

Lecture 8: Moments and maximum likelihood

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

High priority *From a model to an estimator*

Overview. How do we turn a statistical model into an actual estimator? This lecture compares two recurring answers: match features of the data to population moments, or choose parameters that give the observed data the largest likelihood. We begin with familiar means and variances, then see how the same ideas lead to factor models, word embeddings, topic models, and community detection. These examples connect statistical information with computation: a good estimation principle still needs a usable algorithm.

Lecture 7 showed why bias alone does not determine the quality of an estimator. Here we ask what information a method uses, whether that information identifies the target, and how hard the resulting equations are to solve. Large-sample guarantees will be developed later; today we focus on constructing estimators and understanding their limitations.

Mid priority *Estimation and prediction: an opening discussion*

Instructor's Note *Cultural difference: estimation versus prediction*

As an informal picture, statistics often starts by asking what generated the data, whereas computer science often starts by asking what predicts well. Think of these as two emphases:

- **Statistics:** modeling, assumptions about data generation, interpretation, uncertainty, and domain expertise.
- **Computer science:** prediction accuracy, willingness to use black-box models, automation, and the user interface.

The boundaries overlap, but the contrast is useful when reading papers. Leo Breiman's two-cultures article [4] invites exactly this discussion. Is there a diverging trend in recent years, or are the two approaches moving closer together?

Recap from earlier lectures

From Lectures 1, 4, 6 and 7: estimators and models

Reminder only: this block introduces no new exam requirements.

- **Bias and variance.** Mean squared error equals squared bias plus variance. Used in Section 1.1.
- **Identifiability.** Distinct parameters must give distinct model distributions if the parameter is to be recovered. Used in Section 1.1.
- **Sufficiency and factorization.** Factorization separates the data summary carrying parameter information from parameter-free factors. Used in Section 1.1 and Exercise 12.
- **KL divergence.** KL divergence is nonnegative, with equality when the two probability laws agree. Used in Section 3.1.
- **Low-rank approximation.** Leading singular directions minimize the appropriate low-rank approximation error. Used in Section 2.1.

Learning goals

- Construct moment estimators and check existence and identifiability.
- Explain how low-order moments can reveal low-dimensional structure.
- Derive likelihood estimators, including boundary and rank issues.
- Separate statistical guarantees from computational difficulty.

Priority guide

High Core concepts and calculations to master; directly examined.

Mid Supporting concepts useful in other lectures; in scope for examinations.

Low Enrichment supporting understanding; will not be examined.

Further reading Optional self-study, references, extensions, and historical notes.

A star (*) marks optional material; priority labels give the scope of each block. Lecture-specific exemptions and the optional status of Part VI override the general exam guide.

1 Method of moments

1.1 Classical recipe

High priority *Match population and empirical moments*

Let $X_1, \dots, X_n$ be i.i.d. real observations from P_θ , where $\theta \in \Theta \subset \mathbb{R}^k$, and assume $\mathbb{E}_\theta |X_1|^k < \infty$. The **population moments** and **empirical moments** are

$$\mu_j(\theta) = \mathbb{E}_\theta X_1^j, \quad \hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j, \quad 1 \leq j \leq k.$$

Write $h(\theta) = (\mu_1(\theta), \dots, \mu_k(\theta))$ and $\hat{m} = (\hat{\mu}_1, \dots, \hat{\mu}_k)$ for the vector of empirical moments. The **method of moments (MoM)** solves $h(\hat{\theta}) = \hat{m}$ for $\hat{\theta} \in \Theta$. If $h : \Theta \rightarrow h(\Theta)$ is one-to-one and the observed $\hat{m}$ lies in $h(\Theta)$, then

$$\hat{\theta} = h^{-1}(\hat{m}). \tag{1}$$

As with every estimator, the inverse or the chosen solution must be measurable as a function of the data.

What can go wrong? **Moment identification** means that different parameter values have different selected population moments: h is one-to-one. This is stronger than asking the full distributions to be different. Even an identifiable model can have an uninformative choice of moments. A sample moment vector outside $h(\Theta)$ gives no solution, and a noninjective h can give several solutions. An estimator therefore needs both a population identification argument and a rule for handling the actual data.

Looking ahead. The intuition behind **consistency** is that empirical moments approach population moments as n grows. If a law of large numbers applies, the empirical inverse is defined with probability tending to one, and h^{-1} is continuous at $h(\theta)$, the corresponding moment estimate approaches θ in probability. Lecture 9 supplies the convergence language and tools; consistency is not automatic from the name “MoM.”

Exercise prompt. Suppose the selected moment map is $h(\theta) = \theta^2$. What are the possible solutions when the observed moment is 4, first on $\Theta = \mathbb{R}$ and then on $\Theta = [0, \infty)$? What if the observed moment is negative? See Exercise 1 in Section 5 for the full statement and complete solution.

Used again in moment-estimator consistency, asymptotic normal approximations.

High priority Estimate a mean and a variance

Example 1 (Following Example 3.24 of [14]). Suppose the observations are i.i.d. with finite second moment, mean $\mu \in \mathbb{R}$, and variance $\sigma^2 > 0$. Here (μ, σ^2) names the quantities we want to estimate; it does not specify the entire distribution. The moment equations are

$$\mu_1 = \mu, \quad \mu_2 = \mu^2 + \sigma^2.$$

Replacing the moments with empirical averages gives

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2 =: v_n.$$

For $n \geq 2$, the usual sample variance is

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2, \quad v_n = \frac{n-1}{n} S^2.$$

If all observations coincide, $v_n = 0$. It is still a sensible empirical variance, but it lies on the boundary of the enlarged space $\sigma^2 \geq 0$, not inside the original space $\sigma^2 > 0$.

Remark 1. This simple example already tells us several things. First, the moment calculation needs no normality assumption. Second, in the normal family $(\bar{X}, v_n)$ is sufficient, just as $(\bar{X}, S^2)$ is. The mean estimate is unbiased; it is the variance component v_n that is biased slightly downward. Under normal sampling and $n \geq 2$, that component has smaller MSE than S^2 . Exercise 6 gives the full bias-variance calculation, including the normal fourth moments.

For other families, the mean and empirical variance need not be sufficient. For example, in Laplace (double-exponential) and logistic location models they can lose information about location; explicit examples are in Exercise 12. Losing information suggests an opportunity for a better procedure, but does not by itself prove that a particular alternative has smaller MSE at every parameter value.

High priority Estimate the endpoints of a uniform distribution

Example 2 (Following Example 3.25 of [14]). Let $X_i \stackrel{\text{iid}}{\sim} \text{Unif}([a, b])$, with $a < b$. Here $a = \theta_1$ and $b = \theta_2$ in the original notation. Then

$$\mu_1 = \frac{a+b}{2}, \quad \mu_2 = \frac{a^2 + ab + b^2}{3}, \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

The same two moment equations yield

$$\hat{a} = \bar{X} - \sqrt{3v_n}, \quad \hat{b} = \bar{X} + \sqrt{3v_n}. \quad (2)$$

Thus the estimated center is the sample mean and the estimated half-width is $\sqrt{3v_n}$. For $n \geq 2$ under this continuous model, $v_n > 0$ almost surely. The exceptional zero-variance sample again gives a boundary pair.

The extrema $(X_{(1)}, X_{(n)})$ are sufficient for (a, b) ; the moment estimate generally uses information beyond that pair while failing to exploit the support constraint efficiently. In particular, its estimated interval need not contain every observation. Unbiased endpoint estimates built from the extrema and their exact risks are derived in Exercise 2. Their total MSE is of order n^{-2} for fixed $a < b$; the moment endpoints have total MSE of order n^{-1} . The exercise justifies this rate comparison without assuming a CLT. The more detailed comparison remains a useful homework topic.

Instructor's Note *Moments as a fingerprint*

Intuitively speaking, moment matching gives the model a few clues about the data: where its center is, how spread out it is, and perhaps how asymmetric it is. We look for a parameter with the same fingerprint. But a fingerprint made from only two numbers can miss something obvious in a plot—for a uniform model, the leftmost and rightmost observations are especially revealing. For simple problems we solve the equations by hand; for complicated ones we turn to numerical routines or design new algorithms. Finding the equations is only the first step.

Exercise prompt. For the sample $(0, 0, 0, 0, 1)$, does the uniform moment interval contain all five observations? Compare what moment matching asks for with what a positive uniform likelihood requires. See Exercise 2 in Section 5 for the full statement and complete solution.

1.2 Generalized method of moments

Low priority *More moment conditions than parameters*

Let $X_i \in \mathbb{R}^p$, $\theta \in \Theta \subset \mathbb{R}^k$, and choose $g : \mathbb{R}^p \times \Theta \rightarrow \mathbb{R}^m$ so that $\mathbb{E}_{\theta_0} g(X_i, \theta_0) = 0$ at the true parameter θ_0 . The expectations must exist. For scalar data, possible coordinates are

$$g_j(x, \theta) = x^j - \mu_j(\theta), \quad \cos(jx) - V_j(\theta), \quad \text{LeakyReLU}(jx) - \xi_j(\theta),$$

where each subtracted quantity is its corresponding model expectation. LeakyReLU is simply one possible nonlinear feature.¹

The **generalized method of moments (GMM)** estimator minimizes

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} Q_n(\theta), \quad Q_n(\theta) = \bar{g}_n(\theta)^\top W \bar{g}_n(\theta), \quad \bar{g}_n(\theta) = \frac{1}{n} \sum_i g(X_i, \theta), \quad (3)$$

where W is a symmetric positive definite $m \times m$ matrix. There are m equations and k parameters; **overidentification** means $m > k$, not $m > p$. Sample equations may disagree, so GMM finds a weighted compromise. If Θ is nonempty and compact and $g(x, \theta)$ is continuous in θ for each observed x , a minimum is attained. Uniqueness and population identification require separate arguments; recall the earlier moment identification discussion.

