EE/ACM 150 - Applications of Convex Optimization in Signal Processing and Communications Lecture 15

Andre Tkacenko

Signal Processing Research Group
Jet Propulsion Laboratory

May 22, 2012

Caltech

Outline

Maximum Likelihood and the Log-Likelihood Function

Let $\mathbf{y} \in \mathbb{R}^m$ be a random variable whose distribution depends on a parameter $\boldsymbol{x} \in \mathbb{R}^n$. We will denote this distribution by $p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y})$ and refer to it as the *likelihood function*.

Example: If $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the distribution of $\mathbf{v}$ depends on $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\Sigma} \in \mathbb{S}_+^m$.

As many common distributions are log-concave, and often independent observations y_i will be made to form the vector observation $\mathbf{y}$, it is convenient to work with the logarithm of the distribution $p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y})$ called the *log-likelihood function*, given by

$$\ell(\boldsymbol{x}) \triangleq \log p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y}) \ .$$

In *maximum likelihood (ML) estimation*, the parameter $\boldsymbol{x}$ is determined as the argument which maximizes the likelihood (or log-likelihood) function.

Maximum Likelihood (ML) Estimate:

$$\widehat{\boldsymbol{x}}_{\mathrm{ml}} = \operatorname*{argmax}_{\boldsymbol{x}} p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y}) = \operatorname*{argmax}_{\boldsymbol{x}} \ell(\boldsymbol{x}) \ .$$

- Here, $\boldsymbol{y}$ is the observed value.
- We can add constraints that $\boldsymbol{x} \in \mathcal{C}$ explicitly, or define $p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y}) = 0$ for $\boldsymbol{x} \notin \mathcal{C}$.
- Determining the ML estimate is a convex optimization problem if $\log p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y})$ is concave in $\boldsymbol{x}$ for fixed $\boldsymbol{y}$.

Linear Measurements with i.i.d. Noise

In many cases, observations are made following a *linear measurement model* given by

$$y_i = \mathbf{a}_i^T \boldsymbol{x} + v_i \,, \; i = 1, \ldots, m \,.$$

- $\boldsymbol{x} \in \mathbb{R}^n$ is the vector of unknown parameters.
- v_i are *independent and identically distributed (i.i.d.)* measurement noise with density $p_{\mathrm{v}}(v)$.
- y_i is the measurement: $\mathbf{y} \in \mathbb{R}^m$ has density $p_{\mathbf{y}|\boldsymbol{x}}(\boldsymbol{y}) = \prod_{i=1}^m p_{\mathrm{v}}\left(y_i - \mathbf{a}_i^T \boldsymbol{x}\right)$.

ML estimation: An ML estimate $\widehat{\boldsymbol{x}}_{\mathrm{ml}}$ is any solution $\boldsymbol{x}$ of

$$\text{maximize} \quad \ell(\boldsymbol{x}) = \textstyle\sum_{i=1}^m \log p_{\mathrm{v}}\left(y_i - \mathbf{a}_i^T \boldsymbol{x}\right) \quad ,$$

with observations y_i for $i = 1, \ldots, m$.

Examples:

- Gaussian noise: If $v_i \sim \mathcal{N}\left(0, \sigma^2\right)$, then $p_{\mathrm{v}}(v) = \left(2\pi\sigma^2\right)^{-1/2} e^{-v^2/\left(2\sigma^2\right)}$ and so,

  $$\ell(\boldsymbol{x}) = -\left(m/2\right) \log\left(2\pi\sigma^2\right) - \left(1/\left(2\sigma^2\right)\right) ||\mathbf{A}\boldsymbol{x} - \boldsymbol{y}||_2^2 \,.$$

  Thus, the ML estimate is the least-squares (LS) solution.
- Laplacian noise: If $p_{\mathrm{v}}(v) = \left(1/\left(2a\right)\right) e^{-|v|/a}$ for some $a > 0$, then we have,

  $$\ell(\boldsymbol{x}) = -m \log(2a) - (1/a) \, ||\mathbf{A}\boldsymbol{x} - \boldsymbol{y}||_1 \,.$$

  Hence, the ML estimate is the ℓ_1-norm solution.
- Uniform noise: If $p_{\mathrm{v}}(v) = 1/\left(2a\right)$ for $v \in [-a, a]$ and $p_{\mathrm{v}}(v) = 0$ otherwise, then we have

  $$\ell(\boldsymbol{x}) = \begin{cases} -m \log(2a) \,, & ||\mathbf{A}\boldsymbol{x} - \boldsymbol{y}||_\infty \leq a \\ -\infty \,, & \text{otherwise} \end{cases} \,.$$

  So, an ML estimate is any $\boldsymbol{x}$ with $||\mathbf{A}\boldsymbol{x} - \boldsymbol{y}||_\infty \leq a$.

ML Interpretation of Penalty Function Approximation

We can interpret any penalty function approximation problem of the form

$$\text{minimize} \quad \sum_{i=1}^{m} \phi\left(b_i - \mathbf{a}_i^T \mathbf{x}\right) \quad ,$$

as an ML estimation problem with a linear measurement model with i.i.d. noise. Specifically, these two problems are identical when the noise density $p_{\mathrm{v}}(v)$ satisfies

$$p_{\mathrm{v}}(v) = \frac{e^{-\phi(v)}}{\int e^{-\phi(u)}\, du} \iff \phi(v) = -C \log p_{\mathrm{v}}(v) \ ,$$

and the observation is $\boldsymbol{y} = \mathbf{b}$.

