A Markov Categorical Framework for Language Modeling

Yifan Zhang
Princeton University
yifzhang@princeton.edu

Abstract

Autoregressive language models achieve remarkable performance, yet a unified theory explaining their internal mechanisms, how training shapes representations, and why these representations support complex behavior remains incomplete. We introduce an analytical framework that models the single-step generation process as a composition of information-processing stages using the language of Markov categories. This compositional perspective connects three aspects of language modeling that are often studied separately: the training objective, the geometry of the learned representation space, and practical model capabilities. First, our framework gives an information-theoretic rationale for parallel drafting methods such as speculative decoding by quantifying the information surplus a hidden state contains about future tokens beyond the immediate next one. Second, we clarify how the standard negative log-likelihood (NLL) objective learns not only a most likely next token, but also the data's intrinsic conditional uncertainty, formalized through categorical entropy. Our main spectral result is conditional: for a linear-softmax head with bounded output features, a calibrated quadratic upper-bound surrogate to NLL induces, after whitening or variance normalization, a generalized CCA/eigenproblem aligning representation directions with predictive prototypes. This gives a compositional lens for understanding how information flows through a model and how likelihood training can shape its internal geometry.

Project Page: https://github.com/yifanzhang-pro/lm-theory

1 Introduction

Autoregressive language models (AR LMs), particularly those based on the Transformer architecture (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020), have achieved remarkable success, defining the state-of-the-art in natural language generation and demonstrating impressive few-shot learning capabilities. These models operate by sequentially predicting the next token in a sequence based on the preceding context. Formally, given a sequence $\mathbf{w} = w_1 \dots w_L$ with tokens w_i from a finite vocabulary $\mathbb{V}$, the model learns a parameterized probability distribution P_θ that factorizes as:

$$P_\theta(\mathbf{w}) = \prod_{t=1}^{L} P_\theta(w_t|\mathbf{w}_{<t}), \tag{1.1}$$

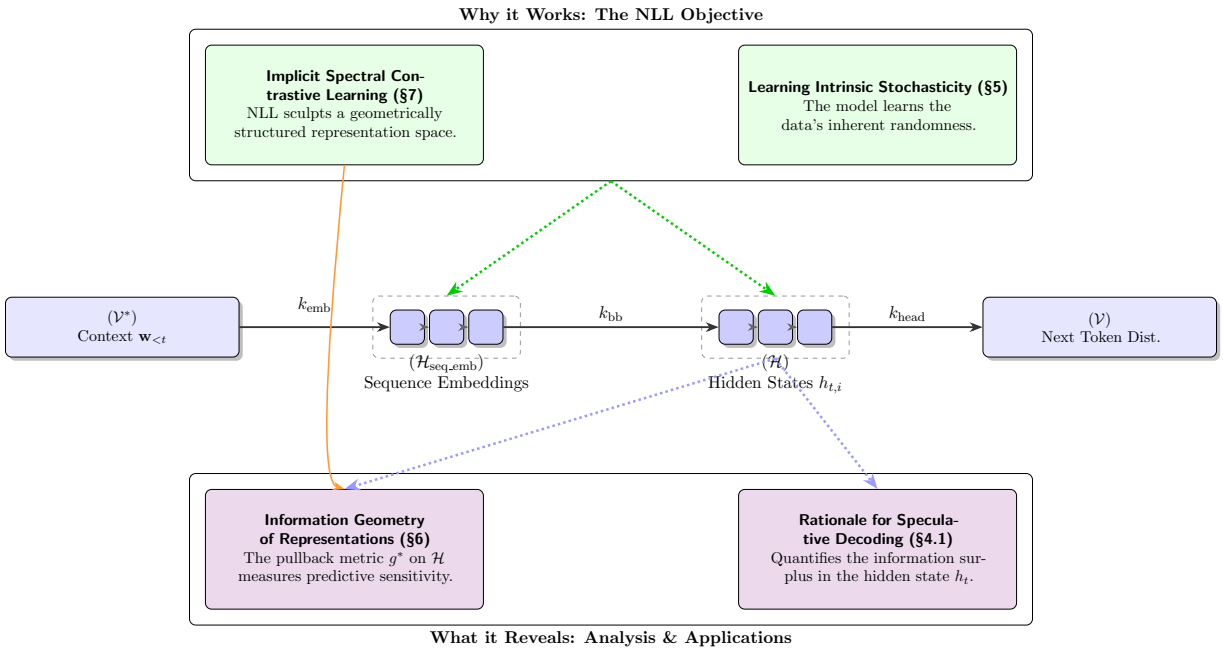

Figure 1 A conceptual overview of our framework. **Center:** The core thesis models the Autoregressive generation step as a composition of Markov kernels $k_{\text{gen}} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in the category `Stoch`. This separates the deterministic context encoding $(k_{\text{emb}}, k_{\text{bb}})$ from the probabilistic output *kernel* k_{head}, which is parameterized by a deterministic map $g_{\text{head}} : \mathcal{H} \to \Delta$. **Top:** This compositional lens reveals the deeper mechanisms of the NLL objective, which we re-frame as minimizing the average KL divergence between the model and true data kernels. Under additional constraints satisfied by linear-softmax LM heads (see §7), we show a conditional spectral connection with a predictive-similarity operator; in all cases, NLL compels the model to learn intrinsic conditional stochasticity (via categorical entropy). **Bottom:** Pulling back the Fisher–Rao metric endows $\mathcal{H}$ with an information geometry that quantifies predictive sensitivity and clarifies the information surplus used by speculative decoding.

