CMPUT 654: Mathematical Foundations of Modern AI Systems Fall 2026

Lecture 3: Probabilistic models, log loss, and compression

September 8, 2026

Lecturer: Csaba Szepesvári *Note prepared by: Csaba Szepesvári*

About these lecture notes. These notes return to the problem introduced in Lecture 1: building a machine that generates natural language. Lecture 2 studied deterministic labels. Here we allow the same context to have different continuations and ask what a probability model should learn from their frequencies. This leads to proper losses, log loss, statistical efficiency, total variation, betting, and compression.

Scope relative to class. The lecture developed the categorical model, empirical frequencies, proper losses, the log-loss and squared-loss regret identities, the statistical efficiency claim, Pinsker's inequality, and the betting and compression interpretations of log loss. The meeting ended during the compression discussion. The class left the regret identities, the properties of KL, Pinsker's inequality, and the Bregman construction as exercises; see Exercises 2 to 4. These notes make the notation precise and complete the compression accounting. The appendices state regularity conditions behind the efficiency result, give a worked example, and discuss its relevance to logistic regression and large language models. They are reference material rather than additional claims about what was covered at the board.

Reading guide. Read through Section 3.7. The exercises develop the calculations that were left open in class. The appendices in Appendices A and B are optional technical references. Concepts and bibliographic remarks appear at the end.

3.1 From deterministic labels to probability distributions

Lecture 2 considered a target function $f : [N] \to \{0, 1\}$. Once a box index x was fixed, its label $f(x)$ was fixed. Natural language does not have this property. A prefix such as

"The meeting begins at"

can be followed naturally by different times, phrases, or punctuation. A language generator must represent more than one acceptable continuation. We begin with a smaller problem than language. There are N possible outcomes, indexed by $[N] = \{1, \dots, N\}$. We identify a function $q : [N] \to \mathbb{R}$ with the vector $(q(1), \dots, q(N))$, so $\mathbb{R}^{[N]} \equiv \mathbb{R}^N$. Under this identification, the set of probability distributions on $[N]$ is the simplex

$$\Delta_N = \left\{ q \in \mathbb{R}^{[N]} : q(x) \geq 0 \text{ for every } x, \quad \sum_{x=1}^{N} q(x) = 1 \right\} \subseteq \mathbb{R}^N. \tag{3.1}$$

Let $p \in \Delta_N$ be the data distribution and suppose

$$X_1, \dots, X_T \overset{\text{iid}}{\sim} p. \tag{3.2}$$

The empirical distribution records relative frequencies:

$$\widehat{p}_T(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{X_t = x\}, \qquad x \in [N]. \tag{3.3}$$

For each x, the law of large numbers gives

$$\widehat{p}_T(x) \longrightarrow p(x) \qquad \text{almost surely as } T \to \infty. \tag{3.4}$$

When N is small and outcomes repeat, the empirical distribution solves the frequency-matching problem directly.

The problem changes when the outcome space or the collection of contexts is large. Many relevant events will occur rarely or not at all. We then introduce a structured model

$$\{p_\theta : \theta \in \Theta\}, \tag{3.5}$$

where one parameter vector controls probabilities at many outcomes or contexts. A transformer reuses the same weights to produce a next-token distribution at every context. The question is how to choose θ from data.

For text, the categorical problem is repeated conditionally. At a prefix $x_{<t} = (x_1, \ldots, x_{t-1})$, a language model reports

$$p_\theta(\cdot \mid x_{<t}) \in \Delta_N.$$

Multiplication gives the autoregressive sequence model

$$p_\theta(x_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}). \tag{3.6}$$

Three terms commonly used for this setup describe different aspects of the learning problem.

Definition 3.1.1 (unsupervised, self-supervised, and generative modeling)**. Unsupervised learning**

The training data are observations without separately supplied task labels. The learner seeks regularities in the distribution of those observations.

Self-supervised learning

The prediction targets are constructed from the observations themselves. For next-token prediction, a sequence $x_{1:T}$ supplies the input–target pairs $(x_{<t}, x_t)$; no additional annotation is required.

Generative modeling

The model specifies a probability distribution from which observations can in principle be sampled. The factorization in Eq. (3.6) defines such a distribution over complete token sequences.

Next-token language modeling is called self-supervised because the tokens provide their own prediction targets, and generative because the fitted conditionals define a joint distribution that can generate sequences. It has also traditionally been called unsupervised because it requires no separately annotated labels. The terms overlap, but they answer different questions: self-supervised describes how targets are obtained, while generative describes what kind of model is learned. The practice of fitting probability models to observations has a long history in statistics and predates this terminology by many decades.

3.2 Losses that recover the probabilities

We still need a rule for choosing a distribution from the observations. In a structured model, this means choosing θ and thereby choosing p_θ. A loss assigns a numerical assessment to each candidate distribution and observed outcome; averaging these assessments gives an objective that can be minimized over q, or over $q = p_\theta$.

The numerical assessment must depend on two things: the distribution we chose and the outcome that occurred. The first is under our control; the second is supplied by the data. The unknown data distribution p cannot be an input to the learning algorithm.

Definition 3.2.1 (loss for categorical probability prediction)**.** A loss for categorical probability prediction is a function

$$\ell : \Delta_N \times [N] \to \mathbb{R} \cup \{+\infty\}. \tag{3.7}$$

Its value $\ell(q, x)$ measures the loss incurred when we choose $q \in \Delta_N$ and observe $x \in [N]$.

Given observations $X_1, \dots, X_T$, empirical risk minimization chooses

$$\widehat{q}_T \in \operatorname*{argmin}_{q \in \Delta_N} \frac{1}{T} \sum_{t=1}^{T} \ell(q, X_t). \tag{3.8}$$

Let X have distribution p. For any fixed q for which the loss has finite expectation, the law of large numbers gives

$$\frac{1}{T} \sum_{t=1}^{T} \ell(q, X_t) \longrightarrow \mathbb{E}[\ell(q, X)] \qquad \text{almost surely.} \tag{3.9}$$

Before asking how fast Eq. (3.8) converges, ask whether the population objective asks for the correct distribution. With a mild abuse of the same notation, write

$$\ell(q, p) = \mathbb{E}[\ell(q, X)] = \sum_{x=1}^{N} p(x)\ell(q, x). \tag{3.10}$$

Thus $\ell(q, x)$ is the loss on one realized outcome, whereas $\ell(q, p)$ is its expectation when the outcome has distribution p.

Definition 3.2.2 (proper and strictly proper loss)**.** A loss ℓ is *proper* if

$$\ell(p, p) \le \ell(q, p) \qquad \text{for all } p, q \in \Delta_N.$$

It is *strictly proper* if equality is possible only when $q = p$.

Properness says that truthful probability reporting minimizes expected loss. Strict properness identifies the entire distribution, not only its most likely outcome. This is the formal version of matching natural frequencies.

Two important examples are the *log loss*

$$\ell_{\log}(q, x) = -\log q(x) \tag{3.11}$$

and the *squared probability loss*, also called the multiclass *Brier loss*,

$$\ell_{\mathrm{B}}(q, x) = \sum_{j=1}^{N} \big(q(j) - \mathbb{I}\{x = j\}\big)^2 = \|q - e_x\|_2^2. \tag{3.12}$$

Here $e_x \in \mathbb{R}^N$ is the vector with a one in coordinate x and zeros elsewhere.

