Lecture #2: Maximimum Likelihood Estimation

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020

Outline

1. A Motivating Example
2. A Statistical Model for a Coin Flip
3. Maximum Likelihood Estimation
4. Convex Optimization: Constrained and Unconstrained
5. Properties of MLE
6. Uncertainty Quantification

A Motivating Example

A Simple Betting Game

I propose to you that we play a betting game: I toss a coin, if the coin lands heads up then you will pay me $\$20$, otherwise I will pay you $\$20$.

Question: What information do you need to determine if this will be a profitable game for you to play?

Estimating the ”Bias” of a Coin

You might want to determine if my coin is a "trick" or "biased" coin before betting your money. A common way to test a coin for bias is to toss this coin N number of times and count the number of heads, H. The fraction

$$\frac{\text{Number of Heads}}{\text{Total Number of Tosses}} = \frac{H}{N}$$

is one way to quantify the probability of the coin to land heads up on any given toss.

Alternatively, we can interpret this fraction to represent the fraction of heads that would appear in a large (infinite) number of such experiments.

Question 1: Is this estimate of the bias valid? I.e. does $\frac{H}{N}$ acurately capture the property of interest?

Question 2: Is this the "best" way to estimate the bias? For example, is the quantity

$$\frac{\text{Number of Heads} + 1}{\text{Total Number of Tosses} + 2} = \frac{H+1}{N+2}$$

an equally valid or better estimate of the bias?

A Statistical Model for a Coin Toss

Likelihood for a Coin Toss

We can formally model the outcome of the single toss of a coin by a Bernoulli distribution

$$Y \sim Ber(\theta)$$

where θ is the probability that the outcome Y will be heads.

Question: what assumptions does this statistical model expose?

After N number of ***independent*** tosses of an ***identical*** coin, the probability (or likelihood) of observing $Y = H$ number of heads is

$$\binom{N}{H} \theta^H (1-\theta)^{N-H}$$

That is, Y is a random variable with a ***binomial*** distribution $Y \sim Bin(N, \theta)$.

We see that the fraction $\frac{H}{N}$ from our empirical experiment is an estimate of the parameter θ of the binomal distribution $Bin(N, \theta)$. Now that we have a statistical model, we can give formal justification for why our estimate is desirable (or undesirable).

Maximum Likelihood Estimation

Parameter Estimation: Maximum Likelihood

Let $Y_1, \ldots, Y_N$ be independently and identically distributed with $Y_n \sim p(Y|\theta)$, where $p(Y|\theta)$ is a distribution parameterized by θ (θ can be a scalar, a vector, a matrix, or a n-tuple of such quantities). The ***joint likelihood*** of N observations, $y_1, \ldots, y_N$, is

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$

Note that we use upper-case letters Y_n to represent random variables and lower-case y_n to represent specific observed values of those variables.

The joint likelihood quantifies how likely (or probable, if Y is discrete) we are to observed the data assuming the model θ. When we consider the joint likelihood as a function of θ (that is, treat the observed data as fixed), the $\mathcal{L}(\theta)$ is called the ***likelihood function***.

The ***maximium likelihood estimate*** of θ is defined as

$$\theta_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ \mathcal{L}(\theta) = \underset{\theta}{\text{argmax}}\ \prod_{n=1}^{N} p(y_n|\theta)$$

Recall that in Lecture #1 we gave some intuitive justification for the validity of the MLE.

