Lecture 10
Data fitting, approximation, and estimation

- norm approximation problems
- least-norm and dual norm problems
- ML and MAP estimation
- application: blind deconvolution
- experiment design

Norm approximation problems

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

- $x \in \mathbf{R}^n$ is variable; $A \in \mathbf{R}^{p \times n}$ and $b \in \mathbf{R}^p$ are problem data
- $\|\cdot\|$ is some norm
- $r = Ax - b$ is called *residual*
- $r_i = a_i^T x - b_i$ is ith residual (a_i^T is ith row of A)
- usually overdetermined, *i.e.*, $b \notin \mathrm{range}(A)$ (*e.g.*, $p > n$, A full rank)

interpretations:

- approximate or fit b with linear combination of columns of A
- b is corrupted measurement of Ax; find 'least inconsistent' value of x for given measurements

examples:

- $\|r\| = \sqrt{r^T r}$: least-squares or ℓ^2 approximation (a.k.a. regression)

- $\|r\| = \sqrt{r^T P r}$, $P \succ 0$: weighted least-squares

- $\|r\| = \max_i |r_i|$: Chebychev, ℓ^∞, or minimax approximation

- $\|r\| = \sum_i |r_i|$: absolute-sum or ℓ^1 approximation

can add (convex) constraints

- max deviation from some prior guess, $e.g.$, $\|x - x_{\text{prior}}\| \leq a$ (can be another norm)

- limits on x_i, $e.g.$, $l_i \leq x_i \leq u_i$

- order-preserving constraints, $e.g.$, $x_1 \leq \cdots \leq x_n$

Least-norm problems

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

- here $b \in \mathrm{range}(A)$ ($e.g.$, A fat, full rank)
- can convert to norm approximation problem by eliminating equality constraints
- x serves as residual here (provided $Ax = b$)

applications:

- extrapolation:
  - b is (perfect, linear) measurement of x
  - $\|x\|$ measures (im)plausibility of x ($i.e.$, x is more 'likely' to be small)
- control:
  - x is actuator input
  - $\|x\|$ measures effort or cost ($e.g.$, energy, fuel)
  - Ax is resulting effect; $Ax = b$ specifies result

can add constraints

Dual norm problems

norm $\|\cdot\|$ and its dual $\|z\|_* = \sup\{ x^T z \mid \|x\| \leq 1 \}$

norm approximation problem:

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

dual of norm approximation problem:

$$\begin{array}{ll} \text{maximize} & \lambda^T b \\ \text{subject to} & A^T \lambda = 0 \\ & \|\lambda\|_* \leq 1 \end{array}$$

least-norm problem:

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

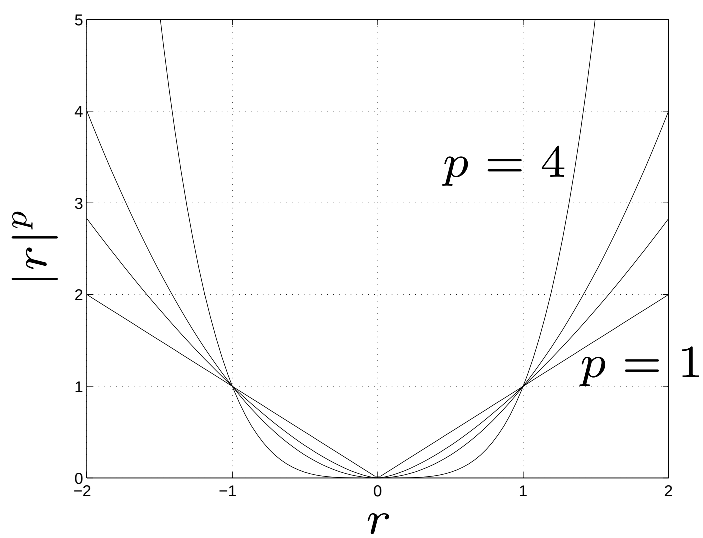
dual of least-norm problem:

$$\begin{array}{ll} \text{maximize} & b^T \lambda \\ \text{subject to} & \|A^T \lambda\|_* \leq 1 \end{array}$$

Interpretation of ℓ^p norm

$$\|r\|_p = \left(\sum_i |r_i|^p\right)^{1/p} \quad \text{(for } p \geq 1\text{)}, \quad \|r\|_\infty = \max_i |r_i|$$

$|r|^p$ for $p = 1,\ 1.5,\ 2,\ 4$:



