Point estimation

↳ Method of Moments

$$g_1(\hat{\theta}) = \underset{p(x;\hat{\theta})}{\mathbb{E}}[X] = \overline{X}_n$$ (in the case of one unknown)

↳ Maximum likelihood

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x_1, \dots, x_n)$$

$$\ell(\theta; x_1, \dots, x_n) = \log L(\theta; x_1, \dots, x_n)$$

Properties of MLE

1. $\hat{\theta} \xrightarrow{P} \theta_*$, where θ_* is the true parameter

Theorem

Let $\hat{\theta}_n$ be the MLE for a random sample (of size n) $X_1, \dots, X_n$ from a parametric population $p(x; \theta_*)$. Then $\hat{\theta}_n \xrightarrow{P} \theta_*$

Proof

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

$$\Downarrow$$

$$\ell(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i; \theta)$$

$$\Downarrow \quad \text{divide by } n$$

$$\ell_n(\theta) \triangleq \frac{1}{n} \ell(\theta; x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(x_i; \theta)$$

$$\ell(\theta) \triangleq$$

$$\text{(WLLN)} \quad \ell_n(\theta) \xrightarrow{p} \mathbb{E}\left[\log p(X; \theta)\right]$$

$$\hat{\theta}_n = \arg\max_{\theta} \ell_n(\theta) \xrightarrow{p} \arg\max_{\theta} \left(\; ! \; \right)$$

Now, we must prove that θ_* is the maximizer of $\ell(\theta)$:

$$\ell_n(\theta) \xrightarrow{p} \ell(\theta)$$

$$\hat{\theta}_n = \arg\max_{\theta} \ell_n(\theta) \xrightarrow{p} \arg\max_{\theta} \ell(\theta)$$

$$= \theta_*$$

$$\ell(\theta) - \ell(\theta_*) = \mathbb{E}\left[\log p(x;\theta)\right] - \mathbb{E}\left[\log p(x;\theta_*)\right]$$

$$= \int \log p(x;\theta)\, p(x;\theta_*)\, dx - \int \log \frac{p(x;\theta_*)}{p(x;\theta_*)}\, dx$$

$$= \int p(x;\theta_*) \log\left(\frac{p(x;\theta)}{p(x;\theta_*)}\right) dx \qquad \left(\log(x) \leq x-1,\ x>0\right)$$

$$\leq \int p(x;\theta_*)\left[\frac{p(x;\theta)}{p(x;\theta_*)} - 1\right] dx$$

$$= \underbrace{\int p(x;\theta)\, dx}_{1} - \underbrace{\int p(x;\theta_*)\, dx}_{1}$$

$$= 0 \Rightarrow \ell(\theta) - \ell(\theta_*) \leq 0$$

$$\Rightarrow \ell(\theta) \leq \ell(\theta_*) \quad \forall\, \theta$$

This means $\theta_* = \arg\max_{\theta} \ell(\theta)$

Thus, $\hat{\theta}_n \xrightarrow{p} \theta_*$ ▨

Asymptotic Distribution of MLE

$\hat{\theta}_n \xrightarrow{p} \theta_*$ and therefore

$\hat{\theta}_n \xrightarrow{d} \theta_*$

Definition (Fisher Information)

The Fisher information of a random variable X with distribution $p(x;\theta)$ is defined as:

$$I(\theta) \triangleq \mathbb{E}\left[\left(\frac{\partial \log p(x;\theta)}{\partial \theta}\right)^2\right]$$

correction after lecture (should be squared)

and if the second derivative of $\log p(x;\theta)$ exists we can also write

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log p(x;\theta)}{\partial \theta^2}\right]$$

Theorem: Asymptotic Distribution of MLE

Let $\hat{\theta}_n$ be the MLE. Then,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_*\right) \xrightarrow{d} N\left(0, I^{-1}(\theta_*)\right)$$

