Statistical estimation

- 7.1 Parametric distribution estimation
- 7.2 Nonparametric distribution estimation
- 7.3 Optimal detector design and hypothesis testing

김민주 이종현

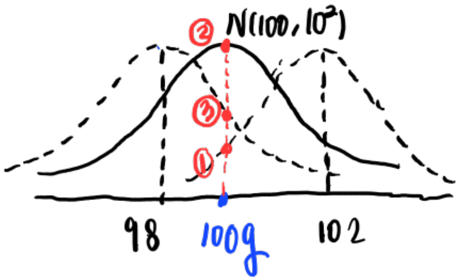
7.1 Parametric distribution estimation

Maximum likelihood estimation

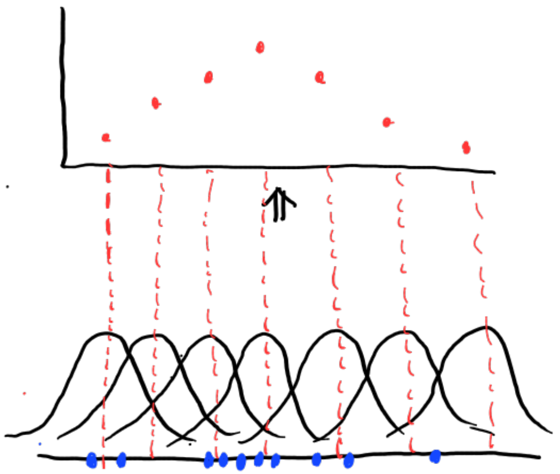
MLE

① <확률> P(X|D)

② <가능도> P(D|X)



< likelihood >



01
02
03

Maximum likelihood estimation

$x \in \mathbf{R}^n, y \in \mathbf{R}^m$ 에 대해

likelihood function: $p_x(y)$

log likelihood function: $l(x) = logp_x(y)$

Maximum likelihood estimation: $\hat{x}_{ml}$ = $argmax_x p_x(x)$=$argmax_x l(x)$

maximize $l(x) = logp_x(y)$

subject to $x \in C$

Maximum likelihood estimation problem이 convex optimization problem이 될 조건

- l이 concave
- linear equality constraints
- convex inequality constraints

01
02
03

Linear measurements with IID noise

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

where $x \in \mathbf{R}^n$: parameter vector, $y_i \in \mathbf{R}$: observed quantities

and v_i's are IID

likelihood function: $p_x(y) = \Pi_{i=1}^{m} p(y_i - a_i^T x)$

log likelihood function: $l(x) = log p_x(y) = \Sigma_{i=1}^{m} log\ p(y_i - a_i^T x)$

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

-$p_x(y)$가 concave면 convex optimization

Gaussian noise

For $v_i \sim N(0, \sigma^2)$, $p(z) = (2\pi\sigma^2)^{-\frac{1}{2}}\ e^{-z^2/2\sigma^2}$

$l(x) = -(m/2) log(2\pi\sigma^2) - \frac{1}{2\sigma^2} ||Ax - y||_2^2$

$x_{ml} = argmin_x ||Ax - y||_2^2$

Uniform noise

For $v_i \sim U[-a, a]$, $p(z) = 1/(2a), -a < z < a$

x_{ml}: $||Ax - y||_\infty \leq a$를 만족하는 any x

01
02
03

Counting problems with Poisson distribution

y: 사건 발생 횟수

$y \sim Poi(\mu)$를 따를 때 $\mathbf{prob}(y = k) = \frac{e^{-\mu}\mu^k}{k!}$ 이고,

$\mu = a^T u + b$

(u: 설명 변수, $a \in \mathbf{R}^n$, $b \in \mathbf{R}$은 parameter)

a, b의 mle를 찾으려면,

$$\prod_{i=1}^{m} \frac{(a^T u_i + b)^{y_i} \exp(-(a^T u_i + b))}{y_i!}$$

$$l(a, b) = \sum_{i=1}^{m} (y_i \log(a^T u_i + b) - (a^T u_i + b) - \log(y_i!))$$

