

## 7. Statistical estimation

- maximum likelihood estimation
- optimal detector design
- experiment design

# 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

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

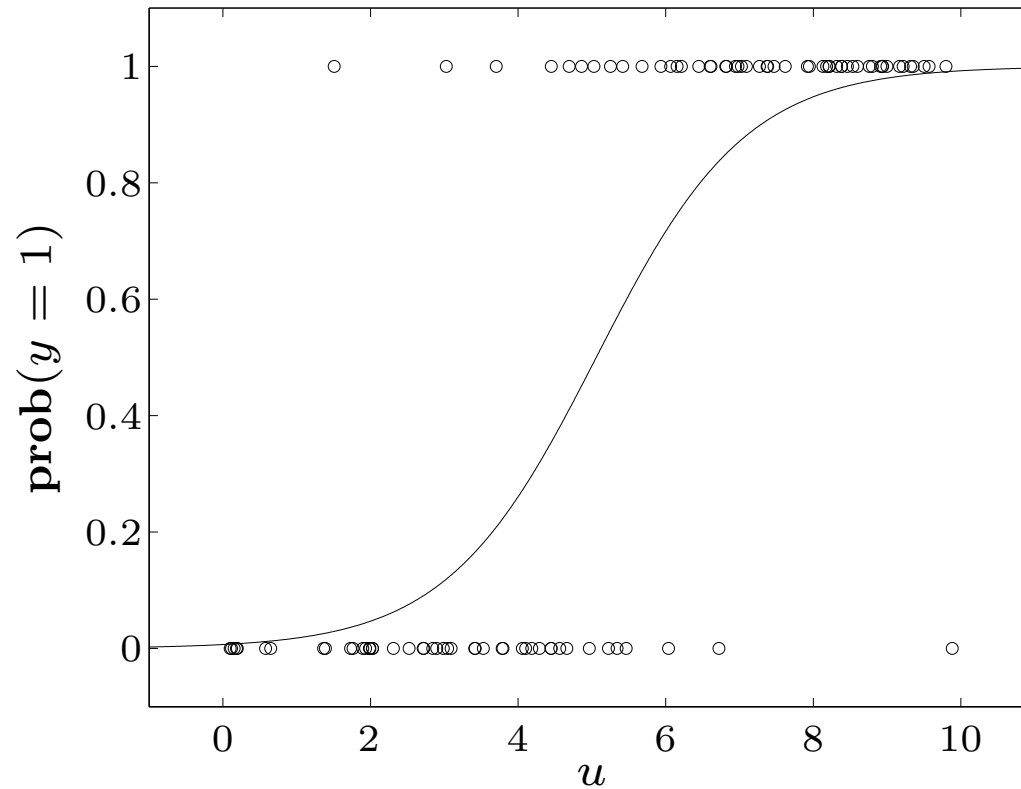
- $a, b$  are parameters;  $u \in \mathbf{R}^n$  are (observable) explanatory variables
- estimation problem: estimate  $a, b$  from  $m$  observations  $(u_i, y_i)$

**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:

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

- $P_{\text{fp}}$  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

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{R}_+^2 \text{)} & (P_{\text{fp}}, P_{\text{fn}}) = ((Tp)_2, (Tq)_1) \\ \text{subject to} & t_{1k} + t_{2k} = 1, \quad k = 1, \dots, n \\ & t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \dots, n \end{array}$$

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)

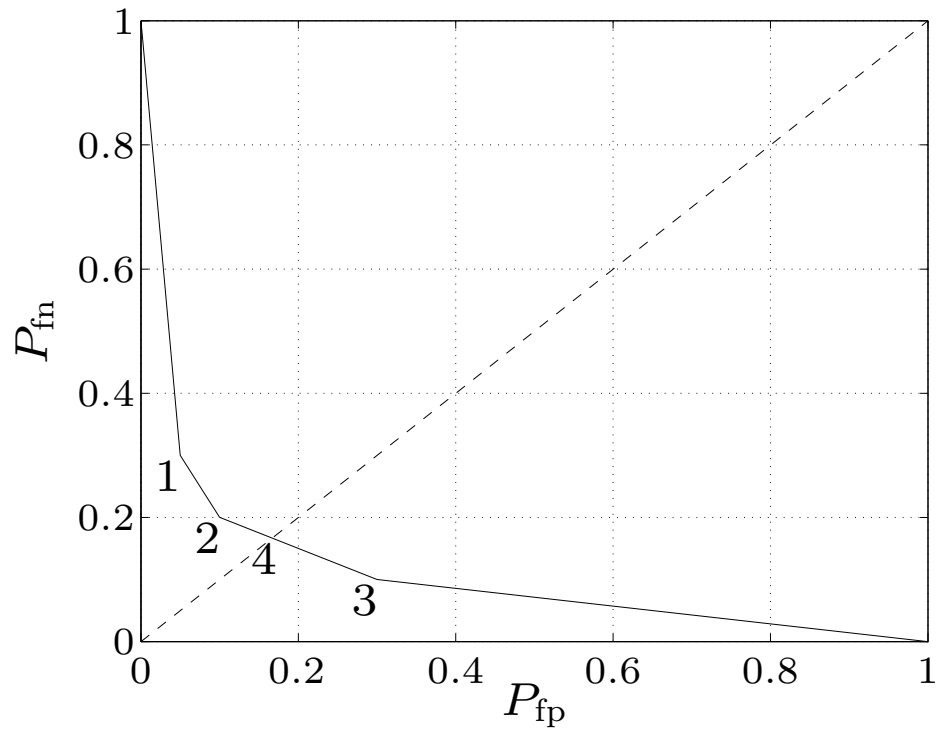
**minimax detector**

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

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

# Experiment design

$m$  linear measurements  $y_i = a_i^T x + w_i$ ,  $i = 1, \dots, m$  of unknown  $x \in \mathbf{R}^n$

