Lec 11: Application, Statistical estimation

- Parametric distribution estimation
  choose from a family of densities $p_x(y)$, indexed by a parameter x

- maximum likelihood estimation

$$\max_x \; p_x(y)$$

y is observed value
$l(x) = \log p_x(y)$ is called log-likelihood function
many densities are log-concave (in x vs. in y)
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

eg. linear measurement with ZLD noise

$$y_i = a_i^T x + v_i$$

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

$$\max \; l(x) = \sum_i \log p(y_i - a_i^T x) \qquad \text{(need } p \text{ to be log-concave)}$$

eg. Gaussian noise $N(0, \sigma^2)$: $p(z) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{z^2}{2\sigma^2}\right)$

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

least-square is maximum likelihood estimation with Gaussian noise

eg. Laplacian noise $p(z) = \frac{1}{2a} \cdot \exp(-|z|/a)$

laplacian has longer tails than gaussian

$$l(x) = \sum_{i=1}^m \log\left[\frac{1}{2a} \cdot \exp(-|y_i - a_i^T x|/a)\right]$$
$$= -m \cdot \log(2a) - \frac{1}{a} \sum_{i=1}^m |y_i - a_i^T x|$$

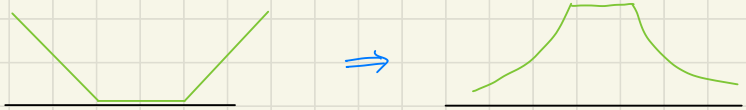
minimizing L_1 norm is mle with Laplacian noise

eg. uniform noise in $[-a, a]$

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

mle estimation is any x with $|a_i^T x - y_i| \le a$ (feasibility problem)

eg. penalty function



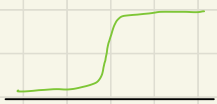
penalty function

noise distribution
(uniform center and exponential tail)

eg. logistic regression

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

$$p = \text{prob}(y=1) = \exp(a^T u + b) / [1 + \exp(a^T u + b)]$$



a, b are parameters, $u \in \mathbb{R}^n$ are observations

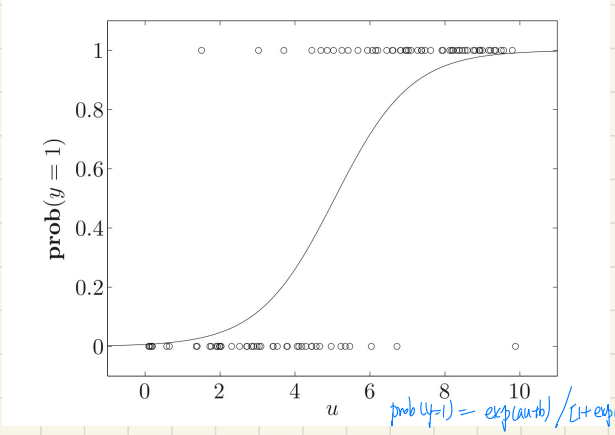
estimation problem: estimate a, b from (u_i, y_i)

$(y_1 = \cdots = y_k = 1 \quad y_{k+1} = \cdots = y_m = 0)$

$$l(a,b) = \log\left(\prod_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} \cdot \prod_{i=k+1}^{m} \frac{1}{1 + \exp(a^T u_i + b)} \right)$$

$$= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log[1 + \exp(a^T u_i + b)]$$

concave in a, b



prob $(y=1) = \exp(au+b) / [1+\exp(au+b)]$

b controls the horizontal point

a controls the steepness

eg. Covariance estimation for Gaussian variables

Suppose $y \in \mathbb{R}^n$ is a gaussian random variable

$\mathbb{E}[y] = 0$ $R = \text{cov}(y) = \mathbb{E}[yy^T]$

$$p_R(y) = \frac{1}{(2\pi)^{n/2} \cdot |R|^{1/2}} \exp\left(-\frac{y^T R^{-1} y}{2}\right) \quad R \in S_{++}^n$$

$$l(R) = \sum_{i=1}^m \log p(y_i)$$

$$= -m \cdot \left[\frac{n}{2} \cdot \log 2\pi + \frac{1}{2} \log\det(R)\right] - \frac{1}{2} \sum_{i=1}^m y_i^T R^{-1} y_i$$

$$= -m\left[\frac{n}{2} \cdot \log 2\pi + \frac{1}{2} \log\det(R)\right] - \frac{m}{2} \text{tr}(R^{-1} Y)$$