Intepretations:

- Penalty functions which strongly (weakly) penalize large residuals will lead to noise densities with small (large) tails.
- Laplacian noise density has much larger tails than Gaussian noise density. Hence, the Laplacian noise based ML estimation method will not penalize large residuals as severely as that corresponding to the Gaussian noise density.
- Residuals corresponding to ML estimation with a Gaussian density will tend to be small, whereas those for a Laplacian density will sparse (many residuals near zero and a few large outliers).

Counting Problems with Poisson Distribution

In many cases, the random variable y has *Poisson distribution* with mean $\mu > 0$:

$$\Pr\{\mathrm{y} = k\} = \frac{e^{-\mu}\mu^k}{k!}\,.$$

Often, y will represent the count or number of events of a Poisson process over some period of time (i.e., photon arrival counts, traffic accidents, etc.). In a simple statistical model, μ is modeled as an affine function of a vector $\boldsymbol{u} \in \mathbb{R}^n$:

$$\mu = \boldsymbol{a}^T\boldsymbol{u} + b\,.$$

- $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are *model parameters*, whereas $\boldsymbol{u} \in \mathbb{R}^n$ is an observed vector of *explanatory variables*.
- The goal is to estimate $\boldsymbol{a}$ and b from m observations (u_i, y_i) for $i = 1, \ldots, m$.

From this, the log-likelihood function is given by the following.

$$\ell(\boldsymbol{a}, b) = \sum_{i=1}^{m}\left(y_i \log\left(\boldsymbol{a}^T u_i + b\right) - \left(\boldsymbol{a}^T u_i + b\right) - \log(y_i!)\right)\,.$$

We can find an ML estimate of $\boldsymbol{a}$ and b by solving the convex optimization problem

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

with variables $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

Logistic Regression

Consider a random variable $\mathrm{y} \in \{0, 1\}$ with

$$\Pr\{\mathrm{y} = 1\} = p = \frac{\exp\left(\boldsymbol{a}^T \boldsymbol{u} + b\right)}{1 + \exp(\boldsymbol{a}^T \boldsymbol{u} + b)}\,, \quad \Pr\{\mathrm{y} = 0\} = 1 - p = \frac{1}{1 + \exp(\boldsymbol{a}^T \boldsymbol{u} + b)}\,.$$

- $\boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R}$ are model parameters; $\boldsymbol{u} \in \mathbb{R}^n$ is an observed explanatory variable vector.
- The goal is to estimate $\boldsymbol{a}$ and b from m observations (u_i, y_i) for $i = 1, \ldots, m$.

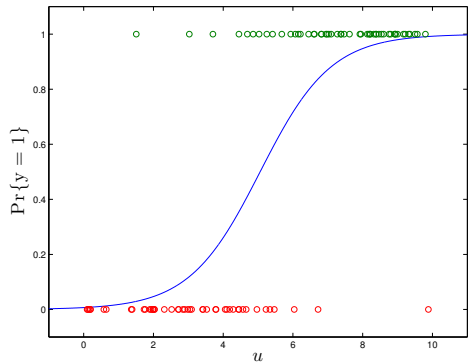
Log-likelihood function: Assuming the data has been ordered such that $y_1 = \cdots = y_k = 1$ and $y_{k+1} = \cdots = y_m = 0$, we get the following.

$$\begin{aligned} \ell(\boldsymbol{a}, b) &= \log\left(\prod_{i=1}^{k} \frac{\exp\left(\boldsymbol{a}^T u_i + b\right)}{1 + \exp(\boldsymbol{a}^T u_i + b)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(\boldsymbol{a}^T u_i + b)}\right)\,, \\ &= \sum_{i=1}^{k} \left(\boldsymbol{a}^T u_i + b\right) - \sum_{i=1}^{m} \log\left(1 + \exp\left(\boldsymbol{a}^T u_i + b\right)\right)\,, \end{aligned}$$

which is concave in $\boldsymbol{a}$ and b.

Example: ($n = 1$, $m = 50$ measurements)

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



Covariance Estimate for Gaussian Random Variables

Suppose $\mathbf{y} \in \mathbb{R}^n$ is a Gaussian random vector with zero mean and covariance matrix $\boldsymbol{R} = E\left[\mathbf{y}\mathbf{y}^T\right]$. In many cases, it may be imperative to estimate $\boldsymbol{R}$ based on independent observations of $\mathbf{y}$. Note the here, the density of $\mathbf{y}$ is given by

$$p_{\mathbf{y}|\boldsymbol{R}}(\boldsymbol{y}) = (2\pi)^{-n/2} \left(\det(\boldsymbol{R})\right)^{-1/2} \exp\left(-\boldsymbol{y}^T \boldsymbol{R}^{-1} \boldsymbol{y}/2\right) .$$

Based on N independent observations $\boldsymbol{y}_1, \ldots \boldsymbol{y}_N \in \mathbb{R}^n$, we would like to find an ML estimate of $\boldsymbol{R}$. The log-likelihood function is given by the following.

$$\ell(\boldsymbol{R}) = -\frac{Nn}{2}\log(2\pi) - \frac{N}{2}\log\det\boldsymbol{R} - \frac{N}{2}\mathrm{tr}\left(\boldsymbol{R}^{-1}\boldsymbol{Y}\right) \text{ , where } \boldsymbol{Y} \triangleq \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{y}_k \boldsymbol{y}_k^T .$$