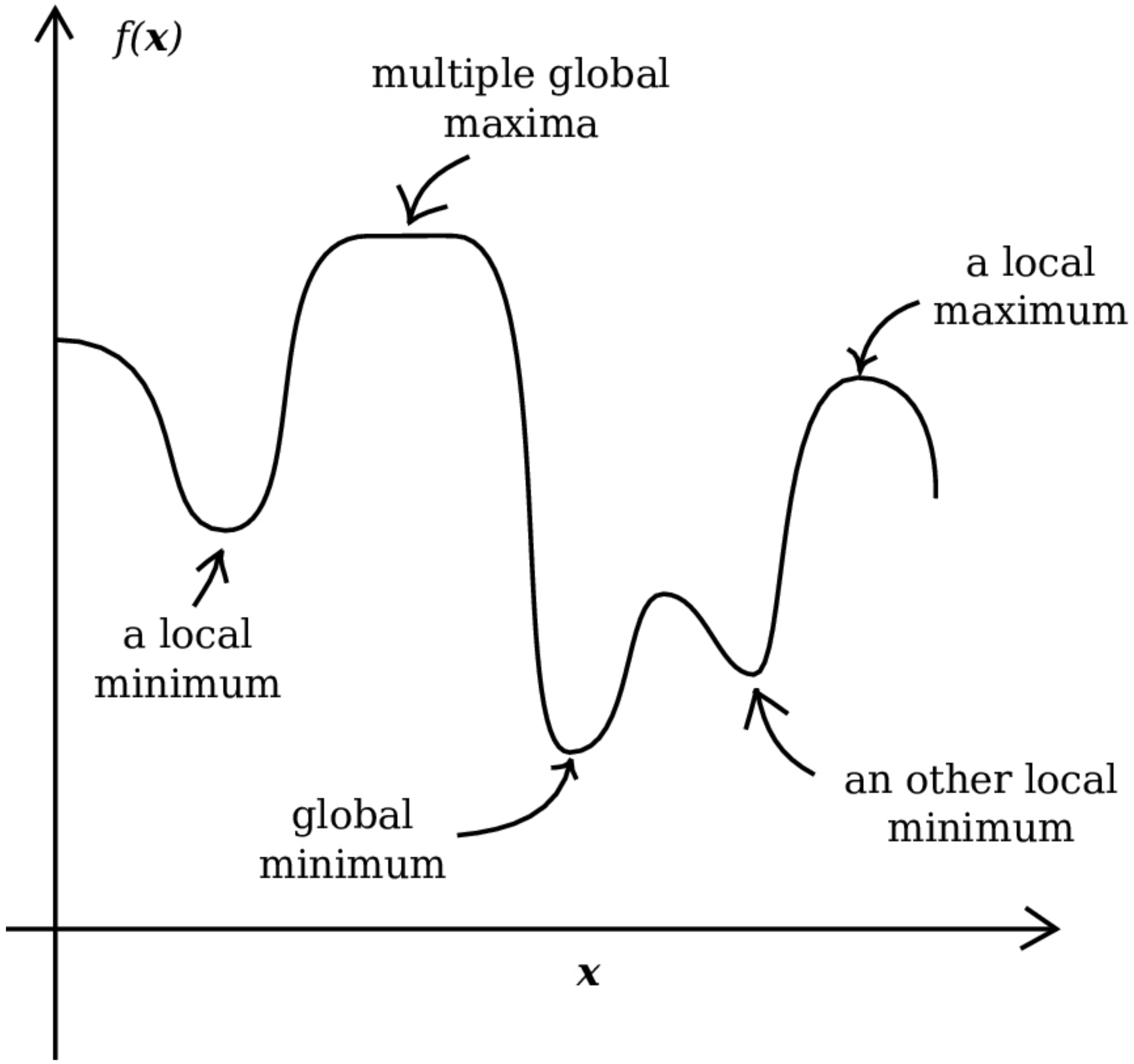
Maximizing Likelihood is Equivalent to Maximizing Log-Likelihood

Frequently, the likelihood function is complex and so it's often preferable to work with the log of the likelihood function. Luckily, ***maximizing the likelihood is equivalent to maximizing the log likelihood*** due to the following fact.

Theorem: For any $f : \mathbb{R}^D \to \mathbb{R}$, we have that $x^* = \underset{\theta}{\text{argmax}}\ f(x)$ if and only if $x^* = \underset{\theta}{\text{argmax}}\ \log(f(x))$.

