CMPUT 654: Mathematical Foundations of Modern AI Systems Fall 2026

Lecture 6: Data collection, dependence, and useful evidence

September 17, 2026

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

Larkin Wisdom prepared a scribe of the September 17 lecture. Csaba Szepesvári prepared and expanded the present notes.

Reading guide. These notes ask how collected data become evidence about deployment. The main text follows the classroom discussion: corpus construction, the role of randomness, concentration in the iid model, exact copies, resampling, and beta-mixing. The public pretraining references, the longer resampling calculation, and the proof behind the mixing theorem were added after class. The exercises develop calculations that were posed but not completed in class. Lectures 1–5 provide the background on sequence prediction, loss, and state-space models.

6.1 A dataset is produced by a procedure

Data is a major part of an ML system. The iid model deliberately suppresses most details of how data is obtained. Before using that model, we should ask where the data came from and what happened to it.

Language-model pretraining commonly starts from web pages, books, code, scientific articles, conversations, and other text collections. A simplified pipeline is

$$\text{sources} \longrightarrow \text{extraction and parsing} \longrightarrow \text{filtering and deduplication}$$
$$\longrightarrow \text{source mixture} \longrightarrow \text{tokenization} \longrightarrow \text{training sequences}.$$

This pipeline raises a recurring question: does deduplication matter, when does it matter, and by how much? Section 6.4 answers it under a simple variance model. Extraction may remove HTML and recover the main text. Filters may use language, quality, safety, or document-type criteria. Exact and approximate deduplication remove repeated material. A mixture rule determines how much of each retained source is used. Tokenization turns text into token IDs. Finally, the tokens must be placed into sequences of the length used for training. Public descriptions of open-weight models document many, though not always all, of these choices; the bibliographical notes give examples.

A common sequence-construction method concatenates documents, places a special boundary token between them, and cuts the resulting stream into fixed-length sequences. Other methods pack shorter documents without allowing attention across document boundaries or truncate long documents. These choices determine which prediction contexts training presents to the model. They can also make training contexts unlike deployment contexts: a training sequence may usually be full, whereas a deployed model must also predict after a short prompt. In a *causally masked transformer*, the prediction at a position cannot use later tokens. A full training sequence therefore contains prediction problems with prefix lengths ranging from short to long. This is a transformer-specific mitigation, not a general answer to mismatch.

The guiding principle is that training data must resemble deployment data in the respects needed for the desired behavior. Exact equality of distributions is not required, and post-training can

alter behavior. It cannot generally recover abilities for which pretraining and post-training supply no relevant evidence. Pretraining only on Hungarian documents and deploying only in English preserves some common structure, but not enough to expect good English behavior.

Three objects should remain separate:

1. the *realized corpus*, which is a fixed collection of documents and sequences;
2. the *collection procedure*, which found, selected, filtered, copied, and divided the sources; and
3. a *probability model*, which describes repeated collection, an ordered source, evaluation cases, or future deployment.

A list of files does not determine a probability model. The proposed unit—a token, window, document, conversation, user, or source—also matters. One book, one thousand independently written books, and one book stored one thousand times can all yield one thousand training sequences but provide different evidence.

6.2 Why introduce randomness?

A probability law specifies which future cases count and how they are weighted. A joint law additionally states how observations are related. It can therefore support quantitative claims about whether empirical averages are stable and whether they estimate deployment quantities.

Randomness can enter in three distinct places:

- *collection randomness*: units are sampled from a stated population or by a randomized experiment;
- *source randomness*: a stochastic process models future documents, interactions, or tokens; and
- *algorithmic randomness*: shuffling, minibatch selection, initialization, and decoding randomize computation.

Algorithmic randomness does not turn repeated source material into independent population evidence. Randomness may also be an effective mathematical description of a deterministic system, but establishing such a description is a separate problem.[1]

Observations that are not independent are called *dependent*. An iid model can be plausible when, for example, a machine independently selects problems from a fixed collection or unrelated users independently initiate conversations under stable conditions. These remain modeling claims about the chosen units and procedure.

