Lecture 9: AI Applications Neural Network Compression and Distance Measures between Probabilistic Language Models

SUZUKI, Atsushi
Jing WANG

Contents

1 Introduction

1.1 Review of the Previous Lecture

In the previous lecture, we rigorously defined and calculated frameworks for the **automatic and quantitative evaluation** of a **single probabilistic language model**, covering perplexity, accuracy of the most likely option in multiple-choice questions, and various metrics for string output (EM/F1, BLEU, ROUGE, chrF, BERTScore, numerical Accuracy).

1.2 Learning Outcomes of This Lecture

By the end of this lecture, students should be able to:

- Explain the motivation and methods to **reduce the scale (model compression)** of a **neural network** while maintaining its properties as a **function**.
- When a **probabilistic language model** is modified, evaluate the **amount of change from the original model** in a **mathematically rigorous** and **quantitative** manner.

2 Preliminaries: Mathematical Notations

We will reiterate the basic notations used in this lecture.

- **Definition:**
  - $(\mathrm{LHS}) \coloneqq (\mathrm{RHS})$: Indicates that the left-hand side is defined by the right-hand side. For example, $a \coloneqq b$ indicates that a is defined as b.
- **Set:**
  - Sets are often denoted by uppercase calligraphic letters. E.g., $\mathcal{A}$.
  - $x \in \mathcal{A}$: Indicates that the element x belongs to the set $\mathcal{A}$.
  - $\{\}$: The empty set.
  - $\{a, b, c\}$: The set consisting of elements a, b, c (roster notation).
  - $\{x \in \mathcal{A} | P(x)\}$: The set of elements in $\mathcal{A}$ for which the proposition $P(x)$ is true (set-builder notation).
  - $|\mathcal{A}|$: The number of elements in the set $\mathcal{A}$ (in this lecture, used principally for finite sets).
  - $\mathbb{R}$: The set of all real numbers.
  - $\mathbb{R}_{>0}$: The set of all positive real numbers.
  - $\mathbb{R}_{\geq 0}$: The set of all non-negative real numbers.

- $\mathbb{Z}$: The set of all integers.
- $\mathbb{Z}_{>0}$: The set of all positive integers.
- $\mathbb{Z}_{\geq 0}$: The set of all non-negative integers.
- $[1, k]_{\mathbb{Z}}$: When k is a positive integer, $[1, k]_{\mathbb{Z}} \coloneqq \{1, 2, \dots, k\}$, i.e., the set of integers from 1 to k. When $k = +\infty$, $[1, k]_{\mathbb{Z}} \coloneqq \mathbb{Z}_{>0}$, i.e., the set of all positive integers.

- **Function:**
  - $f : \mathcal{X} \to \mathcal{Y}$: Indicates that the function f is a map that takes an element from set $\mathcal{X}$ as input and outputs an element from set $\mathcal{Y}$.
  - $y = f(x)$: Indicates that the output of the function f for an input $x \in \mathcal{X}$ is $y \in \mathcal{Y}$.

- **Vector:**
  - In this course, a vector refers to a column of numbers.
  - Vectors are denoted by bold italic lowercase letters. E.g., $\boldsymbol{v}$.
  - $\boldsymbol{v} \in \mathbb{R}^n$: Indicates that the vector $\boldsymbol{v}$ is an n-dimensional real vector.
  - The i-th element of a vector $\boldsymbol{v}$ is denoted by v_i.

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}. \tag{1}$$

  - For two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{emb}}}$, the standard inner product

- **Sequence:**
  - Given a set $\mathcal{A}$, an integer $n \in \mathbb{Z}_{>0} \cup \{+\infty\}$, and a function $\boldsymbol{a} : [1, n]_{\mathbb{Z}} \to \mathcal{A}$, we call $\boldsymbol{a}$ a sequence of length n consisting of elements from the set $\mathcal{A}$. When $n < +\infty$, the sequence is called a finite sequence, and when $n = \infty$, it is called an infinite sequence.
  - Sequences are denoted by bold italic lowercase letters, just like vectors. This is because a finite sequence can be considered an extension of a real vector. In fact, a finite sequence of elements from $\mathbb{R}$ can be regarded as a real vector.
  - For a sequence $\boldsymbol{a}$ of length n with elements from set $\mathcal{A}$, the i-th component a_i for $i \in [1, n]_{\mathbb{Z}}$ is defined as $a_i \coloneqq \boldsymbol{a}(i)$.
  - When $n < +\infty$, a sequence $\boldsymbol{a}$ of length n with elements from set $\mathcal{A}$ is determined by its elements $a_1, a_2, ..., a_n$, so we write it as $\boldsymbol{a} = (a_1, a_2, ..., a_n)$. Similarly, when $\boldsymbol{a}$ is an infinite sequence, we write it as $\boldsymbol{a} = (a_1, a_2, ...)$.

  - The length of a sequence $\boldsymbol{a}$ is denoted by $|\boldsymbol{a}|$.

- **Matrix:**
  - Matrices are denoted by bold italic uppercase letters. E.g., $\boldsymbol{A}$.
  - $\boldsymbol{A} \in \mathbb{R}^{m,n}$: Indicates that the matrix $\boldsymbol{A}$ is an $m \times n$ real matrix.
  - The element in the i-th row and j-th column of a matrix $\boldsymbol{A}$ is denoted by $a_{i,j}$.

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}. \tag{2}$$

  - The transpose of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\top$. If $\boldsymbol{A} \in \mathbb{R}^{m,n}$, then $\boldsymbol{A}^\top \in \mathbb{R}^{n,m}$, and

$$\boldsymbol{A}^\top = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{m,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & \cdots & a_{m,n} \end{bmatrix} \tag{3}$$

    is given.
  - A vector is also a matrix with one column, and its transpose can also be defined.

$$\boldsymbol{v}^\top = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \in \mathbb{R}^{1,n} \tag{4}$$

    is given.

- **Tensor:**
  - In this lecture, the word tensor simply refers to a multi-dimensional array. A vector can be seen as a 1st-order tensor, and a matrix as a 2nd-order tensor. Tensors of 3rd order or higher are denoted by underlined bold italic uppercase letters, like $\underline{\boldsymbol{A}}$.
  - Students who have already learned about abstract tensors in mathematics or physics might feel uncomfortable calling a mere multi-dimensional array a tensor. If we consider that the basis is always fixed to the standard basis and identify the mathematical meaning of a tensor with its component representation (which becomes a multi-dimensional array), then the terminology is (at least) consistent.

3 Neural Network Compression (Model Compression)

3.1 Revisiting the Formulation of AI as Function Learning

The primary goal of AI was to **learn an appropriate relationship (function) between inputs and outputs**. However, even after a good function has been obtained, there is a strong practical motivation to modify it to reduce **implementation resources** (memory, computational complexity, latency) while **preserving the input-output relationship of that function as much as possible**. In this lecture, we will refer to such an act of reducing implementation resources while maintaining the input-output relationship of a function as model **compression**.

3.2 Motivation and Overview of Compression

Even when a good function f has already been obtained, there is motivation to make the **model's representation (parameterization)** more lightweight while preserving the **function values (input-output relationship)**. Representative methods include **low-precision floating point**, **quantization**, and **model distillation** [2, 4–6, 9].

3.3 Rigorous Definition of Low-precision floating point

Definition 3.1 (Generalized Low-Precision Real Number System and Rounding Operator)**.** Let $d_{\mathrm{param}} \in \mathbb{Z}_{>0}$. For each index $i \in \{1, \ldots, d_{\mathrm{param}}\}$, we are given a **finite set** $\mathbb{F}_i \subset \mathbb{R}$ as a **floating-point format** and a **rounding operator** $R_i : \mathbb{R} \to \mathbb{F}_i$.

$$\text{Original parameter space} \qquad \mathcal{F} := \mathbb{F}_1 \times \cdots \times \mathbb{F}_{d_{\mathrm{param}}} \subset \mathbb{R}^{d_{\mathrm{param}}}. \tag{5}$$

Let the post-low-precision format for each component be $\mathbb{F}'_i \subset \mathbb{R}$, and the corresponding rounding operator be $R'_i : \mathbb{R} \to \mathbb{F}'_i$, then

$$\text{Post-low-precision parameter space} \qquad \mathcal{F}' := \mathbb{F}'_1 \times \cdots \times \mathbb{F}'_{d_{\mathrm{param}}} \subset \mathbb{R}^{d_{\mathrm{param}}}. \tag{6}$$

For each format $\mathbb{F}$, we define its **required number of bits** $b(\mathbb{F}) \in \mathbb{Z}_{\geq 0}$ as ”the minimum bit length for which an encoding map $\iota : \mathbb{F} \hookrightarrow \{0, 1\}^{b(\mathbb{F})}$ exists to represent each element of $\mathbb{F}$”. We impose the **resource non-increase constraint** for low-precision conversion as

$$b(\mathbb{F}'_i) \leq b(\mathbb{F}_i) \qquad (i = 1, \ldots, d_{\mathrm{param}}) \tag{7}$$

For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d_{\mathrm{param}}}) \in \mathbb{R}^{d_{\mathrm{param}}}$, we define the **generalized low-precision mapping**

$\Phi_{\mathrm{fp}} : \mathbb{R}^{d_{\mathrm{param}}} \to \mathcal{F}'$ as

$$\Phi_{\mathrm{fp}}(\boldsymbol{\theta}) := \left(R'_1(\theta_1),\ R'_2(\theta_2),\ \ldots,\ R'_{d_{\mathrm{param}}}(\theta_{d_{\mathrm{param}}})\right) \tag{8}$$

The method of replacing $\boldsymbol{\theta}$ with $\Phi_{\mathrm{fp}}(\boldsymbol{\theta})$ for computation during inference (or a part of training) is called **low-precision conversion**.

