

Derivation of Maximum Likelihood Estimators (MLE)

Teaching team of STA314

Feb 9, 2026

Roadmap for today

1. MLE practice problems
2. A few general review bullet points and practice problems in advance of the midterm

MLE for Bernoulli Distribution

Problem: Flipping a coin with outcomes heads (1) and tails (0). p denotes the probability of getting head. (x_1, \dots, x_n) independent samples.

Write out likelihood, then solve for its derivative, get closed form solution.



MLE for Bernoulli Distribution

Problem: Flipping a coin with outcomes heads (1) and tails (0). p denotes the probability of getting head. (x_1, \dots, x_n) independent samples.

Likelihood:

$$L(p \mid x_1, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

Log-Likelihood:

$$\ell(p) = \sum_{i=1}^n [x_i \ln p + (1-x_i) \ln(1-p)]$$

Derivative and Solution:

$$\frac{d\ell(p)}{dp} = \sum_{i=1}^n \left(\frac{x_i}{p} - \frac{1-x_i}{1-p} \right) = 0$$

$$p^* = \frac{1}{n} \sum_{i=1}^n x_i$$

MLE for Linear Regression

Model: $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The errors are i.i.d., normally distributed with mean 0 and variance σ^2 .

Parameter: $\boldsymbol{\beta}$.

Write out likelihood using pdf of normal distribution, then take derivative with respect to $\boldsymbol{\beta}$.



MLE for Linear Regression - Likelihood

Model: $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The errors are i.i.d., normally distributed with mean 0 and variance σ^2 .

Parameter: $\boldsymbol{\beta}$.

Likelihood:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right)$$

Log-Likelihood: Taking the logarithm:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

MLE for Linear Regression - Direct Solution

Step-by-Step Derivation:

1. Focus on minimizing the sum of squared errors (OLS):

$$S(\beta) = \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \beta)^2$$

2. Take the gradient of $S(\beta)$ with respect to β :

$$\nabla_{\beta} S(\beta) = -2X^\top (y - X\beta).$$

3. Set the gradient to zero to find the optimal solution:

$$X^\top X \beta = X^\top y.$$

4. Solve for β (assuming $X^\top X$ is invertible):

$$\beta^* = (X^\top X)^{-1} X^\top y.$$

This is the MLE for linear regression.

Gradient Descent for Linear Regression

Objective: Minimize the sum of squared errors:

$$J(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \beta)^2$$

Gradient:

$$\nabla_{\beta} J(\beta) = -X^\top (y - X\beta).$$

Gradient Descent Update Rule:

$$\beta_{t+1} = \beta_t + \eta X^\top (y - X\beta_t),$$

where η is the learning rate.

Steps: 1. Initialize β_0 randomly. 2. Update β iteratively using the rule above. 3. Stop when the gradient norm is small or the change in $J(\beta)$ is negligible.

Bias/Variance:

- ▶ Metrics:
 - ▶ In regression problems, we have the expected MSE, the training MSE and the test MSE.
 - ▶ In classification problems, we have the expected error rate, the training error rate and the test error rate.
- ▶ Model selection:
 - ▶ The best model yields the smallest expected (test) MSE (error rate).
 - ▶ Among models that have similar expected MSE (error rate), we always prefer the more parsimonious one.
- ▶ Bias and variance trade-off:
 - ▶ A more complex / flexible \hat{f} has smaller bias but larger variance

Practice Question (5 min): Suppose we have a data generating process $Y = f(X) + \epsilon$, where $\mathbb{E}[\epsilon] = 0, \mathbb{E}[\epsilon^2] = \sigma^2$. For a fixed test point x_t , let $\hat{f}(x_t)$ be an estimate of $f(x_t)$ constructed from a training data set \mathcal{D} . Show

$$\mathbb{E}_{\mathcal{D}, \epsilon} \left[(Y - \hat{f}(x_t))^2 \right] = \text{Bias}[\hat{f}(x_t)]^2 + \text{Var}(\hat{f}(x_t)) + \sigma^2$$

Model Selection:

Two approaches to consider for model selection:

1. Estimate the expected MSE by “holding out” a portion of your training data for validation:
 - ▶ Validation set approach
 - ▶ Cross-validation set approach
2. Make an adjustment to the training error to penalize more complicated models:
 - ▶ AIC and BIC
 - ▶ Adjusted R^2
 - ▶ Mallows's C_p

Practice Question (5 min): Let MSE_i be the error on the i^{th} fold in cross-validation. The CV estimate is $MSE_{CV} = \frac{1}{k} \sum_{i=1}^k MSE_i$.

