Hypothesis testing, detection, and classification

Yuxin Chen

Princeton University, Fall 2018

Outline

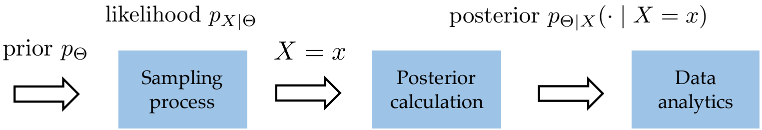
- Bayesian vs. Frequentist statistics
- Bayesian hypothesis testing
- Classical hypothesis testing

Approaches to statistics

Two prominent schools of thought in statistics

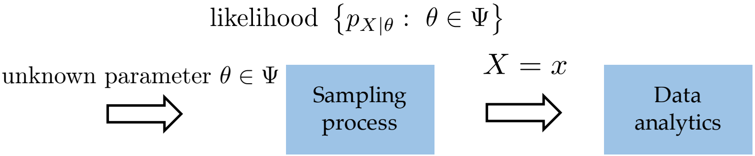
- **Classical (Frequentist):** the unknown variables are treated as deterministic quantities that happen to be unknown
- **Bayesian:** the unknown variables are treated probabilistically with known distributions

Bayesian statistics



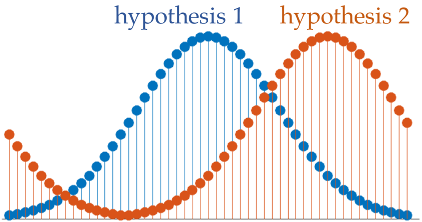
- Let the unknown parameter / quantity Θ be a r.v.
- Postulate a prior distribution $p_\Theta(\theta)$
- Given the observed data x, one can (by Bayes' rule) form the posterior distribution $p_{\Theta|X}(\theta \mid x)$, which captures all information that x can provide about θ
- This means that we only need to deal with *a single* probabilistic model captured by $p_{\Theta|X}(\theta \mid x)$

Frequentist statistics



- The unknown parameter / quantity $\theta \in \Psi$ is a constant
- We are not dealing with a single probabilistic model, but rather with *multiple* candidate probabilistic models $\{p(X \mid \theta) : \theta \in \Psi\}$, one for each possible value of θ

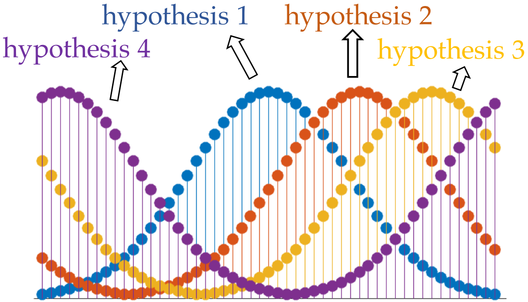
Binary hypothesis testing



We have 2 hypotheses about the unknown quantity, and use the available data to decide which of the two is true

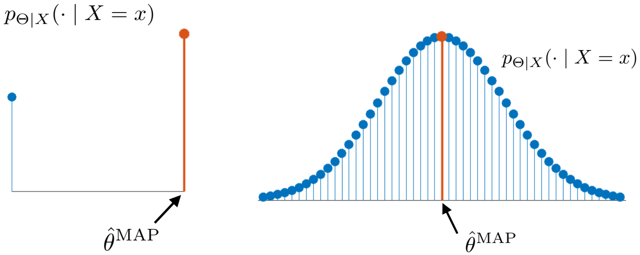
- Example: given a noisy picture, decide whether there is a person in the picture or not
- Example: given a set of trials with two alternative medical treatments, decide which treatment is most effective

Multiple hypothesis testing



We have $m \geq 2$ competing hypotheses about the unknown quantity, and use the available data to decide which of the hypotheses is true

