

## 12. Approximation and fitting

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

12-1

### Norm approximation

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

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

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

- **geometric**:  $Ax^*$  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  $x^*$

- **optimal design**:  $x$  are design variables (input),  $Ax$  is result (output)  
 $x^*$  is design that best approximates desired result  $b$

## Examples

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

$$A^T Ax = A^T b$$

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

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

$$\begin{aligned} &\text{minimize} && t \\ &\text{subject to} && -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1} \end{aligned}$$

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

$$\begin{aligned} &\text{minimize} && \mathbf{1}^T y \\ &\text{subject to} && -y \preceq Ax - b \preceq y \end{aligned}$$

## Penalty function approximation

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

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

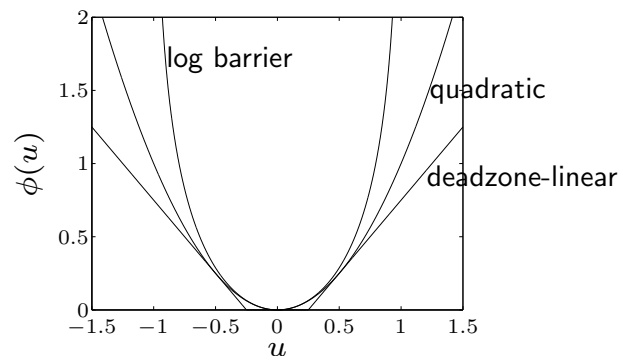
### Examples

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

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

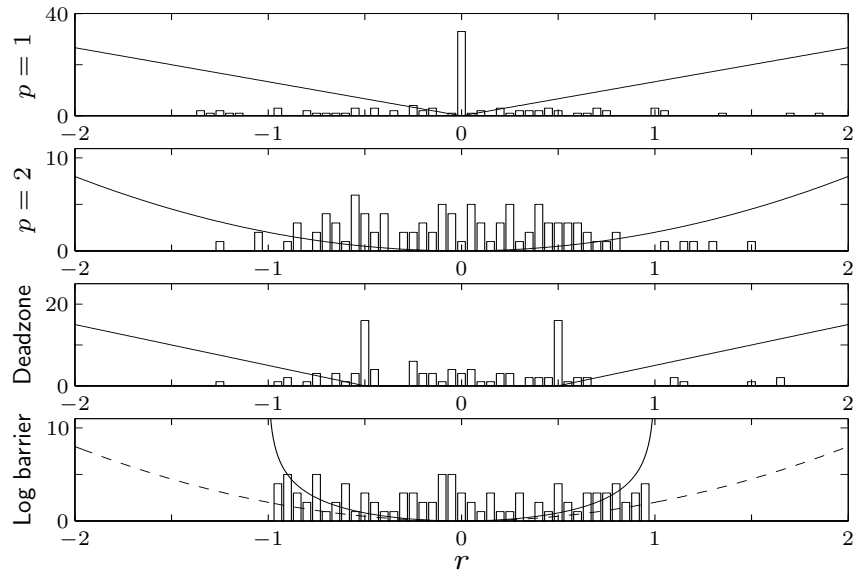
- log-barrier with limit  $a$ :

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



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

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

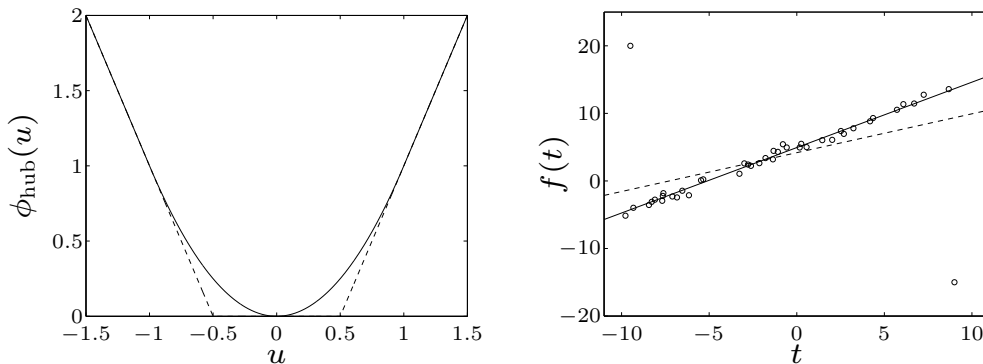


shape of penalty function has large effect on distribution of residuals

**Huber penalty function** (with parameter  $M$ )

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

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

## Least-norm problems

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

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

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

- **geometric:**  $x^*$  is point in affine set  $\{x \mid Ax = b\}$  with minimum distance to 0
- **estimation:**  $b = Ax$  are (perfect) measurements of  $x$ ;  $x^*$  is smallest ('most plausible') estimate consistent with measurements
- **design:**  $x$  are design variables (inputs);  $b$  are required results (outputs)  
 $x^*$  is smallest ('most efficient') design that satisfies requirements

### Examples

- least-squares solution of linear equations ( $\|\cdot\|_2$ ):  
can be solved via optimality conditions

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

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

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

tends to produce sparse solution  $x^*$

**Extension: least-penalty problem** (with convex penalty  $\phi : \mathbf{R} \rightarrow \mathbf{R}$ )

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

## Regularized approximation

$$\text{minimize } (\|Ax - b\|, \|x\|)$$

$A \in \mathbf{R}^{m \times n}$ , 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
- **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  $x$