Will prove later

Invariance Property of MLE

$$\hat{\theta} \rightsquigarrow \text{MoME} \qquad \rightarrow \qquad g(\hat{\theta}) \rightsquigarrow \text{is method of moments?} \qquad \xrightarrow{p} \qquad g(\theta_*)$$

$$\hat{\theta}_n = \arg\max_{\theta} \ell(\theta; X_1, \ldots, X_n) \xrightarrow{p} \theta$$

$$\alpha = \tau(\hat{\theta}_n) = \arg\max_{\alpha} \ell(\alpha; X_1, \ldots, X_n)$$

Theorem (Invariance Property of MLE)

Let $\hat{\theta}_n$ be the MLE. Consider the estimator $\hat{\alpha}_n = \tau(\hat{\theta}_n)$, where τ is any function. Then

$\hat{\alpha}_n$ is the maximum likelihood estimator of $\alpha = \tau(\theta)$.

Proof

Induced likelihood function:

$$L^*(\alpha; x_1, \ldots, x_n)$$

$$= \max_{\{\theta : \tau(\theta) = \alpha\}} L(\theta; x_1, \ldots, x_n)$$

We want to show that the MLE of α under $L^*(\alpha; x_1, \ldots, x_n)$ is exactly $\hat{\alpha}_n = \tau(\hat{\theta}_n)$

$$L^*(\hat{\alpha}_n; x_1, \ldots, x_n) = \max_{\alpha} L^*(\alpha; x_1, \ldots, x_n)$$

$$= \max_{\alpha} \max_{\{\theta : \tau(\theta) = \alpha\}} L(\theta; X_1, \ldots, X_n)$$

$$= L(\hat{\theta}_n; X_1, \ldots, X_n)$$

We now want to show that $\hat{\alpha}_n = \tau(\hat{\theta}_n)$

$$L(\hat{\theta}; X_1, \ldots, X_n) = \max_{\theta : \tau(\theta) = \alpha} L(\theta; X_1, \ldots, X_n)$$

$$= L^*(\tau(\hat{\theta}); X_1, \ldots, X_n)$$

$$\Rightarrow \boxed{\hat{\alpha}_n = \tau(\hat{\theta}_n)}$$

Example: Estimate the power of a noisy signal in dB

Consider a zero-mean white noise model

$$X_i \sim N(0, \sigma^2)$$

It is known that $\mathbb{E}[X^2] = \sigma^2$ is a measure of the power of the signal. Find the MLE of the power in dB, where

$$P = 10 \log_{10}(\mathbb{E}[X^2])$$

$$\hat{\sigma}^2 = \arg\max_{\sigma} \ell(\sigma^2; X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

By the invariance property of MLE, we have

$$\boxed{\hat{P} = 10 \log_{10}(\hat{\sigma}^2)}$$

$$\left. \frac{\partial \ell}{\partial \theta} \right|_{\theta = \hat{\theta}} = 0$$

Finding the roots of the derivative

Consider the example:

$$X_1, \dots, X_n \sim \sum_{k=1}^{K} \pi_k P_k(x)$$

$\pi_k \geq 0$ $\quad \sum_{k=1}^{K} \pi_k = 1$ Constraint

Lagrange multiplier

$$\tilde{\ell}(\pi, \lambda; x_1, \dots, x_n) = \ell(\pi; x_1, \dots, x_n) + \lambda \left(1 - \sum_{k=1}^{K} \pi_k\right)$$

$$\ell(\pi; x_1, \dots, x_n) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k P_k(x_i)\right)$$

$$\frac{\partial \tilde{\ell}}{\partial \pi_j} = 0 \qquad \frac{\partial \tilde{\ell}}{\partial \lambda} = 0$$

$$\frac{\partial \tilde{\ell}}{\partial \pi_j} = \sum_{i=1}^{n} \frac{P_j(x_i)}{\sum_{k=1}^{K} \pi_k P_k(x_i)} - \lambda$$