Independence is also not necessary for learning. For the threshold class $f_k(x) = \mathbb{I}\{x \geq k\}$, the two labeled examples $(k-1, 0)$ and $(k, 1)$ identify an interior threshold exactly. This is a deterministic teaching set: the teacher knows the target and chooses examples that identify it. This differs from active learning, where the learner does not know the target but may choose which labels to request, and from passive learning, where the learner receives examples from a fixed sampling law. For interior thresholds on a finite ordered set, teaching uses two target-dependent examples, active learning permits binary search, and passive learning must wait for the sampling law to reveal points near the threshold. A boundary threshold can be taught with one example. Dependence alone does not explain these differences; the information available when examples are chosen does.

Expert curation for language models can resemble teaching when examples are selected to communicate a desired capability. The analogy concerns data selection: a target-aware example can be more informative than a passive draw. The teaching analogy does not imply that one or a few

examples suffice for a model to learn or transfer a proof. The model class, optimization procedure, and relation between training and deployment tasks still matter.

A dependent process can also provide quantitative coverage. Suppose the deployment law on $[N]$ is P, and, for every interval $I \subseteq [N]$,

$$\mathbb{P}(X_i \in I \mid X_1, \dots, X_{i-1}) \geq \alpha P(I) \quad \text{almost surely.} \tag{6.1}$$

For the threshold learner from Lecture 2, let L denote the deployment probability of its error interval. The event $L > \epsilon$ requires the sample to miss a fixed interval, determined by P, the target, and ϵ, whose P-mass exceeds ϵ. Each observation has conditional probability at least $\alpha\epsilon$ of hitting that interval. Iterated conditioning gives

$$\mathbb{P}(L > \epsilon) \leq (1 - \alpha\epsilon)^n \leq \exp(-\alpha n \epsilon). \tag{6.2}$$

Thus independence is one sufficient explanation for accumulating evidence, not the definition of learnability.

6.3 The iid calculation

Let $X_1, \dots, X_n$ be iid with law P. For a fixed parameter θ, define the *population loss* and *empirical loss* by

$$L(\theta) = \mathbb{E}[\ell(\theta; X)], \qquad \widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; X_i). \tag{6.3}$$

If $0 \leq \ell(\theta; X) \leq 1$, Hoeffding's inequality gives

$$\mathbb{P}\Big(\Big|\widehat{L}_n(\theta) - L(\theta)\Big| > \epsilon\Big) \leq 2e^{-2n\epsilon^2}. \tag{6.4}$$

This is a *concentration* statement: the empirical average is unlikely to be far from its expectation. The systematic study of such behavior is called measure concentration.

The variance explains the first-order scale. Writing $\sigma_\theta^2 = \mathrm{Var}(\ell(\theta; X))$, independence gives

$$\mathrm{Var}(\widehat{L}_n(\theta)) = \frac{\sigma_\theta^2}{n}. \tag{6.5}$$

When $0 < \sigma_\theta^2 < \infty$, the central limit theorem further gives

$$\sqrt{n}\big(\widehat{L}_n(\theta) - L(\theta)\big) \implies \mathcal{N}(0, \sigma_\theta^2).$$

Equivalently, for large n, the deviation $\widehat{L}_n(\theta) - L(\theta)$ is approximately $\mathcal{N}(0, \sigma_\theta^2/n)$. These statements concern a fixed θ. Applying them to a parameter chosen from the same data requires a uniform, stability, information, or other generalization argument.

The copied process $Y_i = U$, where $U \sim \mathrm{Bernoulli}(1/2)$, shows what independence prevents. Every Y_i has the same marginal law, but $n^{-1}\sum_i Y_i = U$ for every n. The empirical average does not concentrate around $1/2$.

6.4 Exact copies and effective sample size

We can now answer the question raised at the start of class: does deduplication matter, when does it matter, and by how much?

Let $X_1, \dots, X_m$ be independent source units with common loss variance σ_θ^2. Suppose unit i occurs $c_i \geq 1$ times in the stored corpus, and let $N = \sum_{i=1}^m c_i$. The stored-data average is

$$\widehat{L}_{\text{copy}}(\theta) = \frac{1}{N} \sum_{i=1}^{m} c_i \ell(\theta; X_i). \tag{6.6}$$

Independence of the m source units gives

$$\text{Var}(\widehat{L}_{\text{copy}}(\theta)) = \sigma_\theta^2 \frac{\sum_{i=1}^m c_i^2}{N^2} = \frac{\sigma_\theta^2}{n_{\text{eff}}}, \qquad n_{\text{eff}} = \frac{N^2}{\sum_{i=1}^m c_i^2}. \tag{6.7}$$

The *effective sample size* is the independent sample size that would give the same variance under this model. Cauchy–Schwarz gives

$$1 \leq n_{\text{eff}} \leq m, \tag{6.8}$$

with $n_{\text{eff}} = m$ exactly when all multiplicities are equal. If one source dominates, n_{eff} can be close to one.