Remark 3.1. Equations (5)–(8) allow for a mixed type where the format differs for each **element (per-parameter)**. In **mixed-precision training**, different $\mathbb{F}_i$ are assigned to gradients, gradient accumulations, weights, and activations to reduce computational resources [9].

3.4 Rigorous Definition of Quantization

Definition 3.2 (General Form of Integer Quantization with Partitions and Dequantization)**.** Given a **parametric function** $f_{(\cdot)}$, a parameter vector $\boldsymbol{\theta} \in \mathbb{F}^{d_{\mathrm{param}}}$, an **ordered partition** $(\mathcal{I}_1, \ldots, \mathcal{I}_k)$ of the index set $\mathcal{I} := \{1, \ldots, d_{\mathrm{param}}\}$, and a **maximum absolute value in the integer domain** $M_j \in \mathbb{Z}_{>0}$ for each partition $(j = 1, \ldots, k)$. For quantization, we choose **modified parameters** $\boldsymbol{\theta}' \in \mathbb{F}^{d_{\mathrm{param}}}$, and for each j, a **quantization scale** $s_j \in \mathbb{R}_{>0}$ and a **quantization zero-point / bias** $b_j \in \mathbb{Z}$.

For each partition $\mathcal{I}_j$, we define a **quantization mapping** $Q_j : \mathbb{R} \to \mathbb{Z}$ and a **dequantization mapping** $\widetilde{Q}_j : \mathbb{Z} \to \mathbb{R}$ as

$$Q_j(x) := \mathrm{clip}_{[-M_j, M_j]}\Big(\mathrm{round}\big(\tfrac{x}{s_j}\big) + b_j\Big), \tag{9}$$

$$\widetilde{Q}_j(n) := s_j\left(n - b_j\right) \qquad (n \in \mathbb{Z}) \tag{10}$$

The vector versions are defined as

$$\text{Quantize}_{(\mathcal{I})_{j=1}^{k}, \boldsymbol{s}, \boldsymbol{b}}(\boldsymbol{\theta}') := \boldsymbol{q} \in \mathbb{Z}^{d_{\mathrm{param}}}, \quad q_i := Q_j(\theta'_i) \ \ (i \in \mathcal{I}_j), \tag{11}$$

$$\text{Dequantize}_{(\mathcal{I})_{j=1}^{k}, \boldsymbol{s}, \boldsymbol{b}}(\boldsymbol{q}) := \widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{d_{\mathrm{param}}}, \quad \widetilde{\theta}_i := \widetilde{Q}_j(q_i) \ \ (i \in \mathcal{I}_j) \tag{12}$$

We write the **inference function with quantized parameters** as

$$f_{\text{Quantize}_{(\mathcal{I})_{j=1}^{k}, \boldsymbol{s}, \boldsymbol{b}}(\boldsymbol{\theta}')} \tag{13}$$

The goal is, for a given $f_{\boldsymbol{\theta}}$, to choose appropriate $(\boldsymbol{\theta}', \boldsymbol{s}, \boldsymbol{b})$ to achieve

$$f_{\text{Quantize}_{(\mathcal{I})_{j=1}^{k}, \boldsymbol{s}, \boldsymbol{b}}(\boldsymbol{\theta}')} \approx f_{\boldsymbol{\theta}} \tag{14}$$

Fundamental operations such as matrix multiplication are performed in the integer domain $\mathbb{Z}$, accompanied by scale composition and rescaling as needed [6].

Remark 3.2. As a method to achieve Equation (14), the approach of setting $\boldsymbol{\theta}' = \boldsymbol{\theta}$ and determining $(\boldsymbol{s}, \boldsymbol{b})$ based on data statistics is called **post-training quantization**. On the other hand, the method of optimizing $(\boldsymbol{\theta}', \boldsymbol{s}, \boldsymbol{b})$ including the training process to directly improve post-quantization performance is called **quantization-aware training** [5, 6].

Remark 3.3. In this lecture, we have focused on **parameter quantization**, but a common design is to quantize **intermediate outputs (activations)** during inference and evaluate the combination (node output and edge weight) with integer multiplication. This allows the arithmetic units to specialize in integer operations, achieving higher speed and lower power consumption [6].

Example 3.1 (Manual Calculation Example of Quantize and Dequantize)**.** Given: $d_{\text{param}} = 4$, $\boldsymbol{\theta}' = (1.20,\ -0.55,\ 0.07,\ 2.31)$, partition $(\mathcal{I}_1, \mathcal{I}_2) = (\{1,2\}, \{3,4\})$, $M_1 = 127$, $M_2 = 7$, $s_1 = 0.01$, $b_1 = 0$, $s_2 = 0.1$, $b_2 = 1$.

Component 1 ($i = 1 \in \mathcal{I}_1$): $q_1 = \text{clip}_{[-127,127]}\big(\text{round}(1.20/0.01) + 0\big) = \text{clip}(120) = 120,$
$\widetilde{\theta}_1 = 0.01 \cdot (120 - 0) = 1.20.$

Component 2 ($i = 2 \in \mathcal{I}_1$): $q_2 = \text{clip}\big(\text{round}(-0.55/0.01)\big) = \text{clip}(-55) = -55,$
$\widetilde{\theta}_2 = 0.01 \cdot (-55 - 0) = -0.55.$

Component 3 ($i = 3 \in \mathcal{I}_2$): $q_3 = \text{clip}_{[-7,7]}\big(\text{round}(0.07/0.1) + 1\big) = \text{clip}(1 + 1) = 2,$
$\widetilde{\theta}_3 = 0.1 \cdot (2 - 1) = 0.1.$

Component 4 ($i = 4 \in \mathcal{I}_2$): $q_4 = \text{clip}_{[-7,7]}\big(\text{round}(2.31/0.1) + 1\big) = \text{clip}(23 + 1) = \text{clip}(24) = 7,$
$\widetilde{\theta}_4 = 0.1 \cdot (7 - 1) = 0.6.$

From the above, $\boldsymbol{q} = (120,\ -55,\ 2,\ 7)$ and $\widetilde{\boldsymbol{\theta}} = (1.20,\ -0.55,\ 0.1,\ 0.6)$. It can be seen that the large value 2.31 is saturated by M_2, increasing the error.

Exercise 3.1 (Quantize/Dequantize Exercise)**.** Let $d_{\text{param}} = 3$, $\boldsymbol{\theta}' = (0.34,\ -1.26,\ 0.51)$, partition $(\mathcal{I}_1, \mathcal{I}_2) = (\{1,3\}, \{2\})$, $M_1 = 15$, $M_2 = 127$, $s_1 = 0.02$, $b_1 = 2$, $s_2 = 0.01$, $b_2 = 0$.

Answer. Component 1 ($i = 1 \in \mathcal{I}_1$): $\text{round}(0.34/0.02) = \text{round}(17) = 17$, $q_1 = \text{clip}_{[-15,15]}(17 + 2) = \text{clip}(19) = 15$, $\widetilde{\theta}_1 = 0.02 \cdot (15 - 2) = 0.26$.
Component 2 ($i = 2 \in \mathcal{I}_2$): $\text{round}(-1.26/0.01) = \text{round}(-126) = -126$, $q_2 = \text{clip}_{[-127,127]}(-126 + 0) = -126$, $\widetilde{\theta}_2 = 0.01 \cdot (-126 - 0) = -1.26$.
Component 3 ($i = 3 \in \mathcal{I}_1$): $\text{round}(0.51/0.02) = \text{round}(25.5) = 26$, $q_3 = \text{clip}_{[-15,15]}(26 + 2) = \text{clip}(28) = 15$, $\widetilde{\theta}_3 = 0.02 \cdot (15 - 2) = 0.26$.
Thus, $\boldsymbol{q} = (15,\ -126,\ 15)$ and $\widetilde{\boldsymbol{\theta}} = (0.26,\ -1.26,\ 0.26)$. The error is large because the 1st and 3rd components were saturated by $M_1 = 15$.