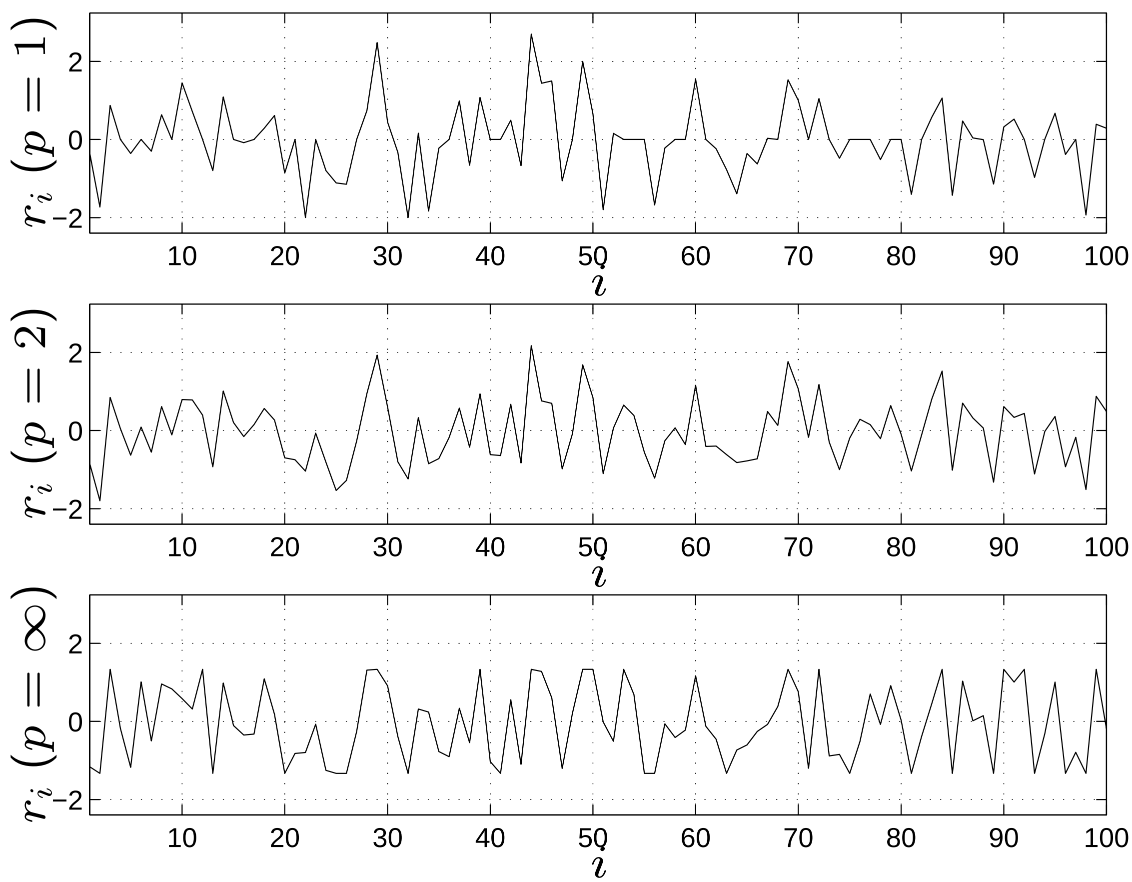
- large p puts more weight on larger residuals
- small p put more weight on small residuals
- $\|r\|_1$ least affected by large residuals
- $\|r\|_\infty$ completely determined by large(st) residuals

