Maximum Likelihood Estimation

[Your Name], Lecturer

Machine Learning for Research Students

May 22, 2025

Outline

Why Estimate Parameters?

- Models depend on unknown parameters θ.
- Data-driven inference for decision-making.
- Examples: regression weights, distribution parameters.

Desirable Estimator Properties

- Consistency: $\hat{\theta}_n \to \theta_0$ as $n \to \infty$.
- Efficiency: achieving the Cramér–Rao lower bound.
- Invariance: transform estimates under reparameterization.

Different Estimation Principles

- Method of Moments
- Least Squares
- Maximum Likelihood Estimation (MLE)

Statistical Model

Assume data $\mathcal{D} = \{x_1, \ldots, x_n\}$ i.i.d. from distribution $p(x; \theta)$.

- Parameter space: $\theta \in \Theta$.
- Goal: infer θ from data.

Likelihood Function

Define the likelihood

$$L(\theta; \mathcal{D}) = \prod_{i=1}^{n} p(x_i; \theta).$$

Intuition: treat θ as variable, data fixed.

Log-Likelihood

$$\ell(\theta) = \log L(\theta; \mathcal{D}) = \sum_{i=1}^{n} \log p(x_i; \theta).$$

Simplifies optimization and numerical stability.

Maximum Likelihood Estimator

$$\hat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta \in \Theta} L(\theta; \mathcal{D}) = \arg\max \ell(\theta).$$

Interpret: best fit parameters making observed data most probable.

Graphical View

- Plot $\ell(\theta)$ vs θ.
- Maximum point gives $\hat{\theta}_{\mathrm{MLE}}$.
- Unique vs multiple peaks.

Likelihood vs Probability

- Probability: $p(x|\theta)$, x random.
- Likelihood: $L(\theta|x)$, θ variable.

Example: Bernoulli Distribution

Data $x_i \in \{0, 1\}$, parameter $\theta = P(X = 1)$.

Example: Bernoulli Distribution

Data $x_i \in \{0, 1\}$, parameter $\theta = P(X = 1)$.

$$L(\theta) = \theta^{\sum x_i}(1 - \theta)^{n - \sum x_i}.$$

Bernoulli MLE

$$\ell(\theta) = \sum x_i \log \theta + (n - \sum x_i) \log(1 - \theta),$$
$$\frac{\partial \ell}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta} = 0.$$

Solve: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Example: Gaussian Distribution

Data $x_i \sim \mathcal{N}(\mu, \sigma^2)$, both unknown.

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right).$$

Gaussian MLE: Mean

$$
\begin{aligned}
\ell(\mu, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2, \\
\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2} \sum (x_i - \mu) = 0 \\
\Rightarrow \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i.
\end{aligned}
$$

Gaussian MLE: Variance

$$
\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum (x_i - \mu)^2 = 0, \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2.
\end{aligned}
$$

Consistency

Under regularity, $\hat{\theta}_{\mathrm{MLE}} \xrightarrow{p} \theta_0$ as $n \to \infty$.

Asymptotic Normality

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0)),$$

where $I(\theta)$ = Fisher information.

Invariance Property

If $\phi = g(\theta)$, then $\hat{\phi} = g(\hat{\theta}_{\mathsf{MLE}})$.

Exponential Family

MLE has closed form solutions in many families:

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

Example: Poisson Distribution

$x_i \sim \mathrm{Pois}(\lambda)$. Likelihood:

$$L(\lambda) = \prod e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \lambda^{\sum x_i}.$$

Poisson MLE

$$\ell(\lambda) = -n\lambda + \left(\sum x_i\right) \log \lambda +,$$

$$\frac{d\ell}{d\lambda} = -n + \frac{\sum x_i}{\lambda} = 0, \quad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Fisher Information

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

Cramér–Rao Lower Bound

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}.$$

Variance Approximation

Approximate:

$$\mathrm{Var}(\hat{\theta}) \approx [-\ell''(\hat{\theta})]^{-1}.$$

Numerical Optimization

- Gradient ascent on $\ell(\theta)$.
- Newton–Raphson: $\theta_{t+1} = \theta_t - [\ell''(\theta_t)]^{-1}\ell'(\theta_t)$.

Regularization and MAP

Bayesian view: Maximum a posteriori (MAP)

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max\{\ell(\theta) + \log p(\theta)\}.$$

Practical Tips

- Check identifiability.
- Initialize carefully.
- Monitor convergence.
- Use penalized likelihood for small samples.

Summary

- MLE: intuitive, general estimation principle.
- Properties: consistency, asymptotic efficiency.
- Computation: closed-form in many cases; numerical otherwise.

References

C. Bishop, *Pattern Recognition and Machine Learning*, 2006.

G. Casella and R. Berger, *Statistical Inference*, 2002.

A. W. van der Vaart, *Asymptotic Statistics*, 1998.