3.5 Rigorous Definition of Knowledge Distillation

Definition 3.3 (Model Distillation)**.** Suppose we are given a function $f_{\boldsymbol{\theta}}$ composed of a **parametric function** $f_{(\cdot)}$ and a parameter vector $\boldsymbol{\theta}$. For the purpose of approximating $f_{\boldsymbol{\theta}}$ at a lower cost, we use another parametric function $g_{(\cdot)}$ with a lower-dimensional parameter space, and find a suitable parameter vector $\boldsymbol{\gamma}$ such that $g_{\boldsymbol{\gamma}}$ becomes a good approximation of $f_{\boldsymbol{\theta}}$. This process is called **model distillation**.

As a result, $g_{\boldsymbol{\gamma}}$ can be used as a low-cost substitute for $f_{\boldsymbol{\theta}}$.

Remark 3.4. Distillation is broadly classified into **model distillation** (transferring the behavior of a large teacher model to a small student) and **dataset distillation** (approximating the learning effect of the original dataset with a small number of synthetic or summarized data). While the formulation in this section focuses on the former, the latter can also be expressed by replacing $f_{\boldsymbol{\theta}}(x)$ in the objective function with ”the teacher’s output for teacher-generated data”.

4 Divergence between Probabilistic Language Models

The core of large language models that handle natural language was the probabilistic language model composed of neural networks. When this is compressed, one becomes concerned about how much it differs from the original probabilistic language model. This chapter deals with metrics that express how much another probabilistic language model has diverged when there is a reference probabilistic language model.

4.1 A Simple Method: Differences in Individual Evaluations such as Multiple-Choice Questions

A simple and commonly used method is to measure the performance of two models on the same task, such as a multiple-choice question, and compare the **difference**. However, if the evaluation is based solely on the answers, it is possible to overlook cases where the inference processes are significantly different even if the final answers are the same [1]. Therefore, one might consider directly comparing the probability distributions constituted by the probabilistic language models.

4.2 Review of Probabilistic Language Models

A probabilistic language model was, mathematically, a conditional probability function. Let’s review it.

Definition 4.1 (Vocabulary and Token Sequence, Notation for Subsequences)**.** The set of possible values a token can take is called the **vocabulary**, denoted by $\mathcal{V}^{a}$. Hereafter, when the vocabulary size is D, we will identify the vocabulary $\mathcal{V}$ with the subset of natural numbers

$[1, D]_{\mathbb{Z}} = \{1, 2, \ldots, D\}$. The set of token sequences of length n can be regarded as the direct product of n copies of $\mathcal{V}$, and this is written as $\mathcal{V}^n$. Note that $\mathcal{V}^0$ is the set consisting of the token sequence $()$. Furthermore, the set of token sequences of finite length is written as $\mathcal{V}^*$. $\mathcal{V}^* = \mathcal{V}^0 \cup \mathcal{V}^1 \cup \mathcal{V}^2 \cup \cdots$.

For a token sequence $\boldsymbol{t} = (t_1, \ldots, t_n) \in \mathcal{V}^n$, we define the notation for **subsequences** as follows:

$$\boldsymbol{t}_{a:b} \coloneqq (t_a, t_{a+1}, \ldots, t_b) \quad (1 \leq a \leq b \leq n), \tag{15}$$

$$\boldsymbol{t}_{<i} \coloneqq \boldsymbol{t}_{1:(i-1)} \quad (i \geq 1), \qquad \boldsymbol{t}_{\leq i} \coloneqq \boldsymbol{t}_{1:i}, \tag{16}$$

$$\boldsymbol{t}_{>i} \coloneqq \boldsymbol{t}_{(i+1):n} \quad (i \leq n), \qquad \boldsymbol{t}_{\geq i} \coloneqq \boldsymbol{t}_{i:n}. \tag{17}$$

The empty subsequence is $\boldsymbol{t}_{<1} = \boldsymbol{t}_{>n} = ()$. Also, to emphasize the **first n elements**, we write $\boldsymbol{t}_{1:n}$.

[a] In previous lectures, the set of nodes in a neural network was also denoted by $\mathcal{V}$, but since this lecture does not explicitly describe the graph structure of neural networks, $\mathcal{V}$ should always be taken to refer to the vocabulary.

Definition 4.2 (Probabilistic Language Model (most general form))**.** Fix a finite vocabulary $\mathcal{V} = \{1, 2, \ldots, D\}$. A **probabilistic language model** is a function $P(\cdot|\cdot)$ that, given any finite-length token sequence $\boldsymbol{t} = (t_1, \ldots, t_n) \in \mathcal{V}^n$, returns the probability mass function of the next token, conditioned on it. More formally, a two-variable function $P(\cdot|\cdot) : \mathcal{V} \times \mathcal{V}^* \to [0, 1]$ is a probabilistic language model if for any $\boldsymbol{t} \in \mathcal{V}^*$,

$$\sum_{v \in \mathcal{V}} P(v \mid \boldsymbol{t}) = 1 \tag{18}$$

holds.

Since a probabilistic language model is mathematically a conditional probability mass function, the quantification of divergence between probabilistic language models can ultimately be reduced to the problem of divergence between general probability mass functions.

4.3 How to Quantify the Divergence of Probability Mass Functions

Given a ”correct” probability mass function P to be referenced and another probability mass function Q, we want to measure how much Q diverges from P. If Q does not diverge much from P, then for a z where $P(z)$ is large, $Q(z)$ also tends to be large. Therefore, for a random variable Z following the probability distribution defined by P, the value of $Q(Z)$ should tend to be large on average (with respect to the probabilistic behavior of Z). Thus, using a monotonically decreasing function ϕ to penalize small probability masses,

$$\mathbb{E}_{Z \sim P}\left[\phi\left(Q(Z)\right)\right] - C \tag{19}$$

it is natural to consider a divergence criterion of this form. Here, the argument of a probability mass function (not a probabilistic language model) is principally denoted by z or Z. In the following, we will establish the axioms that this criterion should satisfy and rigorously show that it uniquely determines the form of ϕ, and thus specifies the KL divergence (relative entropy).

Definition 4.3 (Axiomatization of Expectation-of-Difference Type Divergence Functional)**.** **Setting**: Consider probability mass functions (pmf) on a finite set $\mathcal{S}$. Assume that any pmfs P, Q have the same finite support $\mathcal{S}$, and that for any z where $P(z) > 0$, $Q(z) > 0$ also holds (the converse will be explicitly stated when necessary).

For a monotonically decreasing continuous function $\phi : (0, 1] \to \mathbb{R}$ and a constant $C \in \mathbb{R}$, consider the functional on any pmfs P, Q

$$\Delta_\phi(P \parallel Q) \coloneqq \mathbb{E}_{Z\sim P}\big[\phi\big(Q(Z)\big)\big] - C \tag{20}$$

We impose the following axioms on this:

(A1) **Reflexivity**: $\Delta_\phi(P \parallel P) = 0$.

(A2) **Non-negativity**: $\Delta_\phi(P \parallel Q) \geq 0$, with equality if and only if $P = Q$ (not just equivalent a.e., but identical on $\mathcal{S}$).

(A3) **Continuity**: $\Delta_\phi(P \parallel Q)$ is continuous with respect to the usual topology on the pointwise values of Q.

(A4) **Additivity over independent products**: For two finite sets $\mathcal{S}_1, \mathcal{S}_2$ and their respective pmf pairs (P_1, Q_1) (with common support $\mathcal{S}_1$) and (P_2, Q_2) (with common support $\mathcal{S}_2$), using the **product distributions** (Definition 4.4) $P_1 \otimes P_2$ and $Q_1 \otimes Q_2$,

$$\Delta_\phi(P_1 \otimes P_2 \parallel Q_1 \otimes Q_2) = \Delta_\phi(P_1 \parallel Q_1) + \Delta_\phi(P_2 \parallel Q_2) \tag{21}$$

holds.

Definition 4.4 (Product $\otimes$ of pmfs)**.** For pmfs P_1, Q_1 on a finite set $\mathcal{S}_1$ (with common support $\mathcal{S}_1$) and P_2, Q_2 on $\mathcal{S}_2$ (with common support $\mathcal{S}_2$), the pmf $P_1 \otimes P_2$ on $\mathcal{S}_1 \times \mathcal{S}_2$ is defined as

$$(P_1 \otimes P_2)(x_1, x_2) \coloneqq P_1(x_1)\, P_2(x_2) \tag{22}$$

(and similarly for $Q_1 \otimes Q_2$).