$\|r\|_p$ depends on **amplitude distribution** of residuals

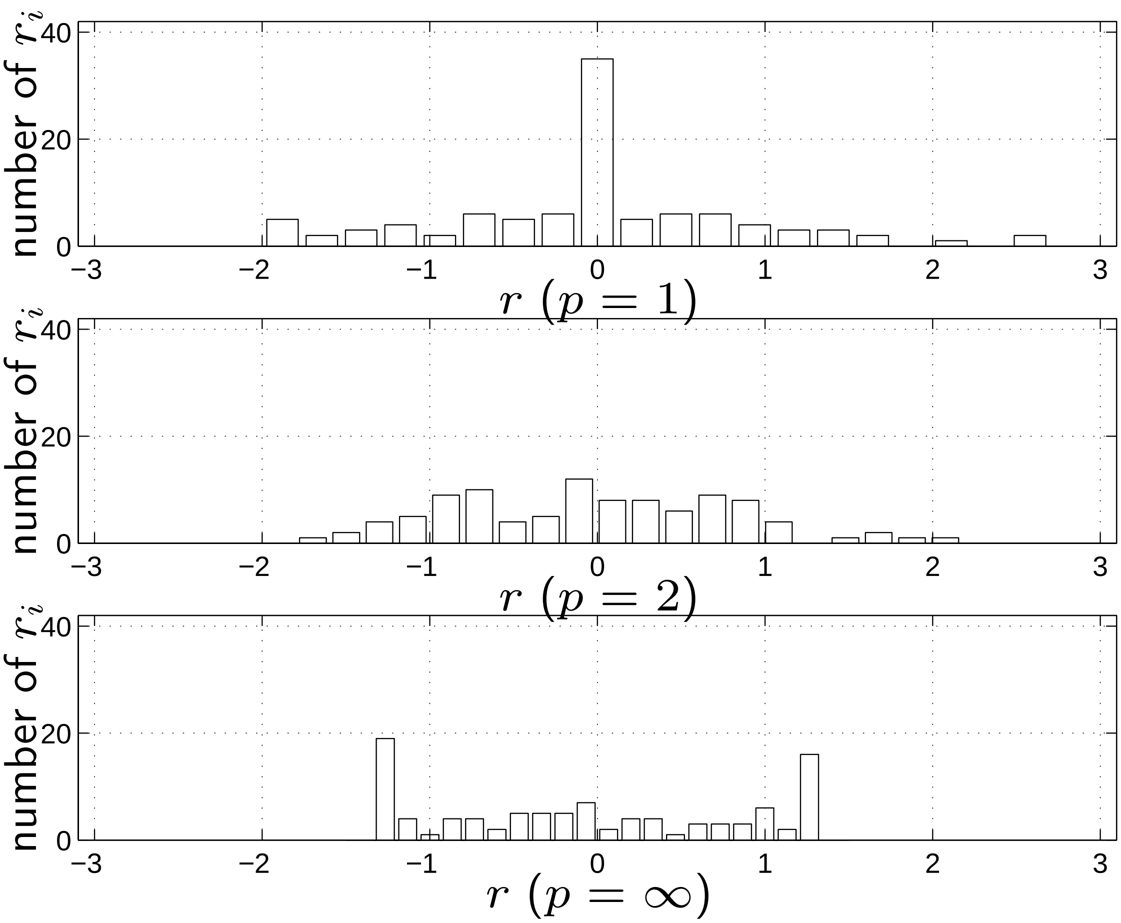
example

- minimize $\|Ax - b\|_p$ for $p = 1,\ 2,\ \infty$
- $A \in \mathbf{R}^{100\times 30}$

resulting residuals:



histogram of amplitude distribution of residuals:



- $p = \infty$ gives 'thinnest' distribution ($i.e.$, smallest interval containing all r_i)
- $p = 1$ residual has widest distribution
- $p = 1$ most very small (or even zero) r_i
- $p = 2$ is in between

Variations and extensions

minimize $\sum_{i=1}^m h(y_i - a_i^T x)$ (or $\max_i h(y_i - a_i^T x)$)

- h is convex
- weights residuals appropriately (for application)

quadratic-linear h

$$h(z) = \begin{cases} z^2 & |z| \le 1 \\ 2|z| - 1 & |z| > 1 \end{cases}$$

- quadratic penalty for small residuals
- linear penalty for large residuals

'dead-zone'