where $\mathbf{w}_{<t} := w_1 \dots w_{t-1}$ is the context sequence, and θ denotes the model parameters, typically optimized by minimizing the negative log-likelihood (NLL) on vast text corpora. The core computational step is the mapping from a context $\mathbf{w}_{<t}$ to the conditional probability distribution $P_\theta(\cdot|\mathbf{w}_{<t})$ over $\mathbb{V}$ for the next token w_t.

Despite their empirical triumphs, a deep theoretical understanding of their internal mechanisms remains incomplete (Manning et al., 2020; Elhage et al., 2021; Yuan, 2023). Current analysis often relies on empirical probes (Hewitt and Manning, 2019) or studies of specific components like attention heads (Olsson et al., 2022). While insightful, these methods can be fragmented and often lack a unified mathematical language to describe the model's compositional and stochastic nature as a whole. A central goal here is not to introduce yet another isolated tool, but to connect well-established tools within a single, compositional language.

Another critical challenge is improving the slow, sequential nature of AR generation. Recent advances in speculative decoding, such as EAGLE (Li et al., 2024), have achieved significant speedups by predicting multiple tokens in parallel, suggesting that the final hidden state h_t contains far more information than is needed for predicting only the single next token w_t. However, a formal understanding of this information surplus is lacking.

This paper addresses this gap by introducing a unifying analytical framework for AR LMs. Our central thesis is that the language of Markov Categories (MCs) (Cho and Jacobs, 2019; Fritz, 2020) provides a natural mathematical setting for connecting several concepts that are usually discussed separately: information flow through the model's components, the geometry of the learned representation space, and the structural effects of the NLL training objective.

While many individual mathematical tools we employ, such as the pullback of the Fisher-Rao metric or the connection between NLL and KL divergence, are well-established, our primary contribution is their novel synthesis and application to dissect AR LMs. The originality of this work lies in using the category `Stoch` to formally model compositional information flow, leading to new insights uniquely enabled by this perspective. Unlike information-theoretic analyses that treat models as monolithic black boxes analyzing external behavior (e.g., the entropy of the output sequence), our framework uses categorical information theory to analyze the internal transformations and the learned geometry of the representation space at each stage of processing.

This paper introduces an analytical framework focused on the internal mechanics of the AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$. We leverage the category `Stoch`, a canonical MC whose objects are standard Borel spaces (like the continuous representation space $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$) and whose morphisms are Markov kernels (Kallenberg and Kallenberg, 1997; Fritz, 2020). We formalize the AR generation step as a composite kernel in `Stoch`:

$$k_{\text{gen},\theta} \coloneqq k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V})). \tag{1.2}$$

Here, k_{emb} and k_{bb} represent the typically deterministic context embedding and backbone transformations that produce the final hidden state $h_t \in \mathcal{H}$, while k_{head} is the *stochastic* kernel induced by a deterministic head map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$, $h \mapsto p_h$; randomness enters only through sampling $W_t \sim p_{h_t}$.

A crucial aspect of our framework is enriching `Stoch` with a statistical divergence D (e.g., D_{KL}) (Baez et al., 2016; Perrone, 2023a,b). This allows for defining intrinsic, categorical information measures like entropy $\mathcal{H}_D^{\text{cat}}$ and mutual information I_D (Perrone, 2023a), which automatically satisfy the Data Processing Inequality (DPI). Leveraging this unified framework, this paper makes the following contributions:

1. We formally model the AR generation step as a composite kernel, $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$. This compositional structure is a powerful tool for reasoning about how information is transformed, preserved, or lost at each distinct stage of processing.

2. We use categorical information measures, the chain rule for mutual information, and DPI-compatible processing bounds to quantify the information surplus used by methods like EAGLE (Li et al., 2024). This surplus is the information in a hidden state H_t about multiple future tokens beyond the immediate next token. It is an information budget for parallel drafting, not by itself a guarantee of realized speedup.

3. We show how the MC framework unifies three critical interpretations of NLL training under one theoretical roof:

   - NLL as KL minimization, equivalent to optimal source coding.
   - NLL forces the model to learn the data's inherent randomness, a process we formalize using categorical entropy. We show that optimizing NLL implies that the model's learned stochasticity converges to that of the data (Theorem 5.2).