Convex Optimization: Constrained and Unconstrained

Introduction to Optimization: Types of Optima



Stationary Points

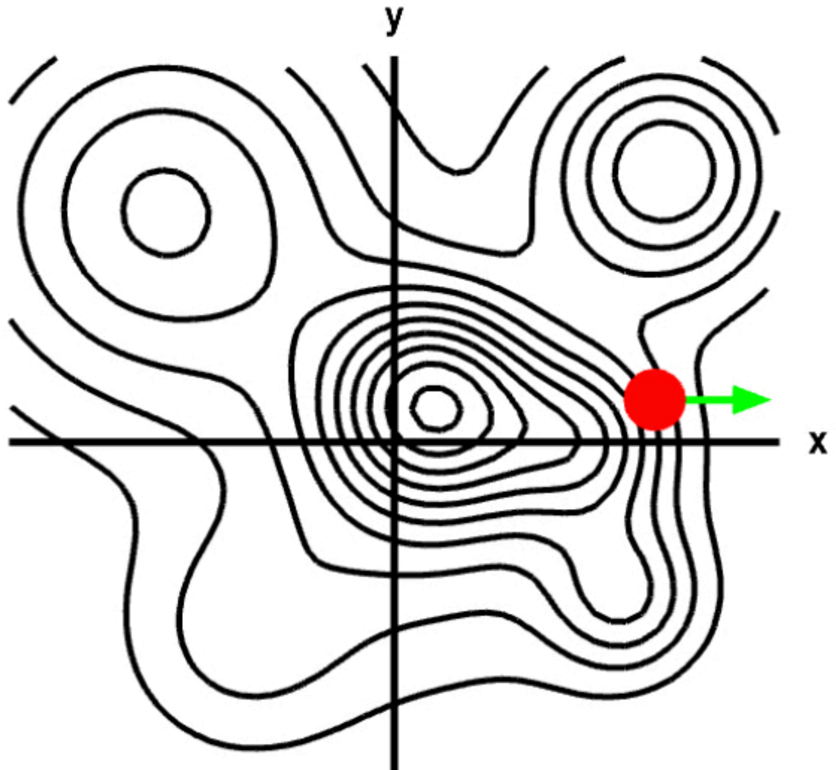
The instaneous rate of change, at $x = x_0$, of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ is given by it's first derivative at $x = x^*$, $\frac{df}{dx}\Big|_{x^*}$.

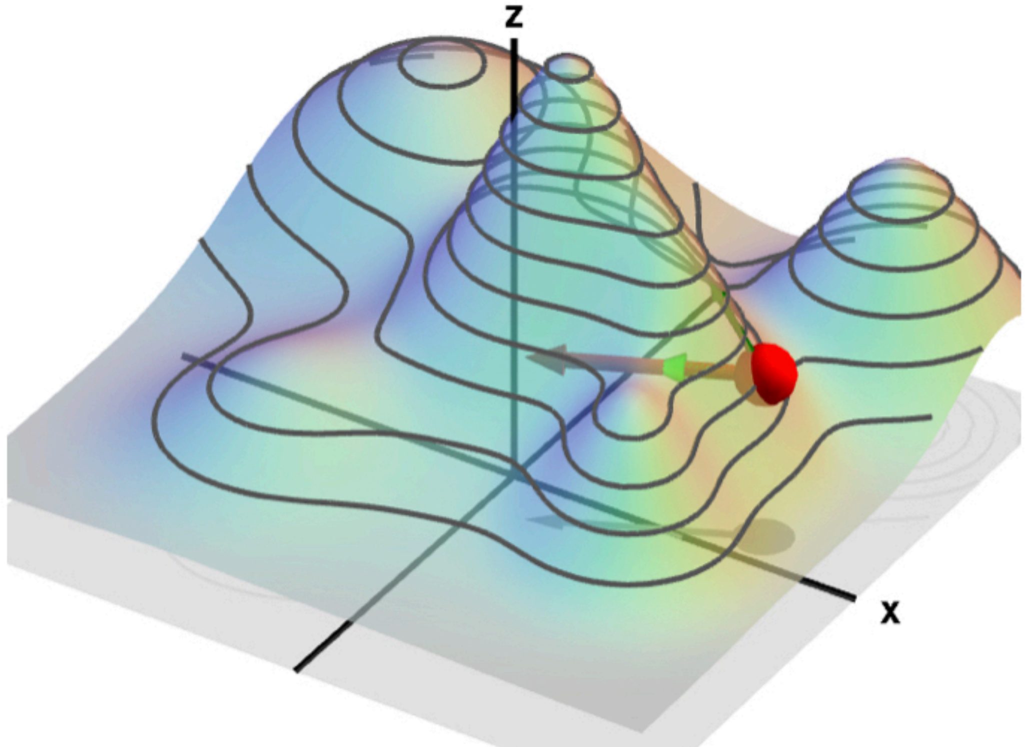
For a multivariate differentiable function $f : \mathbb{R}^D \to \mathbb{R}$, the ***gradient*** of f at a point x^* is a vector consisting of the partial derivatives of f evaluated at x^*:

$$\nabla_x f|_{x^*} = \left[\frac{\partial}{\partial x^{(1)}}\Bigg|_{x^*}, \ldots, \frac{\partial}{\partial x^{(D)}}\Bigg|_{x^*} \right]$$

