Maximum Likelihood Estimation (MLE)

Setup: Given some data $\underline{\mathcal{D} := \{x_1, x_2, \dots, x_n\}}$, $x_i \in \mathbb{R}^d$

Assume a family of distributions / models $p_\vartheta(x)$, $\vartheta \in \Theta$

i.e. $\quad x_i \sim p_\vartheta(x)$ for some ϑ.

Goal: Estimate the true value of ϑ that best explains the observed data.

Definition: ϑ_{MLE} is a maximum likelihood estimate if:

$$\boxed{\vartheta_{MLE} = \arg\max_{\vartheta \in \Theta} \underbrace{p(\mathcal{D} \mid \vartheta)}_{\text{likelihood}}}, \quad p(\mathcal{D} \mid \vartheta_{MLE}) = \max_{\vartheta} p(\mathcal{D} |$$

$$\circledast \quad p(\mathcal{D} \mid \vartheta) = p(x_1, x_2, \dots, x_d \mid \vartheta) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} p(x_i \mid \vartheta) = \prod_{i=1}^{n} p(X = x_i |$$

Remarks: i.) The MLE might not be unique.

ii) The MLE may fail to exist (the maximum likelihood may not be achieved for $\vartheta \in \Theta$)

Pros: i) Usually easy to compute and often is interpretable (e.g. the mean or r.v. is the sample mean)

ii.) Nice assymptotic properties:

- Consistent: as $n \to +\infty$ the MLE converge to the true ϑ in probability.

distributio

- assymptotically normal : as $n \to +\infty$ its ... converges to a normal.

- efficient, i.e. they have the lowest assymptotic variance.

- invariant to re-parametrization :

  θ_{MLE}, $\forall g$: $g(\theta_{MLE})$ is an MLE for $g(\theta)$.

Cons :

i) MLE provides a point estimate (no representation of uncertainty

- Ideally, we'd like to compute the posterior distribution over θ :

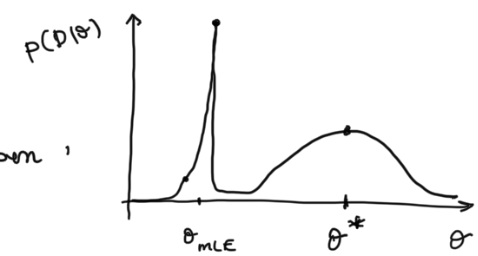
$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} \overset{\text{Bayes}}{\propto} \underbrace{p(\mathcal{D}|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

(Bayesian approach

- Frequentist approach : "Train" an ensemble of models over many initialization $\theta_0 \sim p(\theta_0)$.

ii.) Lack of robustness :

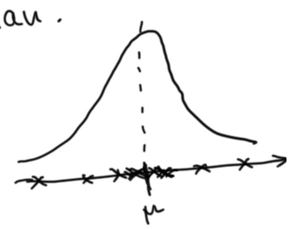
e.g. thing like this may happen,



iii) MLE are prone to overfitting.

iv) Existence and uniqueness may not be guaranteed.

Example : MLE for a univariate Gaussian.



Setup : Given $\mathcal{D} := \{x_1, x_2, \dots, x_n\}$, $x_i \in \mathbb{R}$

Assume $\left\{ \begin{array}{l} x_i \overset{\text{i.i.d.}}{\underset{p}{\sim}} p_\theta(x) = \mathcal{N}(x \mid \mu, \sigma^2) \quad , \quad \theta := \{\mu, \sigma^2\} \\ \Longrightarrow \quad x_i = \mu + \varepsilon \quad , \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad , \quad \sigma^2 > 0 \end{array} \right.$

generative model

<u>Likelihood</u>, $p(\mathcal{D} \mid \theta) = p(x_1, \dots, x_d \mid \mu, \sigma^2)$

$$\overset{\text{i.i.d}}{=} \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu\right.$$

$$p(\mathcal{D} \mid \theta) := \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) := \mathcal{L}(\mu$$

$$\theta_{MLE} := \arg\max_\theta \, p(\mathcal{D} \mid \theta) = \arg\min_\theta \, -\log p(\mathcal{D} \mid \theta)$$

$$-\log p(\mathcal{D} \mid \theta) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Take gradients and set them to zero to compute the critical points of $-\log p(\mathcal{D} \mid \theta) := \mathcal{L}(\mu, \sigma^2)$

- $$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \Longrightarrow \frac{1}{\cancel{2}\sigma^2} \sum_{i=1}^{n} \cancel{2}(x_i - \mu) = 0 \Longrightarrow \sum_{i=1}^{n} x_i - n\mu =$$

$$\Longrightarrow \boxed{\mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

Confirm that this is a minimum!

$$\frac{\partial^2 \mathcal{L}}{\partial \mu^2} = \frac{n}{\sigma^2} > 0 \text{, hence } \mu_{MLE} \text{ is a global}$$

and unique maximizer of the likelihood.