Theorem 4.1 (Characterization of the Logarithm Function from Expectation-of-Difference and Uniqueness of KL)**.** For a monotonically decreasing continuous function ϕ satisfying (A1) – (A4) in Definition 4.3, there exist constants $c \in \mathbb{R}_{>0}$ and $B \in \mathbb{R}$ such that

$$\phi(u) = -c\, \log u + B \qquad (u \in (0, 1]) \tag{23}$$

holds. Furthermore, to satisfy (A1), the constant C must be

$$C = \mathbb{E}_{Z\sim P}\big[\phi\big(P(Z)\big)\big] \tag{24}$$

(a normalization term dependent on P). Therefore,

$$\begin{aligned}\Delta_\phi(P \parallel Q) &= \mathbb{E}_P[-c\log Q(Z) + B] - \mathbb{E}_P[-c\log P(Z) + B] \\ &= c\,\mathbb{E}_P\left[\log\frac{P(Z)}{Q(Z)}\right] = c\,D_{\mathrm{KL}}(P \parallel Q)\end{aligned} \tag{25}$$

Proof. **(1) Derivation of a functional equation from additivity over products**. From (A4), for any finite sets $\mathcal{S}_1, \mathcal{S}_2$ and pmf pairs (P_1, Q_1), (P_2, Q_2),

$$\begin{aligned}\Delta_\phi(P_1 \otimes P_2 \parallel Q_1 \otimes Q_2) &= \mathbb{E}_{(Z_1,Z_2)\sim P_1\otimes P_2}\Big[\phi\Big((Q_1 \otimes Q_2)(Z_1, Z_2)\Big)\Big] - C_{12} \\ &= \Delta_\phi(P_1 \parallel Q_1) + \Delta_\phi(P_2 \parallel Q_2)\end{aligned} \tag{26}$$

for some constant C_{12}. Applying (A1) with $Q_1 = P_1$ and $Q_2 = P_2$ gives

$$C_{12} = \mathbb{E}_{(Z_1,Z_2)\sim P_1\otimes P_2}\Big[\phi\Big((P_1 \otimes P_2)(Z_1, Z_2)\Big)\Big]\,. \tag{27}$$

Note here that the argument of ϕ is limited to the form $Q_1(Z_1)Q_2(Z_2)$ (or $P_1(Z_1)P_2(Z_2)$). Regarding $u \in (0,1]$ as a possible value of $Q_1(Z_1)$ and $v \in (0,1]$ as a possible value of $Q_2(Z_2)$, the fact that the relationship in (26) and (27) holds for any choice of pmf implies an equality depending on u, v:

$$\phi(uv) - \phi(u) - \phi(v) = K \tag{28}$$

holds for a constant K (which depends only on $\phi(1)$). Indeed, by fixing the values of Z_1, Z_2 and comparing (26), the same difference must arise for combinations of arguments of ϕ. By specifying $\phi(1)$ by applying (A1) with $P = Q$ on a one-element set $\mathcal{S} = \{1\}$, we find $K = -\phi(1)$, so

$$\tilde{\phi}(u) \coloneqq \phi(u) - \phi(1) \quad\Rightarrow\quad \tilde{\phi}(uv) = \tilde{\phi}(u) + \tilde{\phi}(v). \tag{29}$$

(2) Elementary solution of the additive equation (using continuity). Let $u = \mathrm{e}^t$, $v = \mathrm{e}^s$ and define $\psi(t) \coloneqq \tilde{\phi}(\mathrm{e}^t)$. Then (29) becomes

$$\psi(t + s) = \psi(t) + \psi(s) \qquad (t, s \in \mathbb{R}) \tag{30}$$

From the continuity in (A3), $\tilde{\phi}$ is continuous, hence ψ is also continuous. From this, we show elementarily that ψ is a linear function:

(2-1) Linearity on rational numbers: For $n \in \mathbb{Z}_{>0}$, repeating (30) n times with $s = t$ gives $\psi(nt) = n\psi(t)$. Also, substituting t/n for t gives $\psi(t) = n\psi(t/n)$, therefore $\psi\big((m/n)t\big) = (m/n)\psi(t)$ holds for $m \in \mathbb{Z}$.

(2-2) Extension to real numbers: Take any $r \in \mathbb{R}$ and $\varepsilon > 0$. There exists a rational sequence $q_k \to r$, and by continuity, $\psi(q_k) \to \psi(r)$. On the other hand, from (2-1), $\psi(q_k) = q_k\,\psi(1)$, so taking the limit gives $\psi(r) = r\,\psi(1)$. Therefore

$$\psi(t) = c'\,t \quad (c' = \psi(1) \in \mathbb{R}). \tag{31}$$

(3) Determination of the form of ϕ. We can write $\tilde{\phi}(u) = \psi(\log u) = c' \log u$. So $\phi(u) = \tilde{\phi}(u) + \phi(1) = c' \log u + B$. To satisfy (A1), it must be that

$$0 = \Delta_\phi(P \parallel P) = \mathbb{E}_P[\phi(P(Z))] - C \quad \Rightarrow \quad C = \mathbb{E}_P[\phi(P(Z))] \tag{32}$$

Hereafter, we replace $c := -c'$ to obtain (23).

(4) Reduction to KL and derivation of $c > 0$ (elementary proof of Gibbs' inequality). Substituting (23) and (32) into (20), we get

$$\begin{aligned}\Delta_\phi(P \parallel Q) &= -c\,\mathbb{E}_P[\log Q(Z)] + B - \Big(-c\,\mathbb{E}_P[\log P(Z)] + B\Big) \\ &= c \sum_{z\in\mathcal{S}} P(z) \log \frac{P(z)}{Q(z)} = c\,D_{\mathrm{KL}}(P \parallel Q).\end{aligned} \tag{33}$$

Here, we show $D_{\mathrm{KL}}(P \parallel Q) \geq 0$ elementarily. For any $a > 0$, $\log a \leq a - 1$ (from the tangent line of the concave function $\log$, or from the derivative $f'(a) = 1 - 1/a$ and $f''(a) = 1/a^2 > 0$ of $f(a) = a - 1 - \log a$ (≥ 0)). Letting $a = \frac{Q(z)}{P(z)}$,

$$-\log \frac{Q(z)}{P(z)} \geq 1 - \frac{Q(z)}{P(z)}. \tag{34}$$

Multiplying both sides by $P(z)$ and summing over $\mathcal{S}$,

$$\sum_z P(z) \log \frac{P(z)}{Q(z)} \geq \sum_z \big(P(z) - Q(z)\big) = 1 - 1 = 0. \tag{35}$$

Equality holds if $\frac{Q(z)}{P(z)} = 1$ at points where $P > 0$, i.e., only when $P = Q$. Therefore, to satisfy (A2), $c > 0$ is necessary. □

Remark 4.1. Theorem 4.1 shows that from natural axioms such as **expectation of difference**, **additivity for product distributions**, and **continuity**, ϕ is uniquely determined to be a **logarithm function** (up to an additive constant), and consequently, the **KL divergence (relative entropy)** is uniquely derived [3, Ch. 2].

4.4 Definition of Kullback – Leibler (KL) divergence

Definition 4.5 (KL Divergence (for pmfs))**.** Let P, Q be two probability mass functions with the same finite set $\mathcal{S}$ as their support, and assume that for any z where $P(z) > 0$, $Q(z) > 0$ also holds (so the ratio is defined). The **KL divergence (relative entropy)** is defined as

$$D_{\mathrm{KL}}(P \parallel Q) := \sum_{z\in\mathcal{S}} P(z) \, \log\!\left(\frac{P(z)}{Q(z)}\right) \in [0, \infty] \tag{36}$$

Non-negativity and the equality condition $D_{\mathrm{KL}}(P \parallel Q) = 0 \;\Leftrightarrow\; P = Q$ follow from Equation (35) [3, 7].

Example 4.1 (Complete Numerical Example of KL)**.** Consider the probability distributions on the vocabulary $\{a, b, c\}$

$$P(a) = 0.5, \; P(b) = 0.3, \; P(c) = 0.2, \qquad Q(a) = 0.4, \; Q(b) = 0.4, \; Q(c) = 0.2$$

From Equation (36),

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x\in\{a,b,c\}} P(x) \, \log\frac{P(x)}{Q(x)} \\ &= 0.5\log\frac{0.5}{0.4} + 0.3\log\frac{0.3}{0.4} + 0.2\log\frac{0.2}{0.2}. \end{aligned} \tag{37}$$

The last term vanishes since $\log 1 = 0$. We calculate the other two terms sequentially:

$$\begin{aligned} 0.5\log\frac{0.5}{0.4} &= 0.5\log(1.25), \\ 0.3\log\frac{0.3}{0.4} &= 0.3\log(0.75). \end{aligned}$$

Using the natural logarithm,

$$\log(1.25) \approx 0.22314, \quad \log(0.75) \approx -0.28768,$$

hence

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &\approx 0.5 \times 0.22314 + 0.3 \times (-0.28768) \\ &\approx 0.11157 - 0.08630 = 0.02527 \text{ [nats]}. \end{aligned} \tag{38}$$

If the base of the logarithm is 2, it is $0.02527/\log 2 \approx 0.03645$ [bits].

Exercise 4.1 (Numerical Calculation of KL)**.** On the vocabulary $\{x_1, x_2, x_3\}$, let

$$P = (0.2, \; 0.5, \; 0.3), \qquad Q = (0.1, \; 0.7, \; 0.2)$$

Calculate $D_{\mathrm{KL}}(P \parallel Q)$ (using natural logarithm).

Answer.