Each $\frac{\partial}{\partial x^{(1)}}\Big|_{x^*}$ compute the instantaneous change of f at $x = x^*$ with respect to $x^{(1)}$.

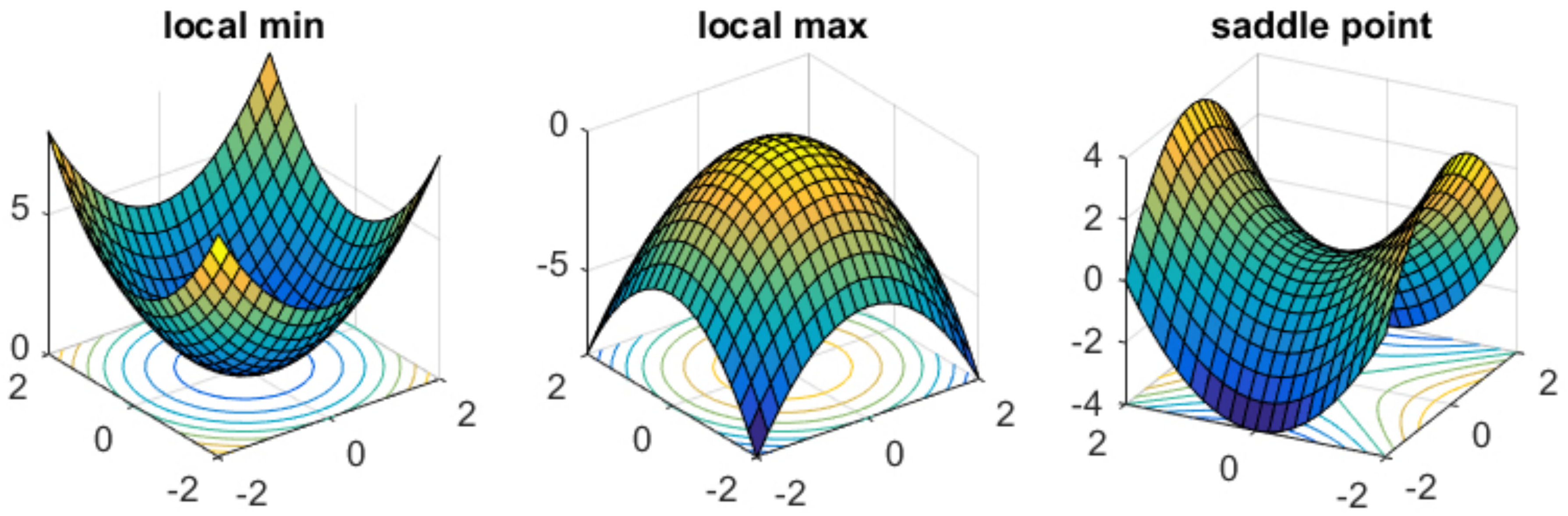
The gradient is orthogonal to the level curve of f at x^* and hence, *when it is not zero*, points in the direction of the greatest instantaneous increase in f.





Characterization of Local Optima

A local optima must be a stationary point, but *a stationary point need not be a local optima*!