Here, $\boldsymbol{Y}$ is the *sample covariance matrix* for the observations $\boldsymbol{y}_1, \ldots \boldsymbol{y}_N \in \mathbb{R}^n$. The log-likelihood function is *not* a concave function of $\boldsymbol{R}$, however it *is* a concave function of $\boldsymbol{S} \triangleq \boldsymbol{R}^{-1}$, which is called the *information matrix*. In this case, the new log-likelihood function $\ell(\boldsymbol{S})$ has the form

$$\ell(\boldsymbol{S}) = -\left(Nn/2\right)\log(2\pi) + \left(N/2\right)\log\det\boldsymbol{S} - \left(N/2\right)\mathrm{tr}(\boldsymbol{S}\boldsymbol{Y}) ,$$

and so, the ML estimate of $\boldsymbol{S}$ (and hence $\boldsymbol{R}$) can be found by solving the problem

$$\begin{array}{ll} \text{maximize} & \log\det\boldsymbol{S} - \mathrm{tr}(\boldsymbol{S}\boldsymbol{Y}) \\ \text{subject to} & \boldsymbol{S} \in \mathcal{S} \end{array} ,$$

where $\mathcal{S}$ represents our prior knowledge of $\boldsymbol{S} = \boldsymbol{R}^{-1}$.

Bayesian Version of ML Estimation

In *maximum a posteriori probability (MAP) estimation*, which can be considered as a Bayesian version of ML estimation, we assume that the underlying parameter $\boldsymbol{x}$ to be estimated is random with some *a priori* or prior probability density $p_{\mathbf{x}}(\boldsymbol{x})$.

If $\mathbf{x}$ (the vector to be estimated) and $\mathbf{y}$ (the observation) are random variables with joint probability density function (pdf) $p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$, then we have the following.

$$p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})\, p_{\mathbf{y}}(\boldsymbol{y}) = p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{x},\boldsymbol{y})\, p_{\mathbf{x}}(\boldsymbol{x}) \ ,$$

where, for instance, $p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$ is the conditional density of $\mathbf{x}$ given $\mathbf{y}$. Specifically, for an observed value $\mathbf{y} = \boldsymbol{y}$, $p_{\mathbf{x}|\mathbf{y}}$ is the *a posteriori* or posterior density of $\mathbf{x}$. For MAP estimation, we seek the vector $\boldsymbol{x}$ which maximizes this quantity.

MAP Estimate:

$$\widehat{\boldsymbol{x}}_{\text{map}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{x},\boldsymbol{y})\, p_{\mathbf{x}}(\boldsymbol{x}) = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) \ .$$

Taking logarithms, we can express the MAP estimate as

$$\widehat{\boldsymbol{x}}_{\text{map}} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, \left(\log p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{x},\boldsymbol{y}) + \log p_{\mathbf{x}}(\boldsymbol{x})\right) \ .$$

- From the last expression for the MAP estimate, we see that the second term penalizes unlikely choices of $\boldsymbol{x}$, according to the prior density $p_{\mathbf{x}}(\boldsymbol{x})$.
- If the log-likelihood function is concave and the prior density for $\mathbf{x}$ is log-concave, the resulting MAP estimation problem will be convex.

Examples of MAP Estimation Problems

Linear measurements with i.i.d. noise: For the linear measurement model

$$y_i = \mathbf{a}_i^T \boldsymbol{x} + v_i \,, \;\; i = 1, \ldots, m \,,$$

where the v_i are i.i.d. with density p_{v}, the MAP estimation problem becomes

$$\text{maximize} \quad \textstyle\sum_{i=1}^m \log p_{\mathrm{v}}\left(y_i - \mathbf{a}_i^T \boldsymbol{x}\right) + \log p_{\mathbf{x}}(\boldsymbol{x}) \quad .$$

If p_{v} and $p_{\mathbf{x}}$ are log-concave, this problem is convex.

MAP estimation with perfect linear measurements: Suppose we have m deterministic linear measurements $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$. In this case, the MAP estimation problem becoems the following.

$$\begin{array}{ll} \text{maximize} & \log p_{\mathbf{x}}(\boldsymbol{x}) \\ \text{subject to} & \mathbf{A}\boldsymbol{x} = \boldsymbol{y} \end{array} \quad .$$

If $p_{\mathbf{x}}$ is log-concave, this is a convex problem.
Assuming the parameters x_i are i.i.d. with density $p_{\mathrm{x}}(x)$, the MAP estimation problem becomes

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \log p_{\mathrm{x}}(x_i) \\ \text{subject to} & \mathbf{A}\boldsymbol{x} = \boldsymbol{y} \end{array} \iff \begin{array}{ll} \text{maximize} & \sum_{i=1}^n \phi(x_i) \\ \text{subject to} & \mathbf{A}\boldsymbol{x} = \boldsymbol{b} \end{array} \quad ,$$

where $\phi(u) = -\log p_{\mathrm{x}}(u)$ and $\mathbf{b} = \boldsymbol{y}$. In other words, the MAP estimation problem can be expressed as a least-penalty problem. Conversely, any least-penalty problem can be expressed as a MAP estimation problem with m perfect measurements $\mathbf{b} = \mathbf{A}\boldsymbol{x}$ and x_i i.i.d. with density

$$p_{\mathrm{x}}(x) = \frac{e^{-\phi(x)}}{\int e^{-\phi(u)}, du} \,.$$

Introduction to Nonparametric Distribution Estimation

In nonparametric distribution estimation, we are interested in estimating the pdf of a continuous random variable or the probability mass function (pmf) of a discrete random variable.