1. Why is the *bias* of the LOOCV estimate generally lower than that of 5-fold CV?
2. In terms of the *variance* of the estimate, $\text{Var}(MSE_{CV})$, why might LOOCV perform worse than 10-fold CV? (hint: Consider the correlation ρ between MSE_i and MSE_j).

Shrinkage Regression:

We can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates by shrinking the coefficient estimates towards zero:

- ▶ **Linear Regression** uses an ordinary least squares penalty:

$$\frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2$$

- ▶ **Ridge Regression** uses an ℓ_2 penalty term as well:

$$\frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

- ▶ **Lasso Regression** uses an ℓ_1 penalty term as well:

$$\frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Neither ridge regression nor the lasso universally dominates the other.

Practice Question (5 min): How does the bias-variance tradeoff play out in terms of comparing a Ridge and a Lasso estimator? Which method is expected to have lower variance in the presence of many irrelevant predictors, and why?

Nonlinear Regression:

To move beyond linearity, we introduced a few methods:

- ▶ Univariate ($p = 1$):
 - ▶ Polynomial Regression
 - ▶ Step Functions
 - ▶ Splines
- ▶ Multivariate ($p > 1$):
 - ▶ (Weighted) K -Nearest Neighbors
 - ▶ Local Regression
 - ▶ Generalized Additive Models

Practice Question (5 min): Consider a cubic spline $S(x)$ on an interval $[a, b]$ with K interior knots ξ_1, \dots, ξ_K .

1. Explain why a piecewise cubic polynomial with no continuity constraints would have $4(K + 1)$ parameters
2. Suppose we impose continuity constraints (i.e., continuity of the function, first derivative, and second derivative). How many degrees of freedom does the model have? (hint: think about the number of constraints we introduced)

Gradient Descent

Suppose we have an optimization problem of the form

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \mathcal{J}(\mathbf{w})$$

When we do not have a direct solution to $\frac{\partial J}{\partial \mathbf{w}} = 0$, we can use **gradient-descent** to iteratively find an optimum. At iteration $k + 1$, for each $j \in 1, \dots, p$:

$$w_j^{(k+1)} \leftarrow w_j^{(k)} - \alpha \frac{\partial J}{\partial \mathbf{w}} \bigg|_{\mathbf{w}=\mathbf{w}^{(k)}}$$

Practice Question (5 min): Let $\mathbf{g}_i = \frac{\partial \mathcal{J}_i}{\partial \mathbf{w}}$ be the gradient for observation i . Assume \mathbf{g}_i are i.i.d. with variance $\text{Var}(\mathbf{g}_i) = \sigma^2 \mathbf{I}$.

1. Show that the variance of the mini-batch update direction, $\bar{\mathbf{g}}_m = \frac{1}{m} \sum_{i \in \mathcal{B}} \mathbf{g}_i$ (where $\mathcal{B} \subset \{1, \dots, n\}$ is a mini-batch of size m), is $\frac{\sigma^2}{m} \mathbf{I}$.
2. Explain the “Computational vs. Statistical” trade-off: Why don’t we just use $m = 1$ to get the most updates possible for the same amount of compute?

Conclusion

- ▶ MLE provides a principled approach for parameter estimation.
- ▶ Some models (e.g., Bernoulli) have closed-form solutions.
- ▶ Optimization techniques can also be used for finding the MLE for methods like linear regression. This becomes a necessity for other models!
- ▶ We have covered a lot of topics! Please review lecture materials, tutorials, and problem sets in advance of the midterm.