Maximum Likelihood Estimation - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Maximum Likelihood Estimation (MLE) is a fundamental principle for parameter estimation in parametric statistical models. Given observations drawn from a distribution family indexed by an unknown parameter vector, MLE selects the parameter that maximizes the probability (density) of observing the data. From first principles, MLE arises from the product rule for independent samples and concavity properties of exponential families; its optimality and large-sample behavior are characterized via the *score*, *Hessian*, and *Fisher information.* We derive (i) the likelihood and log-likelihood; (ii) first- and second-order conditions; (iii) invariance and sufficiency; (iv) asymptotic consistency and normality with the Cramer–Rao lower bound (CRLB) and efficiency; (v) constrained MLE via Lagrange/KKT; and (vi) computational algorithms (Newton, Fisher scoring, EM). Throughout we use bold symbols for vectors/matrices.

1.2 Data and Notation

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the sample space and let

$$\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$$

be i.i.d. observations with common distribution $\mathbb{P}_{\boldsymbol{\theta}_0}$ from a parametric family $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$. Assume $\mathbb{P}_{\boldsymbol{\theta}}$ admits a density (or mass) function $p_{\boldsymbol{\theta}}(\mathbf{x})$ with respect to a dominating measure. Define the data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ and the parameter vector $\boldsymbol{\theta} \in \Theta$.

Dimensions. d: feature dimension; n: sample size; p: parameter dimension.

1.3 Model Formulation: Likelihood and Log-Likelihood

By independence,

$$L(\boldsymbol{\theta};\mathbf{X}) \;=\; \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i), \qquad \ell(\boldsymbol{\theta};\mathbf{X}) \;\triangleq\; \log L(\boldsymbol{\theta};\mathbf{X}) \;=\; \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i). \tag{1}$$

Maximum Likelihood Estimator.

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} \;\in\; \arg\max_{\boldsymbol{\theta}\in\Theta}\; L(\boldsymbol{\theta};\mathbf{X}) \;=\; \arg\max_{\boldsymbol{\theta}\in\Theta}\; \ell(\boldsymbol{\theta};\mathbf{X}). \tag{2}$$

1.4 Score, Hessian, and Fisher Information

Define the *score* and *Hessian*:

$$\mathbf{U}(\boldsymbol{\theta}) \;\triangleq\; \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathbf{X}) \in \mathbb{R}^p, \qquad \mathbf{H}(\boldsymbol{\theta}) \;\triangleq\; \nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\mathbf{X}) \in \mathbb{R}^{p\times p}. \tag{3}$$

The *observed information* is $-\mathbf{H}(\boldsymbol{\theta})$; the *Fisher information* (per sample) is

$$\mathcal{I}(\boldsymbol{\theta}) \;\triangleq\; \mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{1}{n}\,\mathbf{H}(\boldsymbol{\theta})\right] \;=\; \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\,\mathbf{U}(\boldsymbol{\theta})\,\mathbf{U}(\boldsymbol{\theta})^{\top}\right], \tag{4}$$

whenever differentiation under the integral is justified.

Identities (proofs). Assuming regularity (dominated convergence; parameter-independent support):

$$\mathbb{E}_{\boldsymbol{\theta}}\big[\mathbf{U}(\boldsymbol{\theta})\big] = \sum_{i=1}^{n} \int \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)\, p_{\boldsymbol{\theta}}(\mathbf{x}_i)\, d\mathbf{x}_i = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \int p_{\boldsymbol{\theta}}(\mathbf{x}_i)\, d\mathbf{x}_i = \mathbf{0}. \tag{5}$$

$$\mathbb{E}_{\boldsymbol{\theta}}\big[-\mathbf{H}(\boldsymbol{\theta})\big] = \mathbb{E}_{\boldsymbol{\theta}}\big[\mathbf{U}(\boldsymbol{\theta})\,\mathbf{U}(\boldsymbol{\theta})^{\top}\big]. \tag{6}$$

Equation (5) follows by interchanging derivative and integral; Eqn. (6) is obtained by differentiating $\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{U}(\boldsymbol{\theta})] = \mathbf{0}$ with respect to $\boldsymbol{\theta}$ and using independence. ■

1.5 First- and Second-Order Conditions; Concavity

Any interior maximizer satisfies the *likelihood equations*

$$\mathbf{U}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}, \qquad -\mathbf{H}(\widehat{\boldsymbol{\theta}}) \;\succeq\; \mathbf{0}. \tag{7}$$

If ℓ is strictly concave on Θ (e.g., canonical exponential families under full rank), then the solution to Eqn. (7) is unique and globally optimal.