Maximum *a posteriori* (MAP) rule



binary hypothesis test *multiple hypothesis test*

Given the value x of the observation, select a value of θ, denoted $\hat{\theta}$, that maximizes the posterior distribution $p_{\Theta|X}(\theta \mid x)$ (or $f_{\Theta|X}(\theta \mid x)$ if Θ is continuous):

$$\hat{\theta}^{\mathsf{map}} = \begin{cases} \arg\max_{\theta} p_{\Theta|X}(\theta \mid x), & \text{if } \Theta \text{ is discrete} \\ \arg\max_{\theta} f_{\Theta|X}(\theta \mid x), & \text{if } \Theta \text{ is continuous} \end{cases}$$

Maximum *a posteriori* (MAP) rule

Owing to the Bayes' rule, the MAP rule selects $\hat{\theta}^{\mathsf{map}}$ that maximizes over θ:

$$\begin{cases} \frac{p_\Theta(\theta) p_{X|\Theta}(x|\theta)}{p_X(x)}, & \text{if } \Theta \text{ and } X \text{ are discrete} \\ \frac{p_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}, & \text{if } \Theta \text{ is discrete and } X \text{ is continuous} \\ \frac{f_\Theta(\theta) p_{X|\Theta}(x|\theta)}{p_X(x)}, & \text{if } \Theta \text{ is continuous and } X \text{ is discrete} \\ \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}, & \text{if } \Theta \text{ and } X \text{ are continuous} \end{cases} \tag{4.1}$$

Maximum *a posteriori* (MAP) rule

Owing to the Bayes' rule, the MAP rule selects $\hat{\theta}^{\text{map}}$ that maximizes over θ:

$$\begin{cases} \dfrac{p_\Theta(\theta) p_{X|\Theta}(x|\theta)}{\cancel{p_X(x)}}, & \text{if } \Theta \text{ and } X \text{ are discrete} \\ \dfrac{p_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\cancel{f_X(x)}}, & \text{if } \Theta \text{ is discrete and } X \text{ is continuous} \\ \dfrac{f_\Theta(\theta) p_{X|\Theta}(x|\theta)}{\cancel{p_X(x)}}, & \text{if } \Theta \text{ is continuous and } X \text{ is discrete} \\ \dfrac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\cancel{f_X(x)}}, & \text{if } \Theta \text{ and } X \text{ are continuous} \end{cases} \tag{4.1}$$

Maximum likelihood (ML) rule

- Given the value x of the observation, select a value $\hat{\theta}^{\mathsf{ml}}$ of θ that makes the observed data "most likely," i.e.

$$\hat{\theta}^{\mathsf{ml}} = \begin{cases} \arg\max_\theta p_{X|\Theta}(x \mid \theta), & \text{if } X \text{ is discrete} \\ \arg\max_\theta f_{X|\Theta}(x \mid \theta), & \text{if } X \text{ is continuous} \end{cases}$$

- Under *uniform prior* (i.e. if $p_\Theta(\theta)$ is uniform over all possible θ),

$$\text{MAP rule} = \text{ML rule (cf. } (4.1))$$

Optimality of MAP rule

In order to assess the performance of the MAP rule, we need to first determine which performance metric is suitable for our problem

- One natural way is to look at the probability of making an incorrect decision
- We will focus on binary hypothesis tests

Optimality of MAP rule

- Suppose there are two hypotheses $\Theta = \theta_0$ and $\Theta = \theta_1$. The (overall) probability of decision error is defined as

$$
\begin{aligned}
P_e &\stackrel{\triangle}{=} \mathbb{P}\{\hat{\Theta} \neq \Theta\} \\
&= \mathbb{P}\{\Theta = \theta_0,\ \hat{\Theta} = \theta_1\} + \mathbb{P}\{\Theta = \theta_1,\ \hat{\Theta} = \theta_0\} \\
&= \mathbb{P}\{\Theta = \theta_0\}\,\mathbb{P}\{\hat{\Theta} = \theta_1 \mid \Theta = \theta_0\} + \mathbb{P}\{\Theta = \theta_1\}\,\mathbb{P}\{\hat{\Theta} = \theta_0 \mid \Theta = \theta_1\}
\end{aligned}
$$

- We wish to find the decision rule $\hat{\Theta}(Y)$ that minimizes P_e

Theorem 4.1

MAP rule minimizes the (overall) probability of decision error.

