9. Statistical estimation (chap 7)

- maximum likelihood estimation
- logistic regression
- ~~optimal detector design~~
- experiment design (statistics, Pukelsheim '70's, ...
  multi-armed bandits (online decision making)
  (G-optimal design)
  lots of applications ...)

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

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

- 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

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

$v_i = y_i - a_i^T x$

$p(v) = \prod_{i=1}^{m} p(v_i)$

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

- $x \in \mathbf{R}^n$ 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

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

(y is observed value)

examples

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

$$\rightarrow l(x) = \underbrace{-\frac{m}{2}\log(2\pi\sigma^2)}_{\text{const.}} - \frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^m (a_i^T x - y_i)^2}_{\|y-Ax\|_2^2}$$

  ML estimate is LS solution

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

$$l(x) = -m\log(2a) - \frac{1}{a}\underbrace{\sum_{i=1}^m |a_i^T x - y_i|}_{\|y-Ax\|_1}$$

  ML estimate is ℓ_1-norm solution

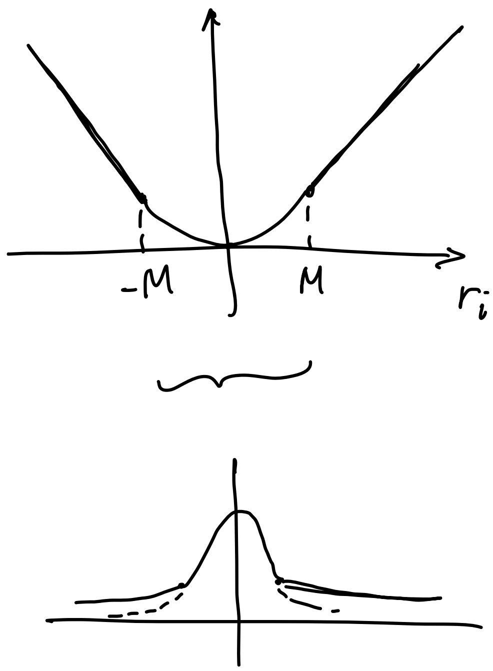
- uniform noise on $[-a,a]$:

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

  $\|y-Ax\|_\infty$

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

noise dist $\sim$ exp(penalty)

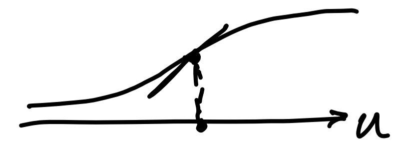


Huber penalty

$$\left[\min_x \sum_i h(y_i - a_i^T x) \right. \equiv$$ max likelihood estimate for what noise?

$$r_i = y_i - a_i^T x$$

Logistic regression



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

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

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

- 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)