Here, we will consider a discrete random variable X with values in the finite set $\{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbb{R}$.

- The distribution of X is characterized by $\mathbf{p} \in \mathbb{R}^n$, where $p_k = \mathrm{Pr}\{\mathrm{X} = \alpha_k\}$, and satisfies
$$\mathbf{p} \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{p} = 1\,.$$
- Conversely, if $\mathbf{p} \in \mathbb{R}^n$ satisfies $\mathbf{p} \succeq \mathbf{0}$ and $\mathbf{1}^T\mathbf{p} = 1$, then it defines a probability distribution for a random variable X, for which $\mathrm{Pr}\{\mathrm{X} = \alpha_k\} = p_k$.
- Therefore, the probability simplex
$$\left\{\mathbf{p} : \mathbf{p} \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{p} = 1\right\}\,,$$
is in one-to-one correspondence with all possible probability distributions for a random variable X taking values in $\{\alpha_1, \ldots, \alpha_n\}$.

Given a combination of prior information and, possibly, observations and measurements, we can often estimate the distribution $\mathbf{p}$ using convex optimization.

Examples of Nonparametric Distribution Estimation

Prior information:

- *Expected values:* As $E[f(\mathrm{X})] = \sum_{i=1}^n p_i f(\alpha_i) = \mathbf{c}^T\mathbf{p}$, where $c_i = f(\alpha_i)$ for $i = 1, \ldots, n$, we can incorporate expected values of functions as either objectives or constraints.
- *Nonlinear functions of expected values:* In some cases, nonlinear functions of expected values can be used to create convex inequality constraints in $\mathbf{p}$. For example,

$$\mathrm{var}(\mathrm{X}) = E\left[\mathrm{X}^2\right] - (E[\mathrm{X}])^2 = -\mathbf{p}^T\mathbf{a}\mathbf{a}^T\mathbf{p} + \mathbf{b}^T\mathbf{p}\,, \text{ where } a_i = \alpha_i^2\,,\ b_i = \alpha_i\,,$$

is concave in $\mathbf{p}$, and so constraints of the form $\mathrm{var}(\mathrm{X}) \geq \beta$ are convex. Similarly,

$$\Pr\{\mathrm{X} \in \mathcal{A} | \mathrm{X} \in \mathcal{B}\} = \Pr\{\mathrm{X} \in \mathcal{A} \cap \mathcal{B}\} \,/\, \Pr\{\mathrm{X} \in \mathcal{B}\} = \mathbf{c}^T\mathbf{p}/\mathbf{d}^T\mathbf{p}\,,$$

where $c_i = 1$ if $\alpha_i \in \mathcal{A} \cap \mathcal{B}$ and 0 otherwise and $d_i = 1$ if $\alpha_i \in \mathcal{B}$ and 0 otherwise, is linear-fractional in $\mathbf{p}$, and so constraints of the form, $\ell \leq \Pr\{\mathrm{X} \in \mathcal{A} | \mathrm{X} \in \mathcal{B}\} \leq u$, are linear inequality constraints, $\ell\mathbf{d}^T\mathbf{p} \leq \mathbf{c}^T\mathbf{p} \leq u\mathbf{d}^T\mathbf{p}$, and hence convex.

Bounding probabilities and expected values: Given prior information $\mathbf{p} \in \mathcal{P}$, where $\mathcal{P}$ is convex, upper and lower bounds on $E[f(\mathrm{X})]$ can be found by solving the convex problems,

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

ML estimation: Suppose we observe N independent samples $x_1, \ldots, x_N$ from the distribution for X. Let k_i denote the number of samples with value α_i, so that $k_1 + \cdots + k_n = N$. The log-likelihood function is then $\ell(\mathbf{p}) = \sum_{i=1}^n k_i \log p_i$, which is concave in $\mathbf{p}$. Thus, the ML estimate of $\mathbf{p}$ can be found by solving the convex problem

$$\begin{array}{ll} \text{maximize} & \ell(\mathbf{p}) = \sum_{i=1}^n k_i \log p_i \\ \text{subject to} & \mathbf{p} \in \mathcal{P} \end{array}\,.$$

Minimum Kullback-Leibler Divergence Problem

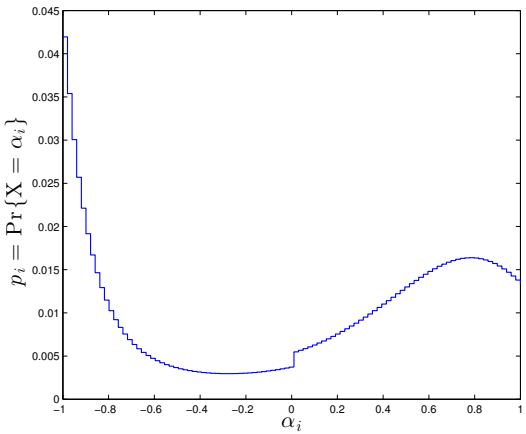
The problem of finding a distribution $\mathbf{p} \in \mathbb{R}^n$ that has the minimum *Kullback-Leiber divergence* from a given prior distribution $\mathbf{q} \in \mathbb{R}^n$, with the prior information that $\mathbf{p} \in \mathcal{P}$, leads to the problem