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= 0.2 \log \frac{0.2}{0.1} + 0.5 \log \frac{0.5}{0.7} + 0.3 \log \frac{0.3}{0.2} \\ &= 0.2 \log 2 + 0.5 \log(5/7) + 0.3 \log(3/2). \end{aligned}$$

Numerically, $\log 2 \approx 0.69315,\ \log(5/7) \approx -0.33647,\ \log(3/2) \approx 0.40547$, so

$0.2 \times 0.69315 + 0.5 \times (-0.33647) + 0.3 \times 0.40547 = 0.13863 - 0.16824 + 0.12164 \approx 0.09203$ [nats].

Proposition 4.1 (Additivity of KL for Product Distributions)**.** For the product $\otimes$ from Definition 4.4, and for pmf pairs (P_1, Q_1) on a finite set $\mathcal{S}_1$ (with common support) and (P_2, Q_2) on $\mathcal{S}_2$ (with common support),

$$D_{\mathrm{KL}}(P_1 \otimes P_2 \parallel Q_1 \otimes Q_2) = D_{\mathrm{KL}}(P_1 \parallel Q_1) + D_{\mathrm{KL}}(P_2 \parallel Q_2) \tag{39}$$

holds [3, Prop. 2.4].

Proof. From Definition 4.5 and (22),

$$\begin{aligned} &D_{\mathrm{KL}}(P_1 \otimes P_2 \parallel Q_1 \otimes Q_2) \\ &= \sum_{(x_1,x_2)\in\mathcal{S}_1\times\mathcal{S}_2} P_1(x_1)P_2(x_2)\ \log\left(\frac{P_1(x_1)P_2(x_2)}{Q_1(x_1)Q_2(x_2)}\right) \\ &= \sum_{x_1,x_2} P_1(x_1)P_2(x_2) \left[\log\left(\frac{P_1(x_1)}{Q_1(x_1)}\right) + \log\left(\frac{P_2(x_2)}{Q_2(x_2)}\right)\right] \\ &= \sum_{x_1} P_1(x_1) \log\left(\frac{P_1(x_1)}{Q_1(x_1)}\right) \sum_{x_2} P_2(x_2) \\ &\qquad + \sum_{x_2} P_2(x_2) \log\left(\frac{P_2(x_2)}{Q_2(x_2)}\right) \sum_{x_1} P_1(x_1) \\ &= D_{\mathrm{KL}}(P_1 \parallel Q_1) + D_{\mathrm{KL}}(P_2 \parallel Q_2), \end{aligned}$$

where we used $\sum_{x_i} P_i(x_i) = 1$. □

4.5 Extension of KL to Language Models (Conditional Distributions) and Two Definitions

When a language model is viewed simply as a conditional probability distribution, the most straightforward definition is the following.

Definition 4.6 (A. KL based on Joint Distribution)**.** Take a fixed length $n \in \mathbb{Z}_{>0}$. For language models P, Q, we define the **joint distributions** they induce, respectively,

$$P^{(n)}(\boldsymbol{t}_{1:n}) \coloneqq \prod_{i=1}^{n} P(t_i \mid \boldsymbol{t}_{<i}), \qquad Q^{(n)}(\boldsymbol{t}_{1:n}) \coloneqq \prod_{i=1}^{n} Q(t_i \mid \boldsymbol{t}_{<i}) \tag{40}$$

(where $\boldsymbol{t}_{<i} = (t_1, \ldots, t_{i-1})$). Then we define

$$D_{\mathrm{KL}}\big(P^{(n)} \parallel Q^{(n)}\big) = \sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n}) \log \frac{P^{(n)}(\boldsymbol{t}_{1:n})}{Q^{(n)}(\boldsymbol{t}_{1:n})} \tag{41}$$

To handle variable-length token sequences and from a computational complexity perspective, the following sequential representation using the chain rule is useful.

Proposition 4.2 (Chain Rule for KL Divergence)**.** For any $n \in \mathbb{Z}_{>0}$,

$$D_{\mathrm{KL}}\big(P^{(n)} \parallel Q^{(n)}\big) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim P^{(i-1)}} \big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel Q(\cdot \mid \boldsymbol{t}_{<i})\big) \big] \tag{42}$$

holds.

Proof. Using definitions (41) and (40),

$$\log \frac{P^{(n)}(\boldsymbol{t}_{1:n})}{Q^{(n)}(\boldsymbol{t}_{1:n})} = \sum_{i=1}^{n} \log \frac{P(t_i \mid \boldsymbol{t}_{<i})}{Q(t_i \mid \boldsymbol{t}_{<i})}.$$

Taking the expectation of both sides with respect to $P^{(n)}$ and exchanging the sum and expectation by linearity,

$$\begin{aligned} D_{\mathrm{KL}}(P^{(n)} \parallel Q^{(n)}) &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{1:n} \sim P^{(n)}} \left[\log \frac{P(t_i \mid \boldsymbol{t}_{<i})}{Q(t_i \mid \boldsymbol{t}_{<i})} \right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim P^{(i-1)}} \left[\sum_{t_i} P(t_i \mid \boldsymbol{t}_{<i}) \log \frac{P(t_i \mid \boldsymbol{t}_{<i})}{Q(t_i \mid \boldsymbol{t}_{<i})} \right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim P^{(i-1)}} \big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel Q(\cdot \mid \boldsymbol{t}_{<i})\big) \big]. \end{aligned}$$

□

Definition 4.7 (B. KL based on the Conditional Next-Token Distribution using a Dataset)**.** Given a validation dataset $\mathcal{D} = \{\boldsymbol{t}^{(j)}\}_{j=1}^{N}$. For each sequence, decompose it into input tokens $\boldsymbol{x}^{(j)}$ and output tokens $\boldsymbol{y}^{(j)}$ such that $\boldsymbol{t}^{(j)} = \boldsymbol{x}^{(j)}\boldsymbol{y}^{(j)}$ ($\boldsymbol{x}^{(j)}$ is always known as the context). Then, for each j, we extend $\boldsymbol{y}^{(j)}$ one token at a time and sum the sequential KL divergences. Furthermore, we also take the average over the sequences:

$$\widehat{D}_{\mathrm{KL}}^{\mathcal{D}}(P \parallel Q) \coloneqq \frac{1}{\sum_{j=1}^{N} |\boldsymbol{y}^{(j)}|} \sum_{j=1}^{N} \sum_{i=1}^{|\boldsymbol{y}^{(j)}|} D_{\mathrm{KL}}\Big(P(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}_{<i}^{(j)}) \parallel Q(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}_{<i}^{(j)})\Big). \tag{43}$$

Proposition 4.3 (Virtual Equivalence of A and B)**.** Assume that all sequence lengths are equal to n ($|\boldsymbol{y}^{(j)}| = n$), that $\mathcal{D}$ is generated i.i.d. according to $P^{(n)}$, and that we take the expectation weighted by the occurrence frequency of each prefix $\boldsymbol{t}_{<i}$. Then

$$\mathbb{E}_{\mathcal{D}\sim(P^{(n)})^{\otimes N}}\left[\widehat{D}_{\mathrm{KL}}^{\mathcal{D}}(P \parallel Q)\right] = \frac{1}{n} D_{\mathrm{KL}}\big(P^{(n)} \parallel Q^{(n)}\big). \tag{44}$$

Proof. **Starting Point (i.i.d. and one sample)**. We assume $\mathcal{D} = \{\boldsymbol{t}^{(j)}\}_{j=1}^{N}$ is generated i.i.d. according to $P^{(n)}$ (assumption of Proposition 4.3). From the definition in Equation (43), by the linearity of expectation,

$$\mathbb{E}_{\mathcal{D}}\left[\widehat{D}_{\mathrm{KL}}^{\mathcal{D}}(P \parallel Q)\right] = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{\sum_{j=1}^{N}|\boldsymbol{y}^{(j)}|}\sum_{j=1}^{N}\sum_{i=1}^{|\boldsymbol{y}^{(j)}|} D_{\mathrm{KL}}\Big(P(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}_{<i}^{(j)}) \parallel Q(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}_{<i}^{(j)})\Big)\right] \tag{45}$$

$$= \frac{1}{Nn}\sum_{j=1}^{N}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{1:n}^{(j)}\sim P^{(n)}}\left[\sum_{u\in\mathcal{V}} P\Big(u \mid \boldsymbol{t}_{<i}^{(j)}\Big)\log\frac{P\Big(u \mid \boldsymbol{t}_{<i}^{(j)}\Big)}{Q\Big(u \mid \boldsymbol{t}_{<i}^{(j)}\Big)}\right], \tag{46}$$

where we used $|\boldsymbol{y}^{(j)}| = n$. Since each sample follows the same distribution (i.i.d.), the same expectation appears N times, independent of j:

$$(46) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{T}_{1:n}\sim P^{(n)}}\left[\sum_{u\in\mathcal{V}} P(u \mid \boldsymbol{T}_{<i})\log\frac{P(u \mid \boldsymbol{T}_{<i})}{Q(u \mid \boldsymbol{T}_{<i})}\right], \tag{47}$$

where $\boldsymbol{T}_{1:n}$ is one sample following $P^{(n)}$.