Proof of Theorem 4.1

Observe that

$$
\begin{aligned}
\min_{\hat{\Theta}} P_e &= 1 - \max_{\hat{\Theta}} \mathbb{P}\{\hat{\Theta}(X) = \Theta\} \\
&= 1 - \max_{\hat{\Theta}} \int_{-\infty}^{\infty} f_X(x)\, \mathbb{P}\{\Theta = \hat{\Theta}(x) \mid X = x\} \mathrm{d}x \\
&= 1 - \int_{-\infty}^{\infty} f_X(x) \max_{\hat{\Theta}(x)} \mathbb{P}\{\Theta = \hat{\Theta}(x) \mid X = x\} \mathrm{d}x
\end{aligned}
$$

- It suffices to optimize $p_{\Theta|X}(\hat{\Theta}(x) \mid x)$ as this is the only part determined by the rule $\hat{\Theta}$
- The probability of error is minimized if one picks the largest $p_{\Theta|X}(\hat{\Theta}(x) \mid x)$ for every x, which is precisely the MAP rule

Example: biased coins

Suppose we have two biased coins with probability of heads p_1 and p_2, respectively. We choose a coin uniformly at random, and we'd like to infer its identity based on the outcome of a single toss.

- Two hypotheses: $\Theta = 1$: 1st coin was chosen
  $\Theta = 2$: 2nd coin was chosen
- Suppose the outcome was a tail. Which hypothesis does the MAP rule accept?

Example: biased coins

- To determine the MAP rule for this problem, we need to compare $p_\Theta(1)p_{X|\Theta}(x \mid 1)$ and $p_\Theta(2)p_{X|\Theta}(x \mid 2)$.
- Since $p_\Theta(1) = p_\Theta(2)$, it suffices to compare the likelihoods $p_{X|\Theta}(x \mid 1)$ and $p_{X|\Theta}(x \mid 2)$ (i.e. it suffices to look at ML rule)
- We can calculate

$$\mathbb{P}(\text{tail} \mid \Theta = 1) = 1 - p_1 \qquad \text{and} \qquad \mathbb{P}(\text{tail} \mid \Theta = 2) = 1 - p_2$$

- Therefore, MAP (and ML) rule decides in favor of coin 1 if $1 - p_1 > 1 - p_2$ (or equivalently, $p_1 < p_2$)

Example: additive Gaussian noise channel

Consider the additive Gaussian noise channel with signal

$$\Theta = \begin{cases} +\sqrt{P} & \text{with probability } \frac{1}{2} \\ -\sqrt{P} & \text{with probability } \frac{1}{2} \end{cases}$$

noise $Z \sim \mathcal{N}(0, \sigma^2)$ (Θ and Z are independent), and output $X = \Theta + Z$

Example: additive Gaussian noise channel

- The MAP rule gives

$$\hat{\Theta}(x) = \begin{cases} +\sqrt{P} & \text{if } \frac{\mathbb{P}\{\Theta=+\sqrt{P}|X=x\}}{\mathbb{P}\{\Theta=-\sqrt{P}|X=x\}} > 1 \\ -\sqrt{P} & \text{otherwise} \end{cases}$$

Since the two hypotheses are equally likely, the MAP rule reduces to the ML rule

$$\hat{\Theta}(x) = \begin{cases} +\sqrt{P} & \text{if } \frac{f_{X|\Theta}(x\,|+\sqrt{P})}{f_{X|\Theta}(x\,|-\sqrt{P})} > 1 \\ -\sqrt{P} & \text{otherwise} \end{cases}$$

Example: additive Gaussian noise channel

- Using the Gaussian PDF, the ML rule reduces to the minimum distance decoder

$$\hat{\Theta}(x) = \begin{cases} +\sqrt{P}, & \text{if } (x-\sqrt{P})^2 < (x-(-\sqrt{P}))^2 \\ -\sqrt{P}, & \text{otherwise} \end{cases}$$