## Scalarized problem

$$\text{minimize } \|Ax - b\| + \gamma\|x\|$$

- 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

$$\text{minimize } \|Ax - b\|_2^2 + \delta\|x\|_2^2$$

can be solved as a least-squares problem

$$\text{minimize } \left\| \begin{bmatrix} A \\ \sqrt{\delta}I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2$$

solution  $x^* = (A^T A + \delta I)^{-1} A^T b$

# Signal reconstruction

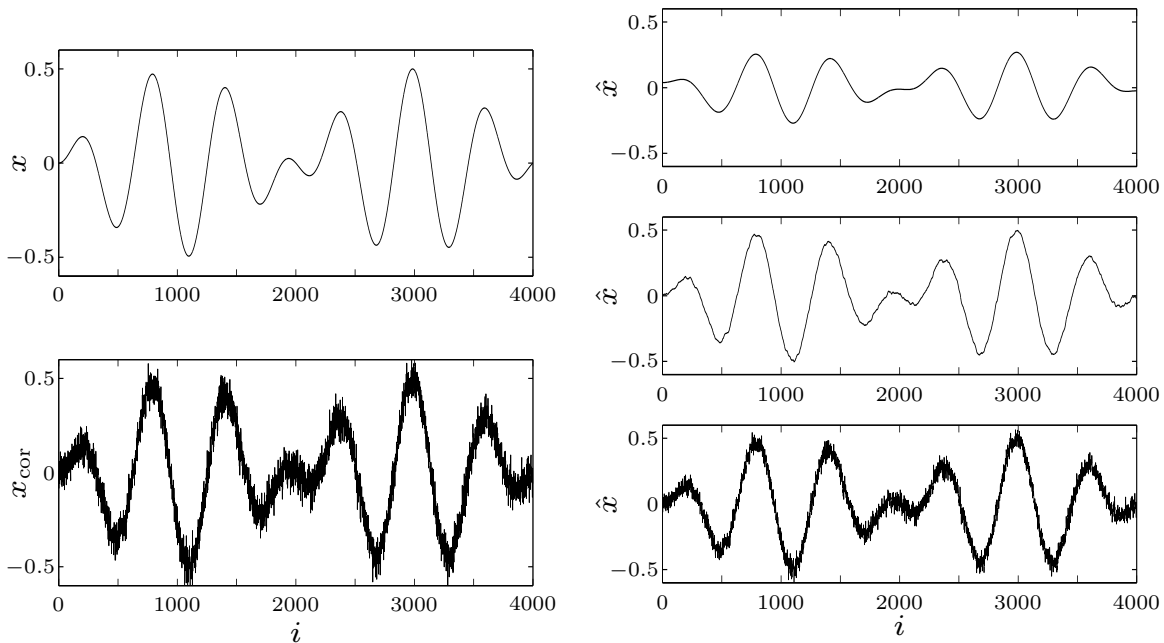
$$\text{minimize } (\|\hat{x} - x_{\text{cor}}\|_2, \phi(\hat{x}))$$

- $x \in \mathbf{R}^n$  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  $x$
- $\phi : \mathbf{R}^n \rightarrow \mathbf{R}$  is regularization function or smoothing objective

**Examples:** quadratic smoothing, total variation smoothing:

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

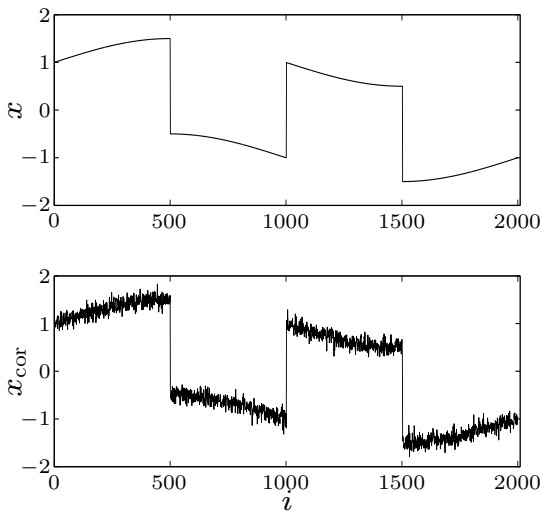
## Quadratic smoothing example



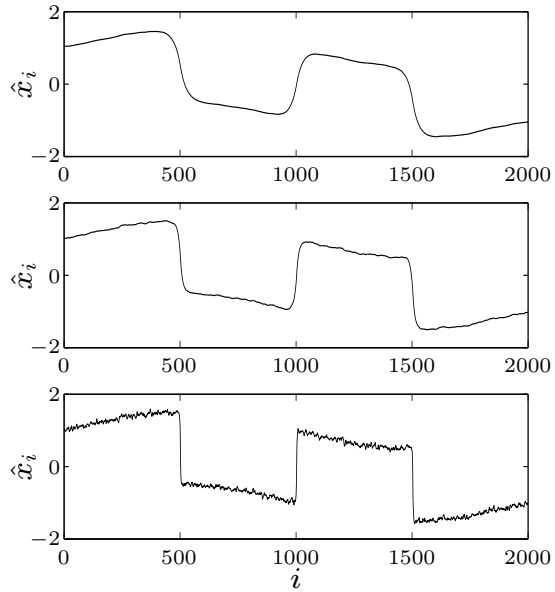
original signal  $x$  and noisy signal  $x_{\text{cor}}$

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

# Total variation reconstruction example

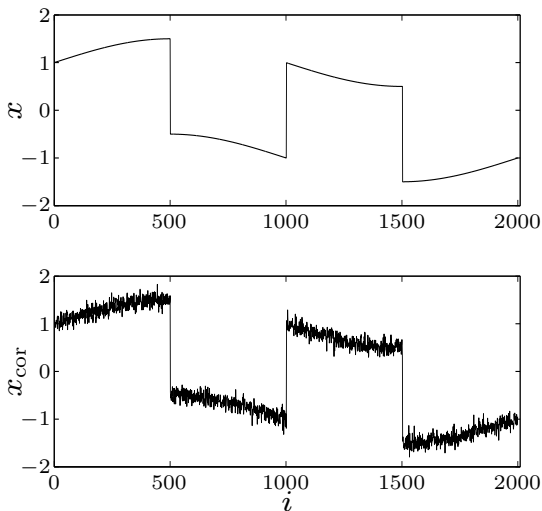


original signal  $x$  and noisy signal  $x_{\text{cor}}$

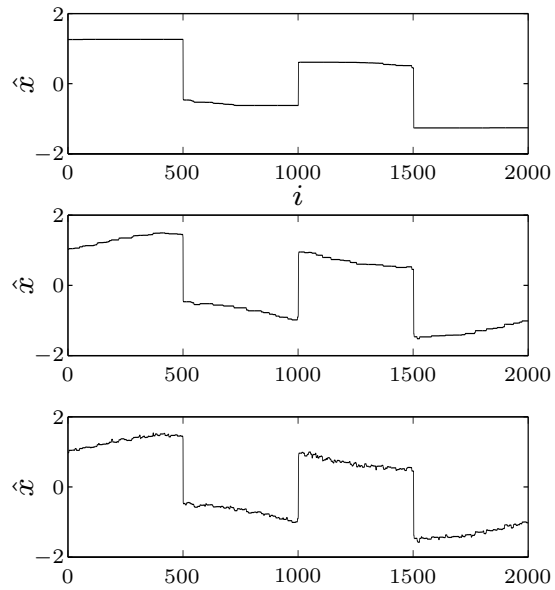


three solutions on trade-off curve  $\|\hat{x} - x_{\text{cor}}\|_2$  versus  $\phi_{\text{quad}}(\hat{x})$