Writing out the marginalization of the prefix. For any function $g(\boldsymbol{t}_{<i})$, the expectation under $P^{(n)}$ is by discrete sum

$$\mathbb{E}_{\boldsymbol{T}_{1:n}\sim P^{(n)}}[g(\boldsymbol{T}_{<i})] = \sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n})\, g(\boldsymbol{t}_{<i}) = \sum_{\boldsymbol{t}_{<i}}\sum_{t_i}\sum_{\boldsymbol{t}_{>i}}\Big\{P^{(n)}(\boldsymbol{t}_{1:n})\, g(\boldsymbol{t}_{<i})\Big\}. \tag{48}$$

From the chain rule expression (40),

$$P^{(n)}(\boldsymbol{t}_{1:n}) = \Big(\prod_{j<i} P(t_j \mid \boldsymbol{t}_{<j})\Big) P(t_i \mid \boldsymbol{t}_{<i}) \Big(\prod_{j>i} P(t_j \mid \boldsymbol{t}_{<j})\Big). \tag{49}$$

On the right-hand side, $g(\boldsymbol{t}_{<i})$ does not depend on $\boldsymbol{t}_{>i}$, so we can compute the inner sum over $\boldsymbol{t}_{>i}$ first:

$$\sum_{\boldsymbol{t}_{>i}}\prod_{j>i} P(t_j \mid \boldsymbol{t}_{<j}) = 1. \tag{50}$$

Therefore,

$$\begin{aligned}(48) &= \sum_{\boldsymbol{t}_{<i}} \sum_{t_i} \left\{ \Big(\prod_{j<i} P(t_j \mid \boldsymbol{t}_{<j}) \Big) P(t_i \mid \boldsymbol{t}_{<i}) \right\} g(\boldsymbol{t}_{<i}) \\ &= \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i})\, g(\boldsymbol{t}_{<i}). \qquad (51)\end{aligned}$$

Here we have set $P^{(i-1)}(\boldsymbol{t}_{<i}) \coloneqq \prod_{j<i} P(t_j \mid \boldsymbol{t}_{<j})$.

Specifying each term of Equation (47). Choosing $g(\boldsymbol{t}_{<i}) \coloneqq \sum_u P(u \mid \boldsymbol{t}_{<i}) \log \frac{P(u|\boldsymbol{t}_{<i})}{Q(u|\boldsymbol{t}_{<i})}$ and applying (51),

$$\mathbb{E}_{\boldsymbol{T}_{1:n}\sim P^{(n)}} \left[\sum_u P(u \mid \boldsymbol{T}_{<i}) \log \frac{P(u \mid \boldsymbol{T}_{<i})}{Q(u \mid \boldsymbol{T}_{<i})} \right] = \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i})\, D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel Q(\cdot \mid \boldsymbol{t}_{<i})\big). \qquad (52)$$

Substituting this into (47),

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\Big[\widehat{D}^{\mathcal{D}}_{\mathrm{KL}}(P \parallel Q)\Big] &= \frac{1}{n} \sum_{i=1}^{n} \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i})\, D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel Q(\cdot \mid \boldsymbol{t}_{<i})\big) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i}\sim P^{(i-1)}} \left[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel Q(\cdot \mid \boldsymbol{t}_{<i})\big) \right]. \qquad (53)\end{aligned}$$

Equivalence with joint distribution KL. Finally, by averaging the decomposition $\log \frac{P^{(n)}(\boldsymbol{t}_{1:n})}{Q^{(n)}(\boldsymbol{t}_{1:n})} = \sum_{i=1}^{n} \log \frac{P(t_i|\boldsymbol{t}_{<i})}{Q(t_i|\boldsymbol{t}_{<i})}$ with respect to $P^{(n)}$ and using a similar decomposition as in (51) for each position i, we get

$$D_{\mathrm{KL}}\big(P^{(n)} \parallel Q^{(n)}\big) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i}\sim P^{(i-1)}} \left[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel Q(\cdot \mid \boldsymbol{t}_{<i})\big) \right]. \qquad (54)$$

Comparing (53) and (54), we obtain

$$\mathbb{E}_{\mathcal{D}\sim(P^{(n)})^{\otimes N}} \Big[\widehat{D}^{\mathcal{D}}_{\mathrm{KL}}(P \parallel Q)\Big] = \frac{1}{n} D_{\mathrm{KL}}\big(P^{(n)} \parallel Q^{(n)}\big) \qquad (55)$$

This completes the proof. □

Example 4.2 (Numerical Example of Sequential KL (fixed length 2))**.** Let the vocabulary be $\{A, B\}$ and length $n = 2$. The conditional distributions of P, Q are

$$P(A \mid ()) = 0.6,\ P(B \mid ()) = 0.4, \quad P(A \mid A) = 0.7,\ P(A \mid B) = 0.2,$$

$$Q(A \mid ()) = 0.5,\ Q(B \mid ()) = 0.5, \quad Q(A \mid A) = 0.6,\ Q(A \mid B) = 0.3$$

(the others are 1 minus these values). First, the KL for $i = 1$:

$$D_{\mathrm{KL}}\big(P(\cdot \mid ()) \parallel Q(\cdot \mid ())\big) = 0.6 \log \frac{0.6}{0.5} + 0.4 \log \frac{0.4}{0.5}.$$

Next, the conditional KL for $i = 2$, averaged under $t_1 \sim P(\cdot \mid ())$:

$$\mathbb{E}_{t_1 \sim P}\big[D_{\mathrm{KL}}\big(P(\cdot \mid t_1) \parallel Q(\cdot \mid t_1)\big)\big] = 0.6 \cdot \left(0.7 \log \frac{0.7}{0.6} + 0.3 \log \frac{0.3}{0.4}\right) + 0.4 \cdot \left(0.2 \log \frac{0.2}{0.3} + 0.8 \log \frac{0.8}{0.7}\right).$$

Summing these gives $D_{\mathrm{KL}}(P^{(2)} \parallel Q^{(2)})$ (Equation (42)).

Exercise 4.2 (Sequential KL Calculation Practice)**.** Let vocabulary be $\{0, 1\}$, $P(1 \mid ()) = 0.3$, $P(1 \mid 1) = 0.6$, $P(1 \mid 0) = 0.2$, $Q(1 \mid ()) = 0.4$, $Q(1 \mid 1) = 0.5$, $Q(1 \mid 0) = 0.3$ (others are differences from 1). For $n = 2$, find $D_{\mathrm{KL}}(P^{(2)} \parallel Q^{(2)})$.

Answer. First token: $0.3 \log(0.3/0.4) + 0.7 \log(0.7/0.6)$. **$P$-average of conditional KL for the second token**:

$$0.3 \cdot \big[0.6 \log(0.6/0.5) + 0.4 \log(0.4/0.5)\big] + 0.7 \cdot \big[0.2 \log(0.2/0.3) + 0.8 \log(0.8/0.7)\big].$$

The sum of both is the answer (Equation (42)). Numerical substitution gives approximately 0.019 [nats].

4.6 Definition of Jensen – Shannon (JS) divergence

Definition 4.8 (JS Divergence based on (Probability Distribution) Joint Distribution)**.** For the **mixture distribution** $M \coloneqq \frac{1}{2}(P + Q)$ of P, Q,

$$D_{\mathrm{JS}}(P \parallel Q) \coloneqq \tfrac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \tfrac{1}{2} D_{\mathrm{KL}}(Q \parallel M) \in [0, \log 2] \tag{56}$$

is called the **Jensen – Shannon divergence**. D_{JS} is symmetric and bounded [8].

Remark 4.2. D_{JS} is a **symmetrized KL**, averaging the **divergence in both directions** via M. The square root $\sqrt{D_{\mathrm{JS}}}$ is known to satisfy the axioms of a **metric** [8].

Example 4.3 (Complete Numerical Example of JS)**.** Using P, Q from Example 4.1. $M = \frac{1}{2}(P + Q)$ is

$$M = (0.45,\ 0.35,\ 0.2).$$

By Equation (56),

$$D_{\mathrm{JS}}(P \parallel Q) = \tfrac{1}{2} \sum_x P(x) \log \frac{P(x)}{M(x)} + \tfrac{1}{2} \sum_x Q(x) \log \frac{Q(x)}{M(x)}. \tag{57}$$

We calculate each term sequentially:

$$\sum_x P(x)\log\frac{P(x)}{M(x)} = 0.5\log\frac{0.5}{0.45} + 0.3\log\frac{0.3}{0.35} + 0.2\log\frac{0.2}{0.2},$$

$$\sum_x Q(x)\log\frac{Q(x)}{M(x)} = 0.4\log\frac{0.4}{0.45} + 0.4\log\frac{0.4}{0.35} + 0.2\log\frac{0.2}{0.2}.$$

One should sum the terms that are not 0 and finally multiply by $1/2$.

Exercise 4.3 (Numerical Calculation of JS)**.** For $P = (0.2, 0.5, 0.3)$, $Q = (0.1, 0.7, 0.2)$ from Example 4.1, find $D_{\mathrm{JS}}(P \parallel Q)$.