$Y = \frac{1}{m} \cdot \sum y_i y_i^T$ is the sample variance

let $S = R^{-1}$ (the information matrix)

$$= -\frac{mn}{2} \cdot \log 2\pi + m/2 \log\det S - \frac{m}{2} \text{tr}(SY)$$

is concave in S

$$\max. \quad \log\det(S) - \text{tr}(SY)$$

1° no constraints on S

$$\max \ \log\det(S) - \text{tr}(SY)$$
$$s.t. \quad S \in S_{++}^n \qquad (S = R^{-1})$$

$$\nabla_S (\log\det(S) - \text{tr}(SY)) = S^{-1} - Y = 0$$

$R = S^{-1} = Y$ if $Y \in S_{++}^n$

the problem is unbounded below if $Y \notin S_{++}^n$

if there's no prior assumptions on R,

then mle estimation of R is sample covariance

- maximum a posteriori probability estimation.
  (a bayesian version of MLE)
  assume that x (to be estimated) and and y (the observation) are random variables with a joint probability $p(x,y)$

  the prior density is given by $p_x(x) = \int p(x,y)\,dy \quad p_y(y) = \int p(x,y)\,dy$

  the conditional density is given by $p_{y|x}(y|x) = \frac{p(x,y)}{p_x(x)}$

  $$\begin{aligned}\hat{x}_{map} &= \arg\max_x p_{x|y}(x|y) \\ &= \arg\max_x p_{x|y}(x|y)\cdot p_y(y) \\ &= \arg\max_x p(x,y) \\ &= \arg\max_x p_{y|x}(y|x)\cdot p(x) \quad \text{taking prior knowledge of } x \text{ into account compared to MLE}\end{aligned}$$

  eg. linear measurement with ZZD noise
  $y_i = a_i^T x + v_i$
  v_i are iid noise with density p_v on $\mathbb{R}$, x has prior density p_x on $\mathbb{R}^n$

  $$p(x,y) = p_x(x)\cdot \prod_{i=1}^m p_v(y_i - a_i^T x)$$
  $$\Downarrow$$
  $$\max \quad \log p_x(x) + \sum_{i=1}^m \log p_v(y_i - a_i^T x)$$

  when v_i is uniform on $[-a,a]$, and the prior distribution of x is gaussian

  $$\max_x \quad \log \frac{1}{(2\pi)^{n/2}\cdot|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\bar{x})^T\Sigma^{-1}(x-\bar{x})\right)$$
  $$\text{s.t.} \quad \|Ax-b\|_\infty \le a$$
  $$\Downarrow$$
  $$\min \quad (x-\bar{x})\cdot\Sigma^{-1}(x-\bar{x})$$
  $$\text{s.t.} \quad \|Ax-b\|_\infty \le a$$

  eg. MAP with perfect measurements
  $x \in \mathbb{R}^n$ is a vector of parameters to be estimated, p_x
  have m perfect, noise-free measurements $y = Ax$

  $$\max. \quad \log p_x(x)$$
  $$\text{s.t.} \quad Ax = y$$

(Binary) hypothesis testing

1. detection (hypothesis testing) problem
   given observation of a random variable $X \in \{1, \dots, n\}$ choose between:
   hypothesis 1: X was generated by distribution $p = (p_1 \cdots p_n)$ $\quad 1^T p = 1 \quad p \geq 0$
   hypothesis 2: X was generated by distribution $q = (q_1 \cdots q_n)$ $\quad 1^T q = 1 \quad q \geq 0$

2. randomized detector
   a nonnegative matrix $T \in \mathbb{R}^{2 \times n}$, with $1^T T = 1$ $\quad \begin{bmatrix} .9 & .8 & .6 \\ .1 & .2 & .4 \end{bmatrix}$
   if we observe $X = k$, we choose hypo 1 with prob T_{1k}, hypo 2 with prob T_{2k}
   if T has all elements 0 or 1, its a deterministic detector

3. detection probability matrix

$$D = T \cdot [p \; q] = [Tp \; Tq] = \begin{bmatrix} 1 - P_{fp} & P_{fn} \\ P_{fp} & 1 - P_{fn} \end{bmatrix}$$