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



original signal  $x$  and noisy signal  $x_{\text{cor}}$

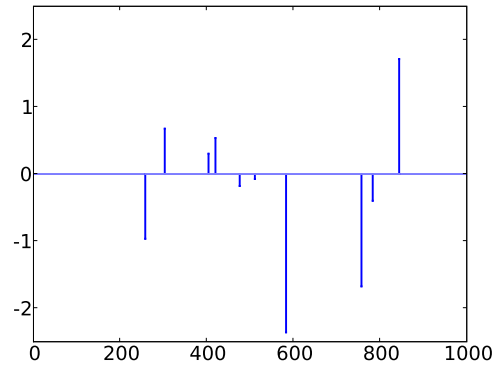


three solutions on trade-off curve  $\|\hat{x} - x_{\text{cor}}\|_2$  versus  $\phi_{\text{tv}}(\hat{x})$

total variation smoothing preserves sharp transitions in signal

## Sparse reconstruction

signal  $\hat{x} \in \mathbf{R}^n$  with  $n = 1000$ , 10 nonzero components



$m = 100$  random noisy measurements

$$b = A\hat{x} + v$$

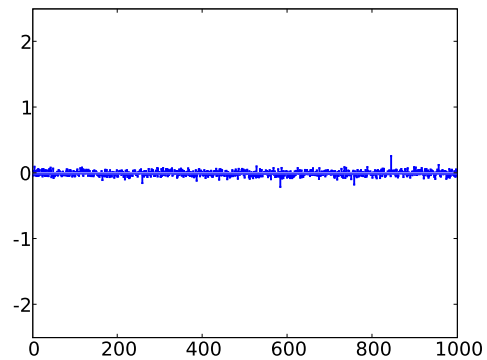
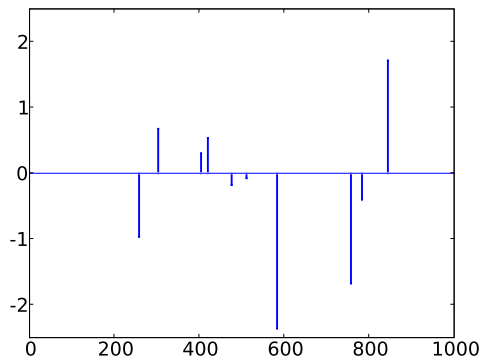
$A_{ij} \sim \mathcal{N}(0, 1)$  i.i.d. and  $v \sim \mathcal{N}(0, \sigma^2 I)$ ,  $\sigma = 0.01$

Approximation and fitting

12-15

## $\ell_2$ -Norm reconstruction

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



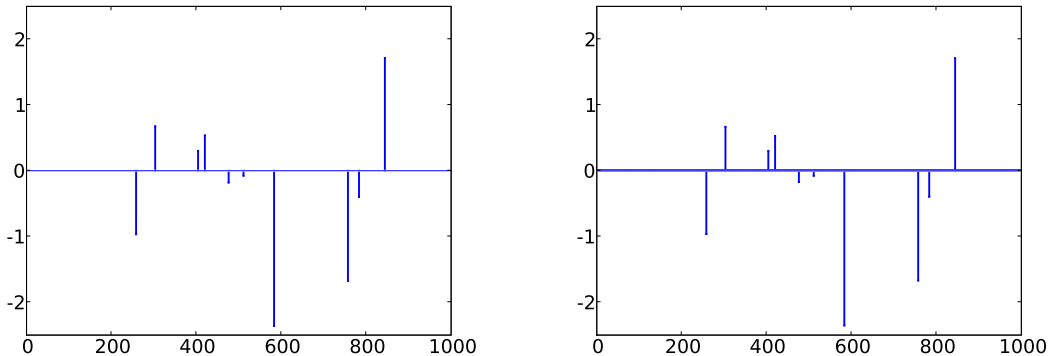
left: exact signal  $\hat{x}$ ; right:  $\ell_2$  reconstruction

Approximation and fitting

12-16

## $\ell_1$ -Norm reconstruction

$$\text{minimize } \|Ax - b\|_2 + \|x\|_1$$



left: exact signal  $\hat{x}$ ; right:  $\ell_1$  reconstruction

- in this example, the sparse signal is recovered *exactly*
- a useful heuristic for estimating sparse signals from noisy measurements

## Robust approximation

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

### Two approaches

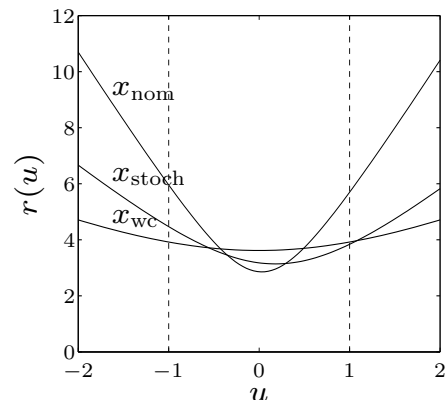
- stochastic: assume  $A$  is random, minimize  $\mathbf{E} \|Ax - b\|$
- worst-case: set  $\mathcal{A}$  of possible values of  $A$ , minimize  $\sup_{A \in \mathcal{A}} \|Ax - b\|$

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

**Example:**  $A(u) = A_0 + uA_1$

- $x_{\text{nom}}$  minimizes  $\|A_0x - b\|_2^2$
- $x_{\text{stoch}}$  minimizes  $\mathbf{E} \|A(u)x - b\|_2^2$   
with  $u$  uniform on  $[-1, 1]$
- $x_{\text{wc}}$  minimizes  $\sup_{-1 \leq u \leq 1} \|A(u)x - b\|_2^2$

figure shows  $r(u) = \|A(u)x - b\|_2^2$



## Stochastic robust least-squares

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

with  $A = \bar{A} + U$ ,  $U$  random,  $\mathbf{E}U = 0$ ,  $\mathbf{E}U^T U = P$