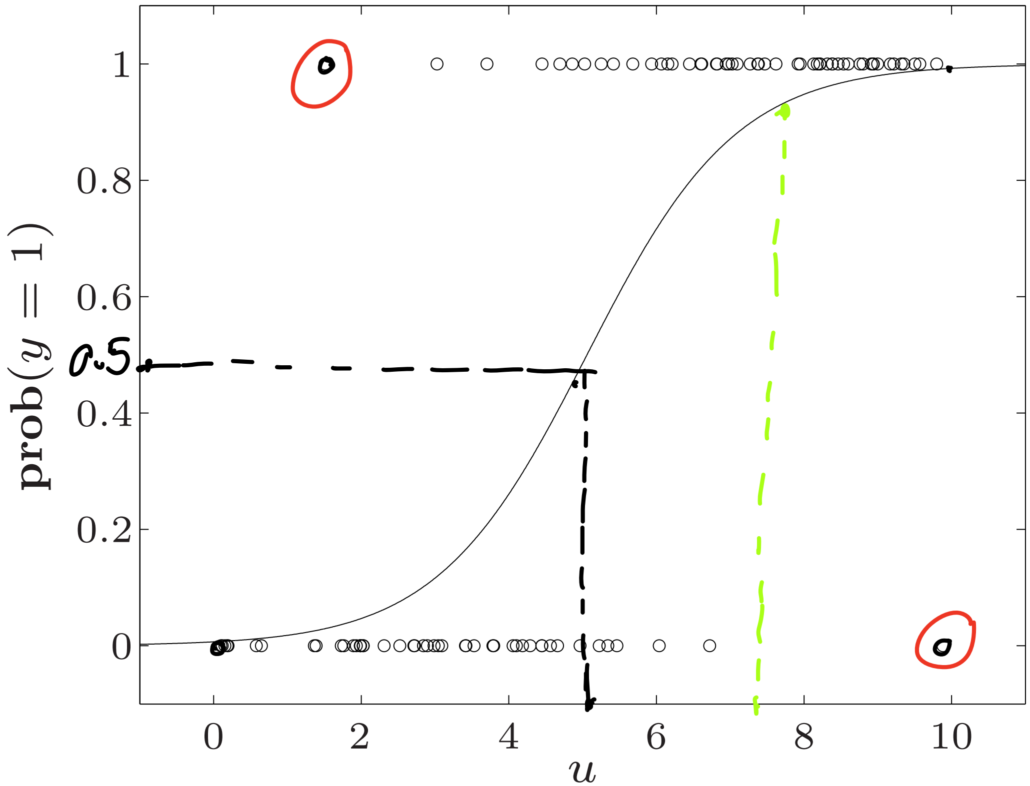
sort the data according to labels

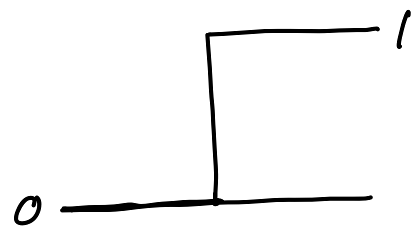
log-likelihood function (for $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$):

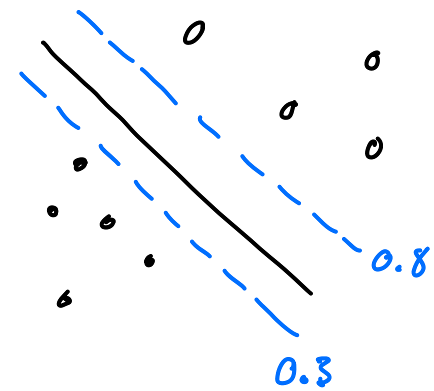
$$\begin{aligned} l(a, b) &= \log \left(\underbrace{\prod_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)}} \underbrace{\prod_{i=k+1}^{m} \frac{1}{1 + \exp(a^T u_i + b)}} \right) \\ &= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \underbrace{\log(\underset{\uparrow\, e^0}{1} + \exp(a^T u_i + b))}_{\text{log-sum-exp}} \end{aligned}$$

concave in a, b

example ($n = 1$, $m = 50$ measurements)







value of logistic loss fct.

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

can add regularizers on a : ... $+ \lambda \|a\|_2^2$

$+ \lambda \|a\|_1$ encourages sparse a

$\Rightarrow$ feature selection

Experiment design

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

deterministic

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

if $\sum a_i a_i^T$ is singular

$\Rightarrow$ a_i do not span $\mathbb{R}^n$

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

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

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

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

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

$x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ $m > n$

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} x + \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

r.v. $\sim N(0,1)$, iid

$y = Ax + w$ $\hat{x} = \underset{x}{\arg\min} \|y - Ax\|_2^2$

$$\hat{x} = \overbrace{(A^T A)^{-1} A^T}^{A^\dagger} y$$

$e = x - \hat{x} = x - (A^T A)^{-1} A^T \underline{y}$ $\leftarrow$ $\mathbb{E}y = Ax + \cancel{\mathbb{E}w}^{0} = Ax$

$\mathbb{E}e = x - (A^T A)^{-1} A^T \mathbb{E}y = x - x = 0$

$e = x - \hat{x} = x - (A^T A)^{-1} A^T (Ax + w) = \cancel{x - (A^T A)^{-1} A^T A x} - (A^T A)^{-1} A^T w = -(A^T A)^{-1} A^T w$

$$\mathbb{E}ee^T = \mathbb{E}(x - \hat{x})(x - \hat{x})^T = \mathbb{E}(A^T A)^{-1} A^T w w^T A (A^T A)^{-1}$$

$$= (A^T A)^{-1} A^T \underbrace{\mathbb{E}ww^T}_{I} A (A^T A)^{-1} = (A^T A)^{-1}$$

$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix}$ $A^T A = \sum_{i=1}^{m} a_i a_i^T$

vector optimization formulation

$E = (\sum a_i a_i^T)^{-1}$ $a_i \in \{v_1, \ldots, v_p\}$
m_1 m_p

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

- 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

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

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

D-optimal design

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

interpretation: minimizes volume of confidence ellipsoids

dual problem

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

$\Leftrightarrow$ all v_k are in ellipsoid centered at 0 defined by W

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

complementary slackness: for λ, W primal and dual optimal

primal $\rightarrow$ dual

$$\lambda_k(\underbrace{1 - v_k^T W v_k}_{\neq 0}) = 0, \quad k = 1,\ldots,p$$

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

$$\begin{cases} \min_{\lambda, X} & \log\det X^{-1} \\ \text{s.t.} & X = \sum_{k=1}^{p} \lambda_k v_k v_k^T \quad ; \quad Z \in S^n \quad \text{(equality)} \\ & \lambda \succeq 0 \quad ; \quad z \in \mathbb{R}^p \quad z \succeq 0 \\ & 1^T \lambda = 1 \quad ; \quad \nu \end{cases}$$

