The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Convex Optimization — Fall 2019

Lecture 22

Aryan Mokhtari Monday, November 18, 2019.

Goal: In this lecture, we talk about some statistical estimation problems that can be written as convex program. In particular, we look MLE for linear models and logistic regression. Then, we talk look at the Experiment Design problem.

1 Maximum Likelihood Estimation (MLE)

In the MLE problem we have a class of probability distribution that are parametrized with a parameter $\mathbf{x}$ and our goal is to choose the optimal parameter that maximizes the probability of our measurements, i.e., maximizes the likelihood function.
Consider $\mathbf{Y}$ as our random variable with probability density $p_{\mathbf{Y}}(\mathbf{y})$. Now assume that the probability distribution is parametrized by $\mathbf{x}$, i.e., $p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$. Hence, for a fixed value of $\mathbf{y}$, $p_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$ is a function of $\mathbf{x}$ which we call it the likelihood function.
The goal of MLE is to find the parameter $\mathbf{x}$ that maximizes $p_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$ for a given $\mathbf{y}$, i.e.,

$$\hat{\mathbf{x}}_{mle} \;=\; \underset{\mathbf{x}}{\operatorname{argmax}} \; p_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$$

As a simple example consider the case that y is a scalar and it has a gaussian distribution with mean x and variance 1, and x has two possible values -1 and 1. In this case for any fixed y, the MLE would choose the probability distribution that assigns a larger probability to $p_Y(y;x)$. For instance, if $y = -1$ then the MLE estimate is $\hat{x}_{mle} = -1$, and for $y = 10$ the MLE estimate is $\hat{x}_{mle} = 1$.

In most cases, instead of maximizing the likelihood function $p_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$ we maximize the log-likelihood function and solve the following problem

$$\hat{\mathbf{x}}_{mle} \;=\; \underset{\mathbf{x}}{\operatorname{argmax}} \; \log \; p_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$$

This is due to two reasons: (i) when probabilities are independent by looking at their log, we maximize sum of their logs instead of their products, (ii) most probability distributions are not concave, but they are log-concave!

1.1 Linear measurements with IID noise

Consider a linear measurement model:

$$y_i = \mathbf{a}_i^\top \mathbf{x} + v_i, \quad \ldots i = 1, \ldots, m$$

1. $\mathbf{x} \in \mathbb{R}^n$ is vector of unknown parameters

2. $y_i \in \mathbb{R}$ are the measured or observed quantities

3. $v_i \in \mathbb{R}$ are the measurement errors (noise) that are independent, identically distributed (i.i.d.)

Assume that the probability density for V_i is $p_{V_i}(v_i) = p_V(v_i) = p(v_i)$. Then, the likelihood of y_i is

$$p_{Y_i}(y_i; \mathbf{x}) = p(y_i - \mathbf{a}_i^\top \mathbf{x})$$

Since, $Y_1, \ldots, Y_m$ are i.i.d. then we can write that

$$p_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}) = \prod_{i=1}^{m} p(y_i - \mathbf{a}_i^\top \mathbf{x})$$

so the log-likelihood function is

$$\log p_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}) = \sum_{i=1}^{m} \log p(y_i - \mathbf{a}_i^\top \mathbf{x})$$

The MLE estimate is any optimal point for the problem

$$\max_{\mathbf{x}} \; \sum_{i=1}^{m} \log p(y_i - \mathbf{a}_i^\top \mathbf{x}) \tag{1}$$

Indeed, when p is a log-concave function with respect to $\mathbf{x}$ then we face a convex program.

Example 1: Let us focus on the case that v_i are Gaussian with zero mean and variance σ^2. In this case, the objective function in problem (1) can be written as

$$\begin{aligned}
\sum_{i=1}^{m} \log p(y_i - \mathbf{a}_i^\top \mathbf{x}) &= \sum_{i=1}^{m} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mathbf{a}_i^\top \mathbf{x})^2}{2\sigma^2}} \right) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{m} \log \left(e^{-\frac{(y_i - \mathbf{a}_i^\top \mathbf{x})^2}{2\sigma^2}} \right) \\
&= -\frac{m}{2} \log 2\pi\sigma^2 - \sum_{i=1}^{m} \frac{(y_i - \mathbf{a}_i^\top \mathbf{x})^2}{2\sigma^2}
\end{aligned}$$

Hence, the MLE estimate is the solution of the following least-squares problem

$$\hat{\mathbf{x}}_{mle} = \underset{\mathbf{x}}{\text{argmin}} \; \sum_{i=1}^{m} (y_i - \mathbf{a}_i^\top \mathbf{x})^2 \;=\; \underset{\mathbf{x}}{\text{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$$

Therefore, in this case if we have enough samples $m > n$ such that $\mathbf{A}^\top \mathbf{A}$ is invertible then the MLE estimate is

$$\hat{\mathbf{x}}_{mle} \;=\; (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \;=\; \left(\sum_{i=1}^{m} (\mathbf{a}_i \mathbf{a}_i^\top) \right)^{-1} \sum_{i=1}^{m} y_i \mathbf{a}_i$$