maximize $\quad \sum_{i=1}^{m} (y_i \log(a^T u_i + b) - (a^T u_i + b))$

ex) y = 교통사고 발생 횟수

$\mu = \underset{\text{교통량}}{a_1 u_1} + \underset{\text{강수량}}{a_2 u_2} + b$

Logistic regression

Y : 병

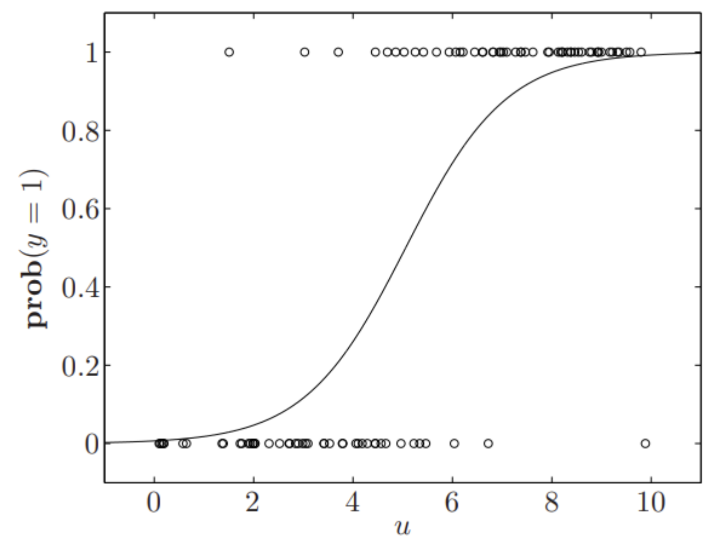
ex) $P(Y=y) = \begin{cases} p, & y=1 \\ 1-p, & y=0 \end{cases}$

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

$$\mathbf{prob}(y = 1) = p, \qquad \mathbf{prob}(y = 0) = 1 - p,$$

$$L(a,b) = \prod_{i=1}^{q} p_i \prod_{i=q+1}^{m} (1 - p_i)$$

$$\begin{aligned} l(a,b) &= \sum_{i=1}^{q} \log p_i + \sum_{i=q+1}^{m} \log(1 - p_i) \\ &= \sum_{i=1}^{q} \log \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} + \sum_{i=q+1}^{m} \log \frac{1}{1 + \exp(a^T u_i + b)} \\ &= \sum_{i=1}^{q} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b)). \end{aligned}$$



7.1.2

MAP estimation

Bayes theorem

$$\underbrace{P(A|B)}_{\text{posterior}} = \frac{\overbrace{P(B|A)}^{\text{likelihood}} \times \overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{B의 prior}}}$$

ex) X는 60% 확률로 거짓말(A), 거짓말 탐지기의 정확도는 90%라 가정 ⇒ $P(A|B) = ?$
(B: 거짓말 탐지기 양성)
(즉, $P(A) = 0.6$, $P(B|A) = 0.9$ 가정)

$$\Rightarrow P(\text{실제거짓}|\text{결과거짓}) = \frac{P(\text{결과거짓}|\text{실제거짓})\, P(\text{실제거짓})}{\underbrace{P(\text{결과거짓})}_{P(\text{결과거짓}|\text{실제 거짓}) + P(\text{결과거짓}|\text{실제 거짓X})}} = \frac{0.9 \times 0.6}{0.9\times0.6 + 0.1\times0.4} = \underbrace{0.93}_{\text{posterior}}$$

01
02
03

MAP estimation

$$p_x(x) = \int p(x,y)\, dy \qquad p_y(y) = \int p(x,y)\, dx.$$

$$p_{y|x}(x,y) = \frac{p(x,y)}{p_x(x)}$$

$$p_{x|y}(x,y) = \frac{p(x,y)}{p_y(y)} = p_{y|x}(x,y)\frac{p_x(x)}{p_y(y)}$$

$$\begin{aligned} \hat{x}_{\text{map}} &= \text{argmax}_x p_{x|y}(x,y) \\ &= \text{argmax}_x p_{y|x}(x,y) p_x(x) \\ &= \text{argmax}_x p(x,y). \end{aligned}$$

$$\hat{x}_{\text{map}} = \text{argmax}_x (\log p_{y|x}(x,y) + \log p_x(x))$$

01
02
03

Linear measurements with IID noise

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

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

$$\text{maximize} \quad \log p_x(x) + \textstyle\sum_{i=1}^{m} \log p_v(y_i - a_i^T x)$$

ex) $v_i \sim U[-a, a]$

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

7.2 Nonparametric distribution estimation

$P(X=a_k)=p_k,\ k=1,\cdots,n$

$$\{p \in \mathbf{R}^n \mid p \succeq 0,\ \mathbf{1}^T p = 1\}$$

Prior information

$$\mathbf{E}\, f(X) = \sum_{i=1}^{n} p_i f(\alpha_i)$$

1) $$\mathbf{prob}(X \in C) = c^T p, \qquad c_i = \begin{cases} 1 & \alpha_i \in C \\ 0. & \alpha_i \notin C. \end{cases}$$

2) $$\mathbf{E}\, X = \sum_{i=1}^{n} \alpha_i p_i = \alpha, \qquad \mathbf{E}\, X^2 = \sum_{i=1}^{n} \alpha_i^2 p_i = \beta, \qquad \sum_{\alpha_i \geq 0} p_i \leq 0.3,$$