$$L(\underbrace{X, \lambda}; Z, z, \nu) = \log\det X^{-1} + \operatorname{Tr} Z\left(X - \sum \lambda_k v_k v_k^T\right) - z^T\lambda + \nu(1^T\lambda - 1)$$

$$= \underbrace{\log\det X^{-1} + \operatorname{Tr} ZX}_{L_1(X;\, Z, z, \nu)} - \underbrace{\operatorname{Tr} Z\left(\sum \lambda_k v_k v_k^T\right) - z^T\lambda + \nu 1^T\lambda}_{L_2(\lambda;\, Z, z, \nu)} - \nu$$

$$g(Z, z, \nu) = \inf_{X, \lambda} L(X, \lambda; Z, z, \nu)$$

$$= \inf_{X} \left(\log\det X^{-1} + \operatorname{Tr} ZX\right) + \inf_{\lambda} \left(\nu 1^T \lambda - \operatorname{Tr} Z\left(\sum \lambda_k v_k v_k^T\right) - z^T\lambda\right) - \nu$$

$$\left(\nu 1 - \begin{bmatrix} \vdots \\ v_k^T Z v_k \\ \vdots \end{bmatrix} - z\right)^T \lambda$$

$$\nabla_X L_1 = 0 \Rightarrow -X^{-1} + Z = 0$$

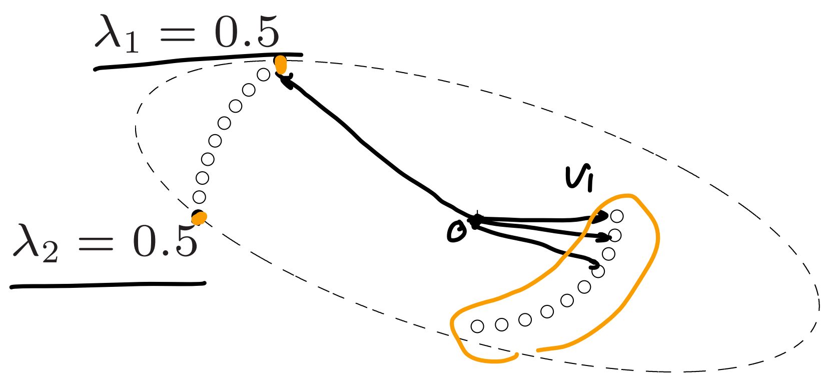
$$\nu 1 - \underbrace{\operatorname{Tr} Z v_k v_k^T}_{= \operatorname{Tr}(v_k^T Z v_k) = v_k^T Z v_k} - z$$

$$\operatorname{Tr}\left(\sum \lambda_k Z v_k v_k^T\right) = \sum \lambda_k \underbrace{\operatorname{Tr}(Z v_k v_k^T)}_{= v_k^T Z v_k} = \sum \lambda_k v_k^T Z v_k$$

$$\inf_{\lambda} L_2(\lambda; Z, z, \nu) = \begin{cases} 0 & \nu - v_k^T Z v_k - z_k = 0 \quad k = 1, \dots, p \\ -\infty & \text{else} \end{cases}$$

$v_1, \cdots, v_{20} \in \mathbb{R}^2$

example $(p = 20)$



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

$y_i = a_i^T x + w_i$ $\quad w_i \sim N(0,1)$ $\quad \|a_i\|_2 \approx SNR$ (large is good)

- pick exp. with good SNR
- pick exp. that are most informative (as orthogonal as possible)

less redundancy / overlap

derivation of dual of page 9–9

first reformulate primal problem with new variable X:

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

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

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

$v_k^T Z v_k \leq \nu$

dual problem

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

change variable $W = Z/\nu$, and optimize over ν to get dual of page 9–9

$\frac{1}{\nu} Z$ $\qquad$ $v_k^T W v_k \leq 1$

$$\left[\begin{array}{ll} \max\limits_{\nu, W} & \log\det(\nu W) - \nu = \log(\nu^n \det W) - \nu = n\underline{\underline{\log\nu}} + \log\det W \underline{\underline{-\nu}} \\ \text{s.t.} & v_k^T W v_k \le 1 \end{array}\right.$$

$$\min_{\nu} \ \nu - n\log\nu$$

$$1 - n/\nu = 0 \Rightarrow \nu = n$$

$$\left[\begin{array}{ll} \max\limits_{W} & \log\det W \\ \text{s.t.} & v_k^T W v_k \le 1 \end{array}\right.$$