   P_{fp} is probability of choosing hypothesis 2 when X generated by distribution 1

eg. multicriterion formulation of detector design.

$$\begin{aligned} \text{min. (w.r.t } \mathbb{R}^2_+) \quad & (P_{fp}, P_{fn}) = ((Tp)_2, (Tq)_1) \\ \text{s.t.} \quad & t_{1k} + t_{2k} = 1 \\ & t_{ik} \geq 0 \end{aligned} \qquad \begin{bmatrix} p_1 & \cdots \\ q_1 & \cdots \end{bmatrix}$$

$$\Downarrow$$

$$\begin{aligned} \min \quad & (Tp)_2 + (Tq)_1 \\ \text{s.t.} \quad & T_{1k} + T_{2k} = 1 \\ & T_{ik} \geq 0 \end{aligned} \qquad [\;\;][\;\;]$$

a LP with a simple analytic solution $\quad (T_{1k}, T_{2k}) = \begin{cases} (1,0) & p_k \geq \lambda \cdot q_k \\ (0,1) & p_k < \lambda q_k \end{cases}$

- a deterministic detector, given by a likelihood ratio test $\left(p_k / q_k \gtrless \lambda \right)$
- when $\lambda = 1$, it's minimizing P(being wrong) $\Leftrightarrow$ maxi P(being right) (MLE)

eg. minimax detector

$$\begin{aligned} \min. \quad & \max\{P_{fn}, P_{fp}\} = \max\{(Tp)_2, (Tq)_1\} \\ \text{s.t.} \quad & t_{1k} + t_{2k} = 1 \quad t_{ik} \geq 0 \end{aligned}$$

a LP; solution is usually not deterministic

• Experiment design

m linear measurements $y_i = a_i^T x + w_i$ of unknown $x \in \mathbb{R}^n$
$w_i \sim \mathcal{N}(0,1)$ iid

$$A = \begin{bmatrix} - & a_1 & - \\ & \vdots & \\ - & a_m & - \end{bmatrix}$$

least-square estimate is: $\hat{x} = (\sum_i a_i a_i^T)^{-1} \sum_i y_i a_i \quad (A^TA)^{-1}A^Ty$

error $e = \hat{x} - x$ has zero mean and covariance: $E = E[ee^T] = (\sum_i a_i a_i^T)^{-1}$

$$\begin{aligned}
\hat{x} &= (A^TA)^{-1}A^Ty \\
&= (A^TA)^{-1}A^T(Ax+w) \\
&= (A^TA)^{-1}A^TAx + (A^TA)^{-1}A^Tw \\
&= x + (A^TA)^{-1}A^Tw
\end{aligned}$$

$E[\hat{x}] = x \quad (w \sim \mathcal{N}(0,1))$

$$\begin{aligned}
E[(\hat{x}-x)(\hat{x}-x)^T] &= E\left[(A^TA)^{-1}A^Tw \cdot (A^TA)^{-1}A^Tw)^T \mid A\right] \\
&= (A^TA)^{-1}A^T E[ww^T] \cdot A (A^TA)^{-1} \\
E &= (A^TA)^{-1} \quad \text{(if } A \text{ large, then it has high signal-to-noise ratio, so covariance small)}
\end{aligned}$$

confidence ellipsoid is given by: $\{x \mid (x-\hat{x})^T \cdot E^{-1}(x-\hat{x}) \le \beta\}$

experiment design: choose a set of possible tests to make E small
{ choose a_i for m_i times $\quad \sum_i m_i = m$ }

eg sensor a_1 has the highest signal-to-noise ratio.
if choose on for two times and no other a_i.
then $\sum a_i a_i^T$ is not invertible (rank 1)

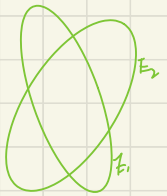
$$\downarrow$$

$$\begin{aligned}
&\min. \ (\text{w.r.t } \le) \quad E = \left(\sum_{k=1}^{p} m_k \cdot v_k v_k^T\right)^{-1} \\
&\text{s.t.} \quad m_k \ge 0 \qquad m_1 + \cdots + m_p = m \\
&\qquad m_k \in \mathbb{Z}
\end{aligned}$$