- explicit expression for objective:

$$\mathbf{E} \|Ax - b\|_2^2 = \mathbf{E} \|\bar{A}x - b + Ux\|_2^2 = \|\bar{A}x - b\|_2^2 + x^T P x$$

- hence, robust LS problem is equivalent to LS problem

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

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

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

## Worst-case robust least-squares

$$\text{minimize } \sup_{A \in \mathcal{A}} \|Ax - b\|_2^2$$

with  $\mathcal{A} = \{\bar{A} + u_1 A_1 + \dots + u_p A_p \mid \|u\|_2 \leq 1\}$

### Worst-case residual

$$f(x) = \sup_{A \in \mathcal{A}} \|Ax - b\|_2^2$$

$f(x)$  is the optimal value of

$$\begin{aligned} & \text{maximize (over } u) && \|P(x)u + q(x)\|_2^2 \\ & \text{subject to} && \|u\|_2^2 \leq 1 \end{aligned}$$

where  $P(x) = [ A_1 x \quad A_2 x \quad \dots \quad A_p x ]$ ,  $q(x) = \bar{A}x - b$

from page 7–13,  $f(x)$  is optimal value of (strong) dual

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

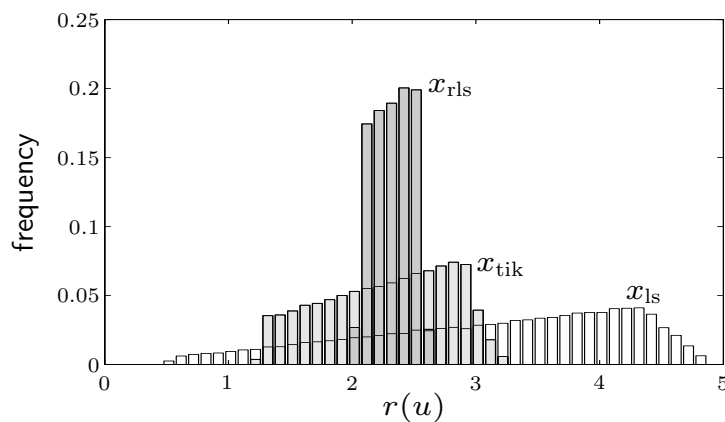
hence, robust LS problem is equivalent to SDP

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

**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  $x$



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

# 13. Geometric problems

- extremal volume ellipsoids
- centering
- classification
- placement and facility location

13-1

## Minimum volume ellipsoid around a set

**Löwner-John ellipsoid** of a set  $C$ : minimum volume ellipsoid  $\mathcal{E}$  s.t.  $C \subseteq \mathcal{E}$

- parametrize  $\mathcal{E}$  as  $\mathcal{E} = \{v \mid \|Av + b\|_2 \leq 1\}$ ; w.l.o.g. assume  $A \in \mathbf{S}_{++}^n$
- $\text{vol } \mathcal{E}$  is proportional to  $\det A^{-1}$ ; to compute minimum volume ellipsoid,

$$\begin{array}{ll} \text{minimize (over } A, b) & \log \det A^{-1} \\ \text{subject to} & \sup_{v \in C} \|Av + b\|_2 \leq 1 \end{array}$$

convex, but evaluating the constraint can be hard (for general  $C$ )

**Finite set**  $C = \{x_1, \dots, x_m\}$

$$\begin{array}{ll} \text{minimize (over } A, b) & \log \det A^{-1} \\ \text{subject to} & \|Ax_i + b\|_2 \leq 1, \quad i = 1, \dots, m \end{array}$$

also gives Löwner-John ellipsoid for polyhedron  $\text{conv}\{x_1, \dots, x_m\}$

## Maximum volume inscribed ellipsoid

maximum volume ellipsoid  $\mathcal{E}$  inside a convex set  $C \subseteq \mathbf{R}^n$

- parametrize  $\mathcal{E}$  as  $\mathcal{E} = \{Bu + d \mid \|u\|_2 \leq 1\}$ ; w.l.o.g. assume  $B \in \mathbf{S}_{++}^n$
- $\text{vol } \mathcal{E}$  is proportional to  $\det B$ ; can compute  $\mathcal{E}$  by solving

$$\begin{aligned} & \text{maximize} && \log \det B \\ & \text{subject to} && \sup_{\|u\|_2 \leq 1} I_C(Bu + d) \leq 0 \end{aligned}$$

(where  $I_C(x) = 0$  for  $x \in C$  and  $I_C(x) = \infty$  for  $x \notin C$ )

convex, but evaluating the constraint can be hard (for general  $C$ )

**Polyhedron**  $\{x \mid a_i^T x \leq b_i, i = 1, \dots, m\}$

$$\begin{aligned} & \text{maximize} && \log \det B \\ & \text{subject to} && \|Ba_i\|_2 + a_i^T d \leq b_i, \quad i = 1, \dots, m \end{aligned}$$

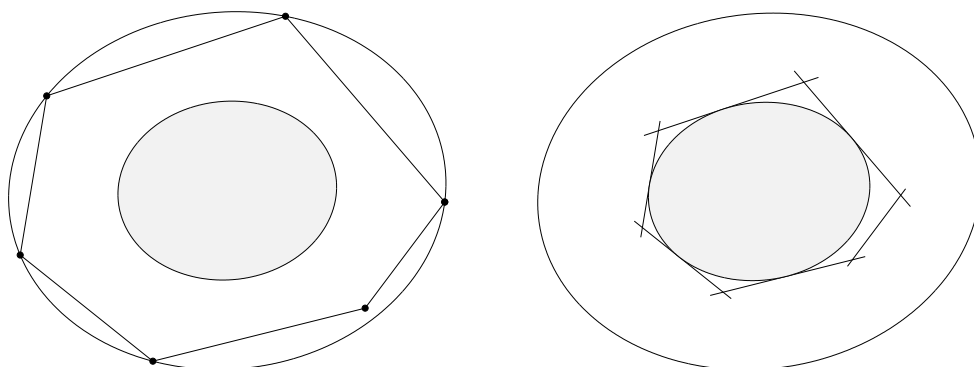
(constraint follows from  $\sup_{\|u\|_2 \leq 1} a_i^T (Bu + d) = \|Ba_i\|_2 + a_i^T d$ )