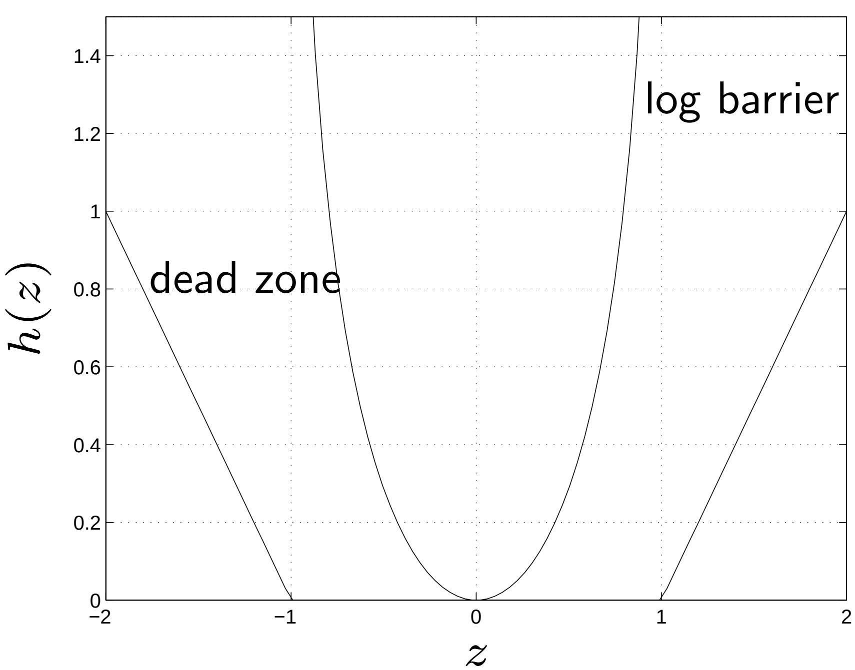
$$h(z) = \begin{cases} |z| - 1 & |z| > 1 \\ 0 & |z| \le 1 \end{cases}$$

- no penalty for small residuals
- linear for larger residuals

log barrier for $|z| \leq 1$

$$h(z) = \begin{cases} -\log(1-z^2) & |z| < 1 \\ \infty & |z| \geq 1 \end{cases}$$

- approximately quadratic for small residuals
- rapidly grows as max residual approaches 1



Maximum likelihood estimation

family of probability densities for y indexed by $x \in \mathbf{R}^n$

$$p_x(y)$$

- x is a parameter
- called *likelihood function* (of x)

maximum likelihood (ML) estimate:
based on observing (a sample of) y, choose as estimate

$$\hat{x} = \operatorname*{argmax}_x p_x(y)$$

variation: **maximum a posteriori (MAP) estimate**

- x is also random
- choose as estimate $\hat{x} = \operatorname{argmax}_x \; p(y|x)$
  maximizes conditional density of y given x

Linear measurements with IID noise

suppose $y_i = a_i^T x + v_i$, v_i IID, density p

$$p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$$

log-likelihood function is defined as

$$\log p_x(y) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

ML estimate is $\hat{x} = \mathrm{argmax}_x \sum_{i=1}^m \log p(y_i - a_i^T x)$

- finding ML estimate is cvx prob if p is log-concave
- can add convex constraints on x (prior assumptions)

if x is random with density q, independent of v_i, MAP estimate is

$$\hat{x} = \underset{x}{\mathrm{argmax}} \left(\sum_{i=1}^m \log p(y_i - a_i^T x) + \log q(x) \right)$$

(last term gives prior probability of x)

Examples

- v_i Gaussian, $p(z) = (2\pi\sigma)^{-1/2} e^{-z^2/2\sigma^2}$
  ML estimate is ℓ^2 estimate $\hat{x} = \mathrm{argmin}_x \|Ax - y\|_2$

- v_i double-sided exponential, $p(z) = (1/2a)e^{-|z|/a}$
  ML estimate is ℓ^1 estimate $\hat{x} = \mathrm{argmin}_x \|Ax - y\|_1$

- v_i is exponential, $p(z) = (1/a)e^{-z/a}$ (for $z \geq 0$)
  ML is found by solving LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T(y - Ax) \\ \text{subject to} & y - Ax \succeq 0 \end{array}$$

- v_i are uniform on $[-a, a]$, $p(z) = 1/(2a)$ on $[-a, a]$
  ML estimate is any x satisfying $\|Ax - y\|_\infty \leq a$

- v_i are uniform on $[-a, a]$, $x \sim \mathcal{N}(\bar{x}, \Sigma)$
  MAP estimate is found by solving (QP)

$$\begin{array}{ll} \text{minimize} & (x - \bar{x})^T \Sigma^{-1} (x - \bar{x}) \\ \text{subject to} & \|Ax - y\|_\infty \leq a \end{array}$$

