# COMPSCI 689
# Lecture 3: Optimization and MLE

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

# Supervised Learning By Density Estimation

Given a data set $\mathcal{D}$ and a probability model $\mathbb{P}(\Theta)$, we use density estimation to approximate $P_*(\mathbf{x}, \mathbf{y})$ by $P(\mathbf{x}, \mathbf{y}|\theta_*)$.
We then use $P(\mathbf{x}, \mathbf{y}|\theta_*)$ in place of $P_*(\mathbf{x}, \mathbf{y})$ in the distributional supervised learning problem:

$$\theta_* = \arg\min_{\theta \in \Theta} \mathcal{L}(\mathcal{D}, P(\theta))$$

$$\hat{f}_* = \arg\min_f \int_{\mathcal{X}} \int_{\mathcal{Y}} L(\mathbf{y}, f(\mathbf{x})) P(\mathbf{x}, \mathbf{y}|\theta_*) d\mathbf{x} d\mathbf{y}$$

$$\hat{f}_*(\mathbf{x}) = \arg\min_{\mathbf{y}'} \int_{\mathcal{Y}} L(\mathbf{y}, \mathbf{y}') P(\mathbf{y}|\mathbf{x}, \theta_*) d\mathbf{y}$$

For example, if $L(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_2^2$, then $\hat{f}_*(\mathbf{x}) = \mathbb{E}_{P(\mathbf{y}|\mathbf{x}, \theta_*)}[\mathbf{y}]$

# Empirical Probability Distribution

- We can instead explicitly model $f(x)$, but this requires making a default choice for approximating $P_*$.
- The standard choice is the *empirical distribution* $P_{\mathcal{D}}$:

$$P_{\mathcal{D}}(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\|\mathbf{x} - \mathbf{x}_n\|)\delta(\|\mathbf{y} - \mathbf{y}_n\|)$$

where $\delta(z)$ is the Dirac delta function $\delta(z) = \begin{cases} \infty & z = 0 \\ 0 & z \neq 0 \end{cases}$

- Expectations under $P_{\mathcal{D}}$ have the following important property:

$$\mathbb{E}_{P_{\mathcal{D}}(\mathbf{x},\mathbf{y})}\left[g(\mathbf{y}, \mathbf{x})\right] = \frac{1}{N} \sum_{n=1}^{N} g(\mathbf{y}_n, \mathbf{x}_n)$$

# Prediction Models

- A prediction model $F$ is a set of functions $F = \{f \mid f\colon \mathcal{X} \to \mathcal{Y}\}$.
- A parametric prediction model $F(\Theta)$ with parameter vector $\theta \in \Theta$ is generated by a fixed function $g$ satisfying $g\colon \mathcal{X} \times \Theta \to \mathcal{Y}$. We write $F(\Theta) = \{f \mid f(\mathbf{x}) = g(\mathbf{x}, \theta), \theta \in \Theta\}$.
- **Example**: Let $F(\mathbb{R}^D) = \{f \mid f(\mathbf{x}) = \theta^T \mathbf{x}, \theta \in \mathbb{R}^D\}$ for $\mathbf{x} \in \mathbb{R}^D$.

# Supervised Learning By Function Optimization

Given a data set $\mathcal{D}$ and a prediction model $F(\Theta)$, we first approximate $P_*$ by $P_{\mathcal{D}}$, and then minimize $L$ with respect to $\theta$.

$$\hat{f}_* = \arg\min_{f \in F(\Theta)} \mathbb{E}_{P_{\mathcal{D}}(\mathbf{x},\mathbf{y})} \left[ L(\mathbf{y}, f(\mathbf{x})) \right]$$

$$\hat{\theta}_* = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} L(\mathbf{y}_n, f(\mathbf{x}_n, \theta))$$

# Optimization and Learning

In either the probability modeling or prediction function modeling frameworks, we identify the optimal model parameters via the solution to an optimization problem.