Unless a base is displayed, log denotes the natural logarithm. Log loss measured with the natural logarithm has units of *nats*; log loss measured with $\log_2$ has units of *bits*. One nat equals $\log_2(e) \approx 1.4427$ bits. We set $-\log 0 = +\infty$.

For $p, q \in \Delta_N$, define

$$H(p) = -\sum_{x=1}^{N} p(x) \log p(x), \qquad D_{\mathrm{KL}}(p\|q) = \sum_{x=1}^{N} p(x) \log \frac{p(x)}{q(x)}. \tag{3.13}$$

Terms with $p(x) = 0$ are defined to be zero, including when $q(x) = 0$. A term with $p(x) > 0$ and $q(x) = 0$ is $+\infty$. Thus $D_{\mathrm{KL}}(p\|q)$ is finite exactly when $q(x) > 0$ for every x with $p(x) > 0$. Relative entropy is not symmetric: in general, $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$.

The logarithm in Eq. (3.11) may initially look arbitrary. The next identity establishes that it is a valid loss for probability estimation.

Proposition 3.2.3 (regret identities)**.** *For every $p, q \in \Delta_N$,*

$$\ell_{\log}(q, p) = H(p) + D_{\mathrm{KL}}(p\|q), \tag{3.14}$$

$$\ell_{\mathrm{B}}(q, p) = 1 - \|p\|_2^2 + \|p - q\|_2^2, \tag{3.15}$$

Consequently, log loss and Brier loss are strictly proper.

The terms $H(p)$ and $1 - \|p\|_2^2$ do not depend on the report q. The avoidable population losses, or *regrets*, are therefore

$$\ell_{\log}(q, p) - \ell_{\log}(p, p) = D_{\mathrm{KL}}(p\|q) \tag{3.16}$$

and

$$\ell_{\mathrm{B}}(q, p) - \ell_{\mathrm{B}}(p, p) = \|p - q\|_2^2. \tag{3.17}$$

The exercise in Exercise 2 asks for the two identities, the nonnegativity and equality condition for KL, and the resulting strict properness of both losses.

There are many strictly proper losses. In fact, differentiable proper losses on a finite simplex are generated by convex functions and their regrets are Bregman divergences; see Appendix A. Properness alone therefore does not select the loss used in language modeling.

3.3 Why single out log loss?

Statistical efficiency gives the first reason to prefer log loss. A structured model, including a transformer, uses the same parameter to determine many probabilities. Different losses can combine the information in the observations differently.

Definition 3.3.1 (statistical efficiency, predictive form)**.** Let $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ be a regular finite-dimensional parametric model. For each sample size T, let $\widehat{\theta}_T = \widehat{\theta}_T(X_1, \ldots, X_T)$ be an estimator. The sequence $(\widehat{\theta}_T)_{T \geq 1}$ is *statistically efficient in* $\mathcal{M}$ if the following holds for every $\theta_* \in \Theta$. When $X_1, \ldots, X_T$ are iid from p_{θ_*},

$$\mathbb{E}\Big[D_{\mathrm{KL}}(p_{\theta_*}\|p_{\widehat{\theta}_T})\Big] = \frac{c_{\mathcal{M}}(\theta_*)}{T} + o(T^{-1}), \tag{3.18}$$

where no regular estimator sequence has a smaller coefficient than $c_{\mathcal{M}}(\theta_*)$.

This definition compares the leading coefficient of the prediction error as $T \to \infty$. It does not claim that the efficient estimator has the smallest error at every finite sample size. The estimator sequence is fixed in advance and must attain the smallest coefficient at every possible true parameter in the model.

Definition 3.3.2 (equivalent losses)**.** Two losses are *equivalent* if

$$\widetilde{\ell}(q, x) = a\ell(q, x) + b(x) \tag{3.19}$$

for every $q \in \Delta_N$ and $x \in [N]$, where $a > 0$ and $b : [N] \to \mathbb{R}$ do not depend on q. Equivalent losses have the same empirical minimizers in every model.

Given a parametric model and observations $X_1, \ldots, X_T$, statisticians call any parameter satisfying

$$\widehat{\theta}_{\mathrm{MLE}} \in \operatorname*{argmax}_{\theta \in \Theta} \prod_{t=1}^{T} p_\theta(X_t) = \operatorname*{argmin}_{\theta \in \Theta} \sum_{t=1}^{T} -\log p_\theta(X_t) \tag{3.20}$$

a *maximum-likelihood estimator*, or MLE. Thus maximum likelihood is the statistical name for empirical log-loss minimization.

Theorem 3.3.3 (model-agnostic efficiency of log loss)**.** *Let $N \geq 2$, let $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ be a regular parametric model on $[N]$, and let $X_1, \ldots, X_T$ be iid from $p_{\theta_*} \in \mathcal{M}$. The empirical log-loss minimizer is statistically efficient in $\mathcal{M}$. If $N \geq 3$, then, among smooth strictly proper categorical losses, only losses equivalent to log loss have the property that their empirical minimizer is statistically efficient in every regular parametric model.*

The guarantee is model-agnostic because the loss is selected without knowing which regular parametric model will be used. If q is optimized directly over Δ_N, log loss and Brier loss both return the empirical distribution; Exercise 1 asks students to verify this. A difference can appear when a shared parameter links several probabilities and the loss must combine information with different noise levels. The two-context calculation in Appendix B.3 gives an explicit example.

The condition $N \geq 3$ in the uniqueness statement is necessary for an unconditional categorical model. A nonconstant regular model in Δ_2 is locally the full Bernoulli model, so every strictly proper loss returns the empirical success proportion. Binary regression with multiple contexts is different because one parameter can link several Bernoulli probabilities; the two-context example has exactly this structure.

The appendix in Appendix B states the regularity assumptions, derives the asymptotic prediction error, and discusses the applicability of the result to neural network models.

The definition of statistical efficiency itself measures prediction error by KL, while the theorem concerns log-loss minimization. This can make the conclusion appear to be built into the definition. We therefore need reasons to care about KL that do not already assume that log loss is the preferred loss. We begin with total variation.

3.4 KL divergence and total variation

We first ask whether small KL guarantees agreement on probabilities and bounded predictions. Total variation expresses exactly this form of agreement.

Definition 3.4.1 (total variation)**.** For $p, q \in \Delta_N$, their total-variation distance is

$$\mathrm{TV}(p, q) = \sup_{A \subseteq [N]} |p(A) - q(A)| = \frac{1}{2} \sum_{x=1}^{N} |p(x) - q(x)|. \tag{3.21}$$

Total variation controls the probability of every event. For every event A,

$$|p(A) - q(A)| \le \mathrm{TV}(p, q). \tag{3.22}$$

More generally, let X have distribution p and Y have distribution q. For every $f : [N] \to [0, 1]$,

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \le \mathrm{TV}(p, q). \tag{3.23}$$

Thus small total variation controls the probability of every event and the expectation of every bounded measurement simultaneously. If a prediction is the expectation of a bounded function under the estimated distribution, its error is therefore controlled by the total variation from the true distribution. The equality in Eq. (3.21) and the bounds in Eqs. (3.22) and (3.23) are proved in Exercise 4. Pinsker's inequality connects this criterion to KL.

Theorem 3.4.2 (Pinsker's inequality)**.** *For every $p, q \in \Delta_N$,*

$$\mathrm{TV}(p, q) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(p \| q)}. \tag{3.24}$$

Pinsker's inequality implies the following distributional guarantee from excess log loss:

$$\ell_{\log}(q, p) - \ell_{\log}(p, p) \le \epsilon \quad \Longrightarrow \quad \mathrm{TV}(p, q) \le \sqrt{\epsilon/2}. \tag{3.25}$$

Its proof is Exercise 4.