ML gives statistical interpretation for norms or weight functions h in terms of noise density p:

$$h(z) = -\log p(z)$$

- if the tails of the noise distribution fall off rapidly (or completely), weight function h rises rapidly (or is ∞)
- if the tails don't fall off rapidly ($e.g.$, exponential), weight function h grows more slowly
- h is approx. constant over intervals of approx. uniform noise distribution

for example, dead-zone estimate with

$$h(z) = \begin{cases} |z| - 1 & |z| > 1 \\ 0 & |z| \leq 1 \end{cases}$$

corresponds to ML with noise density

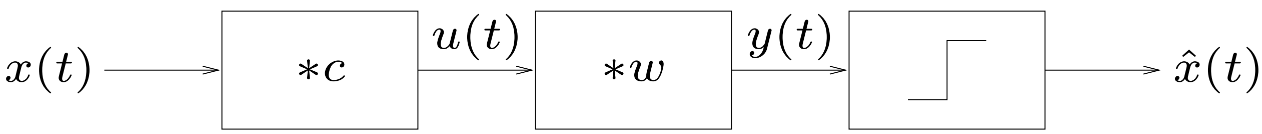
$$p(z) = \begin{cases} (1/4)e^{1-|z|} & |z| > 1 \\ 1/4 & |z| \leq 1 \end{cases}$$

$i.e.$, uniform on $[-1, 1]$, exponential outside $[-1, 1]$

Application: blind deconvolution

thanks to: Alper Erdogan

communications system:



$$u = c * x, \quad y = w * u, \quad \hat{x}(t) = \mathsf{sgn}(y(t + D))$$

- binary signal $x(t) \in \{-1, 1\}$, $t = 1, \ldots, N$
- is convolved by channel impulse response c
- then, by equalizer $w = (w(0), \ldots, w(n-1))$
- binary signal recovered as $\hat{x}(t) = \mathsf{sgn}(y(t + D))$

goal: find equalizer coefficients $w \in \mathbf{R}^n$ s.t. (equalized channel) $h = c * w$ satisfies

$$h(t) \approx \begin{cases} a & t = D \\ 0 & t \neq D \end{cases}$$

- D is some delay
- $a > 0$ is some gain

i.e., w approximately deconvolves c, so $\hat{x}(t) = x(t)$

standard equalization problem: given c, design w

blind equalization problem: given u, design w

- with little knowledge of channel c
- exploiting known structure of signal x

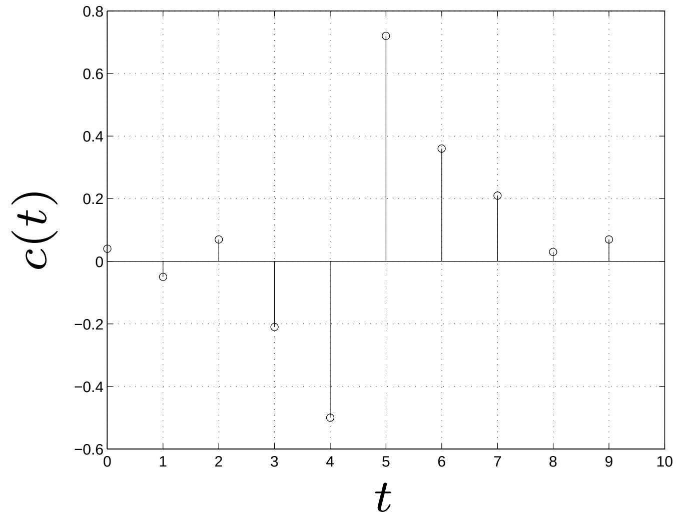
idea: exploit *amplitude distribution* of signals

- amplitude distr of x is concentrated on ± 1
- amplitude distribution of u is 'smeared out' by channel
- if equalizer w is chosen well, amplitude distribution of y is concentrated near $\pm a$

suggests method:
choose w to minimize $\|y\|_\infty = \|w * u\|_\infty$,
subject to some normalization, $e.g.$, $w(0) = 1$