When $m = k$ and $g_j(x, \theta) = x^j - \mu_j(\theta)$, any simultaneous zero of *all* residuals has objective value zero. Positive definiteness makes these exactly the minimizers whenever such a zero exists. Thus ordinary MoM is a special case for any positive definite W , including I_k . Exercise 7 proves this and works out a simple overidentified compromise.

Instructor's Note *Which clues should we trust?*

Imagine several measuring instruments giving slightly different answers. We want reliable instruments to have more say, and we do not want two almost identical instruments to count as two independent pieces of information. This is the role of the weighting matrix. More moments give us more clues, but a very noisy clue can make a poor guide.

For i.i.d. observations with nonsingular finite covariance $\Omega = \text{Cov}_{\theta_0}(g(X_i, \theta_0))$, the asymptotically optimal weight in the regular, correctly specified GMM setting is Ω^{-1} . This is a preview: consistency, derivative/rank assumptions, and estimation of Ω are part of the theory, not consequences of compactness alone. See [9, 8].

¹ For a fixed $0 < \alpha < 1$, $\text{LeakyReLU}(t) = \max\{t, \alpha t\}$: a line of slope 1 for positive inputs and slope α for negative inputs. Its expectation exists when $\mathbb{E}|X| < \infty$. No neural-network machinery is needed for this example.

2 Examples

Mid priority *Low-order moments and large data*

Moments are simple statistics, but they can reveal substantial structure. When the dimension is large, low-order moments combined with spectral or tensor methods may provide a fast route to useful estimates. The examples below retain the same recipe: identify a population relationship, estimate its moments, and solve the resulting algebraic problem.

2.1 Second moments and spectral methods

Mid priority *Factor model and covariance structure*

Suppose $Y_1, \dots, Y_n \in \mathbb{R}^p$ are i.i.d. and

$$Y_i = Bf_i + u_i, \quad B \in \mathbb{R}^{p \times K}, \quad \mathbb{E} f_i = 0, \quad \text{Cov}(f_i) = I_K, \quad \mathbb{E} u_i = 0, \quad \text{Cov}(u_i) = \Sigma.$$

Assume finite second moments, $\text{Cov}(f_i, u_i) = 0$, and known Σ . Independence between factors and noise would suffice for the zero cross-covariance condition. Then

$$\text{Cov}(Y_i) = BB^\top + \Sigma.$$

The covariance alone does not distinguish B from BR for an orthogonal $K \times K$ matrix R . We therefore target $L = BB^\top$ and form

$$A := \widehat{L}^0 = \frac{1}{n} \sum_i Y_i Y_i^\top - \Sigma, \quad \mathbb{E} A = L. \quad (4)$$

If Σ must also be estimated, that is an additional estimation step; the displayed unbiasedness statement assumes it is known.

Use the structure. The target is **positive semidefinite (PSD)** and has **rank** at most K . Write $A = \sum_{j=1}^p \lambda_j u_j u_j^\top$ with $\lambda_1 \geq \dots \geq \lambda_p$ and orthonormal u_j . For $1 \leq K \leq p$, a closest PSD matrix of rank at most K in Frobenius norm is

$$\widehat{L} = \sum_{j=1}^K (\lambda_j)_+ u_j u_j^\top, \quad (t)_+ = \max\{t, 0\}. \quad (5)$$

Clipping negative eigenvalues matters: retaining a negative eigenvalue would contradict the structure of $BB^\top$. Ties at the cutoff may give several equally close matrices. Exercise 3 proves the projection formula and the useful bound $\|\widehat{L} - L\|_F \leq 2\|A - L\|_F$, where $\|C\|_F^2 = \sum_{i,j} C_{ij}^2$. Projection can introduce bias; the bound is not a universal finite-sample MSE improvement theorem.

Instructor's Note *Keep the signal directions*

Think of the eigenvalues as a volume control for different directions. A factor model says only a few directions carry the common signal. The empirical matrix also contains noisy directions; spectral truncation turns those down to zero. It is the same broad instinct as regularization: use structure to avoid chasing every fluctuation in the sample.

Exercise prompt. If $A = \text{diag}(4, -1, -2)$ and $K = 2$, what is its closest PSD matrix of rank at most two? Which part of the rule is needed in addition to keeping two eigenvectors? See Exercise 3 in Section 5 for the full statement and complete solution.

Mid priority *Word embeddings*

In natural language processing, a **word embedding** represents a word or token by a low-dimensional vector. GloVe [12] begins with co-occurrence counts: X_{ij} records how often word j occurs in the context of word i . It fits a weighted least-squares model to log co-occurrence counts to learn word vectors. The connection here is the use of pairwise summaries to uncover low-dimensional structure; a co-occurrence matrix is not, in general, a sample covariance matrix.

Word / context	royal	person	travel
king	18	10	1
queen	17	11	1
train	1	2	20

Illustrative pair counts, not corpus observations: similar count patterns motivate similar representations. GloVe fits weighted log co-occurrence counts.

Linear representations of concepts. A striking empirical pattern is that some semantic relationships appear as approximately parallel vector offsets. Writing $e(w)$ for the learned embedding of word w , the famous analogy is

$$e(\text{king}) - e(\text{queen}) \approx e(\text{man}) - e(\text{woman}). \quad (6)$$

Figure 1 shows the original GloVe visualization, including these word pairs. Such **linear representations** give empirical support to the factor-model viewpoint: reusable concepts can contribute along shared directions. A useful model is $e(w) \approx Ba(w)$, where the columns of B are common semantic directions and $a(w)$ contains the word’s loadings on them. The observed offsets motivate this representation; they do not establish the specific covariance assumptions of our earlier random factor model.

Instructor’s Note *Concepts as reusable directions*

Intuitively speaking, a word vector can mix several ingredients: human, royal, male or female, and so on. Subtract queen from king and the shared royal ingredient cancels, leaving roughly the same contrast as man minus woman. This is the appeal of a factor model: a large collection of words can reuse a much smaller collection of conceptual directions. The model learns the directions from how people use words, without being handed a separate dictionary of those concepts.

A historical connection: n -grams. Earlier language models also used higher-order co-occurrence statistics. An **n -gram** is a block of n consecutive tokens, and an n -gram model predicts the next token from the preceding $n - 1$ tokens. Its block frequencies are empirical versions of the joint moments

$$\mathbb{E} \left[\prod_{j=1}^n \mathbf{1}\{W_j = a_j\} \right] = \Pr [(W_1 = a_1, \dots, W_n = a_n)].$$

Thus “ n -th moments” here means products of n token indicators, not powers of arbitrary numeric word IDs. Shannon’s successive approximations to English used letter and word transition frequencies [13]; Brown et al. [5] developed count-based and class-based n -gram models. These methods connect moment statistics to conditional-probability estimation, which we revisit in Section 3.1.

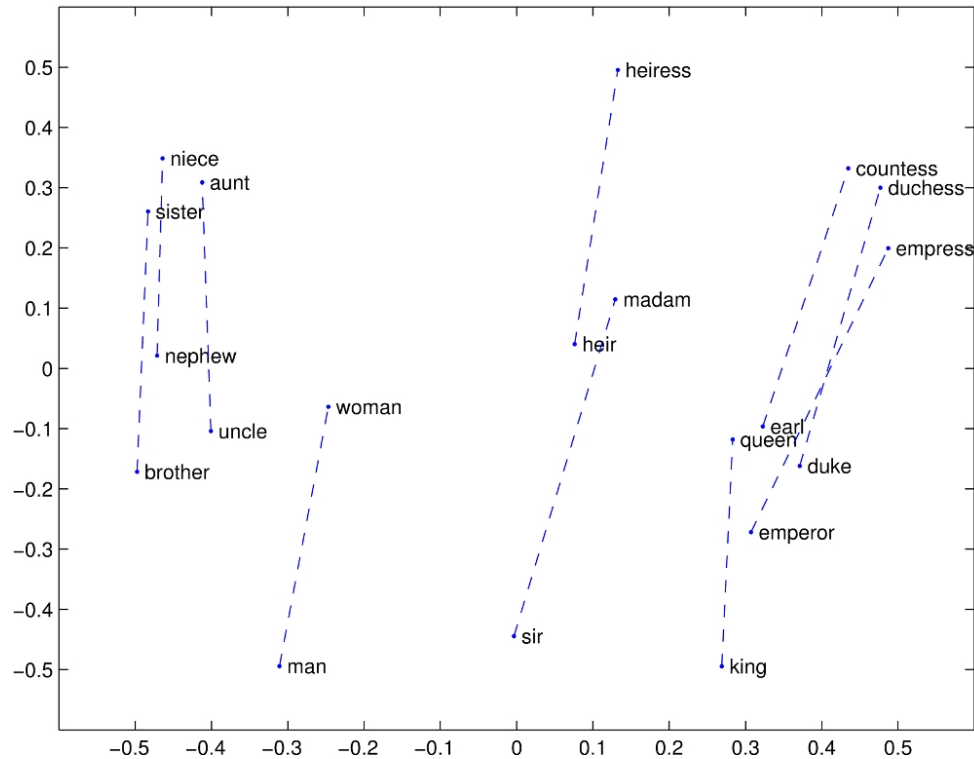
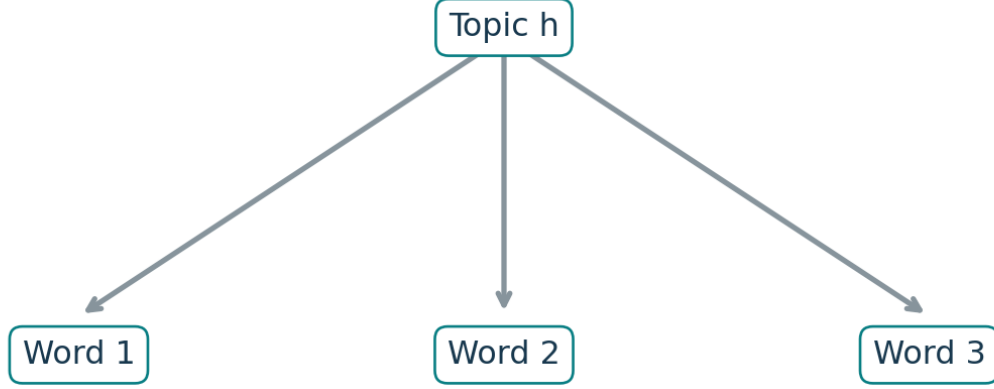