This raises a sharper question: would total variation select the same efficient estimator as KL? In the regular finite-dimensional setting, the answer is yes.

Proposition 3.4.3 (efficiency measured by total variation)**.** *In the setting of Theorem 3.3.3, the empirical log-loss minimizer has the smallest leading expected total-variation error among regular estimators. The total-variation error is of order $T^{-1/2}$. If $N \ge 3$, requiring this property in every regular parametric model again singles out log loss, up to equivalence, among smooth strictly proper categorical losses.*

Pinsker's inequality alone does not prove this proposition because it gives only a one-sided comparison. The local calculation is

$$\mathrm{TV}(p_{\theta_*}, p_{\theta_*+h}) = \frac{1}{2} \sum_{x=1}^{N} \left| \nabla_\theta p_{\theta_*}(x)^\top h \right| + o(\|h\|_2). \tag{3.26}$$

The leading term is a norm on local changes of the distribution. The information bound says that the maximum-likelihood estimator has the smallest asymptotic spread among regular estimators, which also minimizes the expected value of this norm. The appendix in Appendix B.2 gives the argument.

Thus the efficiency result is not an artifact of defining prediction error by log-loss regret. KL remains especially convenient because it is smooth and its local expansion uses Fisher information directly. The TV argument is one reason to take KL seriously; betting and coding provide two more.

3.5 Prediction as betting

Betting gives a second possible justification for KL by connecting stated probabilities to consequential predictions. In the game below, a predictor's probabilities determine bets, and prediction quality is defined by long-run logarithmic wealth growth. The resulting equivalence is exact. Its relevance depends on accepting this game and this definition of successful prediction.

Suppose a market offers mutually exclusive outcomes $[N]$. Fix reference odds $r \in \Delta_N$ with $r(x) > 0$. One dollar staked on outcome x returns $1/r(x)$ dollars if x occurs. A bettor is given the option to split up their wealth and make a bet. Specifically, if the bettor chooses $q \in \Delta_N$, this means that fraction $q(x)$ of the bettor's current wealth will be placed on outcome x (the fractions sum to one, so the bettor is required to invest all current wealth across the possible outcomes.) Then, a random $X \in [N]$ outcome is drawn. If $X = x$ occurs, wealth is multiplied by

$$\frac{q(x)}{r(x)}. \tag{3.27}$$

While the bettor is required to bet, if they choose the reference r for q, no matter the outcome, their wealth will be preserved. As such, the game is quite favorable for the bettors.

Below we consider the case when a bettor bets with the same fractions q in a repeated game as described above where in round $t \in [T]$, the outcome $X_t \in [N]$ comes from some distribution $p \in \Delta_N$. The outcomes are independent of each other and also from the initial wealth $W_0 > 0$ of the bettor. Then, from the rules of the game,

$$W_{t+1} = W_t \cdot \frac{q(X_{t+1})}{r(X_{t+1})}, \qquad t = 0, 1, 2, \dots$$

Let $\rho_T = \frac{1}{T} \log(W_T/W_0)$ be the growth rate of wealth after T steps.

Proposition 3.5.1 (log loss and optimal growth)**.** *Consider $(W_t, X_t)_t$ as above and any $q \in \Delta_N$. Then $\rho_T \to G_p(q; r)$ as $T \to \infty$ almost surely, where*

$$G_p(q; r) = \sum_{x=1}^{N} p(x) \log \frac{q(x)}{r(x)}. \tag{3.28}$$

The sum uses the conventions introduced for relative entropy. In particular, $G_p(q; r) = -\infty$ if $q(x) = 0$ for some x with $p(x) > 0$. Moreover,

$$G_p(p; r) - G_p(q; r) = D_{\mathrm{KL}}(p \| q). \tag{3.29}$$

Therefore truthful betting, $q = p$, uniquely maximizes the long-run growth rate.

Proof. First suppose that $q(x) > 0$ whenever $p(x) > 0$. After T rounds,

$$T\rho_T = \log \frac{W_T}{W_0} = \sum_{t=1}^{T} \log \frac{q(X_t)}{r(X_t)}. \tag{3.30}$$

Divide by T and apply the law of large numbers to obtain Eq. (3.28). The identity in Eq. (3.29) follows by algebra. If $q(x) = 0$ for some x with $p(x) > 0$, that outcome occurs eventually almost surely. The bettor's wealth is then zero, so $\rho_T = -\infty$ from that time onward. This proves the remaining case. The strict equality condition for KL proves uniqueness. □

Within this protocol, "predict well and make money" means maximizing logarithmic wealth growth under this betting protocol. Wealth multiplies across rounds, so its logarithm adds. Expected log wealth also penalizes losing all capital: assigning zero probability to an event that occurs sends log wealth to $-\infty$.

The remaining question is whether one accepts both the full-investment betting protocol and long-run log-growth maximization as a reasonable standard for prediction. If so, KL is exactly the loss in growth caused by using the wrong probabilities. Different betting constraints or objectives can lead to a different criterion.

The realized difference in cumulative log loss has an immediate meaning. For two predictors q and q',

$$\sum_{t=1}^{T} \big(\ell_{\log}(q', X_t) - \ell_{\log}(q, X_t)\big) = \log \frac{W_T^q}{W_T^{q'}}, \tag{3.31}$$

where W_T^q is the wealth after the same outcomes for a bettor who uses q in every round. One unit less cumulative log loss means a factor of e as much wealth for the same outcomes.

3.6 Prediction as compression

Coding gives a third justification for log loss: log loss in bits is the ideal codelength assigned to the observed data. Use logarithms to base two in this section. An ideal code matched to q assigns approximately

$$-\log_2 q(x) \tag{3.32}$$

bits to outcome x. Exact prefix codes require integer lengths and satisfy the Kraft inequality; block or arithmetic coding makes the discrepancy negligible per symbol over a long sequence.

If the true distribution is p, the expected excess ideal codelength from using q instead of p is

$$\sum_{x=1}^{N} p(x)(-\log_2 q(x)) - \sum_{x=1}^{N} p(x)(-\log_2 p(x)) = \sum_{x=1}^{N} p(x) \log_2 \frac{p(x)}{q(x)}. \tag{3.33}$$

This is relative entropy measured in bits. Thus KL is the expected codelength penalty for using probabilities q when the data have distribution p.

For an autoregressive model Q,

$$\begin{aligned} -\log_2 Q(x_{1:T}) &= -\log_2 \prod_{t=1}^{T} Q(x_t \mid x_{<t}) \\ &= \sum_{t=1}^{T} -\log_2 Q(x_t \mid x_{<t}). \end{aligned} \tag{3.34}$$

The average log loss in bits per token is therefore the average codelength per token under the model. If model Q improves on model R by δ bits per token on the same tokenized sequence, it saves $T\delta$ bits on that sequence:

$$\log_2 \frac{Q(x_{1:T})}{R(x_{1:T})} = T\delta. \tag{3.35}$$

If the measured average is b bits per token, its *perplexity* is

$$2^b = \left(\prod_{t=1}^{T} \frac{1}{Q(x_t \mid x_{<t})} \right)^{1/T}. \tag{3.36}$$

This is the geometric mean of the inverse probabilities assigned to the realized tokens. It can be read as an effective branching factor when the next-token uncertainty behaves roughly like a uniform choice among a fixed number of alternatives. For a general nonuniform, context-dependent model, perplexity remains the displayed geometric mean; it is not a literal count of available continuations.

