{subject to} & \begin{bmatrix} I & P(x) & q(x) \\ P(x)^T & \lambda I & 0 \\ q(x)^T & 0 & t \end{bmatrix} \succeq 0 \end{array}$$

example: histogram of residuals



$$r(u) = \|(A_0 + u_1 A_1 + u_2 A_2)x - b\|_2$$

with u uniformly distributed on unit disk, for three values of x



$= x_{\rm nom}$

- $x_{\rm ls}$ minimizes $\|A_0 x - b\|_2$
- $x_{\rm tik}$ minimizes $\|A_0 x - b\|_2^2 + \|x\|_2^2$ (Tikhonov solution)
- $x_{\rm wc}$ minimizes $\sup_{\|u\|_2 \le 1} \|A_0 x - b\|_2^2 + \|x\|_2^2$

$= x_{\rm rls}$