This simplifies to

$$\hat{\Theta}(x) = \begin{cases} +\sqrt{P}, & x > 0 \\ -\sqrt{P}, & x < 0 \end{cases}$$

  - Remark: the decision when $x = 0$ can be arbitrary
  - Remark: the decision is independent of the noise variance σ^2

Example: additive Gaussian noise channel

The minimum probability of error (or the probability of error under the MAP rule) is given by

$$\begin{aligned}
P_e &= \mathbb{P}\{\hat{\Theta}(X) \neq \Theta\} \\
&= \mathbb{P}\{\Theta = \sqrt{P}\}\,\mathbb{P}\{\hat{\Theta}(X) = -\sqrt{P} \mid \Theta = \sqrt{P}\} + \\
&\qquad\qquad \mathbb{P}\{\Theta = -\sqrt{P}\}\,\mathbb{P}\{\hat{\Theta}(X) = \sqrt{P} \mid \Theta = -\sqrt{P}\} \\
&= \tfrac{1}{2}\,\mathbb{P}\{X \leq 0 \mid \Theta = \sqrt{P}\} + \tfrac{1}{2}\,\mathbb{P}\{X > 0 \mid \Theta = -\sqrt{P}\} \\
&= \tfrac{1}{2}\,\mathbb{P}\{Z/\sigma \leq -\sqrt{P}/\sigma\} + \frac{1}{2}\,\mathbb{P}\{Z/\sigma > \sqrt{P}/\sigma\} \\
&= Q(\sqrt{P/\sigma^2})
\end{aligned}$$

where $Q(x) \overset{\text{def}}{=} \mathbb{P}(\xi \geq x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp(-x^2/2)\mathrm{d}x$ with $\xi \sim \mathcal{N}(0,1)$

Example: additive Gaussian noise channel

$$P_e = Q(\sqrt{P/\sigma^2})$$

The probability of error is a decreasing function of P/σ^2 — the signal-to-noise ratio (SNR)

- Useful fact about the Q function: for all $x > 0$,

$$\frac{1}{x + \frac{1}{x}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \leq Q(x) \leq \frac{1}{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

  - If $x \to \infty$, then $Q(x) \approx \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{x^2}{2}\right)$ (the above bounds are tight for very large x)

- If SNR goes to ∞, then $P_e \approx \frac{\sigma}{\sqrt{2\pi P}} \exp\left(-\frac{P}{2\sigma^2}\right)$
- If SNR goes to 0, then $P_e \approx \mathbb{P}(\xi > 0) = 1/2$

Vector hypothesis testing

- Suppose the hypothesis is concerned with a random vector

$$\boldsymbol{\Theta} = \begin{cases} \boldsymbol{\theta}_0, & \text{with prob. } p_0 \\ \boldsymbol{\theta}_1, & \text{with prob. } p_1 = 1 - p_0 \end{cases}$$

- We observe the random vector $\boldsymbol{Y}$, where

$$\begin{aligned} \boldsymbol{Y} \mid \{\boldsymbol{\Theta} = \boldsymbol{\theta}_0\} &\sim f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y} \mid \boldsymbol{\theta}_0) \\ \boldsymbol{Y} \mid \{\boldsymbol{\Theta} = \boldsymbol{\theta}_1\} &\sim f_{\boldsymbol{Y}|\boldsymbol{\Theta}}(\boldsymbol{y} \mid \boldsymbol{\theta}_1) \end{aligned}$$

- **Goal:** find the decision rule $\hat{\boldsymbol{\Theta}}(\boldsymbol{Y})$ that minimizes the probability of error $\mathbb{P}\{\hat{\boldsymbol{\Theta}} \neq \boldsymbol{\Theta}\}$

Vector hypothesis testing

- The MAP rule is still optimal

$$\hat{\mathbf{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0, & \text{if } \frac{p_{\mathbf{\Theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_0|\boldsymbol{y})}{p_{\mathbf{\Theta}|\boldsymbol{Y}}(\boldsymbol{\theta}_1|\boldsymbol{y})} > 1 \\ \boldsymbol{\theta}_1, & \text{otherwise} \end{cases}$$