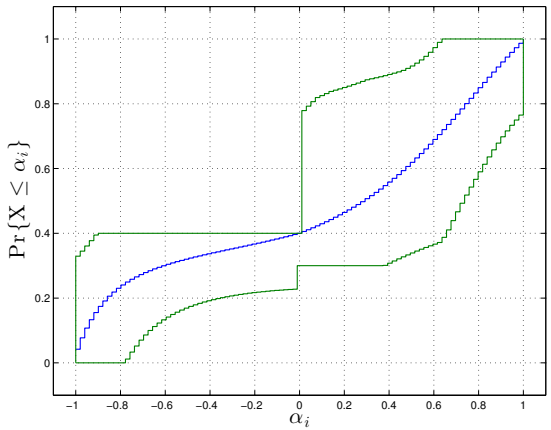
$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n p_i \log(p_i/q_i) \\ \text{subject to} & \mathbf{p} \in \mathcal{P} \end{array} .$$

When the set $\mathcal{P}$ is convex, the problem is convex. In the special case where $\mathbf{q}$ is the uniform distribution, i.e., $\mathbf{q} = (1/n)\,\mathbf{1}$, the resulting solution $\mathbf{p}^\star$ is called the *maximum entropy distribution*.

Example: Consider 100 equidistant points α_i in $[-1, 1]$ with prior information:

$$E[\mathrm{X}] \in [-0.1, 0.1] \ , \ E\left[\mathrm{X}^2\right] \in [0.5, 0.6] \ , \ E\left[3\mathrm{X}^3 - 2\mathrm{X}\right] \in [-0.3, -0.2] \ , \ \Pr\{\mathrm{X} < 0\} \in [0.3, 0.4] \ .$$





- Left: maximum entropy distribution that satisfies the prior information constraints.
- Right: bottom and top curves show minimum and maximum possible values of the cumulative distribution function (cdf) $\Pr\{\mathrm{X} \leq \alpha_i\}$, while the middle curve is the cdf of the maximum entropy distribution.

Hypothesis Testing and Types of Detectors

Hypothesis Testing:

Suppose $\mathbf{x}$ is a random variable with n values $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with a distribution that depends on a parameter $\mathbf{v}$ with m values $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$. This is characterized by a matrix $\mathbf{P} \in \mathbb{R}^{n \times m}$ with

$$P_{k,\ell} = \Pr\{\mathbf{x} = \boldsymbol{x}_k | \mathbf{v} = \boldsymbol{v}_\ell\} \ , \ k = 1, \ldots, n \, , \ \ell = 1, \ldots, m \, .$$

- The values $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ represent m *hypotheses*.
- Based on an observation $\boldsymbol{x}$ of $\mathbf{x}$, we wish to guess the correct value $\boldsymbol{v}_\ell$, a process called *hypothesis testing*.

Types of Detectors:

There are two main types of detectors, namely *deterministic* and *random* detectors.

- For deterministic detectors, if we observe $\boldsymbol{x}_k$, we have $\widehat{\boldsymbol{v}} = \boldsymbol{\psi}(k)$. One obvious choice for $\boldsymbol{\psi}(k)$ is the ML detector, given by

$$\widehat{\boldsymbol{v}} = \boldsymbol{\psi}_{\rm ml}(k) = \boldsymbol{v}_{\ell^\star(k)} \text{ where } \ell^\star(k) = \operatorname{argmax}_\ell P_{k,\ell} \, .$$

- For randomized detectors, the estimate $\widehat{\mathbf{v}}$ is random with a distribution that depends on the observed value of $\mathbf{x}$. It is characterized by a matrix $\mathbf{T} \in \mathbb{R}^{m \times n}$:

$$T_{i,k} = \Pr\{\widehat{\mathbf{v}} = \boldsymbol{v}_i | \mathbf{x} = \boldsymbol{x}_k\} \ , \ i = 1, \ldots, m \, , \ k = 1, \ldots, n \, .$$

  The columns $\mathbf{t}_k$ of $\mathbf{T}$ must satisfy the probability constraints $\mathbf{t}_k \succeq \mathbf{0}$ and $\mathbf{1}^T \mathbf{t}_k = 1$.

Detection Probability Matrix

For the randomized detector characterized by the matrix $\mathbf{T}$, we define the *detection probability matrix* as $\mathbf{D} \triangleq \mathbf{TP} \in \mathbb{R}^{m \times m}$. We have

$$D_{i,\ell} = [\mathbf{TP}]_{i,\ell} = \Pr\{\widehat{\mathbf{v}} = \boldsymbol{v}_i | \mathbf{v} = \boldsymbol{v}_\ell\} \ , \ i = 1, \ldots, m \, , \ \ell = 1, \ldots, m \, .$$

Hence, $D_{i,\ell}$ is the probability of guessing $\widehat{\mathbf{v}} = \boldsymbol{v}_i$, when in fact $\mathbf{v} = \boldsymbol{v}_\ell$ was the true hypothesis.