Example 2: Let us focus on the case that v_i are Laplacian with the probability distribution $p_V(v) = \frac{1}{2b} e^{-\frac{|v|}{b}}$ where $b > 0$. In this case, the objective function in problem (1) can be written as

$$\sum_{i=1}^{m} \log p(y_i - \mathbf{a}_i^\top \mathbf{x}) \;=\; \sum_{i=1}^{m} \log \left(\frac{1}{2b} e^{-\frac{|y_i - \mathbf{a}_i^\top \mathbf{x}|}{b}} \right) \;=\; -m \log 2b - \sum_{i=1}^{m} \frac{|y_i - \mathbf{a}_i^\top \mathbf{x}|}{b}$$

Hence, the MLE estimate is the solution of the following least-squares problem

$$\hat{\mathbf{x}}_{mle} = \underset{\mathbf{x}}{\text{argmin}} \; \sum_{i=1}^{m} |y_i - \mathbf{a}_i^\top \mathbf{x}| \;=\; \underset{\mathbf{x}}{\text{argmin}} \; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_1$$

1.2 Logistic Regression

In the Logistic Regression model, which is used for classification, our measurements y_i are either 1 or 0. The main assumption is that the probability distribution of the random variable Y has the following form:

$$\mathbf{Pr}(y = 1; \mathbf{x}, b) = \frac{\exp(\mathbf{a}^\top \mathbf{x} + b)}{1 + \exp(\mathbf{a}^\top \mathbf{x} + b)}, \qquad \mathbf{Pr}(y = 0; \mathbf{x}, b) = \frac{1}{1 + \exp(\mathbf{a}^\top \mathbf{x} + b)}$$

In this case, $\mathbf{x} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are the model parameters that determine how the probability varies as a function of the explanatory variable $\mathbf{a}$. Assume that we have access to m pairs of $(\mathbf{a}_i, y_i)$ and the goal is to choose $\mathbf{x}$ and b in a way the log-likelihood functions is maximized.
You can think of $\mathbf{a}_1, \dots, \mathbf{a}_m$ as feature vectors and $y_1, \dots, y_m$ as their corresponding labels. WLOG, assume that the first k observations have label 1 and the rest have label 0. Then, the log-likelihood in this case is

$$\begin{aligned}
\log \mathbf{Pr}(\mathbf{y}; \mathbf{x}, b) &= \log \prod_{i=1}^{k} \frac{\exp(\mathbf{a}_i^\top \mathbf{x} + b)}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} \\
&= \sum_{i=1}^{k} \log \frac{\exp(\mathbf{a}_i^\top \mathbf{x} + b)}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} + \sum_{i=k+1}^{m} \log \frac{1}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} \\
&= \sum_{i=1}^{k} \log \exp(\mathbf{a}_i^\top \mathbf{x} + b) + \sum_{i=1}^{k} \log \frac{1}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} + \sum_{i=k+1}^{m} \log \frac{1}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} \\
&= \sum_{i=1}^{k} (\mathbf{a}_i^\top \mathbf{x} + b) + \sum_{i=1}^{m} \log \frac{1}{1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b)} \\
&= \sum_{i=1}^{k} (\mathbf{a}_i^\top \mathbf{x} + b) - \sum_{i=1}^{m} \log(1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b))
\end{aligned}$$

Hence, in this case the MLE estimate is

$$\hat{\mathbf{x}}_{mle}, b_{mle} = \underset{\mathbf{x}, b}{\operatorname{argmax}} \; \sum_{i=1}^{k} (\mathbf{a}_i^\top \mathbf{x} + b) - \sum_{i=1}^{m} \log(1 + \exp(\mathbf{a}_i^\top \mathbf{x} + b))$$

This is a convex program!

2 Experiment Design

Consider the problem of estimating a vector $\mathbf{x} \in \mathbb{R}^n$ from measurements or experiments

$$y_i = \mathbf{a}_i^\top \mathbf{x} + v_i, \quad \dots i = 1, \dots, m$$

Consider the case that the measurement noise has a Gaussian distribution and they are iid. Further, assume that we have enough samples such that the span of vectors $\mathbf{a}_1, \dots, \mathbf{a}_m$ is $\mathbb{R}^n$. As we mentioned above, the MLE estimate, which is the same as the minimum variance estimate, is given by

$$\hat{\mathbf{x}}_{mle} \;=\; (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y} \;=\; \left(\sum_{i=1}^{m} (\mathbf{a}_i \mathbf{a}_i^\top) \right)^{-1} \sum_{i=1}^{m} y_i \mathbf{a}_i$$

It can be easily verified that the error vector $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}_{mle}$ has a zero mean and its covariance matrix is

$$\mathbf{E} = \mathbb{E}[\mathbf{e}\mathbf{e}^\top] = \left(\sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^\top\right)^{-1}$$

The matrix $\mathbf{E}$ characterizes the accuracy of the estimation, or the informativeness of the experiments. For example, the α-confidence level ellipsoid for $\mathbf{x}$ is given by

$$\{\mathbf{z} \mid (\mathbf{z} - \hat{\mathbf{x}})^\top \mathbf{E}^{-1}(\mathbf{z} - \hat{\mathbf{x}}) \leq \beta\}$$

where β is a function of α and n.