- When $p_0 = p_1 = 1/2$, the MAP rule reduces to the ML rule

$$\hat{\mathbf{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0, & \text{if } \frac{f_{\boldsymbol{Y}|\mathbf{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta}_0)}{f_{\boldsymbol{Y}|\mathbf{\Theta}}(\boldsymbol{y}|\boldsymbol{\theta}_1)} > 1 \\ \boldsymbol{\theta}_1, & \text{otherwise} \end{cases}$$

Vector additive Gaussian noise channel

- Consider the vector additive Gaussian noise channel

$$\boldsymbol{Y} = \boldsymbol{\Theta} + \boldsymbol{Z},$$

  where the signal $\boldsymbol{\Theta} \in \mathbb{R}^n$

$$\boldsymbol{\Theta} = \begin{cases} \boldsymbol{\theta}_0, & \text{with prob. } 1/2, \\ \boldsymbol{\theta}_1, & \text{with prob. } 1/2, \end{cases}$$

  and the noise $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$ are independent

- We observe $\boldsymbol{y}$ and wish to find the estimate $\hat{\boldsymbol{\Theta}}(\boldsymbol{Y})$ that minimizes the probability of error $\mathbb{P}\{\hat{\boldsymbol{\Theta}} \neq \boldsymbol{\Theta}\}$

Vector additive Gaussian noise channel

- First assume that $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \sigma^2 \boldsymbol{I}$, i.e. additive white Gaussian noise channel
- The optimal rule is the ML rule. Define the *log-likelihood ratio* as

$$\Lambda(\boldsymbol{y}) \stackrel{\text{def}}{=} \log \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}_0)}{f(\boldsymbol{y} \mid \boldsymbol{\theta}_1)}$$

Then, the ML rule is

$$\hat{\boldsymbol{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0, & \text{if } \Lambda(\boldsymbol{y}) > 0 \\ \boldsymbol{\theta}_1, & \text{otherwise} \end{cases}$$

Vector additive Gaussian noise channel

- Now, the log-likelihood ratio statistic simplifies to

$$\Lambda(\boldsymbol{y}) = \log \frac{\cancel{\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\sigma^2 \boldsymbol{I})}}} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{\theta}_0)^\top(\boldsymbol{y}-\boldsymbol{\theta}_0)}{2\sigma^2}\right)}{\cancel{\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\sigma^2 \boldsymbol{I})}}} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{\theta}_1)^\top(\boldsymbol{y}-\boldsymbol{\theta}_1)}{2\sigma^2}\right)}$$

$$= \frac{1}{2\sigma^2}\left\{(\boldsymbol{y}-\boldsymbol{\theta}_1)^\top(\boldsymbol{y}-\boldsymbol{\theta}_1) - (\boldsymbol{y}-\boldsymbol{\theta}_0)^\top(\boldsymbol{y}-\boldsymbol{\theta}_0)\right\}$$

$$= \frac{1}{2\sigma^2}\left\{\|\boldsymbol{y}-\boldsymbol{\theta}_1\|_2^2 - \|\boldsymbol{y}-\boldsymbol{\theta}_0\|_2^2\right\}$$

- Hence, ML rule reduces to *minimum distance decoder*

$$\hat{\boldsymbol{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0, & \text{if } \|\boldsymbol{y}-\boldsymbol{\theta}_0\|_2 < \|\boldsymbol{y}-\boldsymbol{\theta}_1\|_2 \\ \boldsymbol{\theta}_1, & \text{otherwise} \end{cases}$$

Vector additive Gaussian noise channel

- We can simplify this further to

$$\hat{\mathbf{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0 & \text{if } \boldsymbol{y}^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) < \frac{1}{2}(\boldsymbol{\theta}_1^\top \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_0) \\ \boldsymbol{\theta}_1 & \text{otherwise} \end{cases}$$

- The decision depends only on the value of a scalar

$$W = \boldsymbol{Y}^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)$$

  This is often referred to as a sufficient statistic for the optimal decision rule.