3) $$\mathbf{var}(X) = \mathbf{E}\, X^2 - (\mathbf{E}\, X)^2 = \underbrace{\sum_{i=1}^{n} \alpha_i^2 p_i}_{\text{linear}} - \underbrace{\left(\sum_{i=1}^{n} \alpha_i p_i\right)^2}_{\text{quadratic}}$$

4) $$\mathbf{prob}(X \in A | X \in B) = c^T p / d^T p,$$

where

$$c_i = \begin{cases} 1 & \alpha_i \in A \cap B \\ 0 & \alpha_i \notin A \cap B \end{cases}, \qquad d_i = \begin{cases} 1 & \alpha_i \in B \\ 0 & \alpha_i \notin B. \end{cases}$$

$$l d^T p \leq c^T p \leq u d^T p$$

Bounding probabilities and expected values

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} f(\alpha_i) p_i \\ \text{subject to} & p \in \mathcal{P}. \end{array}$$

Maximum likelihood estimation

$$l(p) = \sum_{i=1}^{n} k_i \log p_i,$$

Maximum entropy

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} p_i \log p_i \\ \text{subject to} & p \in \mathcal{P}. \end{array}$$

Example 7.2 We consider a probability distribution on 100 equidistant points α_i in the interval $[-1, 1]$. We impose the following prior assumptions:

$$\begin{array}{rcl} \mathbf{E}\,X & \in & [-0.1, 0.1] \\ \mathbf{E}\,X^2 & \in & [0.5, 0.6] \\ \mathbf{E}(3X^3 - 2X) & \in & [-0.3, -0.2] \\ \mathbf{prob}(X < 0) & \in & [0.3, 0.4]. \end{array} \tag{7.8}$$

Along with the constraints $\mathbf{1}^T p = 1$, $p \succeq 0$, these constraints describe a polyhedron of probability distributions.

Figure 7.2 shows the maximum entropy distribution that satisfies these constraints. The maximum entropy distribution satisfies

$$\begin{array}{rcl} \mathbf{E}\,X & = & 0.056 \\ \mathbf{E}\,X^2 & = & 0.5 \\ \mathbf{E}(3X^3 - 2X) & = & -0.2 \\ \mathbf{prob}(X < 0) & = & 0.4. \end{array}$$

To illustrate bounding probabilities, we compute upper and lower bounds on the cumulative distribution $\mathbf{prob}(X \le \alpha_i)$, for $i = 1, \ldots, 100$. For each value of i, we solve two LPs: one that maximizes $\mathbf{prob}(X \le \alpha_i)$, and one that minimizes $\mathbf{prob}(X \le \alpha_i)$, over all distributions consistent with the prior assumptions (7.8). The results are shown in figure 7.3. The upper and lower curves show the upper and lower bounds, respectively; the middle curve shows the cumulative distribution of the maximum entropy distribution.

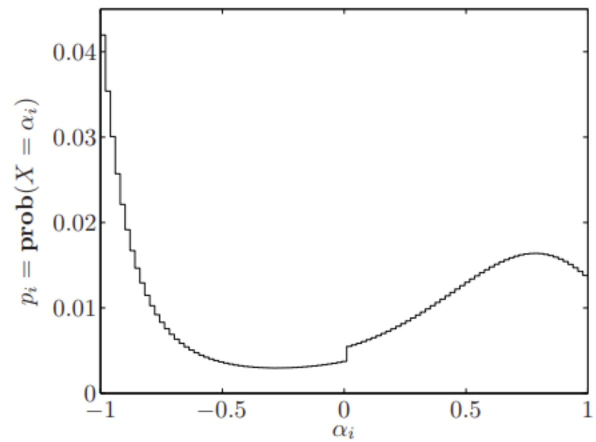


Figure 7.2 Maximum entropy distribution that satisfies the constraints (7.8).

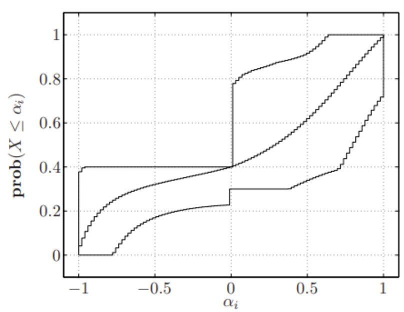


Figure 7.3 The top and bottom curves show the maximum and minimum possible values of the cumulative distribution function, $\mathbf{prob}(X \le \alpha_i)$, over all distributions that satisfy (7.8). The middle curve is the cumulative distribution of the maximum entropy distribution that satisfies (7.8).