Answer. $M = (0.15, 0.6, 0.25)$.

$$\sum_x P\log\frac{P}{M} = 0.2\log\frac{0.2}{0.15} + 0.5\log\frac{0.5}{0.6} + 0.3\log\frac{0.3}{0.25},$$

$$\sum_x Q\log\frac{Q}{M} = 0.1\log\frac{0.1}{0.15} + 0.7\log\frac{0.7}{0.6} + 0.2\log\frac{0.2}{0.25}.$$

The answer is $\frac{1}{2}$ of the sum of both. Numerically, it is approximately 0.016 [nats].

Definition 4.9 (JS based on (Language Model) Joint Distribution)**.** Let M be the average conditional distribution for each prefix, $M(\cdot \mid \boldsymbol{t}_{<i}) := \frac{1}{2}\{P(\cdot \mid \boldsymbol{t}_{<i}) + Q(\cdot \mid \boldsymbol{t}_{<i})\}$. For a fixed length n,

$$\begin{aligned} D_{\mathrm{JS}}\big(P^{(n)} \parallel Q^{(n)}\big) = \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{1:n}\sim P^{(n)}}\left[\log\frac{P^{(n)}(\boldsymbol{t}_{1:n})}{M^{(n)}(\boldsymbol{t}_{1:n})}\right] \\ + \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{1:n}\sim Q^{(n)}}\left[\log\frac{Q^{(n)}(\boldsymbol{t}_{1:n})}{M^{(n)}(\boldsymbol{t}_{1:n})}\right] \end{aligned} \tag{58}$$

is defined.

Proposition 4.4 (Chain rule for JS in language models)**.** For any $n \in \mathbb{Z}_{>0}$,

$$\begin{aligned} D_{\mathrm{JS}}\big(P^{(n)} \parallel Q^{(n)}\big) = \sum_{i=1}^{n} \Big\{ \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{<i}\sim P^{(i-1)}}\big[D_{\mathrm{KL}}(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i}))\big] \\ + \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{<i}\sim Q^{(i-1)}}\big[D_{\mathrm{KL}}(Q(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i}))\big]\Big\} \end{aligned} \tag{59}$$

holds.

Proof. **Step 1: Decomposition of the logarithm** From the definition of joint distribution (40), for any sequence $\boldsymbol{t}_{1:n}$,

$$\log\frac{P^{(n)}(\boldsymbol{t}_{1:n})}{M^{(n)}(\boldsymbol{t}_{1:n})} = \sum_{i=1}^{n}\log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})}. \tag{60}$$

The same equation holds for Q.

Step 2: Writing out the expectation with respect to $P^{(n)}$ For any integrable function $g(\boldsymbol{t}_{1:n})$, by discrete sum,

$$\mathbb{E}_{\boldsymbol{t}_{1:n}\sim P^{(n)}}[g(\boldsymbol{t}_{1:n})] = \sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n})\, g(\boldsymbol{t}_{1:n}). \tag{61}$$

Here, setting $g(\boldsymbol{t}_{1:n}) = \log\bigl(P^{(n)}(\boldsymbol{t}_{1:n})/M^{(n)}(\boldsymbol{t}_{1:n})\bigr)$ and substituting (60),

$$\begin{aligned}\mathbb{E}_{\boldsymbol{t}_{1:n}\sim P^{(n)}}\left[\log\frac{P^{(n)}(\boldsymbol{t}_{1:n})}{M^{(n)}(\boldsymbol{t}_{1:n})}\right] &= \sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n}) \sum_{i=1}^{n} \log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})} \\ &= \sum_{i=1}^{n} \sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n})\, \log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})}. \end{aligned} \tag{62}$$

The last equality is merely an exchange of order of finite sums.

Step 3: Decomposition by prefix and current token For each fixed i, we decompose the sum over $\boldsymbol{t}_{1:n}$ into sums over ”prefix $\boldsymbol{t}_{<i}$”, ”current token t_i”, and ”suffix $\boldsymbol{t}_{>i}$”. That is,

$$\begin{aligned}&\sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n})\, \log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})} \\ &= \sum_{\boldsymbol{t}_{<i}} \sum_{t_i} \sum_{\boldsymbol{t}_{>i}} \left\{ P^{(n)}(\boldsymbol{t}_{1:n})\, \log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})} \right\}. \end{aligned} \tag{63}$$

Using the chain product form (40),

$$P^{(n)}(\boldsymbol{t}_{1:n}) = \Bigl(\prod_{j<i} P(t_j \mid \boldsymbol{t}_{<j})\Bigr) P(t_i \mid \boldsymbol{t}_{<i}) \Bigl(\prod_{j>i} P(t_j \mid \boldsymbol{t}_{<j})\Bigr). \tag{64}$$

Of the right-hand side, $\log\bigl(P(t_i \mid \boldsymbol{t}_{<i})/M(t_i \mid \boldsymbol{t}_{<i})\bigr)$ depends only on t_i and $\boldsymbol{t}_{<i}$, not on $\boldsymbol{t}_{>i}$. Thus, we can perform the innermost sum (over $\boldsymbol{t}_{>i}$) in (63) first:

$$\sum_{\boldsymbol{t}_{>i}} \left\{ \Bigl(\prod_{j>i} P(t_j \mid \boldsymbol{t}_{<j})\Bigr) \right\} = 1. \tag{65}$$

This is the fact that we reach 1 by sequentially using $\sum_{u\in\mathcal{V}} P(u \mid \cdot) = 1$ at each position.

Step 4: Identity of marginalization Substituting (64) and (65) into (63),

$$\begin{aligned}&\sum_{\boldsymbol{t}_{1:n}} P^{(n)}(\boldsymbol{t}_{1:n})\, \log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})} \\ &= \sum_{\boldsymbol{t}_{<i}} \sum_{t_i} \left\{ \Bigl(\prod_{j<i} P(t_j \mid \boldsymbol{t}_{<j})\Bigr) P(t_i \mid \boldsymbol{t}_{<i}) \right\} \log\frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})}. \end{aligned} \tag{66}$$

Here, since $\prod_{j<i} P(t_j \mid \boldsymbol{t}_{<j}) = P^{(i-1)}(\boldsymbol{t}_{<i})$,

$$(66) = \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i}) \sum_{t_i} P(t_i \mid \boldsymbol{t}_{<i}) \log \frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})}. \tag{67}$$

Step 5: Summarizing the P-side contribution Combining (62) and (67),

$$\begin{aligned}\mathbb{E}_{\boldsymbol{t}_{1:n} \sim P^{(n)}}\left[\log \frac{P^{(n)}}{M^{(n)}}\right] &= \sum_{i=1}^{n} \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i}) \sum_{t_i} P(t_i \mid \boldsymbol{t}_{<i}) \log \frac{P(t_i \mid \boldsymbol{t}_{<i})}{M(t_i \mid \boldsymbol{t}_{<i})} \\ &= \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim P^{(i-1)}}\Big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big]. \end{aligned} \tag{68}$$

Step 6: Same for the Q-side Applying the exact same argument to $Q^{(n)}$,

$$\mathbb{E}_{\boldsymbol{t}_{1:n} \sim Q^{(n)}}\left[\log \frac{Q^{(n)}}{M^{(n)}}\right] = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim Q^{(i-1)}}\Big[D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big]. \tag{69}$$

Step 7: Combination Substituting (68) and (69) into the definition (58),

$$\begin{aligned} D_{\mathrm{JS}}\big(P^{(n)} \parallel Q^{(n)}\big) = \tfrac{1}{2} &\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim P^{(i-1)}}\Big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big] \\ &+ \tfrac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{t}_{<i} \sim Q^{(i-1)}}\Big[D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big], \end{aligned}$$

which is the claim (59). This completes the proof. □

Similar to KL, sequential evaluation using data sequences is practically useful.

Definition 4.10 (JS based on (Language Model) Conditional Next-Token Distribution)**.** Given a validation dataset $\mathcal{D} = \{\boldsymbol{t}^{(j)}\}_{j=1}^{N}$, decompose it as $\boldsymbol{t}^{(j)} = \boldsymbol{x}^{(j)}\boldsymbol{y}^{(j)}$, and for each position, define $M(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i}) \coloneqq \frac{1}{2}\{P(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i}) + Q(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i})\}$. Then

$$\begin{aligned} \widehat{D}^{\mathcal{D}}_{\mathrm{JS}}(P \parallel Q) \coloneqq \frac{1}{\sum_{j=1}^{N} |\boldsymbol{y}^{(j)}|} \sum_{j=1}^{N} \sum_{i=1}^{|\boldsymbol{y}^{(j)}|} \Big[&\tfrac{1}{2} D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i}) \parallel M(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i})\big) \\ &+ \tfrac{1}{2} D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i}) \parallel M(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i})\big)\Big] \end{aligned} \tag{70}$$

is defined.