- Further, W is a linear transform of $\boldsymbol{Y}$, and hence

$$\begin{aligned} W \mid \{\mathbf{\Theta} = \boldsymbol{\theta}_0\} &\sim \mathcal{N}\left(\boldsymbol{\theta}_0^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0), \sigma^2(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\right) \\ W \mid \{\mathbf{\Theta} = \boldsymbol{\theta}_1\} &\sim \mathcal{N}\left(\boldsymbol{\theta}_1^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0), \sigma^2(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\right) \end{aligned}$$

Probability of error

Next, we ask: what is the probability of error under this vector Gaussian channel?

- For simplicity, assume two hypotheses have the same power P, i.e. $\boldsymbol{\theta}_0^\top \boldsymbol{\theta}_0 = \boldsymbol{\theta}_1^\top \boldsymbol{\theta}_1 = P$, the MAP rule reduces to

$$\hat{\boldsymbol{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0, & \text{if } W = \boldsymbol{y}^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) < 0 \\ \boldsymbol{\theta}_1, & \text{otherwise,} \end{cases}$$

Probability of error

- Note that

$$W \mid \{\boldsymbol{\Theta} = \boldsymbol{\theta}_0\} \sim \mathcal{N}(\mu_0, V)$$
$$W \mid \{\boldsymbol{\Theta} = \boldsymbol{\theta}_1\} \sim \mathcal{N}(\mu_1, V)$$

where

$$\mu_0 = \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_0, \quad \mu_1 = \boldsymbol{\theta}_1^\top \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_1 = -\mu_0,$$

$$V = \sigma^2 (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) = 2\sigma^2 (P - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_1)$$

- Thus, this is equivalent to the following scalar channel

$$W = \begin{cases} \mu_0 + \xi, & \text{if } \boldsymbol{\Theta} = \boldsymbol{\theta}_0 \\ -\mu_0 + \xi, & \text{if } \boldsymbol{\Theta} = \boldsymbol{\theta}_1 \end{cases}$$

where $\xi \sim \mathcal{N}(0, V)$

Probability of error

- Invoking our results for the scale channel, we know

$$
\begin{aligned}
P_{\mathrm{e}} = Q(\sqrt{\mathsf{SNR}}) &= Q\left(\sqrt{\mu_0^2 \,/\, V}\right) \\
&= Q\left(\sqrt{\frac{\left(P - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_1\right)^2}{2\sigma^2 (P - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_1)}}\right) \\
&= Q\left(\sqrt{\frac{P - \boldsymbol{\theta}_0^\top \boldsymbol{\theta}_1}{2\sigma^2}}\right)
\end{aligned}
$$

- This is minimized by using antipodal signals $\boldsymbol{\theta}_0 = -\boldsymbol{\theta}_1$, which yields

$$
P_e = Q\left(\sqrt{\frac{P}{\sigma^2}}\right)
$$

Vector Gaussian channel with colored noise

- Now suppose that the noise is not white, i.e., $\mathbf{\Sigma}_{\boldsymbol{Z}}$. Then the ML rule reduces to

$$\hat{\mathbf{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0, & \text{if } (\boldsymbol{y}-\boldsymbol{\theta}_0)^\top \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}(\boldsymbol{y}-\boldsymbol{\theta}_0) < (\boldsymbol{y}-\boldsymbol{\theta}_1)^\top \mathbf{\Sigma}_{\boldsymbol{Z}}^{-1}(\boldsymbol{y}-\boldsymbol{\theta}_1) \\ \boldsymbol{\theta}_1, & \text{otherwise} \end{cases}$$

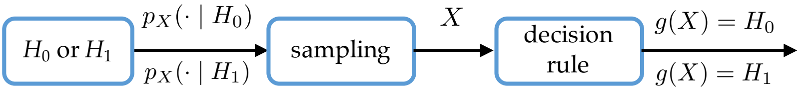
Vector Gaussian channel with colored noise