- measurement errors  $w_i$  are IID  $\mathcal{N}(0, 1)$
- ML (least-squares) estimate is

$$\hat{x} = \left( \sum_{i=1}^m a_i a_i^T \right)^{-1} \sum_{i=1}^m y_i a_i$$

- error  $e = \hat{x} - x$  has zero mean and covariance

$$E = \mathbf{E} e e^T = \left( \sum_{i=1}^m a_i a_i^T \right)^{-1}$$

confidence ellipsoids are given by  $\{x \mid (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq \beta\}$

**experiment design:** choose  $a_i \in \{v_1, \dots, v_p\}$  (a set of possible test vectors) to make  $E$  'small'

## vector optimization formulation

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = \left( \sum_{k=1}^p m_k v_k v_k^T \right)^{-1} \\ \text{subject to} & m_k \geq 0, \quad m_1 + \cdots + m_p = m \\ & m_k \in \mathbf{Z} \end{array}$$

- variables are  $m_k$  ( $\#$  vectors  $a_i$  equal to  $v_k$ )
- difficult in general, due to integer constraint

## relaxed experiment design

assume  $m \gg p$ , use  $\lambda_k = m_k/m$  as (continuous) real variable

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = (1/m) \left( \sum_{k=1}^p \lambda_k v_k v_k^T \right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1 \end{array}$$

- common scalarizations: minimize  $\log \det E$ ,  $\mathbf{tr} E$ ,  $\lambda_{\max}(E)$ ,  $\dots$
- can add other convex constraints, *e.g.*, bound experiment cost  $c^T \lambda \leq B$

## ***D*-optimal design**

$$\begin{array}{ll} \text{minimize} & \log \det \left( \sum_{k=1}^p \lambda_k v_k v_k^T \right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1 \end{array}$$

interpretation: minimizes volume of confidence ellipsoids

## **dual problem**

$$\begin{array}{ll} \text{maximize} & \log \det W + n \log n \\ \text{subject to} & v_k^T W v_k \leq 1, \quad k = 1, \dots, p \end{array}$$

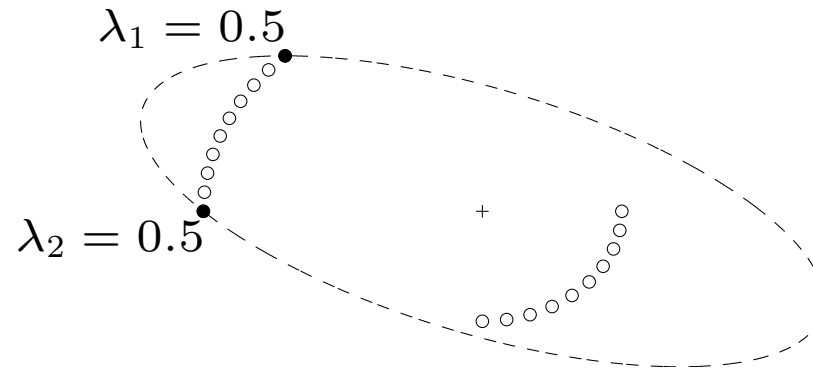
interpretation:  $\{x \mid x^T W x \leq 1\}$  is minimum volume ellipsoid centered at origin, that includes all test vectors  $v_k$

**complementary slackness:** for  $\lambda$ ,  $W$  primal and dual optimal

$$\lambda_k (1 - v_k^T W v_k) = 0, \quad k = 1, \dots, p$$

optimal experiment uses vectors  $v_k$  on boundary of ellipsoid defined by  $W$

**example** ( $p = 20$ )



design uses two vectors, on boundary of ellipse defined by optimal  $W$

## derivation of dual of page 7–13

first reformulate primal problem with new variable  $X$ :

$$\begin{aligned} & \text{minimize} && \log \det X^{-1} \\ & \text{subject to} && X = \sum_{k=1}^p \lambda_k v_k v_k^T, \quad \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1 \end{aligned}$$

$$L(X, \lambda, Z, z, \nu) = \log \det X^{-1} + \mathbf{tr} \left( Z \left( X - \sum_{k=1}^p \lambda_k v_k v_k^T \right) \right) - z^T \lambda + \nu (\mathbf{1}^T \lambda - 1)$$

- minimize over  $X$  by setting gradient to zero:  $-X^{-1} + Z = 0$
- minimum over  $\lambda_k$  is  $-\infty$  unless  $-v_k^T Z v_k - z_k + \nu = 0$

dual problem

$$\begin{aligned} & \text{maximize} && n + \log \det Z - \nu \\ & \text{subject to} && v_k^T Z v_k \leq \nu, \quad k = 1, \dots, p \end{aligned}$$

change variable  $W = Z/\nu$ , and optimize over  $\nu$  to get dual of page 7–13