Figure 1: The GloVe project’s published two-dimensional view of learned word vectors. The king–queen and man–woman offsets point in similar directions, as do several other displayed pairs. The relationship is approximate, not an exact parallelogram identity. Figure: Jeffrey Pennington, Richard Socher, and Christopher D. Manning, GloVe project, “Linear substructures”; reproduced unchanged.

2.2 Latent variable models and tensor decomposition

Low priority *One hidden topic, several observed words*

Consider the single-topic document model of [2]. A document has an unobserved topic $h \in \{1, \dots, K\}$ with $\Pr([h = j]) = w_j > 0$ and $\sum_j w_j = 1$. Conditional on that topic, its words are **independent and identically distributed** with vocabulary probability vector $\mu_h \in \mathbb{R}^d$. Encode word i as the coordinate vector e_i ; thus $x_t = e_i$ exactly when the t -th word is i . The words are generally dependent before conditioning on h .



A shared topic can make words marginally dependent even when they are conditionally independent given that topic.

For three words from the same document, conditional independence gives

$$M_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{j=1}^K w_j \mu_j \mu_j^\top, \quad (7)$$

$$M_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{j=1}^K w_j \mu_j^{\otimes 3}. \quad (8)$$

Here M_2 is a $d \times d$ matrix, or second-order tensor, and M_3 is a $d \times d \times d$ array. The **tensor product** satisfies $[a \otimes b \otimes c]_{rst} = a_r b_s c_t$. Exercise 8 derives both identities. Independent documents, each with at least three words, provide empirical versions by averaging the displayed products across documents.

Our goal is to recover (w_j, μ_j) from (M_2, M_3) and then replace the population moments by empirical ones. For $K = 1$, $w_1 = 1$, $\mu_1 = \mathbb{E} x_1$, and $M_2 = \mu_1 \mu_1^\top$, a rank-one structure reminiscent of the factor model. For several topics, we need more algebra.

Low priority *Whitening and orthogonal tensor decomposition*

Assume $K \leq d$ and that $\mu_1, \dots, \mu_K$ are linearly independent. Then M_2 has rank K . Use its positive eigenvalues to write $M_2 = U D U^\top$, where $U \in \mathbb{R}^{d \times K}$ has orthonormal columns and D is a positive diagonal $K \times K$ matrix. Define

$$W = U D^{-1/2}, \quad W^\top M_2 W = I_K, \quad v_j = \sqrt{w_j} W^\top \mu_j. \quad (9)$$

This rescaling is called **whitening**. The K vectors v_j form an orthonormal basis of $\mathbb{R}^K$. The square-root weights are essential: the unweighted vectors $W^\top \mu_j$ do not generally have unit length.

Transform all three indices of M_3 with W :

$$T := M_3[W, W, W] = \sum_{j=1}^K \lambda_j v_j^{\otimes 3}, \quad \lambda_j = w_j^{-1/2}. \quad (10)$$

For example, $T_{abc} = \sum_{r,s,t} (M_3)_{rst} W_{ra} W_{sb} W_{tc}$. Once the orthogonal tensor components are found, the parameters are recovered, up to relabeling, by

$$w_j = \lambda_j^{-2}, \quad \mu_j = \lambda_j U D^{1/2} v_j.$$

The complete algebra is in Exercise 8.

A **tensor power iteration** starts with a unit vector $u \in \mathbb{R}^K$ and repeatedly normalizes the contraction

$$T(I, u, u) = \sum_j \lambda_j (v_j^\top u)^2 v_j, \quad u_{\text{new}} = \frac{T(I, u, u)}{\|T(I, u, u)\|_2}$$

when the denominator is nonzero. This echoes the matrix power method. Recovering all components uses additional starting points and removal of already found components. The analysis of those steps and of noisy empirical moments is in Sections 4–5 of [2]; the population recovery formula is Theorem 4.3. A single arbitrary starting vector is not a complete recovery guarantee.

Instructor's Note *Unmix before reading the recipe*

Think of each hidden topic as a recipe for choosing words. We only see batches of words produced by an unknown recipe. Pair counts tell us how to stretch and rotate the coordinates so that the hidden directions separate. Triple counts then reveal how much of each recipe was mixed in. The hard work is finding the population mapping; once it is understood, we feed empirical moments into the same construction.

3 Maximum likelihood estimation

High priority *Fix the data and vary the parameter*

Suppose Y has a joint density or pmf f_θ relative to a fixed measure that does not depend on θ . The **likelihood** is $L(\theta; y) = f_\theta(y)$, viewed as a function of θ with y fixed. A **maximum likelihood estimator (MLE)** is a measurable choice

$$\hat{\theta}(y) \in \arg \max_{\theta \in \Theta} L(\theta; y). \quad (11)$$

The maximizing set may be empty or contain several points. The likelihood is not a probability distribution over parameters. Because the logarithm is increasing, we can instead maximize the **log-likelihood** $\ell(\theta; y) = \log L(\theta; y)$, with $\log 0 = -\infty$.

Used again in likelihood consistency, likelihood asymptotics, Fisher information.

High priority *Stationary points and boundaries*

For a differentiable log-likelihood, an interior maximizer satisfies $\nabla_\theta \ell(\theta; y) = 0$. This **first-order condition** finds candidates, not a guarantee of a global maximum. We still need to check curvature or compare objective values, inspect the boundary, and ask whether the maximum is attained. Parameter-dependent support, as in the uniform family, makes a boundary check especially important.

For constrained problems, KKT conditions organize the corresponding stationarity calculations.² STAT 710 studies likelihood theory in more detail, including consistency and asymptotic efficiency under suitable identification and regularity conditions.

² The Karush–Kuhn–Tucker (KKT) conditions combine derivatives with constraint multipliers and complementary slackness: a nonbinding inequality has zero multiplier. Under suitable constraint qualifications they are necessary at a local optimum. Sufficiency for a global optimum requires additional structure, such as the appropriate convexity or concavity. We do not develop constrained optimization here.

Exercise prompt. For i.i.d. Bernoulli observations with parameter $\theta \in [0, 1]$, where is the likelihood maximized if every observation is zero? What changes if the parameter space is $(0, 1)$? See Exercise 4 in Section 5 for the full statement and complete solution.

High priority *Gaussian likelihood and least squares*

Example 3. Let $Y_i = x_i^\top \theta + \varepsilon_i$, where $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ is fixed and known, $\theta \in \mathbb{R}^d$, and the errors are i.i.d. $N(0, \sigma^2)$ with known $\sigma^2 > 0$. Equivalently, $Y \sim N(X\theta, \sigma^2 I_n)$. Then

$$\ell(\theta; y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^\top \theta)^2. \quad (12)$$

Thus every least-squares minimizer is an MLE. If X has full column rank, the unique estimator is

$$\hat{\theta} = (X^\top X)^{-1} X^\top y.$$

If it is rank deficient, the coefficient estimate is not unique, but the fitted vector $X\hat{\theta}$ is unique. Exercise 9 gives the density calculation and the projection argument for both cases.

Instructor's Note *Read the score in the other direction*

A probability model usually starts with a parameter and tells us what data to expect. Likelihood reads that relationship backward: the data are already on the desk, so we ask which parameter gives them the best score. For Gaussian noise, that score falls as squared residuals grow. This is why least squares appears naturally here.

See Lecture 13: Fisher information as likelihood curvature for the same width comparison in information notation.

3.1 MLE and next-token prediction

Mid priority *Start with multinomial maximum likelihood*

Before considering language, suppose $N \geq 1$ i.i.d. categorical observations take values in a finite alphabet $\mathcal{V}$ of size V . Their probability vector is $p = (p_v)_{v \in \mathcal{V}}$ in the closed **probability simplex** $\Delta_V = \{p : p_v \geq 0, \sum_v p_v = 1\}$. Let C_v count category v , so $(C_v)_{v \in \mathcal{V}} \sim \text{Multinomial}(N, p)$. For observed counts c_v , put $\hat{p}_v = c_v/N$. The likelihood is

$$L(p; c) = \frac{N!}{\prod_v c_v!} \prod_v p_v^{c_v}, \quad \ell(p; c) = A(c) + N \sum_v \hat{p}_v \log p_v, \quad (13)$$

where $A(c) = \log(N! / \prod_v c_v!)$ does not depend on p . We use $p_v^0 = 1$, including $0^0 = 1$ in this likelihood. If $c_v > 0$ and $p_v = 0$, the likelihood is zero and its logarithm is $-\infty$.