7.3 Optimal detector design and Hypothesis testing

$P = [\quad]$ θ=j일 때, X의 dist.

$$p_{kj} = \mathbf{prob}(X = k \mid \theta = j).$$

7.3.1 Deterministic and randomized detectorsc

Maximum likelihood detector: $\hat{\theta} = \psi_{\mathrm{ml}}(k) = \underset{j}{\operatorname{argmax}}\, p_{kj}.$

T: $t_{ik} = \mathbf{prob}(\hat{\theta} = i \mid X = k).$

$$t_k \succeq 0, \qquad \mathbf{1}^T t_k = 1.$$

7.3.2 Detection probability matrix

$$D_{ij} = (TP)_{ij} = \mathbf{prob}(\hat{\theta} = i \mid \theta = j),$$

Detection probabilities: $P_i^{\mathrm{d}} = D_{ii} = \mathbf{prob}(\hat{\theta} = i \mid \theta = i).$

Error probabilities: $P_i^{\mathrm{e}} = 1 - D_{ii} = \mathbf{prob}(\hat{\theta} \neq i \mid \theta = i).$ $\quad P_i^{\mathrm{e}} = \sum_{j \neq i} D_{ji}.$

01
02
03

7.3.3 Optimal detector design

Limits on errors and detection probabilities

$$P_j^{\mathrm{d}} = D_{jj} \geq L_j, \qquad D_{ij} \leq U_{ij}$$

Minimax detector design

$$\begin{array}{ll} \text{minimize} & \max_j P_j^{\mathrm{e}} \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1, \ldots, n, \end{array}$$

Bayes detector design

$$q_i = \mathbf{prob}(\theta = i)$$

$$\begin{array}{ll} \text{minimize} & q^T P^{\mathrm{e}} \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1, \ldots, n. \end{array}$$

Bias, mean-square error, and other quantities

$$\mathbf{prob}(\hat{\theta} > \theta \mid \theta = i) = \sum_{j>i} D_{ji},$$

$$\mathbf{prob}(|\hat{\theta} - \theta| > 1 \mid \theta = i) = \sum_{|j-i|>1} D_{ji},$$

Bias: $\mathbf{E}_i(\hat{\theta} - \theta) = \sum^{m} (\theta_j - \theta_i) D_{ji},$

Mean square error: $\mathbf{E}_i(\hat{\theta} - \theta)^2 = \sum_{i=1}^{m} (\theta_j - \theta_i)^2 D_{ji}.$

Average absolute error: $\mathbf{E}_i |\hat{\theta} - \theta| = \sum_{j=1}^{m} |\theta_j - \theta_i| D_{ji}.$

7.3.4 Multicriterion formulation and scalarization

minimize D_{ij} , $i,j=1,\cdots,m$, $i\neq j$
subject to $t_k \succeq 0$, $1^T t_k = 1$, $k=1,\cdots,n$

$W_{ii}=0$, $W_{ij} \sim P^e(\hat{\theta}=i \mid \theta=j)$

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(W^T D) \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \quad k = 1, \ldots, n, \end{array}$$

LP

$$\mathbf{tr}(W^T D) = \mathbf{tr}(W^T T P) = \mathbf{tr}(P W^T T) = \sum_{k=1}^{n} c_k^T t_k,$$

(c_k = WP^T의 열벡터)

$$\begin{array}{ll} \text{minimize} & c_k^T t_k \\ \text{subject to} & t_k \succeq 0, \quad \mathbf{1}^T t_k = 1, \end{array}$$

$$\hat{\theta} = \underset{j}{\operatorname{argmin}} (W P^T)_{jk}.$$

01

02

03

7.3.5 Binary hypothesis testing

$$D = \begin{bmatrix} 1 - P_{\mathrm{fp}} & P_{\mathrm{fn}} \\ P_{\mathrm{fp}} & 1 - P_{\mathrm{fn}} \end{bmatrix}$$

$$\hat{\theta} = \begin{cases} 1 & W_{21} p_k > W_{12} q_k \\ 2 & W_{21} p_k \leq W_{12} q_k \end{cases} \quad \left(\frac{p_k}{q_k} > \frac{W_{12}}{W_{21}} \right)$$

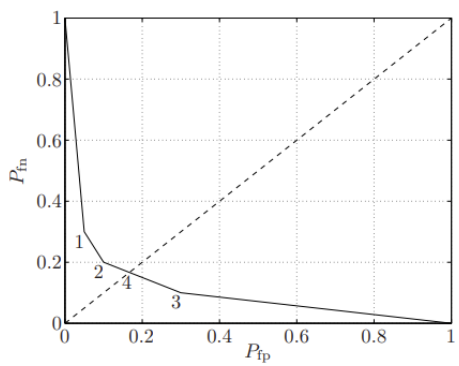


Figure 7.4 Optimal trade-off curve between probability of a false negative, and probability of a false positive test result, for the matrix P given in (7.15). The vertices of the trade-off curve, labeled 1–3, correspond to deterministic detectors; the point labeled 4, which is a randomized detector, is the minimax detector. The dashed line shows $P_{\mathrm{fn}} = P_{\mathrm{fp}}$, the points where the error probabilities are equal.