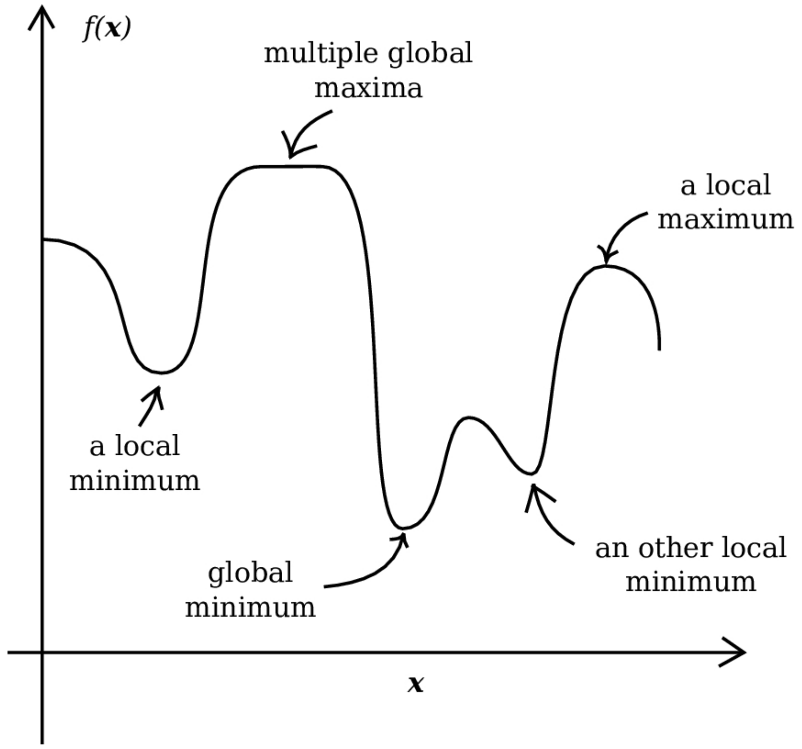
To check that a stationary point is a local max (or local min), we must check that the function is ***concave*** (or ***convex***) at the point.

Recall, that for a twice differentiable function $f : \mathbb{R} \to \mathbb{R}$, f is concave at $x = x^*$ if the second derivative of f is negative; f is convex at $x = x^*$ if the second derivative of f is positive. For a multivariate twice differentiable function $f : \mathbb{R}^D \to \mathbb{R}$, f is concave at $x = x^*$ if the Hessian matrix is semi-negative definite; f is convex at $x = x^*$ if the Hessian is semi-positive definite.

Characterization of Global Optima

For an arbitrary function, we cannot generally determine if a local optimal is a global one! In certain very restricted cases, we can deduce if a local optimal is global:

Theorem: If a continuous function f is convex (or resp. concave) on its domain then every local min (or resp. max) is a global min (or resp. max).



Unconstrained Optimization

Analytically solving an optimization problem without constraints on the domain of the function,

$$x_{\max} = \underset{x}{\operatorname{argmax}} \; f(x)$$

involves:

1. find the expression for $\nabla_x f(x)$.
2. find the stationary points for $\nabla_x f(x)$. That is, solve the equation $\nabla_x f(x) = 0$ for x.
3. determine local optima. That is, check the concavity of f at the stationary points.
4. determine global optima. That is, check if local optima can be characterized as global optima (e.g. check that f is convex everywhere on its domain).

Example: (Univariate) Gaussian Distribution

Likelihood and log-likelihood

Suppose that $Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$. Let θ denote the set of parameters (μ, σ). The likelihood for N observations $y_1, \ldots, y_N$ is

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_n - \mu)^2}{2\sigma^2} \right\} = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{\sum_{n=1}^{N} (y_n - \mu)^2}{2\sigma^2} \right\}.$$

The log likelihood is

$$\ell(\theta) = -\frac{N}{2} \log 2\pi - N \log \sigma - \frac{(y_n - \mu)^2}{2\sigma^2}.$$

Example: (Univariate) Gaussian Distribution

Gradient of log-likelihood

The gradient of ℓ with respect to θ is the vector $\nabla_\theta \ell(\theta) = \left[\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma}\right]$, where the partial derivatives are given by:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + \sigma^{-3} \sum_{n=1}^{N} (y_n - \mu)^2$$

Example: (Univariate) Gaussian Distribution

Stationary points of the gradient

The stationary points of the gradients are solutions to the following system of equations:

$$\begin{cases} \frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \mu) = 0 \\ \frac{\partial \ell}{\partial \sigma} = -\frac{N}{\sigma} + \sigma^{-3} \sum_{n=1}^{N} (y_n - \mu)^2 = 0 \end{cases}$$

Solving this system, we get a *unique* solution at:

$$\begin{cases} \mu = \frac{1}{N} \sum_{n=1}^{N} y_n \\ \sigma = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu)^2} \end{cases}$$

Example: (Univariate) Gaussian Distribution

Characterize local and global optima

The log-likelihood in this case is concave -- the Hessian will be negative semi-definite for μ and $\sigma > 0$. Thus, the log-likelihood is globally maximized at:

$$\begin{cases} \mu_{\text{MLE}} = \frac{1}{N} \sum_{n=1}^{N} y_n \\ \sigma_{\text{MLE}} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu)^2} \end{cases}$$

Check for yourself: write out the matrix of second order parial derivatives of the log-likelihood and check that all the upper-left submatrices have negative determinants.