Deduplicating to one copy of every source gives variance σ_θ^2/m. Hence

$$\frac{\text{Var}(\widehat{L}_{\text{copy}})}{\sigma_\theta^2/m} = m \frac{\sum_i c_i^2}{N^2} = 1 + \text{CV}(c_1, \dots, c_m)^2,$$

where CV is the empirical coefficient of variation using denominator m. Equal duplication wastes storage and computation but does not reduce effective sample size below m. Unequal duplication also reduces effective sample size and changes the empirical weights. If multiplicity is intended to represent deployment frequency, deduplication changes the target; Lecture 7 studies that weighting question.

Near copies require the more general covariance calculation below. Their value depends on the statistic: two variants can be nearly identical for one learned feature and materially different for another.

6.5 Covariance is the general calculation

Let G_t be a scalar loss, score, or projection of a gradient contribution at position t. No independence assumption is needed for

$$\text{Var}\left(\frac{1}{T} \sum_{t=1}^{T} G_t \right) = \frac{1}{T^2} \sum_{s,t=1}^{T} \text{Cov}(G_s, G_t). \tag{6.9}$$

Define

$$S_T = \sum_{t=1}^{T} \text{Var}(G_t), \qquad \Gamma_T = \sum_{s,t=1}^{T} \text{Cov}(G_s, G_t).$$

When $S_T > 0$ and $\Gamma_T > 0$, variance matching gives the statistic-dependent effective sample size

$$T_{\text{eff}}(G) = T \frac{S_T}{\Gamma_T}. \tag{6.10}$$

Independence gives $T_{\text{eff}} = T$. Positive covariance usually reduces it; negative covariance can increase it. There is no single effective sample size for a corpus without specifying the quantity being averaged.

This distinction matters for sequences. A single long trajectory may contain many effective observations of a shared transition parameter but only one draw from its initial-state law. A book may repeatedly inform local syntax while retaining the same notation convention throughout. Overlapping context windows share most of their tokens; Exercise 6 asks for their exact dependence range in a simple model.

6.6 What resampling changes

Fix a realized corpus $x_1, \dots, x_m$. Sampling indices independently with replacement produces conditionally independent draws from its empirical law. It does not create new independent evidence about the population that produced the corpus. As the number of resampled draws grows with the corpus fixed, the random frequencies approach the original empirical frequencies, so the same empirical average is recovered.

Sampling all m indices without replacement is a random shuffle. Sampling $B < m$ without replacement is equivalently a random shuffle followed by truncation. Random minibatches can change the optimization path, but repeated minibatch training still targets the original empirical objective. Sparse subsampling may separate nearby observations in a dependent sequence, at the cost of discarding data and adding variability. The appendix makes these claims precise; Exercise 4 asks you to compare sampling with and without replacement.

6.7 Controlled dependence through beta-mixing

There are many ways to quantify dependence. For a one-sided process $X_1, X_2, \dots$, let $\mathcal{L}(Y)$ denote the distribution of a random element Y. Define the *absolute-regularity*, or *beta-mixing*, coefficient

$$\beta(a) = \sup_{t \geq 1} \|\mathcal{L}(X_{1:t}, X_{t+a:\infty}) - \mathcal{L}(X_{1:t}) \otimes \mathcal{L}(X_{t+a:\infty})\|_{\text{TV}}, \qquad a \geq 1. \tag{6.11}$$

Here $\|P - Q\|_{\text{TV}} = \sup_A |P(A) - Q(A)|$. Small $\beta(a)$ means that an observed past and a future separated from it by a gap of $a - 1$ positions are close to independent in a strong, distributional sense.