The betting and coding calculations are the same likelihood-ratio identity:

$$\underbrace{-\log_2 r(x)}_{\text{reference codelength}} - \underbrace{\big(-\log_2 q(x)\big)}_{\text{model codelength}} = \underbrace{\log_2 \frac{q(x)}{r(x)}}_{\text{log wealth relative to reference}}. \tag{3.37}$$

3.6.1 The model must be counted when it is not already shared

The formula $-\log_2 Q(D)$ assumes that encoder and decoder already possess the same model Q. If a compression claim concerns the data used to construct the model, its description must also be transmitted. A two-part accounting is

$$L_{\text{total}}(D, \theta) = L(\theta) - \log_2 Q_\theta(D), \tag{3.38}$$

where $L(\theta)$ is the number of bits needed to specify the model under an agreed representation.

Without $L(\theta)$, the training-data compression problem has a trivial zero-codelength solution. Define a lookup model Q_D that assigns probability one to the particular training text D. Then

$$-\log_2 Q_D(D) = 0. \tag{3.39}$$

But describing Q_D requires specifying which text receives probability one; that description contains essentially the text itself. This is the precise sense in which, if model size is ignored, "the best compressor of the text is the text as the model." No reusable compression has been demonstrated.

If a trained model is already shared and we encode new held-out text, charging its full training cost again answers a different question. Then $-\log_2 Q_\theta(D_{\text{test}})$ is a legitimate conditional codelength. State which resources are shared before making a compression claim.

3.6.2 What compression establishes

Low log loss and good conditional compression are mathematically equivalent. Therefore the statement "a good language model compresses text" follows from the choice of log loss. If the model is already shared, a shorter codelength on held-out data than a specified baseline is evidence that the model captures reusable predictive regularities. This comparison does not identify which regularities were used. They may include syntax, phrase repetition, topic, formatting, factual associations, discourse structure, or reusable reasoning steps.

The stronger claim that compression by itself explains intelligence requires an additional argument: why the predictive regularities needed for compression also support the capabilities of interest. The coding identity does not supply that argument.

3.7 Take-home messages

1. Deterministic labels are too narrow for language. A context can have many natural continuations, so the target is a conditional distribution.

2. A strictly proper loss is uniquely minimized in expectation by the true distribution. Log loss and Brier loss are both strictly proper.

3. Properness does not select one loss. Among smooth strictly proper losses, log loss is, up to equivalence, the unique choice that guarantees statistical efficiency across regular categorical models with $N \geq 3$. Multiple coupled binary contexts also distinguish the losses.
4. Excess log loss is KL divergence. Pinsker's inequality converts it into control of every event probability through total variation.
5. Under a repeated betting protocol, KL is the loss in long-run log-wealth growth caused by using the wrong probabilities.
6. Log loss in bits is codelength. A difference in cumulative log loss is a code-length difference and a log likelihood ratio.
7. Compression of training data must count the model description. Otherwise a model that stores the data gives an uninformative zero data-codelength claim.
8. Compression of held-out data relative to a specified baseline is evidence that reusable predictive regularity was captured. It does not identify the regularity or explain every capability of the model.

3.8 Exercises

These exercises supply the proofs omitted from the main text and apply the main ideas to classification, evaluation, coding, and structured models.

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

1. **Empirical frequencies and empirical risk.** Let $X_1, \ldots, X_T \overset{\text{iid}}{\sim} p \in \Delta_N$ and let $\widehat{p}_T$ be defined by Eq. (3.3).
   (a) Show that $\widehat{p}_T \in \Delta_N$ and $\mathbb{E}[\widehat{p}_T(x)] = p(x)$.
   (b) Compute $\mathrm{Var}(\widehat{p}_T(x))$ and $\mathrm{Cov}(\widehat{p}_T(x), \widehat{p}_T(y))$ for $x \neq y$.
   (c) Show that empirical log-loss minimization over Δ_N returns $\widehat{p}_T$, with the convention that unobserved outcomes may receive zero probability.
   (d) Show that empirical Brier-loss minimization also returns $\widehat{p}_T$. Explain why statistical efficiency cannot distinguish the two losses in the unrestricted categorical model.
2. **The two regret identities.** Prove Proposition 3.2.3. In particular:
   (a) derive Eq. (3.14) by adding and subtracting $-\sum_x p(x) \log p(x)$;
   (b) derive Eq. (3.15) by expanding the square and taking expectations;
   (c) prove that $D_{\mathrm{KL}}(p\|q) \geq 0$, with equality if and only if $p = q$;
   (d) use the preceding part and the squared Euclidean term to prove that log loss and Brier loss are strictly proper on all of Δ_N;
   (e) explain why the order in $D_{\mathrm{KL}}(p\|q)$ matters, including what can happen when one distribution assigns probability zero to an outcome.

   Hint for (c): Use $\log u \leq u - 1$ for $u > 0$, and treat the case $q(x) = 0 < p(x)$ separately.

3. **Proper losses from Bregman divergences.** Write $\Delta_N^\circ = \{q \in \Delta_N : q(x) > 0 \text{ for every } x\}$ for the relative interior of the simplex. Let $F : \Delta_N^\circ \to \mathbb{R}$ be differentiable and strictly convex, and define

$$\ell_F(q, x) = -F(q) - \langle \nabla F(q), e_x - q \rangle.$$

(a) Show that

$$\ell_F(q, p) - \ell_F(p, p) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle.$$

for $p, q \in \Delta_N^\circ$. Conclude that ℓ_F is strictly proper on Δ_N°.

(b) Recover log loss from $F(q) = \sum_x q(x) \log q(x)$ and Brier loss, up to an additive constant, from $F(q) = \|q\|_2^2$. Show that Brier loss extends to every $q \in \Delta_N$. Extend log loss by setting $-\log 0 = +\infty$, and verify the boundary cases in Proposition 3.2.3.

Optional converse. Let $\ell : \Delta_N^\circ \times [N] \to \mathbb{R}$ be finite, differentiable in its first argument, and strictly proper when $p, q \in \Delta_N^\circ$. Define

$$B(p) = \ell(p, p), \qquad \widetilde{F}(p) = -B(p). \tag{3.40}$$

(c) Prove

$$B(p) = \inf_{q \in \Delta_N^\circ} \ell(q, p).$$

For fixed q, the map $p \mapsto \ell(q, p)$ is affine. Use this fact to prove that B is concave. Then use strict propriety to prove that B is strictly concave, so $\widetilde{F}$ is strictly convex.

(d) For fixed q, define $A_q(p) = -\ell(q, p)$. Prove that A_q is affine and

$$A_q(p) \le \widetilde{F}(p), \qquad A_q(q) = \widetilde{F}(q).$$

Thus A_q is a supporting affine function for $\widetilde{F}$ at q.