Note: If the objective is not concave, then there is no guarantee that the stationary points will be global maxima!

Constrained Optimization

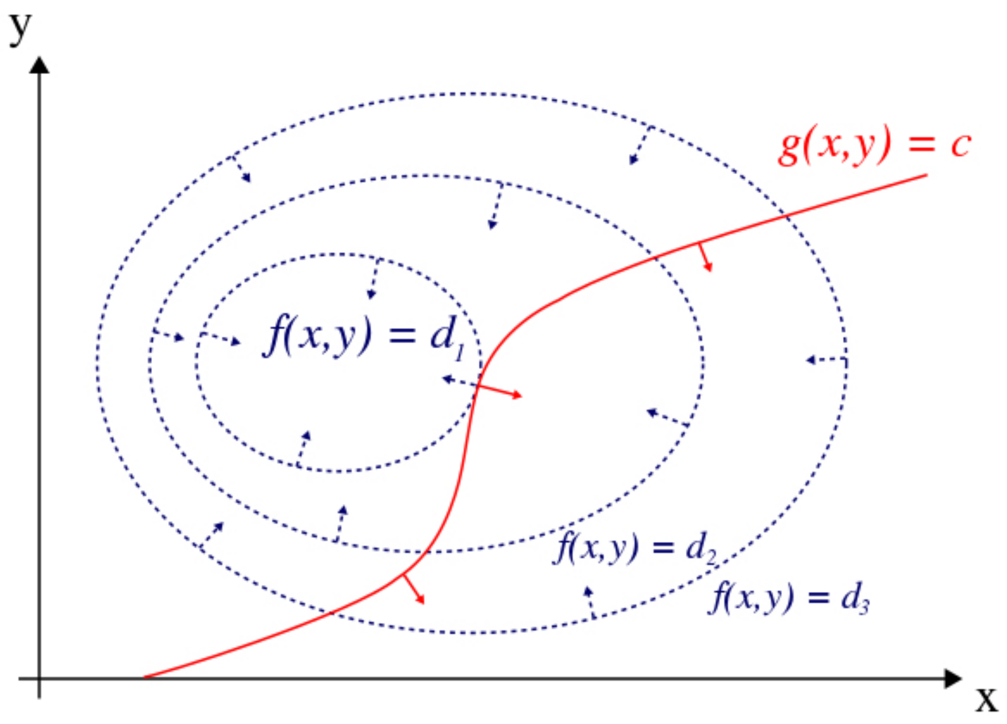
Many times, we are constrained by the application to only consider certain types of values of the input x. Suppose that the ***constraints*** on x are given by the equation $g(x) = 0$. The set of values of x that satisfy the equation are called ***feasible***.

We recall a useful theorem from calculus:

Theorem: For a differentiable function $f : \mathbb{R}^D \to \mathbb{R}$, the local optima of f constrained by $g(x) = 0$ occur at points where the following hold for some $\lambda \in \mathbb{R}$,

$$g(x) = 0, \qquad \nabla_x f(x) = \lambda \nabla_x g(x).$$

The theorem says that the local optima of f satisfying $g(x) = 0$ are where the gradients of f and g are parallel.



Constrained Optimization via Lagrange Multipliers

Unpacking the theorem, we get that solving an optimization problem within the ***feasible region*** of the function, i.e.

$$\max_x \; f(x), \qquad g(x) = 0$$

involves:

1. finding the stationary points of the augmented objective $J(x) = f(x) - \lambda g(x)$, with respect to x and λ.
2. determine global optima. Determine if any of the stationary points maximizes f.

The augmented objective J is called the ***Lagrangian*** of the constrained optimization problem and λ is called the ***Lagrange multiplier***.