## Efficiency of ellipsoidal approximations

$C \subseteq \mathbf{R}^n$  convex, bounded, with nonempty interior

- Löwner-John ellipsoid, shrunk by a factor  $n$ , lies inside  $C$
- maximum volume inscribed ellipsoid, expanded by a factor  $n$ , covers  $C$

**Example** (for two polyhedra in  $\mathbf{R}^2$ )

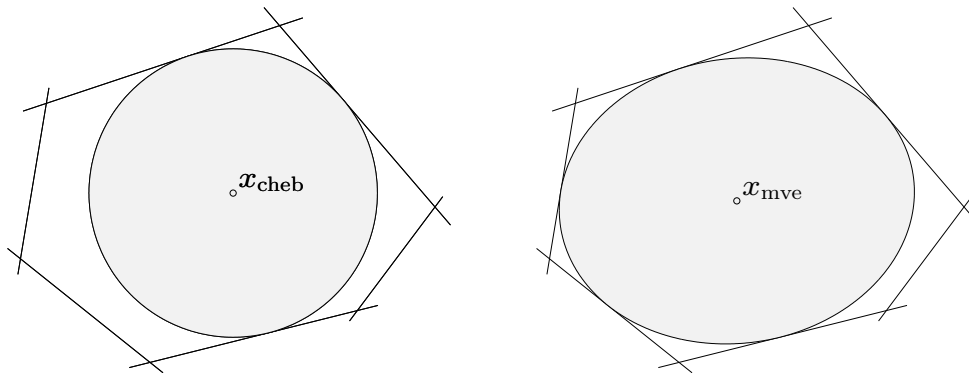


factor  $n$  can be improved to  $\sqrt{n}$  if  $C$  is symmetric

# Centering

some possible definitions of 'center' of a convex set  $C$ :

- center of largest inscribed ball ('Chebyshev center')  
for polyhedron, can be computed via linear programming (page 6–4)
- center of maximum volume inscribed ellipsoid (page 13–3)



MVE center is invariant under affine coordinate transformations

## Analytic center of a set of inequalities

the analytic center of set of convex inequalities and linear equations

$$f_i(x) \leq 0, \quad i = 1, \dots, m, \quad Fx = g$$

is defined as the optimal point of

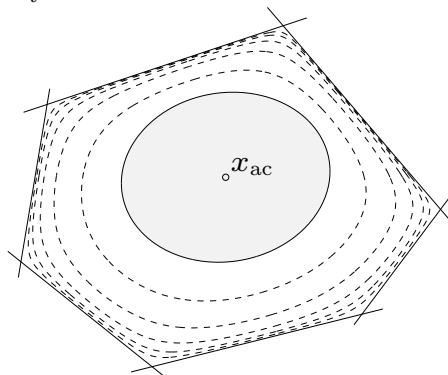
$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^m \log(-f_i(x)) \\ \text{subject to} & Fx = g \end{array}$$

- more easily computed than MVE or Chebyshev center
- not just a property of the feasible set: two sets of inequalities can describe the same set, but have different analytic centers

## Analytic center of linear inequalities $a_i^T x \leq b_i, i = 1, \dots, m$

$x_{ac}$  is minimizer of

$$\phi(x) = - \sum_{i=1}^m \log(b_i - a_i^T x)$$



inner and outer ellipsoids from analytic center:

$$\mathcal{E}_{\text{inner}} \subseteq \{x \mid a_i^T x \leq b_i, i = 1, \dots, m\} \subseteq \mathcal{E}_{\text{outer}}$$

where

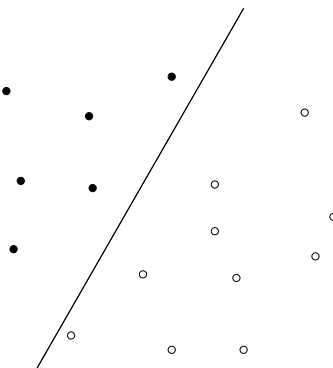
$$\mathcal{E}_{\text{inner}} = \{x \mid (x - x_{ac})^T \nabla^2 \phi(x_{ac})(x - x_{ac}) \leq 1\}$$

$$\mathcal{E}_{\text{outer}} = \{x \mid (x - x_{ac})^T \nabla^2 \phi(x_{ac})(x - x_{ac}) \leq m(m - 1)\}$$

## Linear discrimination

separate two sets of points  $\{x_1, \dots, x_N\}, \{y_1, \dots, y_M\}$  by a hyperplane:

$$a^T x_i + b > 0, \quad i = 1, \dots, N, \quad a^T y_i + b < 0, \quad i = 1, \dots, M$$



homogeneous in  $a, b$ , hence equivalent to

$$a^T x_i + b \geq 1, \quad i = 1, \dots, N, \quad a^T y_i + b \leq -1, \quad i = 1, \dots, M$$

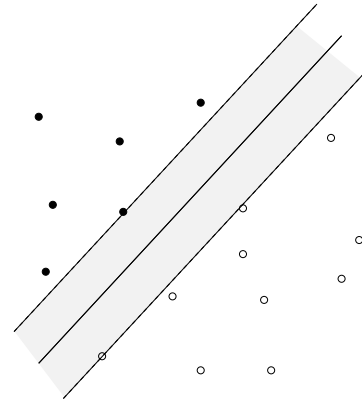
a set of linear inequalities in  $a, b$

## Robust linear discrimination

(Euclidean) distance between hyperplanes

$$\begin{aligned}\mathcal{H}_1 &= \{z \mid a^T z + b = 1\} \\ \mathcal{H}_2 &= \{z \mid a^T z + b = -1\}\end{aligned}$$

is  $\mathbf{dist}(\mathcal{H}_1, \mathcal{H}_2) = 2/\|a\|_2$



to separate two sets of points by maximum margin,

$$\begin{aligned}\text{minimize} & \quad (1/2)\|a\|_2 \\ \text{subject to} & \quad a^T x_i + b \geq 1, \quad i = 1, \dots, N \\ & \quad a^T y_i + b \leq -1, \quad i = 1, \dots, M\end{aligned} \tag{1}$$