For probability vectors r, p , define the **cross-entropy**

$$H(r, p) = - \sum_{v: r_v > 0} r_v \log p_v. \quad (14)$$

It is $+\infty$ if $p_v = 0$ for a category with $r_v > 0$; zero-weight terms contribute zero. All logarithms in this subsection are natural. Equation (13) gives the exact identity

$$-\frac{1}{N} \ell(p; c) = H(\hat{p}, p) - \frac{A(c)}{N}.$$

Therefore **maximizing the multinomial likelihood is exactly minimizing the cross-entropy from the empirical distribution to the model**. The multinomial coefficient changes the objective's value, not its optimizer. If we observe an ordered categorical sample rather than its count vector, that coefficient is absent, and the same optimizer results.

Why the optimizer is the empirical distribution. For a fixed probability vector r , write

$$H(r, p) = H(r) + D_{\text{KL}}(r \| p), \quad H(r) = - \sum_{v: r_v > 0} r_v \log r_v,$$

where $H(r)$ is its **entropy** and $D_{\text{KL}}(r \| p) = \sum_{v: r_v > 0} r_v \log(r_v/p_v)$ is the **Kullback–Leibler (KL) divergence**. To check nonnegativity without overlooking zero counts, let $S = \{v : r_v > 0\}$. If p vanishes anywhere in S , the divergence is infinite. Otherwise $-\log u \geq 1 - u$ for $u > 0$ gives

$$D_{\text{KL}}(r \| p) \geq \sum_{v \in S} r_v \left(1 - \frac{p_v}{r_v}\right) = 1 - \sum_{v \in S} p_v \geq 0. \quad (15)$$

The scalar inequality follows by differentiating $u - 1 - \log u$, whose minimum is zero at $u = 1$. Equality in the divergence requires $p_v/r_v = 1$ throughout S and no probability outside S ; equivalently $p = r$. Consequently the unique multinomial MLE over Δ_V is $\hat{p}_v = c_v/N$, even when some counts are zero. If one insists on strictly positive probabilities, zero counts can instead lead to an unattained supremum at the boundary.

This elementary calculation is the statistical foundation of the loss used below. For textbook treatments, see Manning, Raghavan, and Schütze, Section 12.1.3, and Goodfellow, Bengio, and Courville, Sections 5.5–5.5.1 [7].

Mid priority *Replace one probability vector by a conditional one*

A **token** is one symbol in a language model’s vocabulary: a word, a word fragment, or punctuation, for example. For an observed sequence $w_1, \dots, w_T$, let $w_{<t} = (w_1, \dots, w_{t-1})$ denote the prefix before position t . An **autoregressive language model** gives a categorical distribution $q_\theta(v \mid w_{<t})$ over the next token $v \in \mathcal{V}$. For every prefix, these probabilities are nonnegative and sum to one. A fixed **BOS** token supplies the first context; BOS is the standard shorthand for *beginning of sequence* (here, a sentence). We include an end-of-sequence token among the predicted tokens so that stopping is part of the model.

Often a neural network gives a score $s_\theta(v, h)$ to each candidate v in context h , and **softmax** turns the scores into probabilities:

$$q_\theta(v \mid h) = \frac{\exp s_\theta(v, h)}{\sum_{u \in \mathcal{V}} \exp s_\theta(u, h)}.$$

The parameter vector θ is shared across all positions and contexts. Neural probabilistic language models already used learned word representations and conditional likelihood in [3]; the basic estimation principle does not depend on a particular neural network architecture.

One sequence gives many likelihood factors. The probability chain rule gives

$$q_\theta(w_{1:T}) = \prod_{t=1}^T q_\theta(w_t \mid w_{<t}), \quad -\log q_\theta(w_{1:T}) = \sum_{t=1}^T -\log q_\theta(w_t \mid w_{<t}). \quad (16)$$

We don’t need any assumption about independence, since the decomposition follows from the definition of conditional probability. A model may restrict the context it uses, as an n -gram model does, but the full-prefix chain rule itself is exact. For the toy sequence “the cat sleeps” followed by an end marker, the four prediction tasks are

Observed prefix	Target token	Contribution to negative log-likelihood
BOS	the	$-\log q_\theta(\text{the} \mid \text{BOS})$
the	cat	$-\log q_\theta(\text{cat} \mid \text{the})$
the cat	sleeps	$-\log q_\theta(\text{sleeps} \mid \text{the cat})$
the cat sleeps	end marker	$-\log q_\theta(\text{end} \mid \text{the cat sleeps})$

BOS is implicit in the last three prefixes. Training uses the observed preceding tokens, often called **teacher forcing**; during generation, later predictions instead condition on the tokens generated so far. We score every target position, not merely the final token of the sequence.

Mid priority *The autoregressive loss is conditional cross-entropy*

Suppose a training corpus consists of m sequences, modeled as independent, with lengths $T_1, \dots, T_m$ including the end markers. Let $M = \sum_i T_i$ be the number of predicted tokens. The token-averaged negative log-likelihood is

$$\mathcal{L}(\theta) = -\frac{1}{M} \sum_{i=1}^m \sum_{t=1}^{T_i} \log q_\theta(w_{it} | w_{i,<t}). \quad (17)$$

Maximizing the corpus likelihood is equivalent to minimizing this loss, since taking a negative logarithm reverses the optimization and M is fixed by the corpus. This is an average over tokens; giving every sequence equal weight regardless of length would generally be a different objective.

To identify each summand as cross-entropy, let y_{it} be the probability vector with mass one on the observed target: $y_{it}(v) = \mathbf{1}\{v = w_{it}\}$. Then

$$H(y_{it}, q_\theta(\cdot | w_{i,<t})) = -\sum_v y_{it}(v) \log q_\theta(v | w_{i,<t}) = -\log q_\theta(w_{it} | w_{i,<t}).$$

Thus the common next-token cross-entropy loss is exactly the MLE objective, written as one categorical prediction problem at every position.

What conditional probability are we learning? At the population level, fix a position t and regard $w_{<t}$ and w_t as random tokens drawn from the text distribution. Write $p_*(v | w_{<t})$ for the true conditional probability that the next token w_t equals v , given the prefix $w_{<t}$. Applying the previous entropy identity at each prefix yields

$$\begin{aligned} \mathbb{E}[-\log q_\theta(w_t | w_{<t})] &= \mathbb{E}_{w_{<t}} H(p_*(\cdot | w_{<t}), q_\theta(\cdot | w_{<t})) \\ &= \mathbb{E}_{w_{<t}} H(p_*(\cdot | w_{<t})) \\ &\quad + \mathbb{E}_{w_{<t}} D_{\text{KL}}(p_*(\cdot | w_{<t}) \| q_\theta(\cdot | w_{<t})). \end{aligned}$$

Over all conditional probability distributions, the minimum is attained by $q_\theta(\cdot | w_{<t}) = p_*(\cdot | w_{<t})$ on prefixes with positive probability. A restricted model family tries to approximate this target from finite data. The target is the distribution of text, so there can be several plausible continuations rather than one universally correct next token. Exercise 13 gives a finite count example, the one-token gradient, and a numerical check of the summed loss. Further background is in Deep Learning, Chapter 10 and [3].

Low priority *The miracles of model scaling*

Instructor's Note *One small task, a very large education*

The surprising part is how much can grow out of the same simple assignment: predict the next token. Imagine a student asked to fill in every missing word in an enormous library. A good guess in a story may require tracking the characters; in a scientific explanation it may require knowing the subject; in a proof it may require following the argument. The assignment stays the same, but the variety of situations keeps expanding.

This is the intuition behind the “miracles of model scaling.” With more data and a larger model, countless small prediction tasks can build shared representations useful for many bigger tasks. A prompt changes the context: a question invites an answer, a few worked examples suggest a pattern, and a partial program invites its continuation. Conditional probability is the common language connecting these activities.

Empirically, Kaplan et al. [11] found systematic, approximately power-law trends in held-out cross-entropy as model size, data, and training compute grow. Hoffmann et al. [10] showed the importance of scaling the training data along with the model under a fixed compute budget. Brown et al. [6] demonstrated that a large autoregressive model could use examples supplied in its prompt across many tasks without updating its weights for each task. These are experimental findings in the studied regimes, not a theorem that every capability improves indefinitely. Lower prediction loss is useful evidence of better text modeling; it does not guarantee that every generated statement is true.

3.2 MLE and computational complexity

Mid priority *Community detection from a network*

Suppose n is even and a graph has two equally sized communities. For example, the nodes may represent students and an edge may record an interaction. Write $z_i^* \in \{+1, -1\}$ for the unknown memberships, with $\sum_i z_i^* = 0$. Conditional on these fixed memberships, edges $(Y_{ij})_{i < j}$ are independent, with

$$Y_{ij} \sim \begin{cases} \text{Bern}(p^*) & z_i^* = z_j^*, \\ \text{Bern}(q^*) & z_i^* \neq z_j^*, \end{cases} \quad 0 < q^* \leq p^* \leq \frac{1}{2}.$$

They need not be identically distributed across all pairs. This is a balanced two-community **stochastic block model**. Its parameter space is

$$\Theta = \{(p, q, z) : 0 < q \leq p \leq \frac{1}{2}, \quad z \in \{\pm 1\}^n, \quad z^\top \mathbf{1}_n = 0\}.$$

Swapping the community names changes z to $-z$ without changing the model, so recovery is always understood up to this global sign.