(e) Use differentiability to prove that this supporting affine function is the tangent affine function:

$$-\ell(q, p) = \widetilde{F}(q) + \langle \nabla \widetilde{F}(q), p - q \rangle.$$

Because both sides are affine in p, the identity extends from Δ_N° to its affine hull. Set $p = e_x$ and conclude that

$$\ell(q, x) = -\widetilde{F}(q) - \langle \nabla \widetilde{F}(q), e_x - q \rangle.$$

(f) Conclude that every finite differentiable strictly proper categorical loss on Δ_N° has Bregman regret:

$$\ell(q, p) - \ell(p, p) = D_{\widetilde{F}}(p, q).$$

Boundary remark. The restriction to Δ_N° is used because the displayed gradients need not exist at the boundary. It is not a restriction on proper losses themselves: Brier loss is finite on the whole simplex, and log loss has the extended-value definition given above. A general boundary treatment uses closed extended-valued convex functions and subgradients. Legendre functions provide a standard framework when gradients diverge at the boundary. This theory is beyond the scope of the exercise.

4. **Total variation and Pinsker's inequality.** Let $A = \{x : p(x) \geq q(x)\}$ and put $a = p(A)$, $b = q(A)$. Let X have distribution p and Y have distribution q. For $a, b \in [0,1]$, define the binary KL divergence by

$$\mathrm{kl}(a,b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}, \tag{3.41}$$

with the usual extended-value conventions.

   (a) Prove the equality between the supremum and the ℓ_1 expression in Eq. (3.21), and show that the supremum is attained by A.

   (b) Prove Eq. (3.22).

   (c) Prove Eq. (3.23).

   (d) Let $u_1, \dots, u_m$ and $v_1, \dots, v_m$ be positive numbers. Prove the log-sum inequality

$$\sum_{i=1}^{m} u_i \log \frac{u_i}{v_i} \geq \left(\sum_{i=1}^{m} u_i \right) \log \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} v_i}.$$

   Extend it to nonnegative numbers using the conventions for relative entropy.

   (e) Apply the log-sum inequality separately on A and A^c to prove $D_{\mathrm{KL}}(p\|q) \geq \mathrm{kl}(a,b)$.

   (f) Prove that $\mathrm{kl}(a,b) \geq 2(a-b)^2$, and conclude Pinsker's inequality.

Hints. For (a), write $p(x) - q(x)$ as its positive and negative parts. For (c), use the layer-cake identity $f(x) = \int_0^1 \mathbb{I}\{f(x) \geq t\}\, dt$. For (d), normalize the u_i and apply Jensen's inequality to the logarithm.

5. **Betting and log wealth.** Consider the protocol in Section 3.5.

   (a) Derive the wealth identity Eq. (3.30).

   (b) Explain why maximizing expected wealth after one round is not the same criterion as maximizing expected log wealth.

   (c) For $N = 2$, $p = (0.6, 0.4)$, and even odds $r = (1/2, 1/2)$, compute the expected log-growth rates for $q = p$, $q = (1/2, 1/2)$, and $q = (1, 0)$.

   (d) Generalize Eq. (3.29) to conditional distributions $p_t(\cdot \mid X_{<t})$ and history-dependent bets q_t.

6. **Codelength, likelihood ratios, and model cost.** Suppose two autoregressive models assign probabilities $Q(D)$ and $R(D)$ to a tokenized document D of length T.

   (a) If Q improves on R by 0.02 bits per token, compute the total bit saving and the likelihood ratio $Q(D)/R(D)$ when $T = 1000$.

   (b) Explain why the comparison requires the same tokenization.

   (c) Construct the lookup model Q_D in Eq. (3.39). State what must be communicated to make it a usable code for a recipient who does not already know D.

   (d) Give two distinct compression questions for which charging the model description once, and treating an already shared model as free, are respectively appropriate.

7. **Efficiency from two contexts.** Consider the model in Eq. (3.54) with $T/2$ observations at each context.

(a) Derive the Brier estimator in Eq. (3.58) and compute its variance exactly.
(b) Among unbiased combinations $w\widehat{\theta}_a + (1-w)\widehat{\theta}_b$, find the minimum-variance weight.
(c) Derive the likelihood equation Eq. (3.60). Explain why the MLE approaches the inverse-variance combination.
(d) Show that the variance ratio is $(v_a + v_b)^2/(4v_a v_b)$ and diverges as $\theta_* \downarrow 0$.
(e) Show that squared parameter error equals the average squared error in the two predicted probabilities. Thus interpret the variance ratio as a prediction-error ratio.

8. **Logistic and softmax gradients.** Let $p_\theta(x) = \sigma(f_\theta(x))$ in binary prediction.

(a) Derive
$$\nabla_\theta \ell_{\log} = (p_\theta(x) - y)\nabla_\theta f_\theta(x).$$
(b) Derive the Brier gradient and identify its additional factor $p_\theta(x)(1 - p_\theta(x))$.
(c) For multiclass softmax, prove
$$\frac{\partial q}{\partial z} = \mathrm{diag}(q) - qq^\top.$$
Use it to derive the two gradients in Eq. (3.68).
(d) Explain what happens to the two gradients when the model assigns probability nearly zero to the realized class.

The final two exercises connect probability modeling to classification and evaluation. The exercise in Exercise 9 asks what accurate probabilities imply for classification error. The exercise in Exercise 10 returns to a point from Lecture 2: a population objective does not by itself justify a number reported on a finite test set. The concentration argument is the same one used for deterministic predictors, and it explains when a fresh test set can support a model comparison.

9. **Classification from probabilistic prediction.** Let (X, Y) have an arbitrary distribution on $\mathcal{X} \times \{0, 1\}$, where $\mathcal{X}$ is finite. For every x with $\mathbb{P}(X = x) > 0$, define
$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x).$$
At the remaining inputs, $\eta(x)$ may be defined arbitrarily. For a decision $a \in \{0, 1\}$ and an x with $\mathbb{P}(X = x) > 0$, define the conditional risk
$$r_x(a) = \mathbb{P}(a \neq Y \mid X = x).$$
For a classifier $h : \mathcal{X} \to \{0, 1\}$, define its risk by
$$R_{01}(h) = \mathbb{P}(h(X) \neq Y).$$
A classifier is *Bayes-optimal* if it minimizes R_{01} over all classifiers $h : \mathcal{X} \to \{0, 1\}$. The minimum value
$$R^*_{01} = \min_{h:\mathcal{X}\to\{0,1\}} R_{01}(h)$$
is the *Bayes risk*.

(a) Compute $r_x(0)$ and $r_x(1)$. Show that
$$h^*(x) = \mathbb{I}\{\eta(x) \geq 1/2\}$$
is Bayes-optimal and that the Bayes risk is
$$R^*_{01} = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}].$$

(b) Prove the exact excess-risk identity

$$R_{01}(h) - R_{01}^* = \mathbb{E}\left[|2\eta(X) - 1|\mathbb{I}\{h(X) \neq h^*(X)\}\right].$$

(c) Let $q : \mathcal{X} \to (0,1)$ be a probabilistic predictor, and define

$$L_{\log}(q) = \mathbb{E}[-Y \log q(X) - (1-Y)\log(1-q(X))].$$

Let $L_{\log}^* = L_{\log}(\eta)$, with the usual conventions at $\eta(x) \in \{0,1\}$, and let $h_q(x) = \mathbb{I}\{q(x) \geq 1/2\}$. Prove

$$R_{01}(h_q) - R_{01}^* \leq \sqrt{2\big(L_{\log}(q) - L_{\log}^*\big)}.$$

(d) *Optional extension.* Construct a family of probabilistic predictors whose induced classifiers h_q are Bayes-optimal but whose excess log loss tends to infinity.