# An Optimization Problem

An optimization problem in standard form consists of four primary components:

$$\min_{x \in \mathcal{X}} f(\mathbf{x}) \qquad \text{subject to} \begin{cases} c_i(\mathbf{x}) = 0, & i \in \mathcal{E} \\ c_i(\mathbf{x}) \geq 0, & i \in \mathcal{I} \end{cases}$$

- Variables: $\mathbf{x} = [x_1, ..., x_M] \in \mathcal{X}$
- Objective Function: $f : \mathcal{X} \to \mathbb{R}$
- Equality Constraints: $c_i(\mathbf{x}) = 0, i \in \mathcal{E}$
- Inequality Constraints: $c_i(\mathbf{x}) \geq 0, i \in \mathcal{I}$

# Types of Optimization Problems

There are several different categories of optimization problems distinguished by the following characteristics:

- Discrete vs Continuous Variables
- Constrained vs Unconstrained
- Stochastic vs Deterministic
- Convex vs Non-Convex Objective Functions

# Convex Optimization

A function $f$ is said to be convex if $\mathcal{X}$ is convex and for any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and any scalar $\alpha \in [0, 1]$:

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{x}') \leq \alpha f(\mathbf{x}) + (1 - \alpha)f(\mathbf{x}')$$

A set $\mathcal{X}$ is convex if for any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and any scalar $\alpha \in [0, 1]$:

$$\alpha\mathbf{x} + (1 - \alpha)\mathbf{x}' \in \mathcal{X}$$

# Optimization-based Learning

In optimization-based learning:

- The model parameters are the optimization variables
- There are typically no constraints or simple convex constraints
- The objective function may or may not be convex
- The objective function may or may not be continuous
- The objective function may or may not be differentiable

For now, we will focus on the case of unconstrained objective functions that are continuos and twice differentiable, but not necessarily convex.

# Types of Solutions

Key definitions of types of optimization solutions:

- **Global Minimizer:** A point $\mathbf{x}_* \in \mathcal{X}$ is a global minimizer of $f$ if and only if $f(\mathbf{x}_*) \leq f(\mathbf{x})$ for all $x \in \mathcal{X}$.
- **Local Minimizer:** A point $\mathbf{x}_* \in \mathcal{X}$ is a local minimizer of $f$ if and only if $f(\mathbf{x}_*) \leq f(\mathbf{x})$ for all $x$ in an open set around $\mathbf{x}_*$.
- For most complex, non-convex functions, we can only hope to identify a local minima with no guarantee about it's quality relative to the global minimum.
- **Question:** How can we tell if a point $\mathbf{x}$ in a continuous space is a local minimizer of $f$?

# Characterizing Solutions: Definitions

Key definitions for characterizing optimization solutions:

- **Gradient**: The gradient $\nabla f(\mathbf{x})$ of function $f(\mathbf{x})$ is the vector of partial derivatives of $f$ with respect to each of the optimization variables: $[\nabla f(\mathbf{x})]_i = \frac{\partial}{\partial x_i} f(\mathbf{x})$.
- **Hessian**: The hessian $\nabla^2 f(\mathbf{x})$ of function $f(\mathbf{x})$ is the matrix of mixed partial derivatives of $f$ with respect to each pair of optimization variables: $[\nabla^2 f(\mathbf{x})]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$.
- **Stationary Point**: $\mathbf{x}$ is said to be a stationary point of $f$ if $\nabla f(\mathbf{x}) = 0$.
- **Local Minimizer:** $\mathbf{x}$ is a local minimizer of $f$ if $\mathbf{x}$ is a stationary point of $f$ and the Hessian of $f$ at $\mathbf{x}$ is positive semi-definite.
- **Convexity Condition**: If $f$ is convex and differentiable and $\mathbf{x}_*$ is a stationary point of $f$, then $\mathbf{x}_*$ is a global minimizer of $f$

# Finding Solutions

- Some optimization problems can be solved analytically by solving the system of gradient equations $\nabla f(\mathbf{x}) = 0$ to identify one or more stationary points $\mathbf{x}_*$, and then checking that $\nabla^2 f(\mathbf{x}_*)$ is positive definite (or appealing to convexity).
- When an analytic approach fails, an iterative numerical approach can be used to approximately locate a stationary point.
- More on iterative numerical optimization later...
- For now, let's look at the application of optimization to solve the density estimation problem.