Intro to optimization :

Setup: Given a model with parameters $\theta = (\theta_1, \dots, \theta_d)$ and a likelihood | loss $\mathcal{L}(\theta)$ our goal is to estimate θ^* such that:

$$\boxed{\theta^* = \arg\min_{\theta} \mathcal{L}(\theta)}, \quad \mathcal{L} : \mathbb{R}^d \rightarrow \mathbb{R}$$

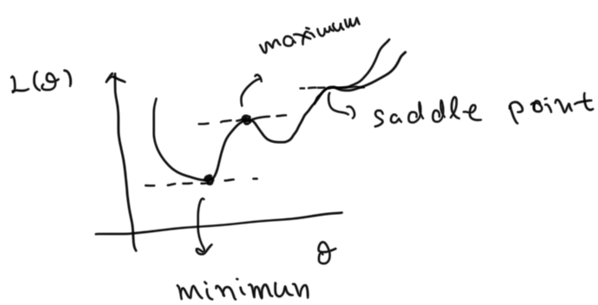
Gradient:

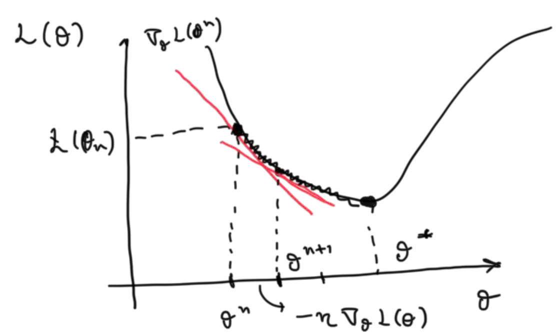
$$\underset{d \times 1}{\nabla_\theta \mathcal{L}(\theta)} = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} \\ \vdots \\ \frac{\partial L}{\partial \theta_d} \end{bmatrix}$$

We need to identify critical points:

$$\nabla_\theta \mathcal{L}(\theta) = 0$$

This condition is met in 3 different scenarios:





Gradient descent:

Starting from initial guess θ^n:

$$\rightarrow \theta^{n+1} = \theta^n - \underline{\underline{\eta}} \nabla_\theta \mathcal{L}(\theta^n)$$

Forward-Euler discretization

ODE gradient flow: $\left\{ \frac{d\theta}{dt} = -\nabla_\theta \mathcal{L}(\theta) \right.$

This is a first-order method

$\theta(0) = \theta_0$

... it relies on a linear approxima
of $\mathcal{L}(\theta)$ around θ.

η: step-size / learning rate

Hessian:

$$\nabla^2_\theta \mathcal{L}(\theta) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \theta_1^2} & \frac{\partial^2 \mathcal{L}}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \theta_1 \partial \theta_d} \\ \vdots & \ddots & & \\ & & \ddots & \frac{\partial^2 \mathcal{L}}{\partial \theta_d^2} \\ \frac{\partial^2 \mathcal{L}}{\partial \theta_d \partial \theta_1} & \cdots & \cdots & \end{bmatrix}$$

$d \times d$

Taylor expansion of $\mathcal{L}(\theta)$ around θ^n:

$$\hat{\mathcal{L}}(\theta) \approx \mathcal{L}(\theta^n) + g_n^T(\theta - \theta^n) + \frac{1}{2}(\theta - \theta^n)^T H_n (\theta - \theta_n) + \cdots$$

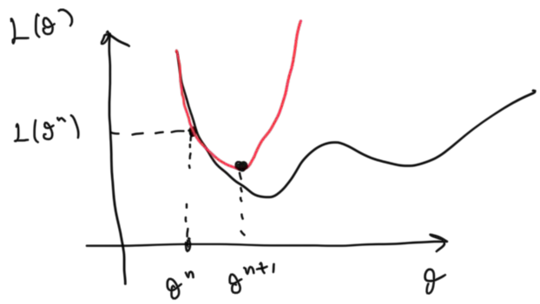
where $g_n := \nabla_\theta \mathcal{L}(\theta^n)$, $H_n := \nabla^2_\theta \mathcal{L}(\theta^n)$.

Compute critical points:

$$\nabla_\theta \hat{\mathcal{L}}(\theta) = 0 \implies g_n^T + H_n \theta^{n+1} - H_n \theta^n = 0$$

second-order method.

$$\implies \boxed{\theta^{n+1} = \theta^n - H_n^{-1} g_n^T \quad : \text{Newton's method}}$$



Pros: Utilizes the geometry of $\mathcal{L}(\theta)$ better than gradient descent.

Cons: Computationally demanding for over-parametrized model. ($d \gg 1$

e.g. in deep learning
$d \sim \mathcal{O}(10^7)$)