Select positions $t_j = 1 + (j-1)a$, for $j = 1, \dots, q$, where $q = 1 + \lfloor (n-1)/a \rfloor$. Assume the selected observations have common marginal law P. The independent-points comparison proved in Appendix B is

$$\left\|\mathcal{L}(X_{t_1}, \dots, X_{t_q}) - P^{\otimes q}\right\|_{\text{TV}} \leq (q-1)\beta(a). \tag{6.12}$$

Theorem 6.7.1 (Finite-class consistency under beta-mixing)**.** *Let $\mathcal{F}$ be finite, let the target $f \in \mathcal{F}$, and, for $X \sim P$, define $L_P(g, f) = P(g(X) \neq f(X))$. If a learner returns a function $\widehat{f} \in \mathcal{F}$ consistent with all labeled observations $(X_i, f(X_i))$, $i \in [n]$, then*

$$\mathbb{P}\big(L_P(\widehat{f}, f) > \epsilon\big) \leq |\mathcal{F}|(1-\epsilon)^q + (q-1)\beta(a). \tag{6.13}$$

Proof. Let E be the event that some $g \in \mathcal{F}$ with $L_P(g, f) > \epsilon$ is consistent at every selected position. Full-sample consistency implies the failure event is contained in E. By Eq. (6.12) and the iid finite-class argument from Lecture 2,

$$\mathbb{P}(E) \leq P^{\otimes q}(E) + (q-1)\beta(a) \leq |\mathcal{F}|(1-\epsilon)^q + (q-1)\beta(a).$$

□

Larger gaps reduce the dependence term but leave fewer selected observations. For example, if $\beta(a) \le e^{-a}$, choosing $a = \lceil \log(2n/\delta) \rceil$ makes the dependence term at most $\delta/2$. The remaining term behaves as though the sample size were roughly $n/\log(n/\delta)$: this blocking argument loses only a logarithmic factor compared with iid sampling. Exercise 7 asks what the same argument gives under polynomial decay.

Finite irreducible aperiodic Markov chains mix geometrically. On a compact Polish state space, irreducibility and the strong Feller property make the whole state space a small set; with an appropriate aperiodicity condition, this gives a Doeblin route to geometric ergodicity and hence exponential beta-mixing. On a noncompact state space, the same argument only makes compact sets small, so a return or drift condition is generally needed. Ergodic chains can mix polynomially or more slowly. The point here is the distinction between the clean compact case and the additional control needed on noncompact spaces, not a catalogue of minimal conditions.

Mixing is highly relevant when learning from one long reinforcement-learning trajectory or another continuing process. It is not automatically a realistic global model of a web corpus. A book can introduce notation at the beginning and use it to the end. An author, topic, task, or other slowly evolving state can create arbitrarily long unconditional dependence even if local residuals forget quickly after conditioning on that state. A state-space account with slow and fast components may then be more informative. Mixing must always be checked for the process and statistic to which a theorem is applied.

6.8 Dependence is also predictive information

The prefix–target pairs in one document are dependent and strongly overlap. That dependence is not merely a defect: earlier tokens contain information about later ones. A sequence model reuses parameters across positions so that many different contexts can constrain shared predictive structure. The same trajectory can therefore provide useful evidence without becoming a collection of iid examples.

Parameter sharing alone does not prove that the shared regularity is correct, that training covers deployment contexts, or that optimization recovers it. The relevant question is whether the observed dependent data constrain the particular shared object needed at deployment.

Evaluation uses the same logic. Define the intended deployment unit and source; keep exact copies, near copies, and overlapping chunks from crossing the training–test boundary; and remember that repeated selection on a test set makes it part of development. A common law P for iid training and test examples is the strongest simple stability model, not a default fact about data.

6.9 Take-home messages

- A corpus is the output of a collection and processing pipeline. Choosing a probability model is an additional analytical decision.
- Independence gives useful quantitative concentration, but deterministic teaching and dependent coverage show that it is not necessary for learning.
- Exact duplication changes empirical weights and can reduce effective sample size. The variance formula states precisely when deduplication helps.

- The covariance of the statistic being averaged determines how much evidence dependent observations provide.
- Resampling or shuffling a fixed corpus does not create new population evidence.
- Mixing and blocking recover weakened iid guarantees under controlled dependence. Persistent state can make unconditional mixing inappropriate.
- Dependence within a sequence is predictive signal. Evaluation is useful only for the deployment claim represented by its construction.