(after squaring objective) a QP in  $a, b$

### Lagrange dual of maximum margin separation problem (1)

$$\begin{aligned}\text{maximize} & \quad \mathbf{1}^T \lambda + \mathbf{1}^T \mu \\ \text{subject to} & \quad 2 \left\| \sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i \right\|_2 \leq 1 \\ & \quad \mathbf{1}^T \lambda = \mathbf{1}^T \mu, \quad \lambda \succeq 0, \quad \mu \succeq 0\end{aligned} \tag{2}$$

from duality, optimal value is inverse of maximum margin of separation

### Interpretation

- change variables to  $\theta_i = \lambda_i / \mathbf{1}^T \lambda$ ,  $\gamma_i = \mu_i / \mathbf{1}^T \mu$ ,  $t = 1 / (\mathbf{1}^T \lambda + \mathbf{1}^T \mu)$
- invert objective to minimize  $1 / (\mathbf{1}^T \lambda + \mathbf{1}^T \mu) = t$

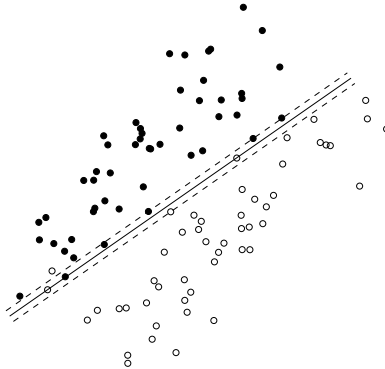
$$\begin{aligned}\text{minimize} & \quad t \\ \text{subject to} & \quad \left\| \sum_{i=1}^N \theta_i x_i - \sum_{i=1}^M \gamma_i y_i \right\|_2 \leq t \\ & \quad \theta \succeq 0, \quad \mathbf{1}^T \theta = 1, \quad \gamma \succeq 0, \quad \mathbf{1}^T \gamma = 1\end{aligned}$$

optimal value is distance between convex hulls

## Approximate linear separation of non-separable sets

$$\begin{aligned} & \text{minimize} && \mathbf{1}^T u + \mathbf{1}^T v \\ & \text{subject to} && a^T x_i + b \geq 1 - u_i, \quad i = 1, \dots, N \\ & && a^T y_i + b \leq -1 + v_i, \quad i = 1, \dots, M \\ & && u \succeq 0, \quad v \succeq 0 \end{aligned}$$

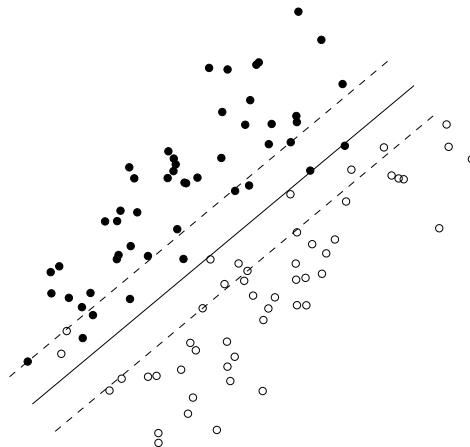
- an LP in  $a, b, u, v$
- can be interpreted as a heuristic for minimizing #misclassified points



## Support vector classifier

$$\begin{aligned} & \text{minimize} && \|a\|_2 + \gamma(\mathbf{1}^T u + \mathbf{1}^T v) \\ & \text{subject to} && a^T x_i + b \geq 1 - u_i, \quad i = 1, \dots, N \\ & && a^T y_i + b \leq -1 + v_i, \quad i = 1, \dots, M \\ & && u \succeq 0, \quad v \succeq 0 \end{aligned}$$

same example as previous page,  
with  $\gamma = 0.1$ :



# Nonlinear discrimination

separate two sets of points by a nonlinear function:

$$f(x_i) > 0, \quad i = 1, \dots, N, \quad f(y_i) < 0, \quad i = 1, \dots, M$$

- choose a linearly parametrized family of functions

$$f(z) = \theta^T F(z)$$

$F = (F_1, \dots, F_k) : \mathbf{R}^n \rightarrow \mathbf{R}^k$  are basis functions

- solve a set of linear inequalities in  $\theta$ :

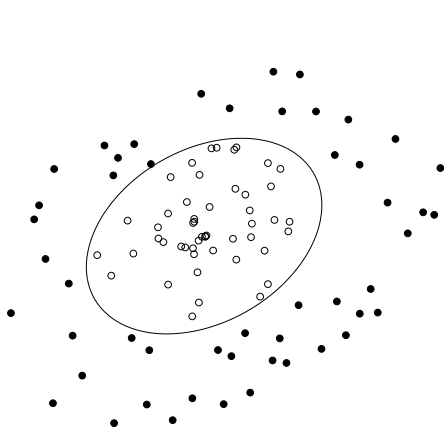
$$\theta^T F(x_i) \geq 1, \quad i = 1, \dots, N, \quad \theta^T F(y_i) \leq -1, \quad i = 1, \dots, M$$

**Quadratic discrimination:**  $f(z) = z^T Pz + q^T z + r$

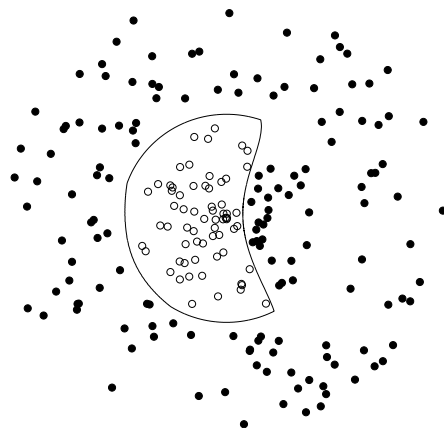
$$x_i^T P x_i + q^T x_i + r \geq 1, \quad y_i^T P y_i + q^T y_i + r \leq -1$$

can add additional constraints (e.g.,  $P \preceq -I$  to separate by an ellipsoid)

**Polynomial discrimination:**  $F(z)$  are all monomials up to a given degree



separation by ellipsoid



separation by 4th degree polynomial

# Placement and facility location

- $N$  points with coordinates  $x_i \in \mathbf{R}^2$  (or  $\mathbf{R}^3$ )
- some positions  $x_i$  are given; the other  $x_i$ 's are variables
- for each pair of points, a cost function  $f_{ij}(x_i, x_j)$