- Letting $\boldsymbol{y}' = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}\boldsymbol{y}$ and $\boldsymbol{\theta}_i' = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}\boldsymbol{\theta}_i$ for $i = 0, 1$, the rule becomes the same as that for the white noise case

$$\hat{\boldsymbol{\Theta}}(\boldsymbol{y}) = \begin{cases} \boldsymbol{\theta}_0 & \text{if } \|\boldsymbol{y}' - \boldsymbol{\theta}_0'\|_2 < \|\boldsymbol{y}' - \boldsymbol{\theta}_1'\|_2 \\ \boldsymbol{\theta}_1 & \text{otherwise} \end{cases}$$

- Thus, the optimal rule is to first multiply $\boldsymbol{Y}$ by $\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}$ to obtain $\boldsymbol{Y}'$ and then to apply the optimal rule for the white noise case with the transformed signals $\boldsymbol{\theta}_i' = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}\boldsymbol{\theta}_i$ $(i = 0, 1)$

Classical hypothesis testing



Let's revisit binary hypothesis testing. In conventional statistics language, the two hypotheses are called

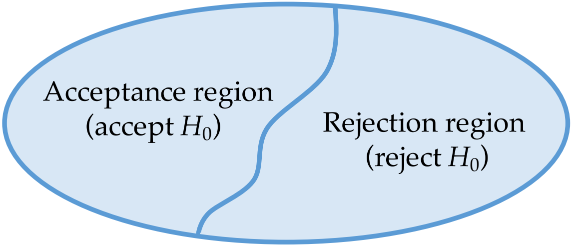
H_0 : null hypothesis

H_1 : alternative hypothesis

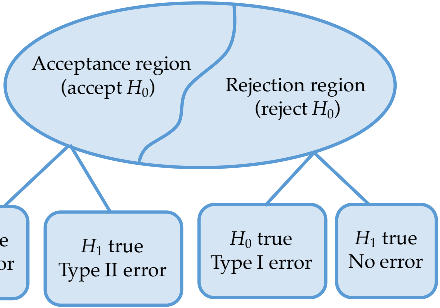
Decision rule

Any decision rule $g(X)$ represents a partition of sample space into two subsets

- Rejection region (H_0 is rejected)
- Acceptance region (H_0 is accepted)



Type I and Type II errors

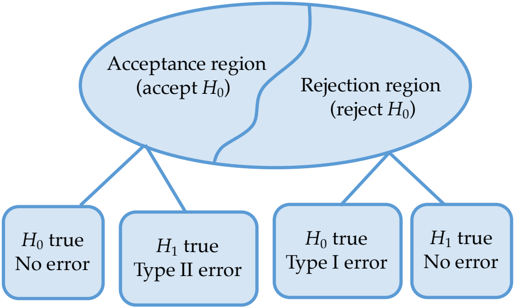


There are two types of decision errors:

Type I error reject H_0 even though H_0 is true

Type II error accept H_0 even though H_0 is false

Error probabilities



Type I error $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$

Type II error $\beta = \mathbb{P}(\text{accept } H_0 \mid H_0 \text{ is false})$

MAP rule revisited

Recall that the MAP rule is

$$P_\Theta(\theta_0)P_{X|\Theta}(x \mid \theta_0) \underset{H_1}{\overset{H_0}{\gtrless}} P_\Theta(\theta_1)P_{X|\Theta}(x \mid \theta_1)$$

$$\iff \qquad \underbrace{\frac{P_{X|\Theta}(x \mid \theta_1)}{P_{X|\Theta}(x \mid \theta_0)}}_{\text{likelihood ratio } L(x)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_\Theta(\theta_0)}{P_\Theta(\theta_1)}$$

The decision rule is based on the likelihood ratio statistics, with the critical value $\xi = \frac{P_\Theta(\theta_0)}{P_\Theta(\theta_1)}$ determined by the prior distribution