1.6 Worked Examples

1.6.1 Bernoulli(p)

Data $x_i \in \{0,1\}$, $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ with $\theta \in (0,1)$.

$$\ell(\theta) = \sum_i \big[x_i \log\theta + (1-x_i)\log(1-\theta)\big], \quad \frac{d\ell}{d\theta} = \frac{\sum_i x_i}{\theta} - \frac{n - \sum_i x_i}{1-\theta}.$$

Set to zero:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i \quad \text{(sample mean)}, \qquad -\frac{d^2\ell}{d\theta^2} = \frac{\sum_i x_i}{\theta^2} + \frac{n - \sum_i x_i}{(1-\theta)^2} > 0.$$

1.6.2 Univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$

Data $x_i \in \mathbb{R}$, parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2.$$

FOC:

$$\partial_\mu \ell = \frac{1}{\sigma^2}\sum_i (x_i - \mu) = 0 \Rightarrow \widehat{\mu} = \overline{x}, \quad \partial_{\sigma^2}\ell = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (x_i - \mu)^2 = 0 \Rightarrow \widehat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2.$$

(Recall the unbiased variance uses $1/(n-1)$; MLE uses $1/n$.)

1.6.3 Multivariate Gaussian $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Data $\mathbf{x}_i \in \mathbb{R}^d$, parameters $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{S}^d_{++}$.

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \text{const}.$$

FOC yield

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i, \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^\top.$$

1.7 Invariance and Sufficiency

Invariance (exact). For any bijection $\boldsymbol{\phi} = g(\boldsymbol{\theta})$,

$$\widehat{\boldsymbol{\phi}}_{\text{MLE}} \;=\; g\Big(\widehat{\boldsymbol{\theta}}_{\text{MLE}}\Big). \tag{8}$$

Proof. By the change-of-variables formula, maximizing $\ell(\boldsymbol{\theta})$ over Θ is equivalent to maximizing the induced log-likelihood in $\boldsymbol{\phi}$; the maximizers correspond via g. ■

Sufficiency (exponential families). If $p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta}) + h(\mathbf{x})\}$, then the log-likelihood depends on data only through $\sum_i \mathbf{T}(\mathbf{x}_i)$, a sufficient statistic. Concavity of $-A(\cdot)$ often implies uniqueness of the MLE.

1.8 Asymptotic Theory: Consistency, Normality, Efficiency

Assume: (i) identifiable model; (ii) i.i.d. sampling; (iii) interior $\boldsymbol{\theta}_0$; (iv) regularity conditions ensuring LLN/CLT and interchange of differentiation and expectation.

Consistency (sketch). The average log-likelihood $n^{-1}\ell(\boldsymbol{\theta}) \to \mathbb{E}_{\boldsymbol{\theta}_0}[\log p_{\boldsymbol{\theta}}(X)]$ uniformly in probability; the limit is uniquely maximized at $\boldsymbol{\theta}_0$ by identifiability (Jensen/KL divergence). Argmax continuity yields $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Asymptotic normality (derivation). Taylor-expand the score about $\boldsymbol{\theta}_0$:

$$\mathbf{0} = \mathbf{U}(\widehat{\boldsymbol{\theta}}) = \mathbf{U}(\boldsymbol{\theta}_0) + \mathbf{H}(\tilde{\boldsymbol{\theta}})\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad \tilde{\boldsymbol{\theta}} \text{ between } \widehat{\boldsymbol{\theta}} \text{ and } \boldsymbol{\theta}_0. \tag{9}$$

