CS109 Lecture 19: LLM Learning Guide

Maximum Likelihood Estimation

Summer 2026

Lecture 19: LLM Learning Guide

Before lecture: read the Lecture 19 slides (parameter estimation / maximum likelihood). Then open your favorite LLM and work through the concepts below **in order**. Each concept has a *Learn* prompt and a *Test me* prompt. Attempt the "Test me" questions yourself before asking the LLM to grade you.

How to get the most out of this:

- Tell the LLM your level: "I'm a student in Stanford's CS109 (probability for computer scientists). Explain at that level."
- After any answer, ask "why?" at least once and ask it to show each step.
- When it states a formula, ask it to define every symbol before continuing.
- Attempt practice yourself first, then say: "Here is my reasoning: _____. Where exactly am I wrong, if anywhere?"

Concept 1: Parametric models and θ

Learn: > "Explain what a parametric model is in probability: a family of distributions (Bernoulli, Poisson, Uniform, Normal, etc.) where fixing the parameter(s) θ pins down one actual distribution. List what θ is for each of the common distributions (p; λ; (α, β); (μ, σ^2)), note that θ can be a vector, and explain how 'parameter estimation' fits into the machine learning pipeline: model the problem, then learn θ from training data."

Test me: > "Name several real scenarios (coin flips, arrival counts, wait times, measurement noise) and make me choose a reasonable parametric model for each and identify exactly what θ is. Push back if my model choice is questionable."

Concept 2: The likelihood function

Learn: > "Define the likelihood function: for i.i.d. data $x_1, \ldots, x_n$, $L(\theta) = \prod_i f(x_i \mid \theta)$, where f is the PMF (discrete) or PDF (continuous). Emphasize that the data is fixed and θ is the variable — likelihood is 'the probability of *this* data as a function of the parameter.' Explain exactly where the i.i.d. assumption is used to turn a joint probability into a product, and compute a tiny example likelihood for two candidate values of θ."

Test me: > "Give me 3 data points from a named distribution and two candidate parameter values, and make me compute (or set up) the likelihood of the data under each and say which parameter the data supports more. Then ask me: what assumption justified multiplying, and is $L(\theta)$ a probability distribution over θ?"

Concept 3: Log-likelihood and argmax

Learn: > "Explain why we maximize the *log*-likelihood instead of the likelihood: products of many small numbers become sums (nicer calculus, no numerical underflow), and since log is monotonically increasing, $\text{argmax}_\theta LL(\theta) = \text{argmax}_\theta L(\theta)$ — prove that argmax claim. Also define argmax carefully, and show the general MLE recipe: write L, take the log, differentiate, set to 0, solve."

Test me: > "Give me a likelihood of the form $p^a(1-p)^b$ and make me: take the log, differentiate, set to zero, and solve for p. Then quiz me: why is taking the log 'legal', and give me one function where the argmax is at a boundary so setting the derivative to zero fails."

Concept 4: The MLE recipe on real distributions

Learn: > "Work the full four-step MLE derivation (likelihood, log-likelihood, differentiate, solve) for two distributions: Bernoulli (using the differentiable PMF $p^x(1-p)^{1-x}$, arriving at $\hat{p} = \frac{\sum x_i}{n}$) and Poisson (arriving at $\hat{\lambda} = \bar{x}$). Then note the pattern: for many classic distributions the MLE is a sample mean or a simple function of it — and mention the geometric ($\hat{p} = n/\sum x_i$) and exponential ($\hat{\lambda} = n/\sum x_i$) as instances I should be able to derive myself."

Test me: > "Give me a small dataset and a distribution (geometric or exponential — don't pick one whose derivation I've just seen) and make me do the entire four-step MLE derivation and compute the numeric estimate. Grade every algebra step like a CS109 grader, and then ask me to sanity-check the answer against the sample mean."

Concept 5: When calculus fails — boundaries and gradient ascent

Learn: > "Explain the two situations where the basic calculus recipe for MLE breaks down. First: likelihoods with boundary maxima, like Uni$(0, \beta)$, where $L(\beta) = \beta^{-n}$ for $\beta \geq \max x_i$, the derivative is never zero, and the MLE is $\hat{\beta} = \max_i x_i$. Second: likelihoods with no closed-form solution, where we use gradient ascent — initialize θ, repeatedly step $\theta \mathrel{+}= \eta \cdot \frac{\partial LL}{\partial \theta}$, and climb to a (local) maximum; explain the role of the step size and of convexity."

Test me: > "First give me a uniform-distribution MLE problem and check that I find the boundary maximum by reasoning rather than differentiating. Then give me a log-likelihood whose derivative can't be solved in closed form and make me write the gradient-ascent pseudocode update loop and explain what could go wrong with a too-large step size."

Concept 6: MLE vs. Bayesian estimation

Learn: > "Compare maximum likelihood estimation with the Bayesian approach from the Beta lectures, using the same data (e.g., a treatment that worked for 14 of 20 patients): the MLE $\hat{p} = 0.7$ is a *point estimate* — the θ that makes the data most likely — while the Bayesian answer is a full posterior distribution (e.g., Beta updated from a Laplace prior) that quantifies remaining uncertainty. Explain when the two roughly agree (lots of data) and when they meaningfully differ (small n, informative prior), and why MLE can badly overfit with tiny samples (e.g., 1 flip)."

Test me: > "Give me a small Bernoulli dataset and make me produce both the MLE and the posterior from a Laplace prior, then explain what each number means and which I'd report in a small-sample medical setting. Then ask me what happens to both answers as the dataset grows by a factor of 100."

Wrap-up prompt (after all six concepts)

> "Give me one multi-part CS109-style problem in which I must (1) identify a parametric model and its θ for a scenario, (2) write the likelihood and log-likelihood of a small i.i.d. dataset, (3) carry out the full calculus derivation of the MLE and compute it numerically, and (4) recognize a case where the maximum is at a boundary and explain how gradient ascent would handle a likelihood with no closed form. Let me solve every part, then grade each part, tell me which concept it tested, and tell me the single concept I should review most."

Note: MLE is the gateway to machine learning — logistic regression and deep learning (coming up next!) are 'just' maximum likelihood estimation on progressively richer models, optimized with gradient ascent. The four-step recipe you practice here is the same one running under the hood.

Bring any sticking points to lecture; the TAs and instructor will help you work through them on the worksheet.