## Placement problem

$$\text{minimize } \sum_{i \neq j} f_{ij}(x_i, x_j)$$

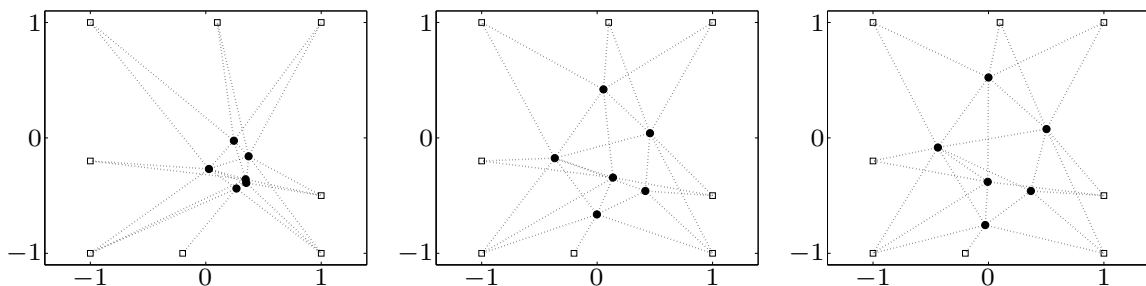
variables are positions of free points

## Interpretations

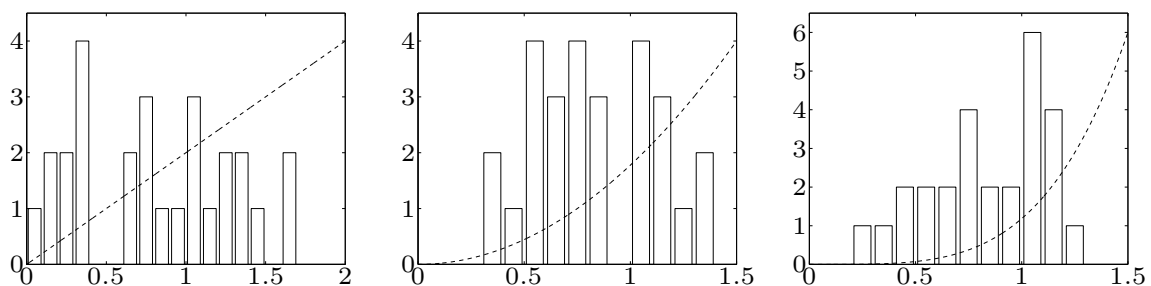
- points represent plants or warehouses;  $f_{ij}$  is transportation cost between facilities  $i$  and  $j$
- points represent cells on an IC;  $f_{ij}$  represents wirelength

**Example:** minimize  $\sum_{(i,j) \in \mathcal{A}} h(\|x_i - x_j\|_2)$ , with 6 free points, 27 links

optimal placement for  $h(z) = z$ ,  $h(z) = z^2$ ,  $h(z) = z^4$



histograms of connection lengths  $\|x_i - x_j\|_2$



# 14. Statistical estimation

- maximum likelihood estimation
- optimal detector design

14-1

## Parametric distribution estimation

- distribution estimation problem: estimate probability density  $p(y)$  of a random variable from observed values
- parametric distribution estimation: choose from a family of densities  $p_x(y)$ , indexed by a parameter  $x$

### Maximum likelihood estimation

$$\text{maximize (over } x) \quad \log p_x(y)$$

- $y$  is observed value
- $l(x) = \log p_x(y)$  is called log-likelihood function
- can add constraints  $x \in C$  explicitly, or define  $p_x(y) = 0$  for  $x \notin C$
- a convex optimization problem if  $\log p_x(y)$  is concave in  $x$  for fixed  $y$

# Linear measurements with IID noise

## Linear measurement model

$$y_i = a_i^T x + v_i, \quad i = 1, \dots, m$$

- $x \in \mathbf{R}^n$  is vector of unknown parameters
- $v_i$  is IID measurement noise, with density  $p(z)$
- $y_i$  is measurement:  $y \in \mathbf{R}^m$  has density  $p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$

**Maximum likelihood estimate:** any solution  $x$  of

$$\text{maximize } l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

( $y$  is observed value)

## Examples

- Gaussian noise  $\mathcal{N}(0, \sigma^2)$ :  $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma^2)}$ ,

$$l(x) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^m (a_i^T x - y_i)^2$$

ML estimate is LS solution

- Laplacian noise:  $p(z) = (1/(2a)) e^{-|z|/a}$ ,

$$l(x) = -m \log(2a) - \frac{1}{a} \sum_{i=1}^m |a_i^T x - y_i|$$

ML estimate is  $\ell_1$ -norm solution

- uniform noise on  $[-a, a]$ :

$$l(x) = \begin{cases} -m \log(2a) & |a_i^T x - y_i| \leq a, \quad i = 1, \dots, m \\ -\infty & \text{otherwise} \end{cases}$$

ML estimate is any  $x$  with  $|a_i^T x - y_i| \leq a$

## Logistic regression

random variable  $y \in \{0, 1\}$  with distribution

$$p = \mathbf{prob}(y = 1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

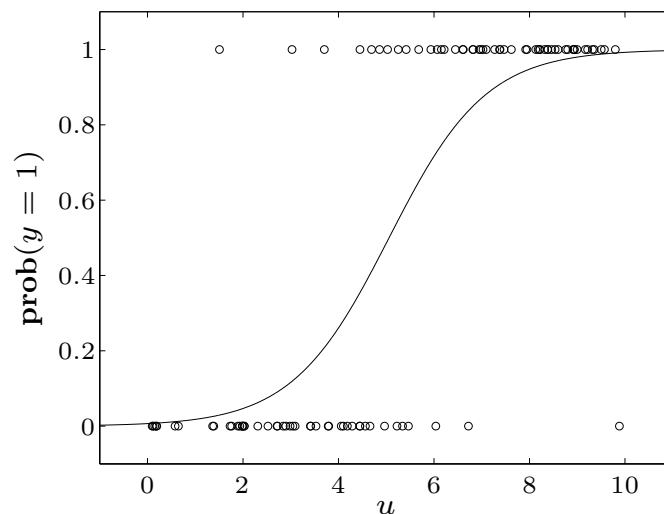
- $a, b$  are parameters;  $u \in \mathbf{R}^n$  are (observable) explanatory variables
- estimation problem: estimate  $a, b$  from  $m$  observations  $(u_i, y_i)$