Divide by n and rearrange:

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \big(-n^{-1}\mathbf{H}(\tilde{\boldsymbol{\theta}})\big)^{-1}\,\frac{1}{\sqrt{n}}\,\mathbf{U}(\boldsymbol{\theta}_0). \tag{10}$$

By LLN, $-n^{-1}\mathbf{H}(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} \mathcal{I}(\boldsymbol{\theta}_0)$; by CLT and Eqn. (6), $n^{-1/2}\mathbf{U}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)\big)$. Hence

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \;\xrightarrow{d}\; \mathcal{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\big). \tag{11}$$

Cramer–Rao Lower Bound and efficiency. For any unbiased estimator $\tilde{\boldsymbol{\theta}}$,

$$\mathrm{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}) \;\succeq\; \frac{1}{n}\mathcal{I}(\boldsymbol{\theta})^{-1}. \tag{12}$$

Moreover, Eqn. (11) shows the MLE is *asymptotically efficient*: its asymptotic covariance attains the Cramer–Rao Lower Bound (CRLB).

Wald/Score/LR tests. Let $\mathbf{I}_n(\widehat{\boldsymbol{\theta}}) = -\mathbf{H}(\widehat{\boldsymbol{\theta}})$. For $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\text{Wald: } W = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top\big(\mathbf{I}_n(\widehat{\boldsymbol{\theta}})\big)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \mathrel{\dot{\sim}} \chi^2_p,$$

$$\text{Score: } S = \mathbf{U}(\boldsymbol{\theta}_0)^\top\big(\mathbf{I}_n(\boldsymbol{\theta}_0)\big)^{-1}\mathbf{U}(\boldsymbol{\theta}_0) \mathrel{\dot{\sim}} \chi^2_p, \quad \text{LR: } \Lambda = 2\big(\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\big) \mathrel{\dot{\sim}} \chi^2_p,$$

(Wilks' theorem), where $\dot{\sim}$ denotes large-sample distribution.

1.9 Constrained MLE

For equality/inequality constraints $g(\boldsymbol{\theta}) = \mathbf{0}$, $h(\boldsymbol{\theta}) \le \mathbf{0}$, maximize $\ell(\boldsymbol{\theta})$ subject to g, h. The KKT conditions are:

$$\nabla\ell(\widehat{\boldsymbol{\theta}}) + \nabla g(\widehat{\boldsymbol{\theta}})^\top\boldsymbol{\lambda} + \nabla h(\widehat{\boldsymbol{\theta}})^\top\boldsymbol{\mu} = \mathbf{0}, \tag{13}$$

$$g(\widehat{\boldsymbol{\theta}}) = \mathbf{0}, \quad h(\widehat{\boldsymbol{\theta}}) \le \mathbf{0}, \quad \boldsymbol{\mu} \ge \mathbf{0}, \quad \boldsymbol{\mu} \odot h(\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{14}$$

In many models, reparameterization (e.g., softmax for probabilities) converts constrained problems to unconstrained ones.

1.10 Computation: Newton, Fisher Scoring, and EM

1.10.1 Newton–Raphson and Fisher Scoring

Iterate from $\boldsymbol{\theta}^{(0)}$:

$$\text{Newton–Raphson (NR):} \quad \boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \big[\mathbf{H}(\boldsymbol{\theta}^{(t)})\big]^{-1}\mathbf{U}(\boldsymbol{\theta}^{(t)}), \tag{15}$$

$$\text{Fisher Scoring (FS):} \quad \boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \big[n\,\mathcal{I}(\boldsymbol{\theta}^{(t)})\big]^{-1}\mathbf{U}(\boldsymbol{\theta}^{(t)}). \tag{16}$$

FS replaces the random Hessian by its expectation; for canonical GLMs this coincides with IRLS.