Example 7.4 We consider a binary hypothesis testing example with $n = 4$, and

$$P = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}. \tag{7.15}$$

The optimal trade-off curve between P_{fn} and P_{fp}, *i.e.*, the receiver operating curve, is shown in figure 7.4. The left endpoint corresponds to the detector which is always negative, independent of the observed value of X; the right endpoint corresponds to the detector that is always positive. The vertices labeled 1, 2, and 3 correspond to the deterministic detectors

$$T^{(1)} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$T^{(2)} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

$$T^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix},$$

respectively. The point labeled 4 corresponds to the nondeterministic detector

$$T^{(4)} = \begin{bmatrix} 1 & 2/3 & 0 & 0 \\ 0 & 1/3 & 1 & 1 \end{bmatrix},$$

which is the minimax detector. This minimax detector yields equal probability of a false positive and false negative, which in this case is 1/6. Every deterministic detector has either a false positive or false negative probability that exceeds 1/6, so this is an example where a randomized detector outperforms every deterministic detector.

ESC 2023 spring WEEK1
About bounds and Experiment Design

학술부: 김민주,이종현

March 9,2023

Three Bounds

- $P(|X| \geq c) \leq E(|Z|^r)/c^r$:Markov Bound
- $P(|X-\mu| \geq 1) \leq \sigma^2$:Chebyshev Bound
- $P(X \geq u) \leq \inf_{\lambda \geq 0} E(e^{\lambda(X-u)})$: Chernoff Bound

can be deriven by convex optimization!

Chebyshev Bound: Objective

X: RV on $S \subseteq R^m, C \subseteq S$
Find the bound of $P(X \in C) = E(1_C(x))$

Chebyshev Bound:Terms

prior knowledge: $E(f_i(X)) = a_i, E(f_0(X)) = 1$
linear combination and expectation:$f(z) = \sum_{i=0}^{n} x_i f_i(z) \rightarrow E(f(X)) = a^T x$
key idea: $f(z) \geq 1_C(z)$ *for all* $z \in C$
$\rightarrow E(f(X)) = a^T x \geq E(1_C(Z)) = P(X)$

Chebyshev Bound Problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & x_0 + a_1 x_1 + \ldots + a_n x_n \\ \text{subject to} \quad & f(z) = \sum_{i=0}^{n} x_i f_i(z) \geq 1,\ \ for\ z \in C \\ & f(z) = \sum_{i=0}^{n} x_i f_i(z) \geq 0,\ \ for\ z \in S \setminus C \end{aligned}$$

- $1 - \inf_{z} f(z) \leq 0$ is convex
- $- \inf_{z \in S \setminus C} f(z) \leq 0$ is convex

$\Rightarrow$ Above problem is convex optimization

Chebyshev Bound Problem:Markov

let $S \subseteq R_+,\ C = [1, \infty),\ f_0(z) = 1,\ f_1(z) = z,\ E(f_1(X)) = \mu \leq 1$

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & x_0 + \mu x_1 \\ \text{subject to} \quad & x_0 \geq 0, x_1 \geq 0 \\ & x_0 + x_1 \geq 1 \end{aligned}$$

$\rightarrow$ This problem becomes simple LP, where it yields:
$P(X \geq 1) \leq \mu$:Markov inequality

Chebyshev Bound Problem:with First and Second moment

let $E(X) = a \in R^m, E(XX^T) = \Sigma \in S^m$
$f(z) = z^T Pz + 2q^T z + r$

$$\begin{aligned} E(f(X)) &= E(X^T PX + 2q^T X + r) \\ &= E(tr(PXX^T) + 2E(q^T X) + r \\ &= tr(\Sigma P) + 2q^T a + r \end{aligned}$$

Chebyshev Bound Problem:with First and Second moment

$f(z) \geq 0$ for all z:

$\rightarrow \begin{bmatrix} P & q \\ q^T & r \end{bmatrix} \succeq 0,\ P \succeq 0$

C we are looking for:

$C = R^m \backslash \mathcal{P},\ \mathcal{P} = \left\{ z | a_i^T z < b_i,\ i = 1, \ldots, k \right\}$