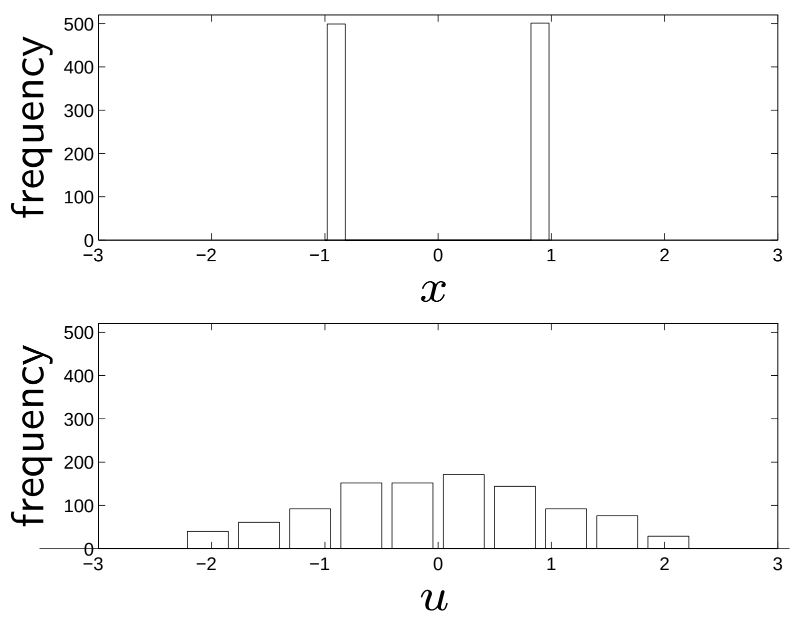
- resulting w tends to 'squeeze' ampl distr of y
- hopefully, ampl distr is not only thin, but concentrated at its extreme points
  $i.e.$, y is (nearly) a binary signal

example. telephone channel model



generate random 1000-bit signal $x \in \{-1, 1\}^{1000}$

amplitude distribution of x and u:

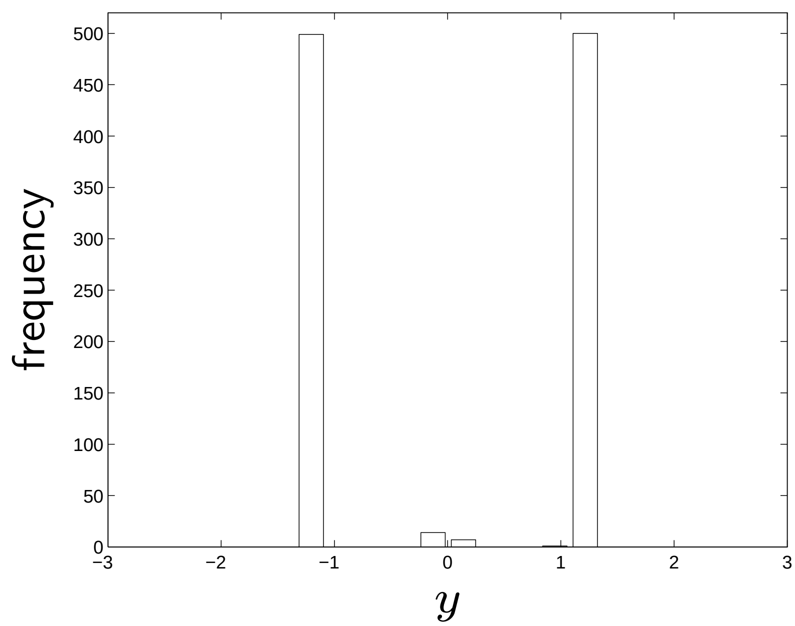


unequalized estimator $\hat{x}(t) = \mathrm{sgn}\, u(t + D)$ has 19% error rate (using $D = 5$)