1.10.2 Expectation–Maximization (EM) for Latent-Variable Models

Suppose $\mathbf{Z}$ are latent variables with complete-data log-likelihood $\ell_c(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z})$. Define

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \;=\; \mathbb{E}\Big[\, \ell_c(\boldsymbol{\theta}) \;\Big|\; \mathbf{X},\, \boldsymbol{\theta}^{(t)} \Big].$$

EM iterates: (E) compute $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$; (M) set $\boldsymbol{\theta}^{(t+1)} \in \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. **Monotonicity (sketch).** By Jensen,

$$\ell(\boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{\theta}^{(t)}) \;\geq\; Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \;\geq\; 0.$$

Thus EM is ascent in the observed log-likelihood (with possible flat regions). ■

1.11 Existence, Uniqueness, and Regularity

- **Existence:** The MLE may fail to exist if the likelihood is unbounded (e.g., Gaussian with $\sigma^2 \downarrow 0$ around a single observation under certain mixture models) or if the feasible set is open and the maximum lies on the boundary at infinity. Compactification or penalization remedies this.
- **Uniqueness:** Guaranteed by strict concavity of ℓ (e.g., regular exponential family with full-rank sufficient statistics); otherwise multiple local maxima can occur.
- **Regularity:** Differentiability of $p_{\boldsymbol{\theta}}$, parameter-independent support, and integrability conditions ensure Eqns. (5)–(11).

1.12 Algorithmic Summary (Generic MLE)

1. **Specify model:** choose $p_{\boldsymbol{\theta}}(\mathbf{x})$ and parameter space $\Theta \subseteq \mathbb{R}^p$.
2. **Form likelihood/log-likelihood:** $L(\boldsymbol{\theta}) = \prod_i p_{\boldsymbol{\theta}}(\mathbf{x}_i)$, $\ell(\boldsymbol{\theta}) = \sum_i \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$.
3. **Differentiate:** compute score $\mathbf{U}(\boldsymbol{\theta})$ and (optionally) Hessian $\mathbf{H}(\boldsymbol{\theta})$ or Fisher information $\mathcal{I}(\boldsymbol{\theta})$.
4. **Solve likelihood equations:** set $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$; use NR/FS (or EM for latent models) until convergence.
5. **Assess uncertainty:** use $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) \approx \big[n\,\mathcal{I}(\widehat{\boldsymbol{\theta}})\big]^{-1}$ and Wald/Score/LR tests.
6. **Transformations:** any reparameterization uses the invariance property Eqn. (8).

1.13 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: ith observation (column vector), $i = 1, \ldots, n$.
- $\mathbf{X} \in \mathbb{R}^{d \times n}$: data matrix with columns $\mathbf{x}_i$.
- $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$: parameter vector; true value $\boldsymbol{\theta}_0$.
- $p_{\boldsymbol{\theta}}(\mathbf{x})$: model density/mass; $L(\boldsymbol{\theta})$, $\ell(\boldsymbol{\theta})$: likelihood/log-likelihood.
- $\mathbf{U}(\boldsymbol{\theta}) \in \mathbb{R}^p$: score; $\mathbf{H}(\boldsymbol{\theta}) \in \mathbb{R}^{p \times p}$: Hessian.
- $\mathcal{I}(\boldsymbol{\theta}) \in \mathbb{R}^{p \times p}$: Fisher information (per sample, positive semidefinite).
- $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$: maximum likelihood estimator; asymptotics per Eqn. (11).

1.14 Summary

From first principles, MLE maximizes the (log-)likelihood Eqn. (1) over the parameter space. Interior solutions satisfy the likelihood equations Eqn. (7); the score has mean zero and the Fisher information equals the variance of the score Eqns. (4)–(6). Under standard regularity, the MLE is consistent and asymptotically normal with covariance $\big(n\,\mathcal{I}(\boldsymbol{\theta}_0)\big)^{-1}$, achieving the Cramer–Rao lower bound asymptotically. The estimator is invariant to reparameterization, and in exponential families depends on data only via sufficient statistics. Computation proceeds via Newton/Fisher scoring, or EM for latent-variable models, providing a unified, principled framework for parametric estimation.