**Log-likelihood function** (for  $y_1 = \dots = y_k = 1, y_{k+1} = \dots = y_m = 0$ ):

$$\begin{aligned} l(a, b) &= \log \left( \prod_{i=1}^k \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} \prod_{i=k+1}^m \frac{1}{1 + \exp(a^T u_i + b)} \right) \\ &= \sum_{i=1}^k (a^T u_i + b) - \sum_{i=1}^m \log(1 + \exp(a^T u_i + b)) \end{aligned}$$

concave in  $a, b$

**Example** ( $n = 1, m = 50$  measurements)



- circles show 50 points  $(u_i, y_i)$
- solid curve is ML estimate of  $p = \exp(au + b)/(1 + \exp(au + b))$

# (Binary) hypothesis testing

## Detection (hypothesis testing) problem

given observation of a random variable  $X \in \{1, \dots, n\}$ , choose between:

- hypothesis 1:  $X$  was generated by distribution  $p = (p_1, \dots, p_n)$
- hypothesis 2:  $X$  was generated by distribution  $q = (q_1, \dots, q_n)$

## Randomized detector

- a nonnegative matrix  $T \in \mathbf{R}^{2 \times n}$ , with  $\mathbf{1}^T T = \mathbf{1}^T$
- if we observe  $X = k$ , we choose hypothesis 1 with probability  $t_{1k}$ , hypothesis 2 with probability  $t_{2k}$
- if all elements of  $T$  are 0 or 1, it is called a deterministic detector

## Detection probability matrix:

$$D = \begin{bmatrix} Tp & Tq \end{bmatrix} = \begin{bmatrix} 1 - P_{\text{fp}} & P_{\text{fn}} \\ P_{\text{fp}} & 1 - P_{\text{fn}} \end{bmatrix}$$

- $P_{\text{fp}}$  is probability of selecting hypothesis 2 if  $X$  is generated by distribution 1 (false positive)
- $P_{\text{fn}}$  is probability of selecting hypothesis 1 if  $X$  is generated by distribution 2 (false negative)

## Multicriterion formulation of detector design

$$\begin{aligned} & \text{minimize} && (P_{\text{fp}}, P_{\text{fn}}) = ((Tp)_2, (Tq)_1) \\ & \text{subject to} && t_{1k} + t_{2k} = 1, \quad k = 1, \dots, n \\ & && t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \dots, n \end{aligned}$$

variable  $T \in \mathbf{R}^{2 \times n}$

## Scalarization (with weight $\lambda > 0$ )

$$\begin{aligned} & \text{minimize} && (Tp)_2 + \lambda(Tq)_1 \\ & \text{subject to} && t_{1k} + t_{2k} = 1, \quad t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \dots, n \end{aligned}$$

an LP with a simple analytical solution

$$(t_{1k}, t_{2k}) = \begin{cases} (1, 0) & p_k \geq \lambda q_k \\ (0, 1) & p_k < \lambda q_k \end{cases}$$

- a deterministic detector, given by a likelihood ratio test
- if  $p_k = \lambda q_k$  for some  $k$ , any value  $0 \leq t_{1k} \leq 1$ ,  $t_{1k} = 1 - t_{2k}$  is optimal (*i.e.*, Pareto-optimal detectors include non-deterministic detectors)

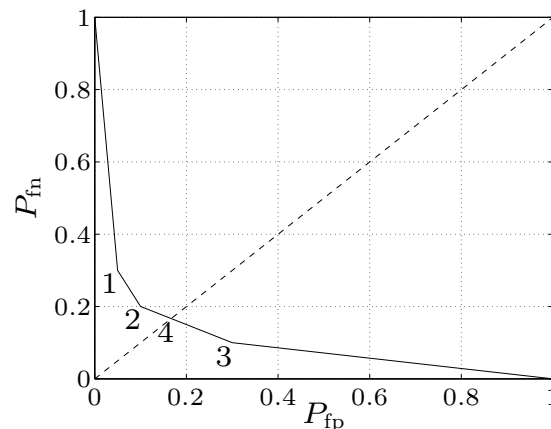
## Minimax detector

$$\begin{aligned} & \text{minimize} && \max\{P_{\text{fp}}, P_{\text{fn}}\} = \max\{(Tp)_2, (Tq)_1\} \\ & \text{subject to} && t_{1k} + t_{2k} = 1, \quad t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \dots, n \end{aligned}$$

an LP; solution is usually not deterministic

## Example

$$P = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}$$



solutions 1, 2, 3 (and endpoints) are deterministic; 4 is minimax detector

# 15. Conclusions

- main ideas
- further topics

15-1

## Modeling

### Mathematical optimization

- problems in engineering design, data analysis and statistics, economics, management, . . . , can often be expressed as mathematical optimization problems
- techniques exist to take into account multiple objectives or uncertainty in the data

### Tractability

- roughly speaking, tractability in optimization requires convexity
- algorithms for nonconvex optimization find local (suboptimal) solutions, or are very expensive
- surprisingly many applications can be formulated as convex problems

## Theoretical consequences of convexity

- local optima are global
- extensive duality theory
  - systematic way of deriving lower bounds on optimal value
  - necessary and sufficient optimality conditions
  - certificates of infeasibility
  - sensitivity analysis
- solution methods with polynomial worst-case complexity theory (with self-concordance)

## Practical consequences of convexity

(most) **convex problems can be solved globally and efficiently**

- interior-point methods require 20 – 80 steps in practice
- basic algorithms (*e.g.*, Newton, barrier method, . . . ) are easy to implement and work well for small and medium size problems (larger problems if structure is exploited)
- more and more high-quality implementations of advanced algorithms and modeling tools are becoming available
- high level modeling tools like CVX ease modeling and problem specification

# How to use convex optimization

to use convex optimization in some applied context

- use rapid prototyping, approximate modeling
  - start with simple models, small problem instances, inefficient solution methods
  - if you don't like the results, no need to expend further effort on more accurate models or efficient algorithms
- work out, simplify, and interpret optimality conditions and dual
- even if the problem is quite nonconvex, you can use convex optimization
  - in subproblems, *e.g.*, to find search direction
  - by repeatedly forming and solving a convex approximation at the current point

## Further topics

some topics we didn't cover:

- methods for very large scale problems
- subgradient calculus, convex analysis
- localization, subgradient, and related methods
- distributed convex optimization
- applications that build on or use convex optimization