- *Correct detection* or *detection probabilities:*

$$P_i^{\mathrm{d}} = \Pr\{\widehat{\mathbf{v}} = \boldsymbol{v}_i | \mathbf{v} = \boldsymbol{v}_i\} = D_{i,i} \ , \ i = 1, \ldots, m \, .$$

- *Error probabilities:*

$$P_i^{\mathrm{e}} = \Pr\{\widehat{\mathbf{v}} \neq \boldsymbol{v}_i | \mathbf{v} = \boldsymbol{v}_i\} = 1 - D_{i,i} = \sum_{\ell \neq i} D_{\ell,i} \ , \ i = 1, \ldots, m \, .$$

- If $\mathbf{D} = \mathbf{I}_m$, then the detector is perfect: no matter what the parameter $\mathbf{v}$ is, we correctly guess $\widehat{\mathbf{v}} = \mathbf{v}$.

Minimax and Bayes Detector Designs

Minimax Detector Design:

$$\begin{array}{ll} \text{minimize} & \max\limits_{i=1,\ldots,m} P_i^{\mathrm{e}} \\ \text{subject to} & \mathbf{t}_k \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{t}_k = 1\,,\ k = 1,\ldots,n \end{array}\,.$$

- The minimax detector minimizes the worst-case probability of error over all m hypotheses.
- It can be reformulated as an LP.

Bayes Detector Design:

In Bayes detector design, the hypotheses have a prior distribution given by $\mathbf{q} \in \mathbb{R}^m$:

$$q_i = \Pr\{\mathbf{v} = \boldsymbol{v}_i\}\ ,\ i = 1,\ldots,m\,.$$

With this, the probability of error of the detector is given by $\mathbf{q}^T\mathbf{p}^{\mathrm{e}}$, where $[\mathbf{p}^{\mathrm{e}}]_i = P_i^{\mathrm{e}}$ for $i = 1,\ldots,m$ is the vector of error probabilities for the hypotheses. The Bayes detector design problem is then as follows.

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

- This problem is an LP that has a simple analytical solution.
- For the special case when $\mathbf{q} = (1/m)\,\mathbf{1}$, the Bayes optimal detector minimizes the average probability of error.

Probability Constraints and Alternate Objectives

Probability Constraints:

- *Detection probability lower limits:*

$$P_i^{\mathrm{d}} = D_{i,i} \geq L_i \,,\ i = 1, \ldots, m \,.$$

- *Error probability upper limits:*

$$D_{k,\ell} \leq U_{k,\ell} \,,\ k \neq \ell \,.$$

Alternate Objectives or Constraints: (valid when ordering of hypothesis values $\mathbf{v} \in \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ has some significance)

- *Bias:*

$$E_i[(\widehat{\mathbf{v}} - \mathbf{v})] = \sum_{k=1}^{m} (\boldsymbol{v}_k - \boldsymbol{v}_i)\, D_{k,i} \,,\ \text{when } \mathbf{v} = \boldsymbol{v}_i \,.$$

- *Mean-square error:*

$$E_i\left[(\widehat{\mathbf{v}} - \mathbf{v})^2\right] = \sum_{k=1}^{m} (\boldsymbol{v}_k - \boldsymbol{v}_i)^2\, D_{k,i} \,,\ \text{when } \mathbf{v} = \boldsymbol{v}_i \,.$$

- *Average absolute error:*

$$E_i[|\widehat{\mathbf{v}} - \mathbf{v}|] = \sum_{k=1}^{m} |\boldsymbol{v}_k - \boldsymbol{v}_i|\, D_{k,i} \,,\ \text{when } \mathbf{v} = \boldsymbol{v}_i \,.$$

Formulations of the Optimal Detector Design Problem

Multicriterion Formulation:

$$\begin{array}{ll} \text{minimize (with respect to } \mathbb{R}_+^{m(m-1)}) & D_{i,\ell}\,,\ i,\ell = 1,\ldots,m\,,\ i \neq \ell \\ \text{subject to} & \mathbf{t}_k \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{t}_k = 1\,,\ k = 1,\ldots,n \end{array} .$$

Scalarization Formulation:

To scalarize the multicriterion problem, we introduce a *loss matrix* $\mathbf{W} \in \mathbb{R}^{m\times m}$ which satisfies

$$W_{i,i} = 0\,,\ i = 1,\ldots,m\,,\ W_{i,\ell} > 0\,,\ i,\ell = 1,\ldots,m\,,\ i \neq \ell\,.$$

We then form the weighted sum objective $\mathrm{tr}\big(\mathbf{W}^T\mathbf{D}\big)$. This leads to the scalarization formulation:

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

As $\mathrm{tr}\big(\mathbf{W}^T\mathbf{D}\big) = \mathrm{tr}\big(\mathbf{W}^T\mathbf{T}\mathbf{P}\big) = \mathrm{tr}\big(\mathbf{P}\mathbf{W}^T\mathbf{T}\big) = \sum_{k=1}^n \mathbf{c}_k^T\mathbf{t}_k$, where $\mathbf{c}_k$ is the k-th column of $\mathbf{W}\mathbf{P}^T$, this objective is *separable* in $\mathbf{t}_k$. Thus, we can solve this problem by separately solving

$$\begin{array}{ll} \text{minimize} & \mathbf{c}_k^T\mathbf{t}_k \\ \text{subject to} & \mathbf{t}_k \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{t}_k = 1 \end{array} ,$$

for $k = 1,\ldots,n$. But this LP has a simple analytic solution: find index i such that $c_{k,i} = \min_\ell c_{k,\ell}$ and then take $\mathbf{t}_k^\star = \mathbf{e}_i$. This optimal solution corresponds to a deterministic detector: when $\mathbf{x} = \boldsymbol{x}_k$ is observed, our estimate is

$$\widehat{\boldsymbol{v}} = \boldsymbol{v}_{\ell^\star(k)} \text{ where } \ell^\star(k) = \operatorname*{argmin}_\ell \Big[\mathbf{W}\mathbf{P}^T\Big]_{\ell,k} .$$