Chebyshev Bound Problem:with First and Second moment

$f(z) \geq 1$ for all $z \in C$:
$a_i^T z \geq b_i \Rightarrow z^T P z + 2q^T z + r \geq 1$
There exist $\tau_1, \ldots, \tau_k \geq 0$ such that

$$\begin{bmatrix} P & q \\ q^T & r-1 \end{bmatrix} \succeq \tau_i \begin{bmatrix} 0 & a_i/2 \\ a_i^T/2 & -b_i \end{bmatrix}, \quad i = 1, \ldots, k$$

참고:Appendix B.2

Chebyshev Bound Problem:with First and second moment

Putting it all together, the Chebyshev bound problem (7.17) can be expressed as

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(\Sigma P) + 2q^T a + r \\ \text{subject to} & \begin{bmatrix} P & q \\ q^T & r-1 \end{bmatrix} \succeq \tau_i \begin{bmatrix} 0 & a_i/2 \\ a_i^T/2 & -b_i \end{bmatrix}, \quad i = 1, \ldots, k \\ & \tau_i \geq 0, \quad i = 1, \ldots, k \\ & \begin{bmatrix} P & q \\ q^T & r \end{bmatrix} \succeq 0, \end{array} \tag{7.19}$$

optimal value α : *upper bound for* P(X$\in$ C)

$1-\alpha$: *lower bound of* $P(X \in \mathcal{P})$

Chernoff Bound

$$P(X \geq u) \leq \inf_{\lambda \geq 0} E(e^{\lambda(X-u)})$$

$$logP(X \geq u) \leq \inf_{\lambda \geq 0} \left\{ -\lambda\mu + logE(e^{\lambda X}) \right\}$$

Terms:
-$logE(e^{\lambda X})$:cumulant generating function(log-mgf)
-ex) $logE(e^{X}) = \lambda^2/2$, when $X \sim N(0,1)$
$\rightarrow P(X \geq u) \leq e^{-u^2/2}$

Chernoff Bound

- $\lambda \in R^m,\ \mu \in R, f : R^m \to R$
- $f(z) = e^{\lambda^T z + \mu}$
- $f(z) \geq 1_C(z)$

$\to P(X \in C) = E(1_C(X)) \leq E(f(X))$
$f(z) \geq 0$ *and* $f(z) \geq 1$ iff $\lambda^T z + \mu \geq 0$

Chernoff Bound

when $-\lambda^T z \leq \mu$:

$$P(X \in C) \leq E(exp(\lambda^T X + \mu)$$
$$logP(X \in C) \leq \mu + logE(exp(\lambda^T X))$$

Chernoff Bound

From this we obtain a general form of Chernoff's bound:

$$\begin{aligned}
\log \mathbf{prob}(X \in C) &\leq \inf\{\mu + \log \mathbf{E} \exp(\lambda^T X) \mid -\lambda^T z \leq \mu \text{ for all } z \in C\} \\
&= \inf_{\lambda} \left(\sup_{z \in C}(-\lambda^T z) + \log \mathbf{E} \exp(\lambda^T X) \right) \\
&= \inf \left(S_C(-\lambda) + \log \mathbf{E} \exp(\lambda^T X) \right),
\end{aligned}$$

where S_C is the support function of C. Note that the second term, $\log \mathbf{E} \exp(\lambda^T X)$, is the cumulant generating function of the distribution, and is always convex (see example 3.41, page 106). Evaluating this bound is, in general, a convex optimization problem.

Chernoff Bound:Gaussian Polyhedron

$$X \sim N(0, I),\ logE(exp(\lambda^T X)) = \lambda^T \lambda/2,\ C = \{x|Ax \preceq b\}$$

$$\begin{aligned}
S_c(y) &= sup\left\{y^T x|Ax \preceq b\right\} \\
&= -inf\left\{-y^T x|Ax \preceq b\right\} \\
&= -sup\left\{-b^T u|A^T u = y,\ u \succeq 0\right\} \\
&= inf\left\{b^T u|A^T u = y,\ u \succeq 0\right\} \\
logP(X \in C) &\leq \underset{\lambda}{inf}(S_C(-\lambda) + logEexp(\lambda^T X)) \\
&= \underset{\lambda}{inf}\inf_u \left\{b^T u + \lambda^T \lambda/2|u \succeq 0,\ A^T u + \lambda = 0\right\}
\end{aligned}$$

Chernoff Bound:Gaussian Polyhedron

This problem has an interesting geometric interpretation. It is equivalent to

$$\begin{array}{ll} \text{minimize} & b^T u + (1/2)\|A^T u\|_2^2 \\ \text{subject to} & u \succeq 0, \end{array}$$

which is the dual of

$$\begin{array}{ll} \text{maximize} & -(1/2)\|x\|_2^2 \\ \text{subject to} & Ax \preceq b. \end{array}$$

In other words, the Chernoff bound is

$$\mathbf{prob}(X \in C) \le \exp(-\mathbf{dist}(0,C)^2/2), \tag{7.22}$$

where $\mathbf{dist}(0, C)$ is the Euclidean distance of the origin to C.