Proposition 4.5 (Virtual Equivalence of Joint and Next-Token JS versions)**.** Assume that $|\boldsymbol{y}^{(j)}| = n$ is constant for all j, that $\mathcal{D}$ is consistent with a mixed generation from $P^{(n)}$ and $Q^{(n)}$ (such that the left and right expectations are evaluated by their respective chains), and that the weighting is chosen such that the empirical measures of prefixes match $P^{(i-1)}$ and $Q^{(i-1)}$

respectively. Then

$$\mathbb{E}\Big[\widehat{D}^{\mathcal{D}}_{\mathrm{JS}}(P \parallel Q)\Big] = \frac{1}{n} D_{\mathrm{JS}}\big(P^{(n)} \parallel Q^{(n)}\big). \tag{71}$$

Proof. **Starting Point (i.i.d. and one sample)**. We assume $|\boldsymbol{y}^{(j)}| = n$ is constant for all j, and by the assumption of Proposition 4.5, $\mathcal{D} = \{\boldsymbol{t}^{(j)}\}_{j=1}^{N}$ has a mixed generation consistent with $P^{(n)}$ and $Q^{(n)}$ (weighted so that the expectations for the P-side and Q-side chains are evaluated accordingly). From Definition 4.10 and the linearity of expectation,

$$\begin{aligned}\mathbb{E}\Big[\widehat{D}^{\mathcal{D}}_{\mathrm{JS}}(P \parallel Q)\Big] = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{i=1}^{n} \mathbb{E}\Big[&\tfrac{1}{2}D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i}) \parallel M(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i})\big) \\ &+\tfrac{1}{2}D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i}) \parallel M(\cdot \mid \boldsymbol{x}^{(j)}, \boldsymbol{y}^{(j)}_{<i})\big)\Big]\end{aligned} \tag{72}$$

Due to i.i.d. and the consistency of the mixed generation, the expectation of each term is the same regardless of j, so

$$\begin{aligned}(72) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{T}_{1:n}\sim P^{(n)}}\Bigg[\sum_{u\in\mathcal{V}} P(u \mid \boldsymbol{T}_{<i}) \log\frac{P(u \mid \boldsymbol{T}_{<i})}{M(u \mid \boldsymbol{T}_{<i})}\Bigg] \\ &+ \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{S}_{1:n}\sim Q^{(n)}}\Bigg[\sum_{u\in\mathcal{V}} Q(u \mid \boldsymbol{S}_{<i}) \log\frac{Q(u \mid \boldsymbol{S}_{<i})}{M(u \mid \boldsymbol{S}_{<i})}\Bigg]\Bigg\},\end{aligned} \tag{73}$$

where $\boldsymbol{T}_{1:n} \sim P^{(n)}$ and $\boldsymbol{S}_{1:n} \sim Q^{(n)}$ are single samples, respectively.

Marginalization by prefix (P-side). As in the KL case, for any $g(\boldsymbol{t}_{<i})$,

$$\mathbb{E}_{\boldsymbol{T}_{1:n}\sim P^{(n)}}[g(\boldsymbol{T}_{<i})] = \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i})\, g(\boldsymbol{t}_{<i}), \tag{74}$$

holds (similar decomposition as in Equation (51)). Choosing $g(\boldsymbol{t}_{<i}) \coloneqq \sum_u P(u \mid \boldsymbol{t}_{<i}) \log\frac{P(u|\boldsymbol{t}_{<i})}{M(u|\boldsymbol{t}_{<i})}$,

$$\mathbb{E}_{\boldsymbol{T}_{1:n}\sim P^{(n)}}\Bigg[\sum_{u} P(u \mid \boldsymbol{T}_{<i}) \log\frac{P(u \mid \boldsymbol{T}_{<i})}{M(u \mid \boldsymbol{T}_{<i})}\Bigg] = \sum_{\boldsymbol{t}_{<i}} P^{(i-1)}(\boldsymbol{t}_{<i})\, D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big). \tag{75}$$

Marginalization by prefix (Q-side). Similarly,

$$\mathbb{E}_{\boldsymbol{S}_{1:n}\sim Q^{(n)}}[h(\boldsymbol{S}_{<i})] = \sum_{\boldsymbol{s}_{<i}} Q^{(i-1)}(\boldsymbol{s}_{<i})\, h(\boldsymbol{s}_{<i}), \tag{76}$$

and choosing $h(\boldsymbol{s}_{<i}) \coloneqq \sum_u Q(u \mid \boldsymbol{s}_{<i}) \log\frac{Q(u|\boldsymbol{s}_{<i})}{M(u|\boldsymbol{s}_{<i})}$,

$$\mathbb{E}_{\boldsymbol{S}_{1:n}\sim Q^{(n)}}\Bigg[\sum_{u} Q(u \mid \boldsymbol{S}_{<i}) \log\frac{Q(u \mid \boldsymbol{S}_{<i})}{M(u \mid \boldsymbol{S}_{<i})}\Bigg] = \sum_{\boldsymbol{s}_{<i}} Q^{(i-1)}(\boldsymbol{s}_{<i})\, D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{s}_{<i}) \parallel M(\cdot \mid \boldsymbol{s}_{<i})\big). \tag{77}$$

Combining contributions per position. Substituting (75) and (77) into (73),

$$\mathbb{E}\left[\widehat{D}_{\mathrm{JS}}^{\mathcal{D}}(P \parallel Q)\right] = \frac{1}{n}\sum_{i=1}^{n}\left\{\tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{<i}\sim P^{(i-1)}}\Big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big]\right.$$
$$\left. + \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{s}_{<i}\sim Q^{(i-1)}}\Big[D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{s}_{<i}) \parallel M(\cdot \mid \boldsymbol{s}_{<i})\big)\Big]\right\}. \tag{78}$$

Equivalence with joint distribution JS. The terms appearing on the right-hand side of Definition 4.9,

$$\mathbb{E}_{\boldsymbol{t}_{1:n}\sim P^{(n)}}\left[\log\frac{P^{(n)}(\boldsymbol{t}_{1:n})}{M^{(n)}(\boldsymbol{t}_{1:n})}\right] = \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{t}_{<i}\sim P^{(i-1)}}\Big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big], \tag{79}$$

$$\mathbb{E}_{\boldsymbol{t}_{1:n}\sim Q^{(n)}}\left[\log\frac{Q^{(n)}(\boldsymbol{t}_{1:n})}{M^{(n)}(\boldsymbol{t}_{1:n})}\right] = \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{t}_{<i}\sim Q^{(i-1)}}\Big[D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big], \tag{80}$$

both follow directly from a prefix decomposition similar to the KL case (an argument of the form of Equation (51)) and $\log\frac{P^{(n)}}{M^{(n)}} = \sum_i \log\frac{P(\cdot|\boldsymbol{t}_{<i})}{M(\cdot|\boldsymbol{t}_{<i})}$ (Equation (60)). Substituting these into Definition (58),

$$D_{\mathrm{JS}}\big(P^{(n)} \parallel Q^{(n)}\big) = \sum_{i=1}^{n}\left\{\tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{<i}\sim P^{(i-1)}}\Big[D_{\mathrm{KL}}\big(P(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big]\right.$$
$$\left. + \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{t}_{<i}\sim Q^{(i-1)}}\Big[D_{\mathrm{KL}}\big(Q(\cdot \mid \boldsymbol{t}_{<i}) \parallel M(\cdot \mid \boldsymbol{t}_{<i})\big)\Big]\right\}. \tag{81}$$

Comparing (78) and (81), we obtain

$$\mathbb{E}\left[\widehat{D}_{\mathrm{JS}}^{\mathcal{D}}(P \parallel Q)\right] = \frac{1}{n}\,D_{\mathrm{JS}}\big(P^{(n)} \parallel Q^{(n)}\big) \tag{82}$$

This completes the proof. □

5 Summary

In this lecture, we organized the motivation for **reducing the scale of a model while preserving its input-output relationship as a function**. Furthermore, we quantified the **difference between probabilistic language models before and after modification** for scale reduction using **KL divergence** and **JS divergence**.

References

[1] Rishiraj Acharya. Why maybe we're measuring llm compression wrong. `https://huggingface.co/blog/rishiraj/kld-guided-quantization`, 2025. Blog article.

[2] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 535–541, 2006.

[3] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley-Interscience, 2 edition, 2006.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop, 2015. arXiv:1503.02531.

[5] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. Journal of Machine Learning Research, 18(187):1–30, 2018. Earlier version: arXiv:1609.07061 (2016).

[6] Benoit Jacob, Skirmantas Kligys, Bo Chen, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2704–2713, 2018.

[7] Solomon Kullback and Richard A. Leibler. On information and sufficiency. Annals of Mathematical Statistics, 22(1):79–86, 1951.

[8] Jianhua Lin. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory, 37(1):145–151, 1991.

[9] Paulius Micikevicius, Sharan Narang, Jonah Alben, et al. Mixed precision training. In International Conference on Learning Representations (ICLR) Workshop, 2018. arXiv:1710.03740.