now, solve (ℓ^∞) problem

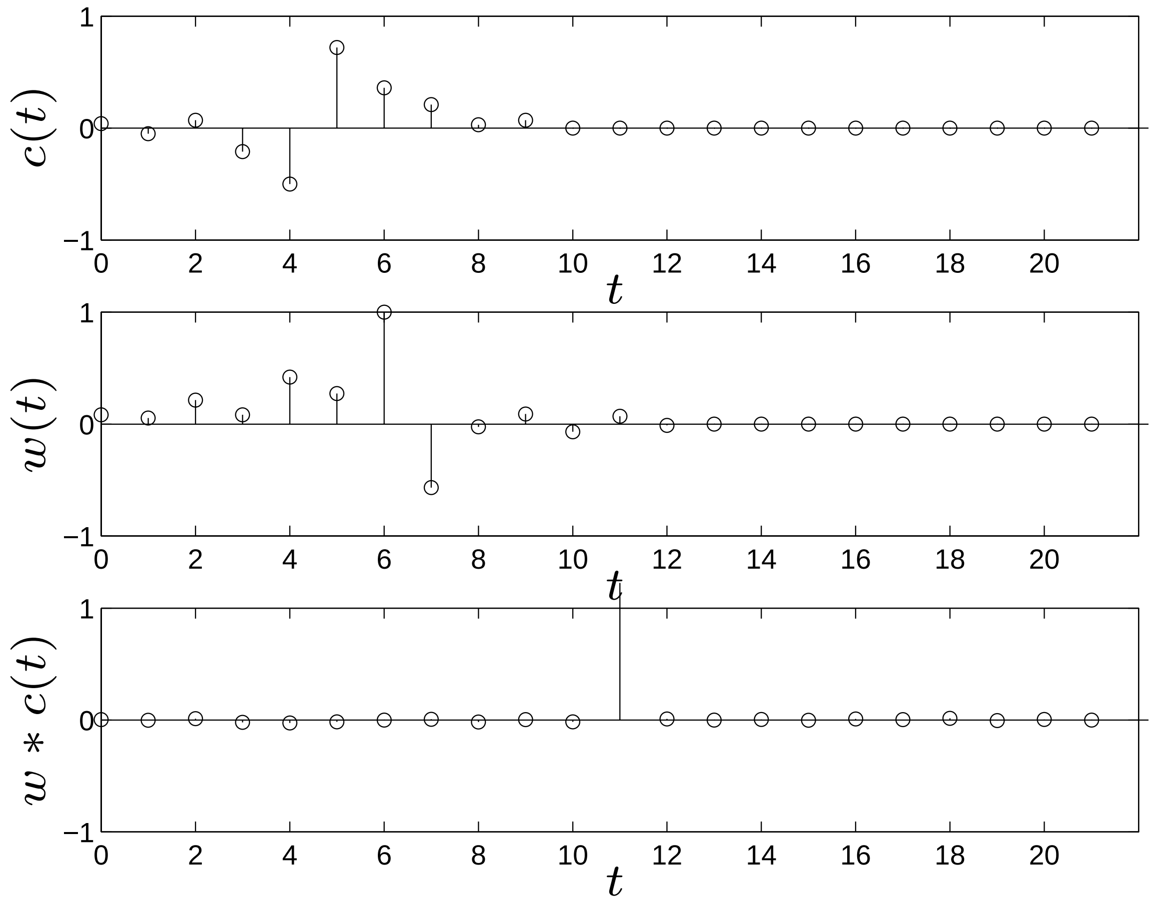
$$\begin{array}{ll} \text{minimize} & \|w * u\|_\infty \\ \text{subject to} & w(0) = 1 \end{array}$$

resulting amplitude distribution of $y = w * u$ is:



- it worked, just as planned!
- error rate 0%:
  $x(t) = \mathrm{sgn}(y(t+11))$ for $t = 1, \ldots, 1000$

channel, equalizer, and equalized channel impulse responses:



i.e., w is good equalizer with $D = 11$
(but . . . we don't know D)

this blind equalization method recovers x up to

- an unknown delay D
- possibly, sign inversion

(neither is a problem in practice)

Robust least-squares

least-squares (ℓ^2) solution of overdetermined equations

$$\hat{x}_{\mathrm{ls}} = \operatorname*{argmin}_x \left(\sum_i (a_i^T x - b_i)^2 \right)^{1/2}$$

suppose a_i are *unknown*, but lie in (known) ellipsoids

$$a_i \in \mathcal{E}_i = \{ \overline{a}_i + P_i u \mid \|u\| \le 1 \}$$

$P_i = P_i^T \succeq 0$ characterizes uncertainty in a_i

define **worst-case residual norm** as

$$\max_{a_i \in \mathcal{E}_i} \left(\sum_{i=1}^n \left(a_i^T x - b_i \right)^2 \right)^{1/2}$$

robust least-squares estimate is given by

$$\hat{x}_{\mathrm{rls}} = \operatorname*{argmin}_x \max_{a_i \in \mathcal{E}_i} \left(\sum_{i=1}^n \left(a_i^T x - b_i \right)^2 \right)^{1/2}$$

- worst-case residual norm is convex in x
- so finding $\hat{x}_{\rm rls}$ is cvx problem
- in fact we can cast it as SOCP . . .

$$\begin{aligned}\max_{a_i \in \mathcal{E}_i} \left|a_i^T x - b_i\right| &= \max_{\|u\| \le 1} \left|\overline{a}_i^T x - b_i + u^T P_i x\right| \\ &= \left|\overline{a}_i^T x - b_i\right| + \|P_i x\|\end{aligned}$$

($u = \pm P_i x / \|P_i x\|$ depending on sgn$(\overline{a}_i^T x - b_i)$)

hence worst-case residual norm is given by

$$\left(\sum_{i=1}^n \left(\left|\overline{a}_i^T x - b_i\right| + \|P_i x\| \right)^2 \right)^{1/2}$$