Mid priority *Write the likelihood before choosing an algorithm*

Recall the earlier likelihood definition: we hold the observed graph fixed while varying its candidate parameters. For a balanced candidate z , let $s_w(z)$ and $s_b(z)$ count observed edges within and between its communities. The numbers of possible pairs are $N_w = n(n-2)/4$ and $N_b = n^2/4$. Then

$$L(p, q, z; Y) = p^{s_w} (1-p)^{N_w - s_w} q^{s_b} (1-q)^{N_b - s_b}, \quad (18)$$

$$\ell(p, q, z; Y) = C_n(p, q) + s \log \frac{q}{1-q} + s_w(z) \log \frac{p(1-q)}{(1-p)q}, \quad (19)$$

where $s = s_w + s_b = \sum_{i < j} Y_{ij}$ is fixed by the data and $C_n(p, q) = N_w \log(1-p) + N_b \log(1-q)$. The pair counts and all algebra are derived in Exercise 5.

For **fixed** $p > q$, the coefficient of s_w is positive. Maximizing the likelihood over balanced labels therefore amounts to

$$\max_{z \in \{\pm 1\}^n : z^\top \mathbf{1}_n = 0} s_w(z). \quad (20)$$

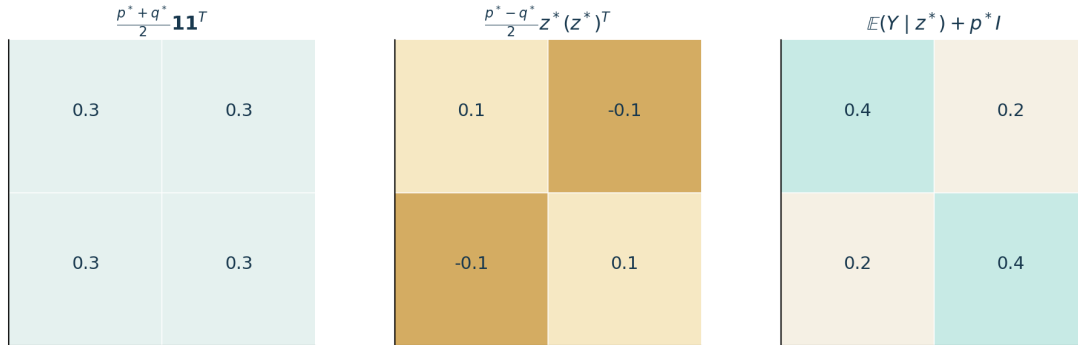
Equivalently, minimize the number of edges crossing an equal-sized split: this is **minimum bisection**. It is not max-cut. Removing the balance condition from this within-edge objective would give the trivial all-in-one-group solution. If $p = q$, every balanced labeling has the same likelihood and the graph contains no information about the memberships.

If p, q are also unknown, we can first maximize over the finite set of balanced labelings and then optimize over p, q . Because $q > 0$ defines an open boundary, the latter optimization may have a supremum without an attained maximum. Exercise 5 writes this profile expression explicitly; one should not assume the joint MLE always exists.

Exercise prompt. Make Y symmetric with zero diagonal. How can $s_w(z)$ be expressed using the total edge count s and the quadratic form $z^\top Y z$? Keep track of the fact that the matrix counts each edge twice. See Exercise 5 in Section 5 for the full statement and complete solution.

Low priority *Minimum bisection and spectral relaxation*

Exact minimum bisection is computationally hard in the worst case (see Abbe, Bandeira, and Hall [1]). It is an NP-hard problem.³ Statistical structure can nevertheless make recovery possible. The cited paper establishes exact-recovery results and efficient algorithms in specified parameter regimes; it does not give a blanket guarantee for any simple spectral rounding rule.



For $p^* = 0.4, q^* = 0.2$, the constant and community components add to $\mathbb{E}(Y | z^*) + p^* I_n$. Subtract $p^* I_n$ to restore the zero diagonal. This diagonal correction shifts eigenvalues without changing eigenvectors.

Since the objective in (20) differs from $z^\top Y z/4$ only by a constant, a natural **spectral relaxation** is

$$\max_{v^\top \mathbf{1}_n = 0, \|v\|_2^2 = n} v^\top Y v. \quad (21)$$

We relax the discrete coordinates but **keep balance**. Choose $Q \in \mathbb{R}^{n \times (n-1)}$ with orthonormal columns spanning $\mathbf{1}_n^\perp$. Find a unit top eigenvector u of $Q^\top Y Q$ and set $v = \sqrt{n} Q u$. Equivalently, work with HYH restricted to $\mathbf{1}_n^\perp$, where $H = I_n - \mathbf{1}_n \mathbf{1}_n^\top / n$. Do not choose the unrestricted top eigenvector of Y , which often primarily reflects average degree.

To round back to a balanced labeling, give $+1$ to the $n/2$ largest coordinates of v and -1 to the rest, resolving ties by vertex index. A sign threshold alone need not produce equal-sized groups. This is a candidate algorithm, not an exact solution to every minimum-bisection instance. Exercise 10 explains the relaxation and shows why the population community direction appears after centering.

Instructor's Note *Hope is not lost*

Imagine two groups of students who interact more within their own group than across groups. Individual friendships are noisy clues, but the whole network can reveal a direction separating the groups. The discrete search asks us to try many assignments; the spectral picture first asks for a continuous direction and then divides the students along it. A hard optimization problem can have an easier statistical version when the data carry a strong enough pattern.

Used again in optional spectral community detection.

³ NP-hard describes worst-case computational difficulty: a general polynomial-time exact algorithm would solve every problem in the class NP in polynomial time. It does not mean every instance is hard, or that a random graph model cannot admit a fast method with high success probability.

3.3 Comparing MoM with MLE

Mid priority *Information, computation, and reliability*

MoM is flexible: we can specify moment relationships without a full parametric distribution, and the resulting equations may be simple to solve. A few moments, however, may discard useful shape or support information. MLE uses a full probability model and, in regular correctly specified families, has strong large-sample efficiency properties.⁴ Neither the existence of a maximizer nor an efficient numerical method comes for free.

High moments can be particularly unstable under heavy tails. Although $\mathbb{E}|X|^j < \infty$ makes the j -th empirical moment meaningful, its variance needs a finite $2j$ -th moment. Thus adding more equations can add more noise. Robustness to misspecification, estimation accuracy, and computational cost all belong in the comparison.

Instructor's Note *Use both tools*

Moment estimates can give us a quick first sketch; likelihood can then refine it using a more detailed model. One practical strategy is to use a moment estimate as the starting value for a likelihood optimization. The two methods need not be rivals. Historically, the study of measurement errors by Laplace, Gauss, and Legendre helped connect probability with estimation and least squares. Fisher's later work on likelihood gave the comparison with moments a more systematic foundation. The story grew in stages, much as these methods do in this lecture.

Key ideas

- Moment methods match empirical summaries to population moments and require identification.
- Maximum likelihood optimizes a data-dependent function of the parameter.
- Multinomial negative log-likelihood is empirical cross-entropy up to a positive factor and an additive constant.
- Existence, boundary behavior and computation matter as much as a formal estimating equation.

⁴ Asymptotic efficiency compares the leading large-sample variance with an appropriate lower bound, among estimators satisfying the relevant regularity conditions. It is not a claim of smallest MSE at every finite sample size, and it does not cover every irregular or misspecified model. The formal theory belongs to STAT 710.

Detailed reference

Notion	Working meaning	Priority
Method of moments	Match empirical and population features; solve for the parameter.	High
Moment identification	Selected moments distinguish parameter values.	High
GMM	Minimize a weighted quadratic moment discrepancy.	Low
Spectral estimation	Use PSD and low-rank structure to constrain an estimate.	Mid
Whitening	Rescale the second moment to the identity on its range.	Low
Likelihood / MLE	Fix the data; score and maximize over parameters.	High
First-order condition	Necessary stationarity for a differentiable interior optimum.	High
Multinomial / cross-entropy	MLE minimizes cross-entropy from empirical category frequencies to the model.	Mid
Autoregressive model	Factor a sequence probability into next-token conditional probabilities.	Mid
Next-token loss	Average one categorical log loss per target token, with shared parameters.	Mid
Stochastic block model	Independent edges with probabilities set by community memberships.	Mid
Minimum bisection	Split equally while minimizing crossing edges.	Mid
Spectral relaxation	Replace discrete labels by a constrained continuous direction.	Low

Looking ahead

Lecture 9 asks whether the estimators constructed here approach their population targets as the sample grows. Laws of large numbers and continuity connect empirical objectives with their population counterparts. Continue to Lecture 9.

You have finished Part II—check yourself. Preview Part III.

4 Further reading

Further reading *Sources and optional algebra*

For the classical examples, see [14], Examples 3.24–3.25. Breiman [4] discusses the two cultures. Hansen [9] and Hall [8] develop GMM; Hansen’s original paper and overview provide further entry points. For history, compare the original notes’ normal-distribution history link with Fisher’s 1922 *On the Mathematical Foundations of Theoretical Statistics*.

For modern applications, see the GloVe paper, Anandkumar et al. (Section 3.1 for the topic model; Section 4.3 for whitening and parameter recovery), and Abbe, Bandeira, and Hall [1] for computational and exact-recovery results in the block model.

Language models and scaling. For the historical count-based view, see Shannon (1948), Section 3, and Brown et al. (1992), Section 2. For learned embeddings and conditional log-likelihood, see Bengio et al. (2003), Section 2. The textbook derivation of maximum likelihood, cross-entropy and conditional likelihood is in [7], Sections 5.5–5.5.1; Chapter 10 treats sequence modeling. The scaling and prompting papers are Kaplan et al. (2020), Hoffmann et al. (2022), and Brown et al. (2020).