Endnotes

1. Deterministic systems can exhibit statistical regularities. Statistical mechanics, dynamical systems, and pseudorandomness give precise versions in particular settings. Their existence does not by itself justify a stochastic model for a corpus.

Bibliographical notes

Public pretraining accounts illustrate which pipeline details are documented. Dolma describes source-specific extraction, filtering, deduplication, and mixture construction [Soldaini et al., 2024]. FineWeb reports large-scale web filtering and deduplication studies [Penedo et al., 2024]. The Llama 3 report describes its data mixture and filtering at a higher level [Dubey et al., 2024]. Sequence packing and document-boundary choices are analyzed explicitly by Ding et al. [2024].

Hoeffding's inequality is from Hoeffding [1963]. The independent-block method for mixing empirical processes is developed by Yu [1994]. The exercise contrasting the central limit theorem with exponential tail bounds is adapted from the distinction developed in Exercises 5.10 and 5.13 of Lattimore and Szepesvári [2020]. Hairer [2002] proves that every compact set is small for an irreducible strong Feller chain. If the state space is compact, this applies to the whole space. Bradley [2005] surveys strong mixing conditions and records that geometric ergodicity of a stationary Markov chain implies exponential beta-mixing. These results give the sufficient route stated in the text; they are not presented as minimal conditions.

Glossary

Absolute regularity or beta-mixing
The total-variation distance in Eq. (6.11) between the joint law of a past and a separated future and the product of their marginal laws.

Collection procedure
The operations that find, select, filter, copy, and divide sources into the realized data.

Concentration
The property that a random quantity, such as an empirical average, is unlikely to be far from a specified central value.

Effective sample size
The independent sample size whose variance matches that of a specified dependent or unequally weighted average.

Empirical loss
The average loss on the realized training observations, as in Eq. (6.3).

Population loss
Expected loss under a stated law representing the cases of interest.

Realized corpus
The fixed collection of documents, tokens, or sequences actually available.

Resampling with replacement
Independent draws from the empirical distribution of a fixed dataset.

6.10 Exercises

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

Exercise profiles. Filled mountains show difficulty on a four-mountain scale: one is a direct application, two uses a familiar method, three combines ideas, and four is a substantial challenge. Filled cogs show technical load on a three-cog scale: one is light, two is sustained, and three is dense. Filled hourglasses show estimated time: one is under 20 minutes, two is 20–45 minutes, three is 45–90 minutes, and four is more than 90 minutes. Pale pictograms show the unused part of each scale. The work row uses the canonical order proof, calculation, construction, computation, and interpretation; pale icons mark work types not used. ◎ marks the intended learning outcome. Time estimates cover the non-optional parts. Ratings are not prerequisites or grading weights. A superscript asterisk marks an exercise based on supplementary lecture material; it is not a difficulty rating and does not make the exercise optional when assigned.

Exercise 1: Teaching, active learning, and passive learning.

◎ Distinguish three information structures by deriving their requirements for identifying a threshold.

Let $N \geq 6$, let $\mathcal{K} = \{2, \dots, N-1\}$, and, for $k \in \mathcal{K}$, define $f_k(x) = \mathbb{I}\{x \geq k\}$ on $[N]$. The goal is to identify k exactly from labeled examples.

(a) A teacher knows k and may choose labeled examples. Prove that $\{(k-1, 0), (k, 1)\}$ identifies k uniquely within $\mathcal{K}$. Explain why no single labeled example identifies an interior threshold $k \in \{3, \dots, N-2\}$. Give one-example teaching sets for the two boundary thresholds $k = 2$ and $k = N-1$.

(b) An active learner does not know k, but may choose each query after observing the previous labels. Give a binary-search strategy that identifies every $k \in \mathcal{K}$ using at most $\lceil \log_2(N-2) \rceil$ queries. Prove that no active strategy can have a smaller worst-case number of binary-label queries.

(c) A passive learner receives $X_1, \dots, X_n$ independently and uniformly from $[N]$, together with their labels. Fix $k \in \{3, \dots, N-2\}$. Prove that the observed labels identify k uniquely if and only if both $k-1$ and k occur in the sample, and that the probability of this event is

$$1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n.$$

Deduce that $n \geq N \log(2/\delta)$ suffices for this probability to be at least $1 - \delta$.