MAP and ML Detector Design Problems

MAP Detector:

For a Bayes detector design with prior distribution $\mathbf{q}$, the mean probability of error is

$$\mathbf{q}^T \mathbf{p}^{\mathrm{e}} = \sum_{\ell=1}^{m} q_\ell \sum_{k \neq \ell} D_{k,\ell} = \sum_{k,\ell=1}^{m} W_{k,\ell} D_{k,\ell} \,,$$

if we define the weight matrix $\mathbf{W}$ as

$$W_{k,\ell} = q_\ell \,,\ k,\ell = 1,\ldots,m \,,\ k \neq \ell \,,\ W_{k,k} = 0 \,,\ k = 1,\ldots,m \,.$$

Thus, we have,

$$\left[\mathbf{W}\mathbf{P}^T\right]_{\ell,k} = \sum_{i \neq \ell} q_i P_{k,\ell} = \sum_{i=1}^{m} q_i P_{k,i} - q_\ell P_{k,\ell} \,.$$

Note that the first term is independent of ℓ. So, when $\mathbf{x} = \boldsymbol{x}_k$ is observed, the optimal detector is

$$\widehat{\boldsymbol{v}} = \boldsymbol{v}_{\ell^\star(k)} \text{ where } \ell^\star(k) = \operatorname*{argmax}_{\ell} \left(P_{k,\ell} q_\ell\right) \,.$$

Since $P_{k,\ell} q_\ell$ given the probability that $\mathbf{v} = \boldsymbol{v}_\ell$ and $\mathbf{x} = \boldsymbol{x}_k$, this detector is a MAP detector.

ML Detector:

For the special case $\mathbf{q} = (1/m)\,\mathbf{1}$, which corresponds to a uniform prior distribution on $\mathbf{v}$, this MAP detector reduces to an ML detector:

$$\widehat{\boldsymbol{v}} = \boldsymbol{v}_{\ell^\star(k)} \text{ where } \ell^\star(k) = \operatorname*{argmax}_{\ell} P_{k,\ell} \,.$$

Hence, an ML detector minimizes the average or mean probability of error.

Likelihood-Ratio Test and Neyman-Pearson Lemma

For the special case in which $m = 2$, we have *binary hypothesis testing*. In this case, $\mathbf{x}$ is generated from one of two distributions, $\mathbf{p} \in \mathbb{R}^n$ or $\mathbf{q} \in \mathbb{R}^n$. Also, we have

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

- P_{fp} is the probability of a *false positive*, or *false alarm probability*.
- P_{fn} is the probability of a *false negative*, or *missed detection probability*.
- The optimal trade-off curve between P_{fp} and P_{fn} is called the *receiver operating characteristic* (ROC).

For a weight matrix $\mathbf{W}$, an optimal detector, assuming $\mathbf{x} = \boldsymbol{x}_k$ is observed, is given by

$$\widehat{\boldsymbol{v}} = \begin{cases} \boldsymbol{v}_1\,, & p_k/q_k > W_{1,2}/W_{2,1} \\ \boldsymbol{v}_2\,, & p_k/q_k \leq W_{1,2}/W_{2,1} \end{cases},$$

which is a *likelihood ratio threshold test* with likelihood ratio p_k/q_k and threshold $W_{1,2}/W_{2,1}$. Choosing different thresholds leads to different Pareto optimal detectors that give different levels of false positive and false negative error probabilities. This result is known as the *Neyman-Pearson lemma*.

Detector Design Example

Consider a binary hypothesis testing example with $n = 4$ and

$$\mathbf{P} = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}$$

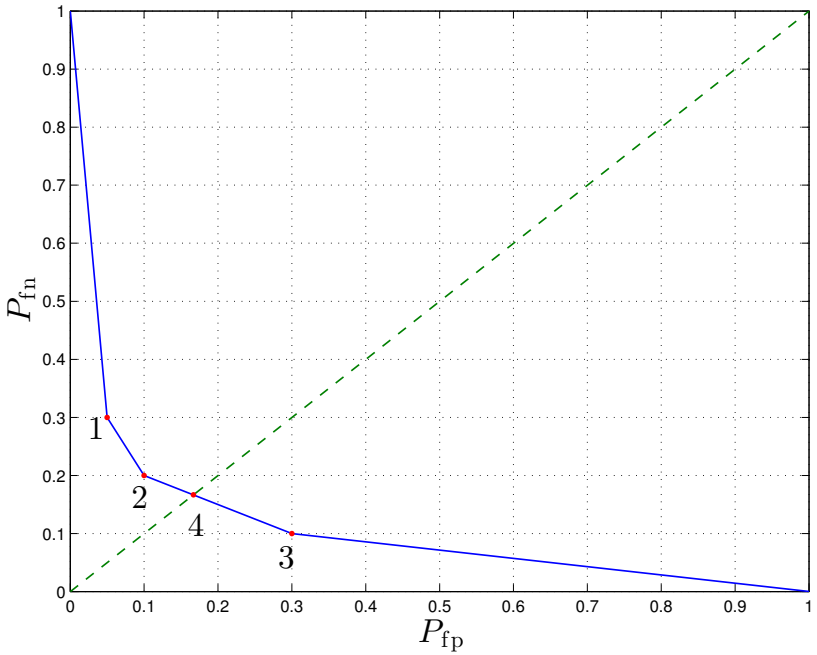
Four Pareto optimal detector matrices are given below.