Hint for (c): You may use Pinsker's inequality for Bernoulli distributions:

$$|p - q| \leq \sqrt{\frac{1}{2}\left(p \log \frac{p}{q} + (1-p)\log \frac{1-p}{1-q}\right)}.$$

What this exercise is meant to teach. A probabilistic predictor and a classifier answer different questions. The excess classification risk weights a wrong decision by twice the distance of $\eta(x)$ from the decision boundary $1/2$. Small excess log loss controls excess classification risk, but Bayes-optimal classification does not imply accurate probabilities or small log loss.

10. **Evaluation on a fresh test set.** Let $\mathcal{X}$ be finite. Let $D = ((X_i, Y_i))_{i=1}^m$ be an i.i.d. test sample from a distribution on $\mathcal{X} \times \{0,1\}$, and let (X, Y) be an independent draw from the same distribution. For a classifier $h : \mathcal{X} \to \{0,1\}$, define

$$R(h) = \mathbb{P}(h(X) \neq Y), \qquad \widehat{R}_D(h) = \frac{1}{m}\sum_{i=1}^m \mathbb{I}\{h(X_i) \neq Y_i\}.$$

(a) Show that every classifier $h : \mathcal{X} \to \{0,1\}$ satisfies, for every $\epsilon > 0$,

$$\mathbb{P}\big(|\widehat{R}_D(h) - R(h)| > \epsilon\big) \leq 2e^{-2m\epsilon^2}.$$

(b) Fix a finite set $\mathcal{H} = \{h_1, \ldots, h_K\}$ of classifiers. Use the union bound to prove

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} |\widehat{R}_D(h) - R(h)| > \epsilon\right) \leq 2Ke^{-2m\epsilon^2}.$$

Give a sufficient sample size for the displayed maximum to be at most ϵ with probability at least $1 - \delta$.

(c) Let $f : (\mathcal{X} \times \{0,1\})^m \to \mathcal{H}$ and let $\widehat{h} = f(D)$. Show that

$$\mathbb{P}\left(\left|\widehat{R}_D(\widehat{h}) - R(\widehat{h})\right| > \epsilon\right) \leq 2Ke^{-2m\epsilon^2}.$$

(d) The previous part uses a simultaneous guarantee over all classifiers that the selection rule may return. Show what can happen without such a guarantee: a classifier defined using the test sample can have zero test error but large population risk. In particular, consider the following: Let $\mathcal{X} = [N]$, let X be uniform on $[N]$, and let $Y = 1$ almost surely. Define the following function of the sample D:

$$h_D(x) = \mathbb{I}\{x \in \{X_1, \ldots, X_m\}\}.$$

Compute $\widehat{R}_D(h_D)$ and $R(h_D)$. Show that $N \geq 2m$ implies $R(h_D) \geq 1/2$, and explain why this does not contradict part (a) or part (b).

Hint for (a): Apply Hoeffding's inequality to the independent random variables $\mathbb{I}\{h(X_i) \neq Y_i\}$.

A Bregman divergences and proper losses

This appendix records the general construction mentioned in class. For a differentiable convex function $F : \Delta_N \to \mathbb{R}$, define its Bregman divergence by

$$D_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle. \tag{3.42}$$

Convexity gives $D_F(p, q) \geq 0$; strict convexity gives equality only for $p = q$. Define

$$\ell_F(q, x) = -F(q) - \langle \nabla F(q), e_x - q \rangle. \tag{3.43}$$

Taking expectation under p gives

$$\ell_F(q, p) - \ell_F(p, p) = D_F(p, q). \tag{3.44}$$

Hence every strictly convex differentiable F produces a strictly proper loss. The optional converse in Exercise 3 proves that every finite differentiable strictly proper categorical loss on Δ_N° has this form. Nondifferentiable losses require supporting subgradients in place of ∇F, and the boundary requires the appropriate extended-value conventions.

For

$$F_{\log}(q) = \sum_x q(x) \log q(x),$$

Eq. (3.43) gives $-\log q(x)$ and $D_{F_{\log}}(p, q) = D_{\mathrm{KL}}(p \| q)$. For

$$F_{\mathrm{B}}(q) = \|q\|_2^2,$$

it gives $\|q\|_2^2 - 2q(x)$, which differs from $\|q - e_x\|_2^2$ only by the constant one.

This representation explains why properness does not determine a unique loss. The generator F chooses the geometry of probability error.

B Statistical efficiency: assumptions, example, and neural networks

This appendix explains the efficiency theorem in three steps. The regular likelihood calculation gives the optimal first-order prediction error. A two-context example shows that the gap between log loss and Brier loss can be arbitrarily large. The final subsection identifies which parts of the argument apply directly to neural networks and which parts require a fixed-dimensional regular model.

B.1 The regular likelihood result

The regular likelihood calculation gives expected predictive KL error $d/(2T) + o(T^{-1})$ for maximum likelihood and explains why no regular estimator has a smaller leading term.

Let $Z_1, \dots, Z_T$ be iid from P_{θ_*} in a model $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$. We use the standard assumptions that θ_* is an interior identifiable parameter, the log density is sufficiently differentiable near θ_*, differentiation and expectation can be interchanged, the Fisher information is finite and nonsingular, and the likelihood estimator is consistent. Exact assumptions vary among versions of the theorem.

For Z distributed according to P_θ, define the likelihood score and Fisher information by

$$s_\theta(Z) = \nabla_\theta \log p_\theta(Z), \qquad I(\theta) = \mathbb{E}[s_\theta(Z) s_\theta(Z)^\top]. \tag{3.45}$$

The score $s_\theta(Z) \in \mathbb{R}^d$ is a column vector. The matrix $I(\theta) \in \mathbb{R}^{d \times d}$ has parameter coordinates along both axes. Under "mild" regularity conditions,

$$\sqrt{T}(\widehat{\theta}_{\mathrm{MLE}} - \theta_*) \Longrightarrow \mathcal{N}(0, I(\theta_*)^{-1}). \tag{3.46}$$

The local KL expansion is

$$D_{\mathrm{KL}}(P_{\theta_*} \| P_{\theta_* + h}) = \frac{1}{2} h^\top I(\theta_*) h + o(\|h\|_2^2). \tag{3.47}$$

Combining Eq. (3.46) and Eq. (3.47) gives

$$2T D_{\mathrm{KL}}(P_{\theta_*} \| P_{\widehat{\theta}_{\mathrm{MLE}}}) \Longrightarrow \chi^2_d. \tag{3.48}$$

With the additional uniform-integrability conditions needed to pass to expectations,

$$\mathbb{E}\Big[D_{\mathrm{KL}}(P_{\theta_*} \| P_{\widehat{\theta}_{\mathrm{MLE}}})\Big] = \frac{d}{2T} + o(T^{-1}). \tag{3.49}$$

The results in Eqs. (3.48) and (3.49) give a leading prediction-error rate, its constant, and an approximate sampling distribution for confidence regions. An efficiency ratio also approximates the sample-size ratio needed to reach the same local accuracy.

Among regular estimators, the information bound makes $I(\theta_*)^{-1}$ the smallest possible asymptotic covariance. If another loss has estimating function $\psi_{\theta_*}(Z)$, equality requires

$$\psi_{\theta_*}(Z) = M s_{\theta_*}(Z) \quad \text{almost surely} \tag{3.50}$$

for a nonsingular matrix $M \in \mathbb{R}^{d \times d}$ whose rows and columns index parameter coordinates. Log loss has $\psi_\theta = -s_\theta$ in every model. A proper loss that is not equivalent to log loss can satisfy Eq. (3.50) in a particular model. It lacks this guarantee uniformly across models.