Note: Constrained optimization with inequality constraints can similarly be formulated in terms of finding stationary points of an augmented objective like the Lagrangian; this follows from the ***Karush–Kuhn–Tucker theorem***.

Example: Binomial Distribution

Likelihood and Log-Likelihood

Suppose that $Y \sim Bin(N, \theta)$. To make the connection with constrained optimization (and to motivate the multinomial case), let's write θ as a vector $[\theta_0, \theta_1]$, where θ_1 is the probability of a head and $\theta_0 + \theta_1 = 1$.

The likelihood for a single observations is

$$\mathcal{L}(\theta) = \frac{N!}{y!(N-y)!}\theta_1^y \theta_0^{N-y}.$$

The log likelihood is

$$\ell(\theta) = \log(N!) - \log(y!) - \log(N-y)! + y \log \theta_1 + (N-y) \log \theta_0.$$

We are interested in solving the following constrained optimization problem:

$$\max\ \ell(\theta), \quad \theta_0 + \theta_1 = 1$$

whose Lagrangian is give by:

$$J(\theta, \lambda) = \ell(\theta) - \lambda(\theta_0 + \theta_1 - 1).$$

Example: Binomial Distribution

Gradient of log-likelihood

The gradient of the Lagrangian J with respect to (θ, λ) is the vector $\nabla_{(\theta,\lambda)} J = \left[\frac{\partial \ell}{\partial \theta_0}, \frac{\partial \ell}{\partial \theta_1}, \frac{\partial \ell}{\partial \lambda}\right]$, where the partial derivatives are given by:

$$\frac{\partial \ell}{\partial \theta_0} = \frac{(N - y)}{\theta_0} - \lambda$$

$$\frac{\partial \ell}{\partial \theta_1} = \frac{y}{\theta_1} - \lambda$$

$$\frac{\partial \ell}{\partial \lambda} = \theta_0 + \theta_1 - 1$$

Example: Binomial Distribution

Stationary points of the Lagrangian

The stationary points of the Lagrangian are solutions to the following system of equations:

$$\begin{cases} \frac{\partial \ell}{\partial \theta_0} & = \frac{(N-y)}{\theta_0} - \lambda = 0 \\ \frac{\partial \ell}{\partial \theta_1} & = \frac{y}{\theta_1} - \lambda = 0 \\ \frac{\partial \ell}{\partial \lambda} & = \theta_0 + \theta_1 - 1 = 0 \end{cases}$$

Solving this system, we get a *unique* solution at:

$$\begin{cases} \theta_0 & = \frac{N-y}{\lambda} \\ \theta_1 & = \frac{y}{\lambda} \\ \theta_0 + \theta_1 & = 1 \end{cases}$$

In other words, $\lambda = N$ and $\theta_1 = \frac{y}{N}$ and $\theta_0 = \frac{N-y}{N}$.

Example: Binomial Distribution

Characterize global optima

Since the log-likelihood $\ell(\theta)$ is concave and the constraint $\theta_0 + \theta_1 = 1$ is affine (linear up to a constant), we know then that $\ell(\theta)$ is maximized on the line at the stationary point. Hence,

$$\begin{cases} \theta_0^{\text{MLE}} = \frac{N-y}{N} \\ \theta_1^{\text{MLE}} = \frac{y}{N} \end{cases}$$

Note: If the objective function we are maximizing is not concave and the equality constraint is not affine, then we have no guarantee that the stationary points of the langrangian either locally or globally optimizes our objective!

What Is a Good Estimator?

We see that if we assume a binomial model, $Bin(N, \theta)$, for the number of heads in N trials, then the fraction $\frac{H}{N}$ is the maximum likelihood estimate of θ.

Question 1: Is the MLE a good estimator of θ?

Question 2: Is this the "best" way to estimate the θ? For example, is the quantity

$$\frac{\text{Number of Heads} + 1}{\text{Total Number of Tosses} + 2} = \frac{H+1}{N+2}$$

an equally valid or better estimate of θ?

These questions depend on our list of desiderata for our estimator.

Properties of MLE

Desiderata of Estimators

Let $\widehat{\theta}$ be an estimator of the parameter θ of a statistical model. We ideally want:

1. **(Consistency)** when the sample size N increases, in the limit, $\widehat{\theta}$ approaches the true value of θ.