Experiment Design

Consider estimating a vector x in $y_i = a_i^T x + w_i,\ i = 1, \dots, m,\ w_i \sim N(0,1)$

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

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

Note: Think about Regression Analysis:

- $\hat{x} = (A^T A)^{-1} A^T y,\ A = (a_1^T, \dots, a_m^T)$(rows)
- $cov(\hat{\beta}) = (X^T X)^{-1} \sigma^2$

Experiment Design

why care about E?:

$\rightarrow$ *Because of* α confidence level ellipsoid

$$\zeta = \left\{ z \middle| (z - \hat{x})^T E^{-1} (z - \hat{x}) \leq \beta \right\}$$

Experiment Design

Possible choice of a_i: $v_1, \ldots, v_p$
p개의 v중 m개를 뽑은게 a가 된다.(중복 허용하는듯)
m_j: v_j가 몇번 뽑혔는가
$\rightarrow m_1 + \ldots + m_p = m$
Objective: E를 가장 작게하는 choice들을 찾아라.

$$E = \left(\sum_{i=1}^{m} a_i a_i^T\right)^{-1} = \left(\sum_{i=1}^{p} m_i v_i v_i^T\right)^{-1}$$

ex)v_1이 a_1, a_2에서2개 뽑혔다
$\rightarrow m_j = 2$

Experiment Design

$$\begin{aligned} \underset{w.r.t. S_+^n}{\text{minimize}} \quad & E = \left(\sum_{i=1}^{p} m_i v_i v_i^T \right)^{-1} \\ \text{subject to} \quad & m_i \geq 0, m_1 + \ldots + m_p \\ & m_i \in \mathbf{Z} \end{aligned}$$

interpretation:
$E \preceq \tilde{E} \rightarrow$ the first ellipsoid contained in second
first experiment design estimate the variance of $q^T \hat{x}$ better

Relaxed Experiment Design

when m is sufficiently large compared to n,
$\rightarrow$ use relaxed experiment design
$\lambda_i = m_i / m$

$$E = \frac{1}{m}\left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^{-1}$$

Relaxed Experiment:

$$\underset{w.r.t. S_+^n}{\text{minimize}} \quad E = \frac{1}{m}\left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^{-1}$$

$$\text{subject to} \quad \lambda \succeq 0, \mathbf{1}^T \lambda = 1$$

This is a convex optimization since E is anS_+^n convex optimization of λ

Usefulness of Relaxed Experiment Design

앞에 꺼보다 constraint가 하나 적으므로, RED는 ED의 lower bound제공
λ_i가 integer m_i에 비례(1/m)한다는 걸 빼고본다면,

$$mi = \textbf{round}(m\lambda_i), \;\; i = 1, \ldots, p$$
$$\tilde{\lambda}_i = (1/m)\textbf{round}(m\lambda_i)$$

- λ:비례 뺌,$\tilde{\lambda}$:비례 함
- λ_i 와 $\tilde{\lambda}_i$의 차이는 1/m에 비례

→ m이 크면, 비례제약이 없는 λ를 구하고 $\tilde{\lambda}$를 구해도 상관이 없다.
→ 제약 하나를 없앨 수 있어서 relaxed experiment인것.

D-optimal design

The most widely used scalarization is called *D-optimal design*, in which we minimize the determinant of the error covariance matrix E. This corresponds to designing the experiment to minimize the volume of the resulting confidence ellipsoid (for a fixed confidence level). Ignoring the constant factor $1/m$ in E, and taking the logarithm of the objective, we can pose this problem as

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1, \end{array} \tag{7.26}$$

which is a convex optimization problem.

About Elipsoid

A related family of convex sets is the *ellipsoids*, which have the form

$$\mathcal{E} = \{x \mid (x - x_c)^T P^{-1} (x - x_c) \leq 1\}, \tag{2.3}$$

where $P = P^T \succ 0$, *i.e.*, P is symmetric and positive definite. The vector $x_c \in \mathbf{R}^n$ is the *center* of the ellipsoid. The matrix P determines how far the ellipsoid extends in every direction from x_c; the lengths of the semi-axes of $\mathcal{E}$ are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of P. A ball is an ellipsoid with $P = r^2 I$. Figure 2.9 shows an ellipsoid in $\mathbf{R}^2$.