(d) Explain why the two-example teaching set is not evidence that dependence alone makes learning faster. State precisely what target information the teacher has that the active and passive learners do not.

Hints For (b), maintain the set of thresholds consistent with all answers. For the lower bound, a depth-q binary decision tree has at most 2^q leaves. For (c), use inclusion–exclusion and then a union bound.

Exercise 2: Hoeffding bounds and the normal approximation.

◎ Distinguish the central-limit scale from fixed-size deviations and compare the conclusions supplied by the central limit theorem and Hoeffding's inequality.

Let $X_1, \dots, X_n$ be iid Bernoulli(p) variables with $p \in (0,1)$, and let $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$.

(a) Derive $\mathbb{E}[\widehat{p}_n]$ and $\mathrm{Var}(\widehat{p}_n)$, and state the central limit theorem for $\sqrt{n}(\widehat{p}_n - p)$.

(b) Apply Hoeffding's inequality to show that, for every $\epsilon > 0$,

$$\mathbb{P}(\widehat{p}_n - p \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

(c) Fix $t > 0$ and set $\epsilon_n = t\sqrt{p(1-p)/n}$. Determine the limit of $\mathbb{P}(\widehat{p}_n - p \geq \epsilon_n)$, and show that the Hoeffding bound in part (b) becomes $e^{-2t^2 p(1-p)}$. Explain what the two conclusions say at the $n^{-1/2}$ scale.

(d) Now fix $\epsilon \in (0, 1-p)$, and let $Z_n \sim \mathcal{N}(p, p(1-p)/n)$. You may use

$$\lim_{x\to\infty} \frac{1}{x^2}\log(1-\Phi(x)) = -\frac{1}{2}.$$

Show that

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}(Z_n - p \geq \epsilon) = -\frac{\epsilon^2}{2p(1-p)}.$$

Explain why the central limit theorem alone does not justify using this Gaussian rate for $\mathbb{P}(\widehat{p}_n - p \geq \epsilon)$, and contrast this with the finite-sample conclusion in part (b).

Exercise 3: Exact duplication and effective sample size.

Derive how unequal duplication changes variance and effective sample size, and state what this calculation does not establish about learning performance.

Let $X_1, \dots, X_m$ be independent, and suppose $Z_i = \ell(\theta; X_i)$ has variance $\sigma_\theta^2 > 0$ for every i. Unit i is stored $c_i \geq 1$ times. Set $N = \sum_{i=1}^m c_i$ and $\widehat{L} = N^{-1}\sum_{i=1}^m c_i Z_i$.

(a) Derive $\mathrm{Var}(\widehat{L})$ and the effective sample size n_{eff} in Eq. (6.7).

(b) Prove $1 \leq n_{\mathrm{eff}} \leq m$. State exactly when the upper bound is attained.

(c) Calculate n_{eff} when (i) $c_i = c$ for every i, and (ii) $c_1 = r$ and $c_i = 1$ for $i \geq 2$. Find the limit in (ii) as $r \to \infty$ with m fixed.

(d) Let $\bar{c} = N/m$ and $\mathrm{CV}^2 = m^{-1}\sum_i (c_i - \bar{c})^2/\bar{c}^2$. Prove

$$\frac{\mathrm{Var}(\widehat{L})}{\sigma_\theta^2/m} = 1 + \mathrm{CV}^2\,.$$

Explain what this ratio does and does not establish about deduplication.

Hints For (b), apply Cauchy–Schwarz to $\sum_i c_i$. For the lower bound, compare $(\sum_i c_i)^2$ and $\sum_i c_i^2$.

Sampling from a finite corpus

Exercise 4: Sampling with and without replacement.

Derive the laws induced by sampling indices with and without replacement and identify what these procedures do not change about the source data.

Fix a corpus $x_1, \dots, x_m \in \mathbb{R}$ and let $\bar{x} = m^{-1}\sum_i x_i$.