In the experiment design problem, the goal is to choose the vectors $\mathbf{a}_i$ in a smart way that we minimize the confidence ellipsoid with a limited budget for sampling.

Setup: Suppose that the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$, which characterize the measurements, can be chosen among p possible test vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$. The goal is to sample m measurements while we minimize the error covariance matrix $\mathbf{E}$. If we consider m_i as the number of times that $\mathbf{v}_i$ is chosen, then the problem can be written as

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

Note that in this problem the variables are $m_1, \ldots, m_p$. Note this is a vector optimization problem over the positive semidefinite cone.

This problem in general is hard, since we have an integer constraint! To resolve this issue we will relax it.

If one experiment design results in $\mathbf{E}$, and another in $\tilde{\mathbf{E}}$,

1. with $\mathbf{E} - \tilde{\mathbf{E}} \preceq \mathbf{0}$, then certainly the first experiment design is as good as or better than the second. (The confidence ellipsoid for the first experiment design is contained in the confidence ellipsoid of the second.)
2. but if $\mathbf{E} - \tilde{\mathbf{E}} \npreceq \mathbf{0}$ or $\tilde{\mathbf{E}} - \mathbf{E} \npreceq \mathbf{0}$, then we can not conclude which one is better than the other one.

2.1 The relaxed experiment design problem

If we eliminate the constraint $m_k \in \mathbf{Z}$ and use the notation $\lambda_k = m_k/m$ then the problem can be relaxed and written as

$$\begin{array}{ll} \text{minimize(w.r.t. } \mathbf{S}_+^n) & \mathbf{E} = \frac{1}{m}\left(\sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top\right)^{-1} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda}^\top \mathbf{1} = 1 \end{array}$$

This problem is a convex program.

Several scalarizations have been proposed for the relaxed experiment design problem that we study them in the following subsections.

2.2 D-optimal design

The most widely used scalarization is called D-optimal design in which we minimize the log det of the covariance matrix which is proportional to the volume of confidence ellipsoids.

$$\begin{array}{ll} \text{minimize} & \log\det\left(\sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top\right)^{-1} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda}^\top \mathbf{1} = 1 \end{array}$$

If we write the dual of this problem we obtain that

$$\begin{array}{ll} \text{maximize} & n\log n + \log\det \mathbf{W} \\ \text{subject to} & \mathbf{v}_k^\top \mathbf{W} \mathbf{v}_k \leq 1, \quad k = 1, \ldots, p \end{array}$$

The interpretation is that $\mathbf{v}_k^\top \mathbf{W} \mathbf{v}_k \leq 1$ is minimum volume ellipsoid centered at origin, that includes all test vectors $\mathbf{v}_i$.

Complementary slackness implies that for optimal primal dual variables we have

$$\lambda_k^*(1 - \mathbf{v}_k^\top \mathbf{W}^* \mathbf{v}_k) = 0, \quad k = 1, \ldots, p$$

Therefore, D-optimal experiment uses vectors $\mathbf{v}_k$ on boundary of ellipsoid defined by $\mathbf{W}^*$.

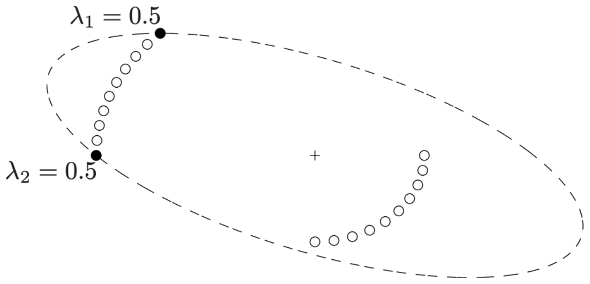

Figure 7.9 Experiment design example. The 20 candidate measurement vectors are indicated with circles. The D-optimal design uses the two measurement vectors indicated with solid circles, and puts an equal weight $\lambda_i = 0.5$ on each of them. The ellipsoid is the minimum volume ellipsoid centered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

2.3 E-optimal design

In E-optimal design, we minimize the norm of the error covariance matrix, i.e., the maximum eigenvalue of $\mathbf{E}$. Geometrically can be interpreted geometrically as minimizing the diameter of the confidence ellipsoids.

$$\begin{array}{ll} \text{minimize} & \left\| \left(\sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top\right)^{-1} \right\|_2 \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda}^\top \mathbf{1} = 1 \end{array}$$