# The Likelihood Function

- Given a parametric probability model $\mathbb{P}(\theta)$ defined using a parametric density $P(\mathbf{Z}|\theta)$ and a data set $\mathcal{D} = [\mathbf{z}_1, ..., \mathbf{z}_N]$, the likelihood function of $\mathbb{P}(\theta)$ is defined as:

$$L(\mathcal{D}, \theta) = \prod_{n=1}^{N} P(\mathbf{Z} = \mathbf{z}_n|\theta)$$

- It is often easier to work with the log of $L$, the log likelihood function:

$$l(\mathcal{D}, \theta) = \sum_{n=1}^{N} \log P(\mathbf{Z} = \mathbf{z}_n|\theta)$$

- The (log) likelihood function should be interpreted as a function of $\theta$ for fixed $\mathcal{D}$. It tells us how probable the data are given a particular choice of the parameters $\theta$.

# Maximum Likelihood Estimation

- The maximum likelihood principle asserts that the optimal parameters $\theta$ are the parameters that make the observed data the most likely.

- Maximum Likelihood Estimation is a method for selecting the parameters $\theta$ of a parametric probability model $P(\mathbf{Z}|\theta)$ by maximizing the (log) likelihood function:

$$\theta_* = \arg\max_\theta L(\mathcal{D}, \theta) = \arg\max_\theta l(\mathcal{D}, \theta)$$
$$= \arg\min_\theta -\sum_{n=1}^{N} \log P(\mathbf{Z} = \mathbf{z}_n|\theta) = \arg\min_\theta \mathcal{L}(\mathcal{D}, \theta)$$

- The function $\mathcal{L}(\mathcal{D}, \theta)$ is called the negative log likelihood function.

# Example: Bernoulli Distribution

- Let $P(Z = z|\theta) = \theta^z(1-\theta)^{(1-z)}$ for $\mathcal{Z} = \{0, 1\}$ and $\theta \in [0, 1]$.
- Suppose we have a data set $\mathcal{D} = [z_1, ..., z_N]$ and we want to find the MLE of $\theta$.
- The log likelihood function is:

$$l(\mathcal{D}, \theta) = \sum_{n=1}^{N} (z_n \log \theta + (1 - z_n) \log(1 - \theta))$$

- The MLE is $\theta_* = \frac{1}{N} \sum_{n=1}^{N} z_n$.

# Example: Normal Mean

- Let $P(Z = z|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right)$
- Suppose we have a data set $\mathcal{D} = [z_1, ..., z_N]$ and we want to find the MLE of $\mu$.
- The log likelihood function is:

$$l(\mathcal{D}, \mu, \sigma) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z_n - \mu)^2$$

- The MLE is $\mu_* = \frac{1}{N}\sum_{n=1}^{N} z_n$.

# Example: Normal Standard Deviation

- Let $P(Z = z|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right)$
- Suppose we have a data set $\mathcal{D} = [z_1, ..., z_N]$ and we want to find the MLE of $\sigma$.
- The log likelihood function is:

$$l(\mathcal{D}, \mu, \sigma) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z_n - \mu)^2$$

- The MLE is $\sigma_* = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(z_n - \mu_*)^2}$.

# Example: Conditional Normal

- Let $P(Y = y|X = x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - wx)^2\right)$
- Suppose we have a data set $\mathcal{D} = [z_1, ..., z_N]$ and we want to find the MLE of $w$.
- The conditional log likelihood function is:

$$l(\mathcal{D}, w, \sigma) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - wx_n)^2$$

- The MLE is $w_* = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}$.