B.2 Why total variation gives the same first-order conclusion

Total variation and KL have different local scales, but they select the same first-order efficient estimator in a regular finite categorical model. For a local parameter change h, define

$$g_{\theta_*}(h) = \frac{1}{2} \sum_{x=1}^{N} \left| \nabla_\theta p_{\theta_*}(x)^\top h \right|. \tag{3.51}$$

Differentiability gives

$$\mathrm{TV}(p_{\theta_*}, p_{\theta_* + h}) = g_{\theta_*}(h) + o(\|h\|_2). \tag{3.52}$$

If $\sqrt{T}(\widehat{\theta}_T - \theta_*) \Longrightarrow Z$, then, under the corresponding uniform-integrability condition,

$$\sqrt{T}\,\mathbb{E}\Big[\mathrm{TV}(p_{\theta_*}, p_{\widehat{\theta}_T})\Big] \longrightarrow \mathbb{E}[g_{\theta_*}(Z)]. \tag{3.53}$$

The convolution theorem for regular estimators says that their limiting fluctuation equals the efficient Gaussian fluctuation plus independent additional noise. Since g_{θ_*} is a convex, symmetric norm on identifiable local changes, additional noise cannot reduce its expected value. The maximum-likelihood estimator therefore minimizes the leading expected total-variation error.

A minimum-loss estimator ties this bound only when it has the same first-order influence function as maximum likelihood. Thus the equality condition in Eq. (3.50) yields the same model-specific ties and the same model-agnostic uniqueness conclusion as predictive KL when $N \geq 3$. The difference is the rate: expected local KL is of order T^{-1}, while expected local total variation is of order $T^{-1/2}$.

B.3 A two-context calculation

This example shows the practical content of efficiency. Log loss combines two estimates using their inverse variances; Brier loss weights them equally. The resulting variance ratio can be arbitrarily large.

Let half the observations have context a and half have context b, with

$$P_\theta(Y = 1 \mid X = a) = \theta, \qquad P_\theta(Y = 1 \mid X = b) = \frac{1}{2} + \theta, \qquad 0 < \theta < \frac{1}{2}. \tag{3.54}$$

The two empirical success proportions give unbiased estimates

$$\widehat{\theta}_a = \widehat{p}_a, \qquad \widehat{\theta}_b = \widehat{p}_b - \frac{1}{2}. \tag{3.55}$$

At the true parameter, write

$$v_a = \theta_*(1 - \theta_*), \qquad v_b = \frac{1}{4} - \theta_*^2. \tag{3.56}$$

Then

$$\mathrm{Var}(\widehat{\theta}_a) = \frac{2v_a}{T}, \qquad \mathrm{Var}(\widehat{\theta}_b) = \frac{2v_b}{T}. \tag{3.57}$$

Brier minimization weights the two probability residuals equally:

$$\widehat{\theta}_{\mathrm{B}} = \frac{\widehat{\theta}_a + \widehat{\theta}_b}{2}, \qquad \mathrm{Var}(\widehat{\theta}_{\mathrm{B}}) = \frac{v_a + v_b}{2T}. \tag{3.58}$$

The minimum-variance unbiased linear combination instead uses inverse-variance weights:

$$\widehat{\theta}_{\mathrm{oracle}} = \frac{\widehat{\theta}_a/v_a + \widehat{\theta}_b/v_b}{1/v_a + 1/v_b}, \qquad \mathrm{Var}(\widehat{\theta}_{\mathrm{oracle}}) = \frac{1}{T}\frac{1}{\frac{1}{2}(1/v_a + 1/v_b)}. \tag{3.59}$$

It is an oracle because v_a and v_b depend on the unknown parameter.

The likelihood equation is

$$0 = \frac{\widehat{\theta}_a - \widehat{\theta}_{\mathrm{MLE}}}{v_a(\widehat{\theta}_{\mathrm{MLE}})} + \frac{\widehat{\theta}_b - \widehat{\theta}_{\mathrm{MLE}}}{v_b(\widehat{\theta}_{\mathrm{MLE}})}, \tag{3.60}$$

where $v_a(\theta) = \theta(1 - \theta)$ and $v_b(\theta) = 1/4 - \theta^2$. Thus the MLE is a fixed point of the inverse-variance combination, using variances at its fitted value. Consistency replaces those fitted variances by v_a and v_b to first order.

The asymptotic variance ratio is

$$\frac{V_{\mathrm{B}}}{V_{\log}} = \frac{(v_a + v_b)^2}{4 v_a v_b} \geq 1. \tag{3.61}$$

As $\theta_* \downarrow 0$, this ratio is asymptotic to $1/(16\theta_*)$ and can be arbitrarily large. For every fixed $\theta_* > 0$ the model is regular, but seeing the large-sample behavior near zero requires enough observations of the rare outcome; in particular, $T\theta_*$ should be large.

The parameter comparison has a direct predictive meaning here. Both conditional probabilities move one-for-one with θ, so averaging equally over the two contexts gives

$$\frac{1}{2} \sum_{x \in \{a,b\}} \big(p_{\widehat{\theta}}(1 \mid x) - p_{\theta_*}(1 \mid x)\big)^2 = (\widehat{\theta} - \theta_*)^2. \tag{3.62}$$

The variance ratio is therefore also the leading squared-probability-error ratio. The example shows why structure matters: two contexts estimate the same unknown quantity, and likelihood uses their different noise levels.

B.4 Logistic regression and neural networks

The gradient comparison applies exactly to logistic regression and neural networks. The fixed-dimensional efficiency theorem requires additional assumptions that overparameterized neural networks generally do not satisfy.

For binary logistic prediction, let

$$p_\theta(x) = \sigma(f_\theta(x)), \qquad g_\theta(x) = \nabla_\theta f_\theta(x). \tag{3.63}$$

The log-loss and Brier gradients are

$$\nabla_\theta \ell_{\log} = (p_\theta(x) - y) g_\theta(x), \tag{3.64}$$

$$\nabla_\theta \ell_{\mathrm{B}} = 4 p_\theta(x)(1 - p_\theta(x))(p_\theta(x) - y) g_\theta(x). \tag{3.65}$$

Brier retains an additional Bernoulli-variance factor. It can suppress the contribution of a context at which the model is saturated, including a confidently wrong observation.

For a K-class neural network, let

$$z_\theta(x) \in \mathbb{R}^K, \qquad q_\theta(\cdot \mid x) = \mathrm{softmax}(z_\theta(x)). \tag{3.66}$$

The vector $z_\theta(x)$ contains the logits. Define

$$J_x = \frac{\partial z_\theta(x)}{\partial \theta^\top} \in \mathbb{R}^{K \times d}, \qquad C_q = \mathrm{diag}(q) - q q^\top \in \mathbb{R}^{K \times K}. \tag{3.67}$$

The rows of J_x index logits and its columns index parameter coordinates. The rows and columns of C_q index classes; C_q is both the softmax Jacobian and the covariance matrix of a categorical outcome with probabilities q. The gradients and the context's Fisher-information contribution are

$$\begin{aligned} \nabla_\theta \ell_{\log} &= J_x^\top (q - e_y), \\ \nabla_\theta \ell_{\mathrm{B}} &= 2 J_x^\top C_q (q - e_y), \\ I_x(\theta) &= J_x^\top C_q J_x. \end{aligned} \tag{3.68}$$

The matrix $I_x(\theta) \in \mathbb{R}^{d\times d}$ has parameter coordinates along both axes. Thus J_x determines which shared directions the context informs. Brier inserts an additional C_q and can suppress low-variance probability directions.