Tensor coordinates and change of basis. A tensor $A \in \mathbb{R}^{n_1 \times \cdots \times n_r}$ is an array with expansion

$$A = \sum_{i_1=1}^{n_1} \cdots \sum_{i_r=1}^{n_r} A_{i_1 \dots i_r} e_{i_1} \otimes \cdots \otimes e_{i_r}.$$

For matrices $V_\ell \in \mathbb{R}^{n_\ell \times m_\ell}$, define the transformed array of size $m_1 \times \cdots \times m_r$ by

$$(A[V_1, \dots, V_r])_{j_1 \dots j_r} = \sum_{i_1=1}^{n_1} \cdots \sum_{i_r=1}^{n_r} A_{i_1 \dots i_r} \prod_{\ell=1}^r (V_\ell)_{i_\ell j_\ell}.$$

The old indices are summed out; the new indices remain. For $r = 2$ this is $V_1^\top A V_2$. An SVD $A = U S V^\top$ therefore gives $A[U, V] = S$: in these orthogonal coordinates, the map simply rescales along singular directions. For rectangular A , S is rectangular too. Using the reduced SVD $A = U_r D V_r^\top$ with the r positive singular values, let $\tilde{U} = U_r D^{-1/2}$ and $\tilde{V} = V_r D^{-1/2}$. Then

$$\tilde{U}^\top A \tilde{V} = I_r.$$

The inverse square roots, not square roots, perform whitening. This is on the rank- r subspace; zero singular values cannot be inverted. See Exercise 11 for the complete calculation.

Convex relaxation. Another approach to hard optimization is to keep a larger, convex feasible set. For balanced bisection, write $Z = z z^\top$. The constraints imply $Z \succeq 0$, $Z_{ii} = 1$, $Z \mathbf{1}_n = 0$, and rank one. Dropping the rank-one constraint gives the semidefinite program

$$\max_Z \text{tr}(Y Z) \quad \text{subject to} \quad Z \succeq 0, \quad Z_{ii} = 1, \quad Z \mathbf{1}_n = 0.$$

A semidefinite program optimizes a linear function of a matrix subject to linear constraints and positive semidefiniteness. Its value bounds the discrete optimum from above; rounding still requires analysis. The celebrated semidefinite approximation for max-cut, mentioned in the original notes, is a related example for a different cut objective.

5 Exercises and complete solutions

The first five exercises correspond to the short prompts in the lecture. The remaining exercises supply calculations and proof routes used above. All solutions are included; try the questions before reading them.

High priority *Moment identification*

Exercise 1. Solve $h(\theta) = m$ for $h(\theta) = \theta^2$, with $m = 4$ and $m < 0$, on $\Theta = \mathbb{R}$ and on $\Theta = [0, \infty)$. What happens at $m = 0$? Explain why a model can be identifiable while this moment is not.

Solution. On $\mathbb{R}$, $m = 4$ gives $\theta = \pm 2$; on $[0, \infty)$ only 2 remains. Negative m gives no real solution in either space, and $m = 0$ gives $\theta = 0$. On $[0, \infty)$ the inverse is $\sqrt{m}$ for $m \geq 0$; on $\mathbb{R}$ no single inverse exists. For a concrete identifiable model, $X \sim N(\theta, 1)$ distinguishes different means, but the selected moment $\mathbb{E}(X^2) - 1 = \theta^2$ cannot distinguish θ and $-\theta$. Using the first moment instead would identify the mean. An empirical second moment minus one can be negative, explaining the feasibility issue in a genuine moment equation. $\square$

High priority *Uniform endpoints: moment and support information*

Exercise 2. (a) Calculate the moment interval for $(0, 0, 0, 0, 1)$. (b) Derive (2). (c) For $n \geq 2$, find unbiased estimators of a, b from the extrema and calculate their total MSE. Justify the rate comparison with MoM.

Solution. (a) Here $\bar{X} = 1/5$ and $v_n = 1/5 - 1/25 = 4/25$, so

$$[\hat{a}, \hat{b}] = [(1 - 2\sqrt{3})/5, (1 + 2\sqrt{3})/5].$$

The upper endpoint is approximately $0.893 < 1$. Thus moment matching can assign zero likelihood to the observed sample, since a positive uniform likelihood requires $a \leq X_{(1)} \leq X_{(n)} \leq b$.

(b) The center $c = (a + b)/2$ is μ_1 , and the positive half-width $r = (b - a)/2$ satisfies $r^2 = 3(\mu_2 - \mu_1^2)$. Substituting the empirical moments gives $c = \bar{X}$ and $r = \sqrt{3v_n}$, proving the formula.

(c) Put $d = b - a$, $U = (X_{(1)} - a)/d$, and $V = (X_{(n)} - a)/d$. For uniforms on $[0, 1]$, $\Pr[(U > t) = (1 - t)^n$ and $\Pr[(V \leq t) = t^n$. Integrating their densities gives

$$\mathbb{E}U = \frac{1}{n+1}, \quad \mathbb{E}V = \frac{n}{n+1}, \quad \text{Var}(U) = \text{Var}(V) = \frac{n}{(n+1)^2(n+2)}.$$

For completeness, their joint density is $n(n-1)(v-u)^{n-2}$ on $0 < u < v < 1$. Conditional on $V = v$, the other $n-1$ observations are uniform on $[0, v]$, so $\mathbb{E}(U | V = v) = v/n$. Since $\mathbb{E}V^2 = n/(n+2)$, this yields $\mathbb{E}UV = 1/(n+2)$ and $\text{Cov}(U, V) = 1/[(n+1)^2(n+2)]$. Consequently

$$\tilde{a} = \frac{nX_{(1)} - X_{(n)}}{n-1}, \quad \tilde{b} = \frac{nX_{(n)} - X_{(1)}}{n-1}$$

are unbiased, and each has variance

$$\frac{d^2}{(n-1)^2} \frac{n^3 + n - 2n}{(n+1)^2(n+2)} = \frac{nd^2}{(n-1)(n+1)(n+2)}.$$

Their total MSE is twice this expression, of order n^{-2} .

To compare rates without a CLT, let $r = d/2 > 0$ and $\hat{r} = \sqrt{3v_n}$. The sum of the two squared endpoint errors is $2(\bar{X} - c)^2 + 2(\hat{r} - r)^2$. Its expectation is at least $2\text{Var}(\bar{X}) = d^2/(6n)$. For an upper bound,

$$(\hat{r} - r)^2 = \frac{9(v_n - d^2/12)^2}{(\hat{r} + r)^2} \leq \frac{9}{r^2}(v_n - d^2/12)^2.$$

Write $Z_i = X_i - c$ and $v_n = n^{-1} \sum Z_i^2 - \bar{Z}^2$. The first term has variance $\text{Var}(Z_1^2)/n$. Also, independence and $\mathbb{E}Z_i = 0$ give

$$\mathbb{E}\bar{Z}^4 = \frac{n\mathbb{E}Z_1^4 + 3n(n-1)(\mathbb{E}Z_1^2)^2}{n^4} = O(n^{-2}).$$

Using $(u - v)^2 \leq 2u^2 + 2v^2$, we obtain $\mathbb{E}(v_n - d^2/12)^2 = O(n^{-1})$. All moments here are finite because $|Z_i| \leq r$. Thus the total MoM MSE is bounded above and below by positive constants times $1/n$ for fixed $a < b$. This is a rate comparison, not a claim of a particular finite- n risk ratio. $\square$

Mid priority PSD low-rank approximation and the factor covariance

Exercise 3. (a) Apply (5) to $A = \text{diag}(4, -1, -2)$, $K = 2$. (b) Verify the factor-model covariance and unbiasedness of A . (c) Prove the closest-matrix formula and the stated factor-two error bound.

Solution. (a) The answer is $\text{diag}(4, 0, 0)$. Merely retaining the first two eigenvalues gives $\text{diag}(4, -1, 0)$, which is not PSD. Clipping removes the retained negative eigenvalue. Rank at most two allows the resulting estimate to have rank one.

(b) Expanding $\mathbb{E}YY^\top$ gives $B\mathbb{E}ff^\top B^\top + \mathbb{E}uu^\top$ plus the two cross terms. Those vanish by the assumed zero means and cross-covariance. Therefore $\mathbb{E}YY^\top = BB^\top + \Sigma$ and linearity of expectation gives $\mathbb{E}A = BB^\top$.

(c) If $C \succeq 0$ has rank at most K , write its eigenvalues as $c_1 \geq \dots \geq c_p \geq 0$, with $c_j = 0$ for $j > K$, and eigenvectors v_j . We need the trace bound $\text{tr}(AC) \leq \sum_j \lambda_j c_j$. Here is a direct justification. For any r orthonormal vectors v_j ,

$$\sum_{j=1}^r v_j^\top A v_j = \sum_i \lambda_i a_i \leq \sum_{i=1}^r \lambda_i, \quad a_i = \sum_{j=1}^r (u_i^\top v_j)^2 \in [0, 1], \quad \sum_i a_i = r.$$

The inequality follows by moving weight toward the larger ordered λ_i . Multiplying these bounds by $c_r - c_{r+1} \geq 0$ (with $c_{p+1} = 0$) and summing over r gives the trace bound. It follows that

$$\|A - C\|_F^2 = \text{tr}(A^2) + \text{tr}(C^2) - 2\text{tr}(AC) \geq \sum_j (\lambda_j - c_j)^2.$$

The minimum occurs at $c_j = (\lambda_j)_+$ for $j \leq K$ and zero otherwise. Aligning eigenvectors with those of A attains the bound. Since the true L is an admissible competitor, $\|A - \hat{L}\|_F \leq \|A - L\|_F$. The triangle inequality yields $\|\hat{L} - L\|_F \leq \|\hat{L} - A\|_F + \|A - L\|_F \leq 2\|A - L\|_F$. The earlier rotation ambiguity follows from $(BR)(BR)^\top = BRR^\top B^\top = BB^\top$. $\square$

High priority *A maximum can lie on the boundary or fail to exist*

Exercise 4. For Bernoulli data with $s = \sum_i X_i$, maximize the likelihood over $[0, 1]$. Discuss both constant samples and the open parameter space $(0, 1)$.