(a) Draw $I_1, \dots, I_B$ independently and uniformly from $[m]$, and define $C_i = \sum_{b=1}^B \mathbb{I}\{I_b = i\}$. Give the joint law of $(C_1, \dots, C_m)$, and prove that $B^{-1}\sum_i C_i x_i$ is conditionally unbiased for $\bar{x}$.

(b) Take $B = m$. Find the marginal law of C_i and its limit as $m \to \infty$. Find the limiting probability that item i is omitted. Why does this calculation not show that the resampled items are independent population observations?

(c) Suppose instead that indices are sampled without replacement. Prove that sampling all m indices gives a uniformly random permutation. Show that the first $B < m$ indices are a uniformly random ordered sample without replacement, or equivalently a uniformly random B-element subset followed by a uniformly random ordering.

Exercise 5: Dependence retained by resampling.

Use total covariance to distinguish conditional independence of resampled draws from unconditional dependence inherited from the corpus.

Let $X_1, \dots, X_m$ be square-integrable and possibly dependent. Independently of $X_{1:m}$, draw I_1, I_2 independently and uniformly from $[m]$, and set $Y_b = X_{I_b}$. Prove

$$\operatorname{Cov}(Y_1, Y_2) = \operatorname{Var}\left(\frac{1}{m} \sum_{i=1}^{m} X_i \right).$$

Interpret the result.

Hints Condition on $X_{1:m}$ and use the law of total covariance.

Exercise 6: Dependence created by overlapping windows.

Compute the beta-mixing coefficient induced by overlapping iid windows and distinguish token independence from window independence.

Let $\mathcal{Z}$ be a finite or countable alphabet, let $p : \mathcal{Z} \to [0,1]$ satisfy $\sum_{z \in \mathcal{Z}} p(z) = 1$, and let $(Z_t)_{t \geq 1}$ be iid $\mathcal{Z}$-valued random variables with law p. Let $L \geq 2$ be an integer, and define the $\mathcal{Z}^L$-valued windows $W_t = (Z_t, \dots, Z_{t+L-1})$. For an integer $a \geq 1$, define the beta-mixing coefficient

$$\beta_W(a) = \sup_{t \geq 1} \left\| \mathcal{L}(W_{1:t}, W_{t+a:\infty}) - \mathcal{L}(W_{1:t}) \otimes \mathcal{L}(W_{t+a:\infty}) \right\|_{\mathrm{TV}},$$

(a) Show that the window process is $(L-1)$-dependent: collections of windows separated by at least L starting positions are independent. Deduce that $\beta_W(a) = 0$ for $a \geq L$.

(b) For $1 \leq a < L$, show that

$$\beta_W(a) = 1 - \left(\sum_{z \in \mathcal{Z}} p(z)^2 \right)^{L-a}.$$

Explain from this formula why the coefficient is close to one when the collision probability $\sum_{z \in \mathcal{Z}} p(z)^2$ is small or when the number $L - a$ of shared positions is large, even though the underlying token process is independent.

Hints For (b), identify the underlying variables that occur in both window collections.

Exercise 7: The cost of polynomial beta-mixing.

Convert a polynomial mixing rate into a sample-size bound and compare its dependence on accuracy with iid and geometric-mixing bounds.

Assume the setting of Theorem 6.7.1 and $\beta(a) \le Ca^{-r}$ for constants $C, r > 0$. Let $\epsilon, \delta \in (0,1)$ and $A = \log(2|\mathcal{F}|/\delta)$.

(a) Starting from Eq. (6.13), give sufficient conditions on the selected-sample size q and gap a that make each term at most $\delta/2$.

(b) Use $n \ge 1 + (q-1)a$ and a suitable choice of q, a to show, up to universal constants and integer rounding, that it suffices to take

$$n \gtrsim \frac{A}{\epsilon} + \left(\frac{A}{\epsilon}\right)^{1+1/r} \left(\frac{C}{\delta}\right)^{1/r}.$$

(c) Compare the dependence on ϵ with iid sampling and with geometric beta-mixing. Is the degradation exponential in $1/\epsilon$?

HINTS Use $(1-\epsilon)^q \le e^{-\epsilon q}$. First choose $q \asymp A/\epsilon$, then choose a so that $qCa^{-r} \lesssim \delta$, and finally take n of order qa.