   More formally, let $\{p_\theta; \theta \in \Theta\}$ be a family of candidate distributions and X^θ be an infinite sample from p_θ. Define $\widehat{g}_N(X^\theta)$ to be an estimator for some parameter $g(\theta)$ that is based on the first N samples. Then we say that the sequence of estimators $\{\widehat{g}_N(X^\theta)\}$ is (weakly) consistent if $\lim_{N\to\infty} \widehat{g}_N(X^\theta) = g(\theta)$ in probability for all $\theta \in \Theta$.

Desiderata of Estimators

1. **(Unbiasedness)** on average, over all possible sets of observations from the distribution, the estimator nails the true value of θ.

   More formally, we want $\mathbb{E}_{X^\theta}\widehat{\theta}(X^\theta) = \theta$.

Desiderata of Estimators

3. **(Minimum Variance)** Note that since our estimator $\widehat{\theta}$ depends on the random sample X^θ, it follows that $\widehat{\theta}$ also a random variable. The distribution of $\widehat{\theta}$ is called the ***sampling distribution***. Given that our estimator is unbiased, we want it to have minimum variance with respect to the sampling distribution.

Properties of MLE

Assumptions: In order for nice properties of the MLE to hold, we need to make some assumptions, including

(A) the model is **well-specified** -- the observed data is drawn from the same model class as the model being fitted;

(B) the estimation problem is **well-posed** -- there are not two different set of parameters that generate the same data.

With these assumptions, we have that:

1. **(Consistency)** The MLE of *iid* observations is consistent. The asymptotic sampling distribution of the MLE is a Gaussian.
2. **(Unbiasedness)** The MLE can be biased.
3. **(Minimum Variance)** The MLE is not the estimator with the lowest variance.

Asympotically, however, the MLE is unbiased and has the lowest variance (for unbiased estimators).

Uncertainty Quantification

Confidence Intervals

Since the MLE depends on the sample, it is important to quantify how certain we are about the estimate.

Confidence intervals of estimates θ_{MLE} are ways of summarizing the sampling distribution by describing it's coverage. Specifically, a 95% confidence interval for θ is a ***random interval*** $\left(L_{\theta_{\text{MLE}}}, U_{\theta_{\text{MLE}}}\right)$, where L and U are bounds constructed from the estimate θ_{MLE}, that contains the fixed true parameter θ with 95% probability.

Let $\delta = \theta_{\text{MLE}} - \theta$ be the distribution of the error of the estimator θ_{MLE}, then the following is a confidence interval for θ:

$$\left[\theta_{\text{MLE}} - \delta_{0.25}, \theta_{\text{MLE}} + \delta_{0.975}\right]$$

where $\delta_{0.25}, \delta_{0.975}$ are the 2.5% and 97.5% thresholds of δ respectively.

We can take advantage of the asymptotic normality of the MLE and approximate the distribution of δ as a Gaussian distribution.

Interpretation of Confidence Intervals

It is very easy to misinterpret confidence intervals!

A Simplified Rule: When in doubt, treat the confidence interval just as an **indication of the precision of the measurement.**

If you estimated some quantity in a study with a confidence interval of $[17 - 6, 17 + 6]$ and someone else estimated it with a confidence interval of $[23 - 5, 23 + 5]$, then there is little reason to think that the two studies are inconsistent.

On the other hand, if your estimate gives $[17 - 2, 17 + 2]$ and the other estimate is $[23 - 1, 23 + 1]$, then there is evidence that these studies differ.

Bootstrap Confidence Intervals

In practice, we may not know how to approximate the sampling distribution of θ_{MLE}. We can approximate the sampling distribution by ***bootstraping***, i.e. we simulate samples X^{θ} with size N from p_{θ} by sampling observations with size N from the observed data (also with size N).

We denote MLE obtained on a bootstrap sample by $\theta_{\text{MLE}}^{\text{bootstrap}}$. When N is sufficiently large, $\theta_{\text{MLE}}^{\text{bootstrap}}$ approximates the distribution of θ_{MLE}.

Thus, we can approximate the 95% confidence interval of θ using $\theta_{\text{MLE}}^{\text{bootstrap}}$.