Solution. For $0 < \theta < 1$, $L(\theta) = \theta^s(1 - \theta)^{n-s}$. If $s = 0$, $L = (1 - \theta)^n$ decreases, so its unique maximizer on $[0, 1]$ is 0. If $s = n$, $L = \theta^n$ increases and the unique maximizer is 1. The endpoint values are understood directly from the Bernoulli model, without differentiating a logarithm at zero. For $0 < s < n$, the derivative of the log-likelihood is $s/\theta - (n - s)/(1 - \theta)$, which vanishes at s/n . Its second derivative is $-s/\theta^2 - (n - s)/(1 - \theta)^2 < 0$; the endpoints have likelihood zero, so s/n is the unique maximizer. Thus $\hat{\theta} = \bar{X}$ in all cases on $[0, 1]$. On $(0, 1)$ this remains an MLE for mixed samples, but a constant sample has only a supremum of 1 approached at an excluded endpoint. There is no MLE then. $\square$

Mid priority *Network likelihood and balanced cuts*

Exercise 5. Derive the pair counts, likelihood, and profile expression in the balanced block model. Express $s_w(z)$ in terms of $z^\top Y z$ and s . Explain the cases $p > q$ and $p = q$.

Solution. Each group has $n/2$ nodes, so $N_w = 2\binom{n/2}{2} = n(n - 2)/4$ and $N_b = (n/2)^2$. Multiplying independent Bernoulli probabilities gives (18). Taking logarithms gives

$$N_w \log(1 - p) + N_b \log(1 - q) + s_w \log \frac{p}{1 - p} + s_b \log \frac{q}{1 - q}.$$

Substitute $s_b = s - s_w$ to obtain (19). Let $\mathcal{Z} = \{z \in \{\pm 1\}^n : z^\top \mathbf{1}_n = 0\}$ and $c(p, q) = \log[p(1 - q)/((1 - p)q)] \geq 0$. Then

$$\sup_{(p, q, z) \in \Theta} \ell(p, q, z; Y) = \sup_{0 < q \leq p \leq 1/2} \left\{ C_n(p, q) + s \log \frac{q}{1 - q} + c(p, q) \max_{z \in \mathcal{Z}} s_w(z) \right\}.$$

This equality concerns optimal values; it does not assert attainment at an open boundary. For example, the empty graph has likelihood increasing to 1 as $p = q \downarrow 0$, with no maximum at allowed positive p, q .

Since each edge appears twice in the matrix product, $z^\top Y z = 2 \sum_{i < j} z_i z_j Y_{ij}$. Hence

$$s_w(z) = \sum_{i < j} \frac{1 + z_i z_j}{2} Y_{ij} = \frac{s}{2} + \frac{1}{4} z^\top Y z.$$

For $p > q$, $c(p, q) > 0$, so maximizing likelihood maximizes s_w and minimizes $s_b = s - s_w$ under balance. For $p = q$, $c = 0$ and every labeling has the same likelihood. Dropping balance from s_w allows all labels equal, placing every observed edge within one group; that trivial maximization is not the max-cut problem. $\square$

High priority *Normal sample variance: a small bias can help*

Exercise 6. For i.i.d. $N(\mu, \sigma^2)$ observations and $n \geq 2$, compute the bias, variance, and MSE of $v_n = (n - 1)S^2/n$ and compare with S^2 .

Solution. Put $Z = (X - \mu \mathbf{1})/\sigma \sim N(0, I_n)$. Choose an orthogonal matrix whose first row is $\mathbf{1}^\top/\sqrt{n}$. Its transformed coordinates $Z'_1, \dots, Z'_n$ are independent standard normals, since the joint Gaussian density remains proportional to $e^{-\|z\|^2/2}$ under this orthogonal change of variables. Removing the first coordinate gives $nv_n/\sigma^2 = \sum_{j=2}^n (Z'_j)^2$. For a standard normal G with density ϕ , $\phi'(t) = -t\phi(t)$. Integration by parts gives $\mathbb{E} G^2 = -\int t\phi'(t)dt = 1$ and $\mathbb{E} G^4 = -\int t^3\phi'(t)dt = 3\mathbb{E} G^2 = 3$; the boundary terms vanish exponentially. Thus $\text{Var}(G^2) = 2$, and independence yields

$$\mathbb{E} v_n = \frac{n-1}{n}\sigma^2, \quad \text{Var}(v_n) = \frac{2(n-1)}{n^2}\sigma^4.$$

Its bias is $-\sigma^2/n$ and

$$\text{MSE}(v_n) = \frac{2n-1}{n^2}\sigma^4, \quad \text{MSE}(S^2) = \frac{2}{n-1}\sigma^4.$$

Their difference, in the order S^2 minus v_n , is $\sigma^4(3n-1)/[n^2(n-1)] > 0$. This comparison uses normal fourth moments; finite second moments alone do not give these MSE formulas. For any i.i.d. finite-variance sample, the identity $\sum (X_i - \bar{X})^2 = \sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2$ still gives $\mathbb{E} S^2 = \sigma^2$. $\square$

Low priority *GMM as an exact match or a compromise*

Exercise 7. (a) Prove that a nonempty set of zeros of $\bar{g}_n$ is exactly the minimizer set for positive definite W . (b) For a sample mean $\bar{x}$, take $\bar{g}_n(\theta) = (\bar{x} - \theta, 2\bar{x} - 2\theta + 1)$ and $W = \text{diag}(1, w)$, $w > 0$, on $\Theta = \mathbb{R}$. Find the GMM minimizer. Is this toy set of moment conditions correctly specified?

Solution. (a) Positive definiteness gives $r^\top W r \geq 0$, with equality exactly when $r = 0$. An existing zero reaches the global minimum zero, and any minimizer must be another zero. This requires all coordinates to vanish. (b) Let $t = \theta - \bar{x}$. Then $Q = t^2 + w(1-2t)^2$ has derivative $2t - 4w(1-2t)$ and positive second derivative $2+8w$. Its unique minimizer is $\hat{\theta} = \bar{x} + 2w/(1+4w)$. The two sample equations ask for incompatible values $\bar{x}$ and $\bar{x} + 1/2$; the estimate moves toward the second as w increases. The incompatibility also holds at the population level, so this is deliberately a misspecified toy objective, not an example of consistent GMM under correct specification. It illustrates compromise, not an optimal weighting theorem. $\square$

Low priority *Topic moments and the missing square-root weights*

Exercise 8. Derive the topic moments and all whitening/recovery formulas. Explain why $K \leq d$ alone is not enough for this construction.

Solution. Given $h = j$, $\mathbb{E}(x_t | h = j) = \mu_j$. Conditional independence makes each entry of $\mathbb{E}(x_1 \otimes x_2 | h = j)$ equal to the product of two corresponding entries of μ_j , and likewise for three entries. Averaging over h gives (7)–(8).

Let $M = [\mu_1, \dots, \mu_K]$ have full column rank and $D_w = \text{diag}(w_1, \dots, w_K)$. Then $M_2 = M D_w M^\top$ has rank K and range equal to the column space of M . For its reduced eigendecomposition, $W = U D^{-1/2}$ satisfies $W^\top M_2 W = I_K$. Let $V = [v_1, \dots, v_K] = W^\top M D_w^{1/2}$. We get $V V^\top = I_K$. Since V is square, it is invertible with $V^{-1} = V^\top$, so also $V^\top V = I_K$: the columns are orthonormal. The factor $\sqrt{w_j}$ is exactly what makes this identity work.

Transformation of each rank-one third-order tensor gives

$$M_3[W, W, W] = \sum_j w_j (W^\top \mu_j)^{\otimes 3} = \sum_j w_j^{-1/2} v_j^{\otimes 3}.$$

Therefore $w_j = \lambda_j^{-2}$. Since $U U^\top \mu_j = \mu_j$, $U D^{1/2} v_j = \sqrt{w_j} \mu_j$, proving $\mu_j = \lambda_j U D^{1/2} v_j$. Contracting two tensor indices with u gives $T(I, u, u) = \sum_j \lambda_j (v_j^\top u)^2 v_j$ as claimed. These statements verify the population algebra, not the convergence of a chosen numerical iteration. If two topics have identical probability vectors,

their columns are dependent even when $K \leq d$; the rank drops and this K -dimensional whitening construction fails. $\square$

High priority *Gaussian least squares and rank deficiency*

Exercise 9. Derive the Gaussian log-likelihood and characterize its maximizers for both full-column-rank and rank-deficient X .