The gradient identities above apply exactly to neural networks. The fixed-dimensional efficiency theorem does not generally apply to overparameterized networks, which can have nonidentifiable parameters, singular Fisher information, growing dimension, and many interpolating solutions. A language model still reuses its parameters across a vast number of contexts; overparameterization relative to a finite training sample does not make all conditional distributions independently adjustable. The additional C_q weighting may therefore matter for language models, but the regular theorem does not establish its effect on their generalization. A theory for that regime must account for representation, optimization, regularization, and function-space rather than raw parameter-space behavior.

C Concepts and terminology

This glossary collects the notation and terms used in the lecture.

Probability simplex
Δ_N is the set of nonnegative functions on $[N]$ whose values sum to one. Identifying functions on $[N]$ with vectors gives $\Delta_N \subseteq \mathbb{R}^{[N]} \equiv \mathbb{R}^N$.

Empirical distribution
The vector of relative frequencies in a sample.

Probabilistic model
A family of distributions, often written $\{p_\theta : \theta \in \Theta\}$. Shared parameters connect probabilities across outcomes or contexts.

Autoregressive model
A sequence model that factors a joint probability into conditional next-token probabilities as in Eq. (3.6).

Unsupervised learning
Learning from observations without separately supplied task labels.

Self-supervised learning
Learning from prediction targets constructed from the observations themselves. In next-token prediction, the prefix is the input and its following token is the target.

Generative modeling
Learning a probability distribution over observations. Such a model can assign probabilities and, in principle, generate new observations by sampling.

Population loss; empirical loss
Population loss is expected loss under the data distribution. Empirical loss is its sample average.

Maximum likelihood
The choice of model parameter that maximizes the probability assigned to the observed sample. It is equivalently the minimizer of empirical log loss.

Proper scoring loss
A loss whose population value is minimized by the true probability distribution. Strict properness makes that minimizer unique.

Equivalent losses
Losses related by a positive scaling and an outcome-dependent additive term, as in Eq. (3.19). They have the same empirical minimizers.

Bayes loss; generalized entropy
For a proper loss, $B(p) = \ell(p, p) = \inf_q \ell(q, p)$ is the smallest population loss achievable when the data distribution is p. Under the loss convention, B is concave.

Supporting affine function
An affine function A supports a convex function F at q if $A(p) \le F(p)$ for every p and $A(q) = F(q)$. If F is differentiable, this is its tangent affine function at q.

Bregman divergence
For differentiable convex F, $D_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle$. It is the gap between $F(p)$ and the tangent affine approximation to F made at q.

Savage representation
The characterization of proper categorical losses by convex functions and their supporting affine functions. In the differentiable case, the associated regrets are Bregman divergences.

Log loss; cross-entropy
The realized loss $-\log q(x)$. Its population value is cross-entropy and its regret is $D_{\mathrm{KL}}(p\|q)$.

Brier loss
Squared Euclidean distance between the reported probability vector and the one-hot realized outcome. Its regret is $\|p - q\|_2^2$.

Entropy
$H(p) = -\sum_x p(x) \log p(x)$, the minimum population log loss under p.

KL divergence
$D_{\mathrm{KL}}(p\|q) = \sum_x p(x) \log(p(x)/q(x))$. It is nonnegative, asymmetric, and infinite if q assigns zero probability to an event of positive p-probability. A term with $p(x) = 0$ contributes zero.

Binary KL divergence
$\mathrm{kl}(a, b)$ is the KL divergence from Bernoulli(a) to Bernoulli(b), as defined in Eq. (3.41).

Total variation
The largest discrepancy between probabilities assigned to the same event by two distributions; on a finite space it is one half of their ℓ_1 distance.

Statistical efficiency
Attainment of the smallest leading asymptotic estimation or prediction error within a specified regular model and class of estimators. These notes use predictive KL as the error measure.

Model-agnostic efficiency
An efficiency guarantee that holds across regular parametric models without adapting the loss to the particular model.

Log-growth betting
A repeated betting criterion that maximizes the long-run exponential growth rate of wealth. Its regret under the stated protocol is KL divergence.

Bits and nats
Log loss uses bits with logarithm base two and nats with the natural logarithm. One nat is $\log_2(e) \approx 1.4427$ bits; one bit is $\log(2) \approx 0.6931$ nats.

Perplexity
The exponential of average log loss: 2^b when the average is b bits per token, or e^n when it is n nats per token. It is the geometric mean of the inverse probabilities assigned to the realized tokens.

Conditional codelength
The number of bits needed to encode data when a probability model is already shared. For a sequence, it is approximately $-\log_2 Q(D)$.

Two-part code
A code that transmits both a model description and data encoded under that model. Its total length has the form $L(\theta) - \log_2 Q_\theta(D)$.

Bibliographic remarks

The following sources provide the background for the results stated here.

Proper scoring rules. Gneiting and Raftery [2007] develop proper scoring rules, entropy functions, Bregman divergences, and examples including logarithmic, quadratic, spherical, ranked-probability, and energy scores. Theorems 1 and 2 give the general and finite-categorical representations in terms of convex functions and supporting affine functions. The optional converse in Exercise 3 is the differentiable finite-simplex version. Nondifferentiable scores require subgradients, and scores such as log loss require care at the boundary. Our convention treats smaller losses as better; the paper uses rewards, with larger values preferred.

Inference from proper scores. Dawid et al. [2016] develop estimating equations, asymptotic covariance, and efficiency for estimators obtained from proper scoring rules. Maximum likelihood is the log score case. The equality between an efficient estimating function and a nonsingular transform of the likelihood score is the basis for Eq. (3.50). The regular likelihood expansions and the convolution theorem used in Appendix B are treated by van der Vaart [1998, Chapters 7–8].

Betting. Kelly [1956] connects information rate with exponential capital growth under repeated betting. The protocol in Section 3.5 is the finite-outcome form needed here.

Compression. Shannon [1948] established the fundamental connection among probability, entropy, and codelength. The ideal length $-\log_2 q(x)$ may be nonintegral; prefix coding, block coding, and arithmetic coding make the operational statement precise in different ways. Lattimore and Szepesvári [2020, Chapter 14] give a short treatment of prefix-free codes, entropy bounds, arithmetic coding, and KL as the expected extra codelength caused by using the wrong distribution.

Alternative scores for language generation. Shao et al. [2024] study training language generators with Brier and spherical scores. Their results show that log loss is not the only trainable proper objective. The efficiency result identifies one general statistical property of log loss; it is not a claim that every finite-sample or misspecified comparison favors log loss.

References

A. Philip Dawid, Monica Musio, and Laura Ventura. Minimum Scoring Rule Inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016. https://doi.org/10.1111/sjos.12168.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. https://doi.org/10.1198/016214506000001437.

John L. Kelly, Jr. A New Interpretation of Information Rate. *Bell System Technical Journal*, 35(4):917–926, 1956. https://doi.org/10.1002/j.1538-7305.1956.tb03809.x.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. https://tor-lattimore.com/downloads/book/book.pdf.

Chenze Shao, Fandong Meng, Yijin Liu, and Jie Zhou. Language Generation with Strictly Proper Scoring Rules. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44474–44488, 2024. https://proceedings.mlr.press/v235/shao24c.html.

Claude E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948. https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.