- gradient ascent / descent
- Newton
- EM algorithm

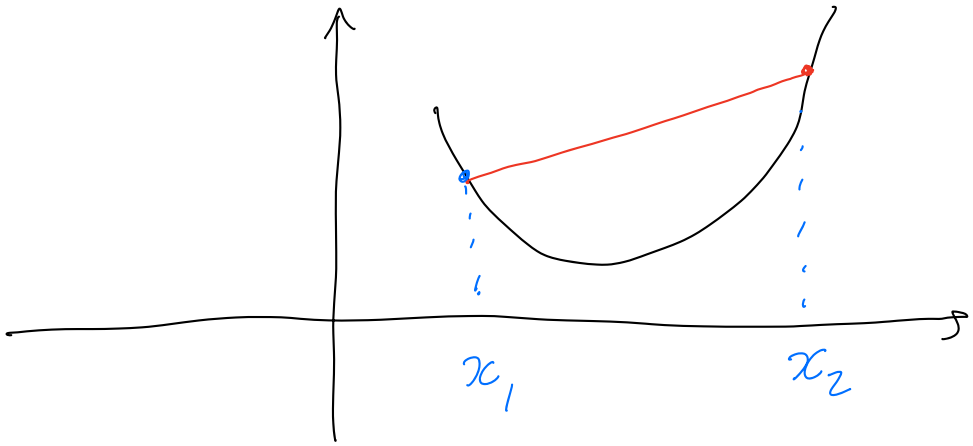
Basics of convex analysis

Def. Convex Function

A function $f: \mathbb{R}^D \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathbb{R}^D$, we have

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

where $t \in [0, 1]$



Why is it important?

Any optima of a convex function is the global one.

$$\theta_* = \arg\min_{\theta} f(\theta)$$ is the global optima always

$$\frac{\partial f(\theta)}{\partial \theta} = 0$$

In the context of MLE:

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x_1, \ldots, x_n)$$

$$= \arg\min_{\theta} -L(\theta; x_1, \ldots, x_n)$$

is convex?

Theorem (Convexity)

Let f be a twice differentiable scalar input scalar-valued function. Then, f is convex if and only if

$$\frac{\partial^2 f}{\partial \theta^2} \geq 0 \quad \text{for all } \theta$$

Examples:

$$f(\theta) = \theta \xrightarrow{\partial/\partial\theta} 1 \xrightarrow{\partial/\partial\theta} 0$$

$$f(\theta) = \theta^2 \xrightarrow{\partial/\partial\theta} 2\theta \xrightarrow{\partial/\partial\theta} 2$$

$$f(\theta) = \theta^4 \xrightarrow{\partial/\partial\theta} 4\theta^3 \xrightarrow{\partial/\partial\theta} 12\theta^2$$

$$f(\theta) = -\log(\theta) \xrightarrow{\partial/\partial\theta} -\frac{1}{\theta} \xrightarrow{\partial/\partial\theta} \frac{1}{\theta^2}$$

$\theta > 0$

Multivariate Extension

the input is a vector of size D

In the case that $f: \mathbb{R}^D \to \mathbb{R}$ we define

1. Gradient of f w.r.t. θ

$$\nabla_\theta f(\theta) = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \vdots \\ \frac{\partial f}{\partial \theta_D} \end{bmatrix}$$

2. Hessian (matrix of second partial derivatives)

$$H(\theta) \in \mathbb{R}^{D \times D} = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_D} \\ \vdots & & & \vdots \\ \frac{\partial^2 f}{\partial \theta_D \partial \theta_1} & \cdots & & \frac{\partial^2 f}{\partial \theta_D^2} \end{bmatrix}$$

Theorem:

Let f be a vector input scalar valued function that is twice differentiable: We say that f

is convex if and only if:

$$H(\theta) \succeq 0$$

$$\Updownarrow$$

$$\theta^T H(\theta)\, \theta \geq 0 \iff \text{all eigenvalues are non-negative}$$

($\theta^T H(\theta)\, \theta$: inner product)

Definition (Canonical Exponential Families)

A probability distribution $p(x;\theta)$ belongs to the exponential family of probability distributions if the pdf/pmf can be expressed as:

$$p(x;\theta) = h(x) \exp\left(\eta(\theta)^T T(x) - A(\theta)\right)$$

($T(x)$: sufficient statistic)

where h, η, T, and A are known functions with

$$A(\theta) = \log \int h(x) \exp\left(\eta(\theta)^T T(x)\right) dx$$

We say that $p(x;\theta)$ is a canonical exponential

family member if $\eta(\theta) = \theta$.