Solution. Independence gives $L(\theta; y) = (2\pi\sigma^2)^{-n/2} \exp\{-\|y - X\theta\|_2^2/(2\sigma^2)\}$. Taking logs proves (12), including the factor $1/\sigma^2$. Let p be the orthogonal projection of y onto the column space of X . For any θ , the two terms in $y - X\theta = (y - p) + (p - X\theta)$ are orthogonal. Thus

$$\|y - X\theta\|_2^2 = \|y - p\|_2^2 + \|p - X\theta\|_2^2.$$

The minimum is attained exactly when $X\theta = p$, proving uniqueness of fitted values and existence of a minimizer even with deficient rank. Equivalently the normal equations are $X^\top X\theta = X^\top y$. Full column rank gives the stated inverse formula; otherwise the full solution set is $\theta_0 + \ker X$ for any one solution θ_0 . $\square$

Low priority *The population community direction*

Exercise 10. Solve the balanced spectral relaxation and compute its population operator under the block model. Explain why unrestricted eigenvectors and unbalanced rounding can give the wrong interpretation.

Solution. Every $v \perp \mathbf{1}_n$ can be written $v = Qa$, with $\|a\| = \|v\|$. Diagonalize the symmetric matrix $Q^\top YQ$. Its quadratic form is maximized on the sphere of radius $\sqrt{n}$ by a top eigenvector scaled by $\sqrt{n}$, proving the recipe after (21). This argument does not require the largest eigenvalue to be positive.

For fixed balanced z^* , write $a = (p^* + q^*)/2$ and $b = (p^* - q^*)/2$. Directly checking off-diagonal and diagonal entries gives

$$P := \mathbb{E}(Y \mid z^*) = a\mathbf{1}_n\mathbf{1}_n^\top + bz^*(z^*)^\top - p^*I_n.$$

The constant vector has eigenvalue $na - p^*$; the community vector z^* has eigenvalue $nb - p^*$. Every vector orthogonal to both has eigenvalue $-p^*$. For $p^* > q^*$, the community direction is the largest one *inside* $\mathbf{1}_n^\perp$, with gap $nb > 0$ when that subspace has other directions. In the full space, the constant direction is larger because $a > b$ when $q^* > 0$.

The centered matrix HPH has eigenvalue zero on $\mathbf{1}_n$. If $nb - p^* < 0$, that zero could outrank the community eigenvalue, so even for the centered matrix we must explicitly restrict to $\mathbf{1}_n^\perp$. A balanced real vector can also have unequal counts of positive and negative coordinates. Selecting the largest $n/2$ coordinates enforces the required balance; arbitrary sign rounding does not. Sampling noise and the error introduced by rounding require further analysis before any exact-recovery claim. $\square$

Further reading *Tensor transformations and inverse square roots*

Exercise 11. Verify the dimensions in the tensor transformation, its matrix form, and the reduced-SVD whitening identity.

Solution. In each factor $(V_\ell)_{i_\ell j_\ell}$ the row index runs from 1 to n_ℓ and is contracted with the old tensor index; the column index runs from 1 to m_ℓ and remains free. The output therefore has shape $m_1 \times \cdots \times m_r$. In the two-index case,

$$(A[V_1, V_2])_{jk} = \sum_{s,t} (V_1)_{sj} A_{st} (V_2)_{tk} = (V_1^\top A V_2)_{jk}.$$

If $A = USV^\top$ with orthogonal U, V , multiplication gives $U^\top A V = S$. For the reduced SVD $A = U_r D V_r^\top$,

$$(U_r D^{-1/2})^\top A (V_r D^{-1/2}) = D^{-1/2} U_r^\top U_r D V_r^\top V_r D^{-1/2} = I_r.$$

Using positive square roots instead would give D^2 , not I_r . Keeping only positive singular values makes every inverse legitimate. $\square$

High priority *Two moments can miss shape information*

Exercise 12. Find two samples with the same mean and empirical variance but a likelihood ratio that depends on the location parameter, first for a Laplace location family and then for a logistic location family, both with scale fixed at one. Explain the implication for sufficiency.

Solution. Take samples $x = (-1, -1, 2)$ and $y = (-\sqrt{3}, 0, \sqrt{3})$. Both have sum zero and sum of squares six, so both have mean zero and $v_3 = 2$. For Laplace density $f_\mu(t) = \frac{1}{2}e^{-|t-\mu|}$, their likelihood ratio is

$$\frac{L_x(\mu)}{L_y(\mu)} = \exp \left\{ -\sum_i |x_i - \mu| + \sum_i |y_i - \mu| \right\}.$$

At $\mu = 0$ this equals $e^{-4+2\sqrt{3}} \neq 1$, while for $\mu > 2$ both sums equal 3μ and the ratio is 1.

For the logistic density $f_\mu(t) = e^{t-\mu}/(1+e^{t-\mu})^2$, set $s = e^{-\mu} > 0$. Since both sample sums are zero, the ratio is

$$\frac{L_x(\mu)}{L_y(\mu)} = \left\{ \frac{\prod_i (1 + se^{y_i})}{\prod_i (1 + se^{x_i})} \right\}^2.$$

It cannot be constant: each polynomial has constant term 1, but its roots are the multiset $\{-e^{-x_i}\}$ or $\{-e^{-y_i}\}$, and these multisets differ (the first has a repeated root and the second does not). So the polynomials are not identical or proportional.

To see that the continuous-data issue is not confined to two points, write a nonconstant sample of size three as $\bar{x}\mathbf{1} + r(\cos\phi e + \sin\phi f)$, where e, f are an orthonormal basis of $\mathbf{1}^\perp$ and $r^2 = 3v_3 > 0$. In these polar coordinates, the volume factor is proportional to r and does not depend on ϕ or μ . Conditional on $(\bar{X}, v_3)$, the angle therefore has density proportional to the joint sample density. The displayed samples lie on the same circle and give a ratio that varies with μ . Continuity makes this variation persist on arcs and for nearby centers and radii, a set of statistics with positive probability. Thus the conditional angle distribution depends on μ , proving that the first two empirical moments are not sufficient in these location families. Sufficiency for a larger location-scale family would imply sufficiency for its fixed-scale subfamilies, so it fails there too. No finite-sample MSE dominance follows just from this observation. $\square$

Mid priority *Categorical counts and an autoregressive loss*

Exercise 13. (a) For counts $(3, 1, 0)$ in three categories, write the multinomial likelihood, its empirical cross-entropy, and its MLE on the closed simplex. What happens if all model probabilities must be positive? (b) Interpret these counts as next-token counts after a single context. If a second context has counts $(0, 0, 2)$, write the token-averaged loss for the combined six observations. (c) A four-token sequence, including its end marker, has conditional probabilities 0.6, 0.5, 0.25, 0.8 for its observed targets. Compute its likelihood, total negative log-likelihood, and average cross-entropy. (d) For one target w and softmax scores (s_v) , differentiate its loss with respect to s_v .

Solution. (a) The likelihood is

$$L(p) = \frac{4!}{3!1!0!} p_1^3 p_2 = 4p_1^3 p_2, \quad H(\hat{p}, p) = -\frac{3}{4} \log p_1 - \frac{1}{4} \log p_2.$$

There is no contribution from category 3 because its count is zero. The closed-simplex MLE is $(3/4, 1/4, 0)$, by the proof in Section 3.1. With strict positivity, the empirical vector is excluded. For any $0 < \varepsilon < 1$, set $p^{(\varepsilon)} = ((1-\varepsilon)3/4, (1-\varepsilon)/4, \varepsilon)$. Its likelihood approaches the maximum as $\varepsilon \downarrow 0$, but no positive vector attains it. The supremum is not a maximum.

(b) Let the two contexts be h_1, h_2 . Their separate empirical vectors are $\hat{p}_{h_1} = (3/4, 1/4, 0)$ and $\hat{p}_{h_2} = (0, 0, 1)$. The loss is

$$\begin{aligned}\mathcal{L}(\theta) &= -\frac{1}{6}\{3 \log q_\theta(1 \mid h_1) + \log q_\theta(2 \mid h_1) + 2 \log q_\theta(3 \mid h_2)\} \\ &= \frac{4}{6}H(\hat{p}_{h_1}, q_\theta(\cdot \mid h_1)) + \frac{2}{6}H(\hat{p}_{h_2}, q_\theta(\cdot \mid h_2)).\end{aligned}$$

The weights are $4/6$ and $2/6$, not equal weights for the two contexts. Separate unconstrained context tables attain both empirical vectors; a shared parameterization may not be able to choose both independently. A context never observed has zero weight in this empirical objective, so the likelihood alone does not determine its prediction.

(c) The chain rule gives likelihood $0.6 \cdot 0.5 \cdot 0.25 \cdot 0.8 = 0.06$. The total negative log-likelihood is $-\log 0.06 \approx 2.813411$, also equal to the sum of the four negative log probabilities. The average loss is $-\log 0.06/4 \approx 0.703353$ in natural-log units per token. No independence assumption between successive tokens is needed.

(d) Since $q_v = e^{s_v} / \sum_u e^{s_u}$, the one-target loss is $-s_w + \log \sum_u e^{s_u}$. Therefore

$$\frac{\partial(-\log q_w)}{\partial s_v} = q_v - \mathbf{1}\{v = w\}.$$

Gradient descent raises the observed target’s score and lowers the other scores when regarded as free scores. In a neural model, the chain rule propagates these contributions to the shared parameters, and every token in the loss supplies such a contribution. $\square$

References

- [1] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.
- [2] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, Matus Telgarsky, et al. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.
- [3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.
- [4] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- [5] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992.
- [6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [8] A.R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.
- [9] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.
- [10] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

- [11] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [12] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [13] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.
- [14] Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 2003.

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 14, 2026. Added issue-reporting links to the webpage and PDF. The mathematics and priority assignments are unchanged in this update.