Likelihood ratio test (LRT)

$$L(x) = \frac{P_{X|\Theta}(x \mid \theta_1)}{P_{X|\Theta}(x \mid \theta_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \xi$$

for some threshold ξ

Probabilities of Type I and Type II errors can be calculated as functions of ξ:

$$\alpha(\xi) \qquad \text{and} \qquad \beta(\xi),$$

where choosing ξ trades off these two error probabilities

Example

Suppose the likelihood under two hypotheses are

$$\begin{aligned} H_0 \quad &\rightarrow \quad \mathcal{N}(0,1) \\ H_1 \quad &\rightarrow \quad \mathcal{N}(1,1) \end{aligned}$$

Then the likelihood ratio statistic is

$$L(x) = \frac{f_X(x \mid H_1)}{f_X(x \mid H_0)} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} = \exp\left(x - \frac{1}{2}\right)$$

So the LRT is

$$L(x) \underset{H_0}{\overset{H_1}{\gtrless}} \xi \qquad \Longleftrightarrow \qquad x \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} + \log \xi$$

Optimality of LRT

Encouragingly, LRT is optimal in the following sense:

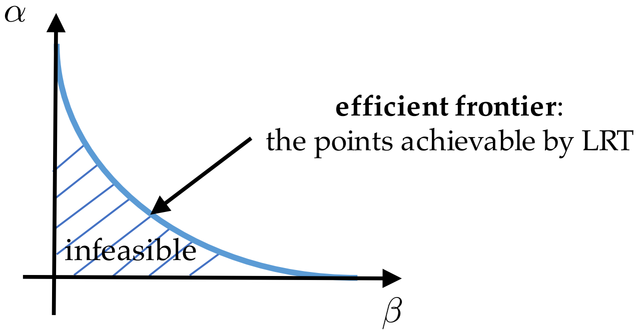
For a given Type-I error (i.e. α), LRT achieves the smallest possible Type II error (i.e. β)

Theorem 4.2 (Neyman-Pearson)

For any threshold ξ of LRT, suppose the resulting Type I and Type II errors are α and β, respectively. Then for any other test whose Type I error is smaller than α, its Type II error must exceed β

Optimality of LRT

Efficient frontier: the set of error probability pairs (α, β) such that we cannot simultaneously improve α and β.



Neyman-Pearson says that the probabilities of errors of LRTs lie on the efficient frontier

Proof of Neyman-Pearson

Conisder a hypothetical Bayesian hypothesis test problem in which the decision boundary obeys

$$\frac{P_{X|\Theta}(x \mid \theta_1)}{P_{X|\Theta}(x \mid \theta_0)} = \xi = \frac{P_\Theta(\theta_0)}{P_\Theta(\theta_1)}$$

$$\iff \qquad P_\Theta(\theta_0) = \frac{\xi}{1+\xi}, \qquad P_\Theta(\theta_1) = \frac{1}{1+\xi}$$

Clearly, the MAP rule is the LRT with threshold ξ, namely,

$$L(x) = \frac{P_{X|\Theta}(x \mid \theta_1)}{P_{X|\Theta}(x \mid \theta_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \xi$$

Proof of Neyman-Pearson

The Bayesian probability of error is

$$P_{\mathrm{e},\mathsf{MAP}} = \mathbb{P}_\Theta(\theta_0)\alpha + \mathbb{P}_\Theta(\theta_1)\beta$$

Since $P_{\mathrm{e},\mathsf{MAP}}$ is Bayesian-optimal, we cannot simultaneously improve α and β (otherwise we get a test that achieves strictly lower Bayesian probability of error than the MAP rule, which is impossible). This concludes the proof.

Reference

[1] "*Lecture notes for Statistical Signal Processing*," A. El Gamal.

[2] "*Introduction to probability (2nd Edition)*," D. Bertsekas, J. Tsitsiklis, *Athena Scientific*, 2008.

[3] "*Probability and Computing (2nd Edition)*," M. Mitzenmacher, E. Upfal, *Cambridge University Press*, 2017.

[4] "*Statistical Inference (2nd Edition)*," G. Casella, R. Berger, *Cengage*, 2002.