. . . an explicit (but complicated) convex function of x

can find robust least-squares estimate via SOCP:

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & \|t\| \le s \\ & u_i + \|P_i x\| \le t_i \\ & \left|\overline{a}_i^T x - b_i\right| \le u_i \end{array}$$

(variables are x, s, t, u)

Experiment design

N linear measurements $y_1, \ldots, y_N$ of $x \in \mathbf{R}^p$:

$$y_k = a_k^T x + w_k, \quad k = 1, \ldots, N$$

- measurement noises w_k are IID $\mathcal{N}(0, 1)$
- least-squares estimator:

$$\widehat{x} = \left(\sum_{k=1}^{N} a_k a_k^T \right)^{-1} \sum_{i=1}^{N} y_k a_k$$

- error covariance

$$\Sigma = \mathbf{E}(\widehat{x} - x)(\widehat{x} - x)^T = \left(\sum_{k=1}^{N} a_k a_k^T \right)^{-1}$$

choose $a_k \in \{v_1, \ldots, v_m\}$ to make Σ small

- v_i are given test vectors
- small Σ can mean trace, determinant, etc.
- Σ depends only on *numbers* $n_1, \ldots, n_m$ of each type of test performed

in general get (hard) integer problem

Relaxation/approximation

- define $\lambda_i = n_i/N$
  ($i.e.$, fraction of measurements with $a_k = v_i$)
- suppose we have $N \gg m$
- allow (relax) λ_i to be real, $\lambda_i \geq 0$, $\sum_{i=1}^m \lambda_i = 1$

error covariance is

$$\Sigma(\lambda) = \left(\sum_{k=1}^N a_k a_k^T \right)^{-1} = \frac{1}{N} \left(\sum_{i=1}^m \lambda_i v_i v_i^T \right)^{-1}$$

optimal experiment design:

choose $\lambda_i \geq 0$, $\sum_{i=1}^m \lambda_i = 1$, to make $\Sigma(\lambda)$ 'small'

- minimize $\lambda_{\max}(\Sigma(\lambda))$ (E-optimal)
- minimize $\mathbf{Tr}\, \Sigma(\lambda)$ (A-optimal)
- minimize $\det \Sigma(\lambda)$ (D-optimal)

E-**optimal design**: minimize $\lambda_{\max}(\Sigma(\lambda))$

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \sum_{i=1}^m \lambda_i v_i v_i^T \succeq tI \\ & \sum_{i=1}^m \lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m \end{array}$$

. . . an SDP

A-**optimal design**: minimize $\mathbf{Tr}\,\Sigma(\lambda)$

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\left(\sum_{i=1}^m \lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \sum_{i=1}^m \lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m \end{array}$$

. . . convex (can be cast as SDP)

D-**optimal design**: minimize $\det \Sigma(\lambda)$

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{i=1}^m \lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \sum_{i=1}^m \lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1, \ldots, m \\ & \sum_{i=1}^m \lambda_i v_i v_i^T \succ 0 \end{array}$$

. . . convex

can add other convex constraints, $e.g.$,

- bounds on cost or time of measurements:

$$c_i^T \lambda \leq b_i$$

- no more than 90% of the measurements is concentrated in less than 10% of the test vectors

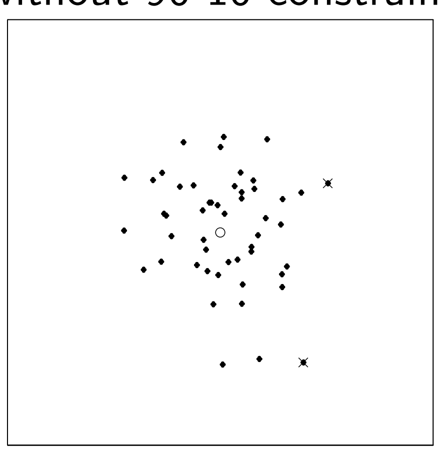
$$\sum_{i=1}^{\lfloor 0.1m \rfloor} \lambda_{[i]} \leq 0.9$$

($\lambda_{[i]}$ is ith largest component of λ)

equivalent to linear inequalities, with auxiliary x, t

$$\lfloor 0.1m \rfloor t + \sum_{i=1}^{m} x_i \leq 0.9, \quad t + x_i \geq \lambda_i, \quad x \geq 0$$

without 90-10 constraint



with 90-10 constraint