(hard to solve due to the integer constraint)

$$\Downarrow$$

relaxed experiment design



min. (w.r.t S_+^n) $E = 1/m \left(\sum_{k=1}^{p} \lambda_k \cdot v_k v_k^T \right)^{-1}$

s.t. $\lambda \geq 0$

$1^T \lambda = 1$

Common scalarization: min. $\log\det(E)$, $\mathrm{tr}(E)$, $\lambda_{max}(E)$

eg. D-optimal design

min. $\log\det \left(\sum_{k=1}^{p} \lambda_k v_k v_k^T \right)^{-1}$

s.t. $\lambda \geq 0$ $1^T \lambda = 1$ (minimizing the volume of confidence ellipsoid)

$\Downarrow$

min. $\log\det(X^{-1})$

s.t. $X = \sum_{k=1}^{p} \lambda_k v_k v_k^T$ (z) $(v_k \in \mathbb{R}^n)$

$\lambda \geq 0$ (α) $1^T \lambda = 1$ (β)

$$\begin{aligned} L(X, \lambda, \alpha, z, \beta) &= \log\det(X^{-1}) + \mathrm{tr}\left\{ z \left(X - \sum_{k=1}^{p} \lambda_k v_k v_k^T \right) \right\} - \alpha^T \lambda + \beta (1^T \lambda - 1) \\ &= \log\det(X^{-1}) + \mathrm{tr}(zX) - \mathrm{tr}\left\{ z \cdot \left(\sum_{k=1}^{p} \lambda_k v_k v_k^T \right) \right\} - \alpha^T \lambda + \beta (1^T \lambda - 1) \\ &= \log\det(X^{-1}) + \mathrm{tr}(zX) - \sum_{k=1}^{p} \lambda_k \cdot v_k^T z v_k - \sum_{k=1}^{p} \alpha_k \lambda_k + \sum_{k=1}^{p} \beta \cdot \lambda_k - \beta \\ &= \log\det(X^{-1}) + \mathrm{tr}(zX) + \sum_{k=1}^{p} \lambda_k \left(-v_k^T z v_k - \alpha_k + \beta \right) - \beta \end{aligned}$$

$$g(\alpha, z, \beta) = \inf_{\substack{X \in S_+^n \\ \lambda}} \left\{ L(X, \alpha, z, \beta) \right\}$$

$$= \begin{cases} \log\det(X^{-1}) + \mathrm{tr}(zX) - \beta & -v_i^T z v_i - \alpha_i + \beta = 0 \\ -\infty & \text{otherwise} \end{cases}$$

$\forall X \quad \log\det(X^{-1}) + \mathrm{tr}(zX) = 0$

$-X^{-1} + z = 0 \quad z = X^{-1}$

$$\therefore g(\alpha, z, \beta) = \begin{cases} \log\det z + n - \beta & -v_i^T z v_i - \alpha_i + \beta = 0 \\ -\infty & \text{otherwise} \end{cases}$$

$\Downarrow$

max. $\log\det z + n - \beta$

s.t. $\beta - v_i^T z v_i \geq 0$

$\Downarrow$ $w = (1/\beta) \cdot z$

max. $\log\det W + n\log\beta + n - \beta$

s.t. $v_i^T W v_i \le 1$

$\Leftrightarrow$ max $n\log\beta - \beta$ over β : $\beta = n$

dual of D-optimal

max. $\log\det W + n\log n$

s.t. $v_i^T W v_i \le 1$

$\{x \mid x^T W x \le 1\}$ is the minimum volume ellipsoid at origin that includes all test vector v_i

Complementary slackness:

for α, W that's a primal-optimal – dual-optimal pair

$\alpha_i (1 - v_k^T W v_k) = 0$

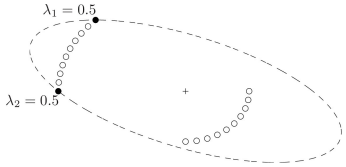


Figure 7.9 Experiment design example. The 20 candidate measurement vectors are indicated with circles. The D-optimal design uses the two measurement vectors indicated with solid circles, and puts an equal weight $\lambda_i = 0.5$ on each of them. The ellipsoid is the minimum volume ellipsoid centered at the origin, that contains the points v_i.