It turns out that $\log p(x;\theta)$ is concave if $p(x;\theta)$ belongs to the canonical exponential family:

- Gaussian
- Binomial
- Gamma
- ... many more

Gradient descent:

We want to solve

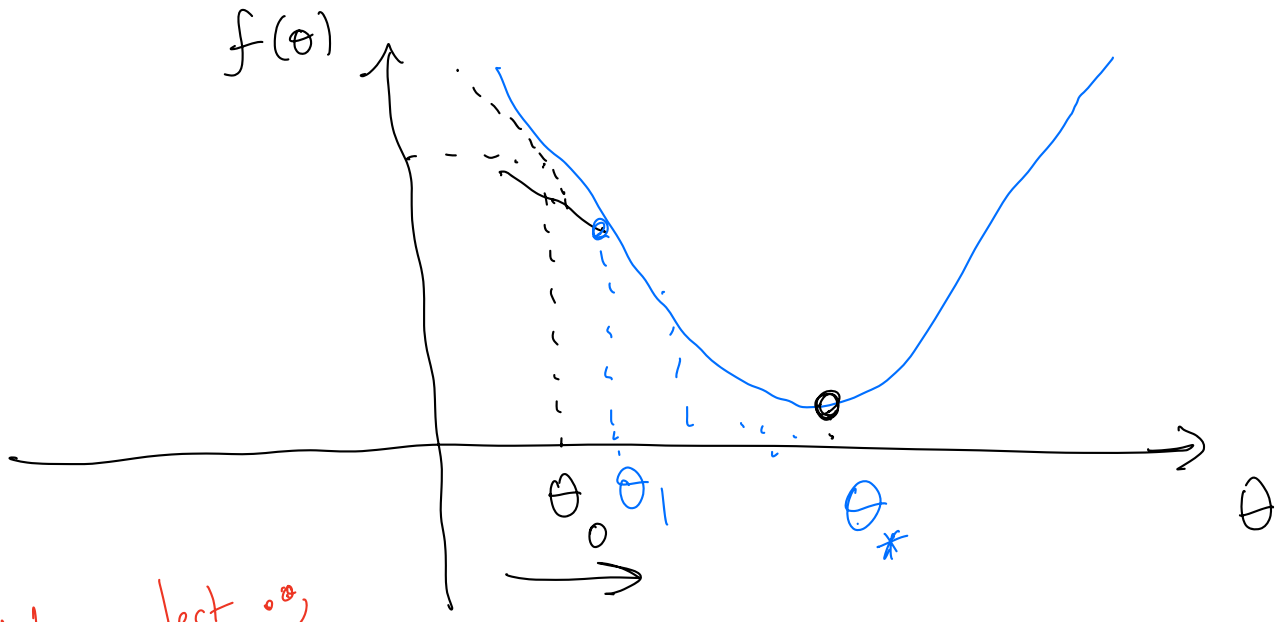
$$\hat{\theta}_n = \arg\min_{\theta} \underbrace{-\ell(\theta; x_1, \ldots, x_n)}_{f(\theta)}$$

We start off with a guess θ_0 and update our guess iteratively as:

$$\theta' \leftarrow \theta - \eta \nabla_\theta f(\theta)$$

η: learning rate / step size

$\nabla_\theta f(\theta)$: gradient



Not on test :)

If there are constraints, we can apply a projection to the update. For example, if $\theta \in \mathcal{H}$, we can update as follows:

$$\theta' = \Pi\left(\theta - \eta \nabla_\theta f(\theta)\right)$$

called the projection operator