Scalarization:E-optimal design

E-optimal design

In *E-optimal design*, we minimize the norm of the error covariance matrix, *i.e.*, the maximum eigenvalue of E. Since the diameter (twice the longest semi-axis) of the confidence ellipsoid $\mathcal{E}$ is proportional to $\|E\|_2^{1/2}$, minimizing $\|E\|_2$ can be interpreted geometrically as minimizing the diameter of the confidence ellipsoid. E-optimal design can also be interpreted as minimizing the maximum variance of $q^T e$, over all q with $\|q\|_2 = 1$.

The E-optimal experiment design problem is

$$\begin{array}{ll} \text{minimize} & \left\| \left(\sum_{i=1}^p \lambda_i v_i v_i^T \right)^{-1} \right\|_2 \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1. \end{array}$$

The objective is a convex function of λ, so this is a convex problem.

The E-optimal experiment design problem can be cast as an SDP

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \sum_{i=1}^p \lambda_i v_i v_i^T \succeq tI \\ & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1, \end{array} \qquad (7.27)$$

with variables $\lambda \in \mathbf{R}^p$ and $t \in \mathbf{R}$.

Proof to SDP cast

$\sum_{i=1}^{p} \lambda_i v_i v_i^T = A$ and this is sym

use

$\|A\|_2 \le t$ iff $A^T A \preceq t^2 I$ (책 170p)

$\|A^{-1}\|_2 \le t$ iff $(A^{-1})^T A^{-1} \preceq t^2 I$.

$(A^{-1})^T = (A^T)^{-1} = A^{-1}$

$\therefore A^{-2} \preceq t^2 I \Rightarrow \sum_{i=1}^{p} \lambda_i v_i v_i^T \succeq t I$.

So the prob becomes max t for $\sum_{i=1}^{p} \lambda_i v_i v_i^T \succeq t I$.

Scalarization:A-optimal

A-optimal design

In A*-optimal experiment design*, we minimize $\mathbf{tr}\, E$, the trace of the covariance matrix. This objective is simply the mean of the norm of the error squared:

$$\mathbf{E}\, \|e\|_2^2 = \mathbf{E}\, \mathbf{tr}(ee^T) = \mathbf{tr}\, E.$$

The A-optimal experiment design problem is

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\left(\sum_{i=1}^p \lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1. \end{array} \tag{7.28}$$

This, too, is a convex problem. Like the E-optimal experiment design problem, it can be cast as an SDP:

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T u \\ \text{subject to} & \begin{bmatrix} \sum_{i=1}^p \lambda_i v_i v_i^T & e_k \\ e_k^T & u_k \end{bmatrix} \succeq 0, \quad k = 1, \ldots, n \\ & \lambda \succeq 0, \quad \mathbf{1}^T \lambda = 1, \end{array}$$

u_k is the diagonal element of the objective

Proof to SDP

see https://www.tandfonline.com/doi/full/10.1080/03610918.2015.1030414

D-optimal's Geometric Meaning

Optimal experiment design and duality

The Lagrange duals of the three scalarizations have an interesting geometric meaning.

The dual of the D-optimal experiment design problem (7.26) can be expressed as

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

with variable $W \in \mathbf{S}^n$ and domain $\mathbf{S}^n_{++}$ (see exercise 5.10). This dual problem has a simple interpretation: The optimal solution $W^\star$ determines the minimum volume ellipsoid, centered at the origin, given by $\{x \mid x^T W^\star x \leq 1\}$, that contains the points $v_1, \ldots, v_p$. (See also the discussion of problem (5.14) on page 222.) By complementary slackness,

$$\lambda_i^\star(1 - v_i^T W^\star v_i) = 0, \quad i = 1,\ldots,p, \tag{7.29}$$

i.e., the optimal experiment design only uses the experiments v_i which lie on the surface of the minimum volume ellipsoid.

D-optimal's Geometric Meaning

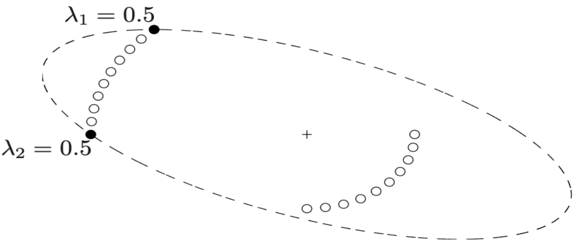


Figure 7.9 Experiment design example. The 20 candidate measurement vectors are indicated with circles. The D-optimal design uses the two measurement vectors indicated with solid circles, and puts an equal weight $\lambda_i = 0.5$ on each of them. The ellipsoid is the minimum volume ellipsoid centered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

Coding

https://jump.dev/JuMP.jl/stable/tutorials/conic/experiment_design/#A-optimal-design