A Resampling a fixed corpus

Fix numbers $x_1, \dots, x_m$, let $\bar{x} = m^{-1}\sum_i x_i$, and draw $I_1, \dots, I_B$ independently and uniformly from $[m]$. Define counts $C_i = \sum_{b=1}^{B} \mathbb{I}\{I_b = i\}$ and

$$\bar{X}^* = \frac{1}{B}\sum_{b=1}^{B} x_{I_b} = \frac{1}{B}\sum_{i=1}^{m} C_i x_i.$$

Conditionally on the corpus,

$$(C_1, \dots, C_m) \sim \text{Multinomial}\left(B; \frac{1}{m}, \dots, \frac{1}{m}\right), \qquad \mathbb{E}[\bar{X}^* \mid x_{1:m}] = \bar{x}.$$

For $B = m$, each $C_i \sim \text{Binomial}(m, 1/m)$ and converges in law to Poisson(1) as $m \to \infty$. Thus the expected fraction of the corpus omitted by one bootstrap resample approaches e^{-1}. If $B = Em$ and E grows, then C_i/E concentrates near one: repeated resampling reconstructs the original empirical weights.

Now let $X_1, \dots, X_m$ be square-integrable and possibly dependent, and set $Y_b = X_{I_b}$. Two resampled observations are independent conditional on the corpus, but unconditionally

$$\text{Cov}(Y_1, Y_2) = \text{Var}(\bar{X}) = \frac{1}{m^2}\sum_{r,s=1}^{m} \text{Cov}(X_r, X_s). \tag{6.14}$$

The law of total variance gives

$$\text{Var}(\bar{X}^*) = \text{Var}(\bar{X}) + \frac{1}{B}\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})^2\right]. \tag{6.15}$$

The second term is nonnegative. Averaging out the resampling randomness—the Rao–Blackwell step—returns $\bar{X}$. Resampling cannot erase the first term, which contains the original dependence.

Small or structured subsamples can still change a statistical procedure. For example, taking well-separated positions from a time series can lower pairwise dependence, but it also uses fewer positions. Whether the tradeoff helps depends on the estimator and the dependence structure; it is not a consequence of replacement sampling itself.

B Proof of the independent-points comparison

Proposition B.1 (Independent-points comparison)**.** *Under the assumptions preceding Eq. (6.12),*

$$\left\|\mathcal{L}(X_{t_1},\dots,X_{t_q}) - P^{\otimes q}\right\|_{\mathrm{TV}} \le (q-1)\beta(a).$$

Proof. Let $Z_j = X_{t_j}$, $Q_j = \mathcal{L}(Z_1,\dots,Z_j)$, and $d_j = \|Q_j - P^{\otimes j}\|_{\mathrm{TV}}$. Restricting the distributions in Eq. (6.11) to $(Z_1,\dots,Z_j)$ gives

$$\|Q_j - Q_{j-1}\otimes P\|_{\mathrm{TV}} \le \beta(a).$$

Taking the product of both measures with the same probability law preserves total variation distance. Since $Q_1 = P$, the triangle inequality gives

$$d_1 = 0, \qquad d_j \le \beta(a) + d_{j-1}.$$

Iteration yields $d_q \le (q-1)\beta(a)$. □

References

Richard C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. Probability Surveys, 2:107–144, 2005. https://doi.org/10.1214/154957805100000104.

Hantian Ding, Zijian Wang, Giovanni Paolini, Varun Kumar, Anoop Deoras, Dan Roth, and Stefano Soatto. Fewer truncations improve language modeling. 2024. https://arxiv.org/abs/2404.10830.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and others. The Llama 3 herd of models. 2024. https://arxiv.org/abs/2407.21783.

Martin Hairer. Exponential mixing for a stochastic PDE driven by degenerate noise. Nonlinearity, 15(2):271–279, 2002. https://arxiv.org/abs/math-ph/0103039.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301):13–30, 1963. https://doi.org/10.1080/01621459.1963.10500830.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, and others. The FineWeb datasets: Decanting the web for the finest text data at scale. 2024. https://arxiv.org/abs/2406.17557.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, and others. Dolma: An open corpus of three trillion tokens for language model pretraining research. 2024. https://arxiv.org/abs/2402.00159.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. Annals of Probability, 22(1):94–116, 1994. https://doi.org/10.1214/aop/1176988849.