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

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

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

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



- The vertices labeled 1, 2, and 3 correspond to deterministic detectors.
- The point labeled 4 corresponds to a random detector.
- Here, the dashed line shows $P_{\text{fn}} = P_{\text{fp}}$. The point at which this line intersects with the optimal trade-off curve, namely point 4, corresponds to the minimax detector.

Worst-Case Detection Probability Matrix

- When $\mathbf{P}$ is not known exactly, but $\mathbf{P} \in \mathcal{P}$, where $\mathcal{P}$ is a known set of possible distributions, we can attempt to accommodate this uncertainty to develop a *robust* detector.
- Here, error and detection probabilities will be judged by their worst-case values over $\mathbf{P} \in \mathcal{P}$.
- This leads us to define a *worst-case detection probability matrix* $\mathbf{D}^{\mathrm{wc}}$ as follows.

$$D^{\mathrm{wc}}_{k,\ell} = \sup_{\mathbf{P}\in\mathcal{P}} D_{k,\ell}\,,\ k,\ell = 1,\ldots,m\,,\ k \neq \ell\,, D^{\mathrm{wc}}_{k,k} = \inf_{\mathbf{P}\in\mathcal{P}} D_{k,k}\,,\ k = 1,\ldots,m\,.$$

- The off-diagonal entries of $\mathbf{D}^{\mathrm{wc}}$ give the largest error probabilities, while the diagonal entries give the smallest detection probabilities, over $\mathbf{P} \in \mathcal{P}$.
- We define the worst-case probability of error as

$$P_i^{\mathrm{wce}} \triangleq 1 - D^{\mathrm{wc}}_{i,i}\,,$$

  which is the largest probability of error, when $\mathbf{v} = \boldsymbol{v}_i$, over all possible distributions in $\mathcal{P}$.
- The *robust minimax detector* is defined as the detector that minimizes the worst-case probability of error, over all hypotheses, i.e., minimizes the objective

$$\max_i P_i^{\mathrm{wce}} = \max_{i=1,\ldots,m} \sup_{\mathbf{P}\in\mathcal{P}} \left\{1 - [\mathbf{TP}]_{i,i}\right\} = 1 - \min_{i=1,\ldots,m} \inf_{\mathbf{P}\in\mathcal{P}} \left\{[\mathbf{TP}]_{i,i}\right\}\,.$$

Examples of Robust Minimax Detectors

Robust minimax detector for finite $\mathcal{P}$:

When $\mathcal{P} = \{\mathbf{P}_1, \ldots, \mathbf{P}_k\}$, the robust minimax detector can be found by solving

$$\begin{array}{ll} \text{maximize} & \min\limits_{i=1,\ldots,m} \inf\limits_{\mathbf{P}\in\mathcal{P}} [\mathbf{TP}]_{i,i} = \min\limits_{i=1,\ldots,m} \min\limits_{\ell=1,\ldots,k} [\mathbf{TP}_\ell]_{i,i} \\ \text{subject to} & \mathbf{t}_i \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{t}_i = 1\,,\ i = 1,\ldots,n \end{array}\,.$$

- The objective is concave piecewise-linear, so the problem can be posed as an LP.
- The resulting robust minimax detector is also optimal for the polyhedron $\mathrm{conv}(\mathcal{P})$.

Robust minimax detector for polyhedral $\mathcal{P}$

When $\mathcal{P}$ is a polyhedron of the form

$$\mathcal{P} = \left\{\mathbf{P} = \left[\begin{array}{ccc} \mathbf{p}_1 & \cdots & \mathbf{p}_m \end{array}\right] : \mathbf{A}_k\mathbf{p}_k = \mathbf{b}_k\,,\ \mathbf{1}^T\mathbf{p}_k = 1\,,\ \mathbf{p}_k \succeq \mathbf{0}\right\}\,,$$

the robust minimax detector design problem can be shown to be expressed as the LP

$$\begin{array}{ll} \text{maximize} & \gamma \\ \text{subject to} & \mathbf{b}_i^T\boldsymbol{\nu}_i + \mu_i \geq \gamma\,,\ i = 1,\ldots,m \\ & \mathbf{A}_i^T\boldsymbol{\nu}_i + \mu_i\mathbf{1} \preceq \widetilde{\mathbf{t}}_i\,,\ i = 1,\ldots,m \\ & \mathbf{t}_i \succeq \mathbf{0}\,,\ \mathbf{1}^T\mathbf{t}_i = 1\,,\ i = 1,\ldots,n \end{array}\,,$$

with variables $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_m$, $\mu_1, \ldots, \mu_n$, and $\mathbf{T}$ (with columns $\mathbf{t}_1, \ldots, \mathbf{t}_n$ and rows $\widetilde{\mathbf{t}}_1, \ldots, \widetilde{\mathbf{t}}_m$).