- Under a linear-softmax head with bounded output features, we show that a calibrated quadratic upper-bound surrogate to NLL yields a regression-to-predictive-prototypes objective. After whitening or variance normalization, the associated alignment problem becomes a generalized CCA/eigenproblem. This gives a precise conditional sense in which likelihood training can align representation directions with a predictive-similarity operator (Theorem 7.10).

By formalizing how information is transformed (Section 4), how predictive sensitivity is encoded in representation geometry (Section 6), and how the NLL objective implicitly structures representations (Section 7), we can move towards more principled approaches to model design, interpretation, and control.

The paper is organized as follows. Sections 2 and 3 introduce the MC framework and our compositional model of LMs. Section 4 uses the framework to analyze information flow, providing a rationale for speculative decoding. Section 5 connects the NLL objective to learning the data's intrinsic stochasticity. Sections 6 and 7 present our main theoretical result, showing how NLL performs implicit spectral learning by shaping the geometry of the representation space. Sections 8 and 9 discuss related work and then conclude.

2 Background

This section reviews the essential mathematical concepts forming the foundation of our framework: the definition of Markov Categories and the specific category `Stoch`, followed by the enrichment of `Stoch` with statistical divergences leading to categorical information measures.

2.1 Markov Categories and `Stoch`

Markov Categories provide an axiomatic framework for probability and stochastic processes using category theory (Fritz, 2020).

Definition 2.1 (Markov Category (Fritz, 2020))**.** A Markov category $(\mathcal{C}, \otimes, \mathsf{I})$ is a symmetric monoidal category in which each object X is equipped with a commutative comonoid structure $(\Delta_X : X \to X \otimes X, !_X : X \to \mathsf{I})$, compatible with the monoidal product. The discard maps are natural: for every morphism $f : X \to Y$, $!_Y \circ f = !_X$. Equivalently, in the causal case relevant here, the monoidal unit I is terminal. The copy maps are *not* natural for arbitrary stochastic morphisms; morphisms satisfying $\Delta_Y \circ f = (f \otimes f) \circ \Delta_X$ are the deterministic morphisms.

Morphisms $k : X \to Y$ are interpreted as stochastic processes. Composition $h \circ k$ is sequential processing, while $k \otimes h$ is parallel processing. The comonoid maps Δ_X (copy) and $!_X$ (discard) abstractly model duplication and deletion of information. Copying before and after a genuinely stochastic map need not agree; this mismatch is exactly what categorical entropy measures. The causality axiom enforces probability normalization ($\int k(x, \mathrm{d}y) = 1$) in concrete examples like `Stoch`. States (probability distributions) on an object X are represented as morphisms $p : \mathsf{I} \to X$.

The key example for our purposes is the category `Stoch`.

Definition 2.2 (Category `Stoch` (Fritz, 2020; Perrone, 2023a))**.** The Markov category `Stoch` is defined by:

- **Objects**: Standard Borel spaces $(X, \mathscr{B}(X))$. These are general measure spaces that include finite sets (like a vocabulary $\mathbb{V}$), countable sets, and continuous spaces like Euclidean space $\mathbb{R}^d$ or other Polish spaces. This ensures the framework can handle both discrete tokens and continuous representations. The monoidal unit I is a singleton space $(\{\star\}, \{\emptyset, \{\star\}\})$.
- **Morphisms**: Markov kernels $k : X \to Y$. A map $k : X \times \mathscr{B}(Y) \to [0, 1]$ where $k(x, \cdot)$ is a probability measure on Y for each $x \in X$, and $k(\cdot, A)$ is a measurable function on X for each $A \in \mathscr{B}(Y)$.
- **Composition**: Given $k : X \to Y$ and $h : Y \to Z$, the composite $h \circ k : X \to Z$ is $(h \circ k)(x, C) \coloneqq \int_Y h(y, C)\, k(x, \mathrm{d}y)$ (Chapman-Kolmogorov). Identity $\mathrm{id}_X(x, A) = \delta_x(A)$.
- **Monoidal Product ($\otimes$)**: Product space $(X \times Y, \mathscr{B}(X) \otimes \mathscr{B}(Y))$ with the product σ-algebra. Product kernel $(k \otimes h)((x, y), \cdot) \coloneqq k(x, \cdot) \otimes h(y, \cdot)$ (product measure).
- **Symmetry**: Swap map $\sigma_{X,Y} : X \otimes Y \to Y \otimes X$ is $\sigma_{X,Y}((x, y), \cdot) = \delta_{(y,x)}$.
- **Comonoid Structure**: Copy $\Delta_X : X \to X \otimes X$ is $\Delta_X(x, \cdot) = \delta_{(x,x)}$. Discard $!_X : X \to \mathsf{I}$ maps to the unique point measure on I, $!_X(x, \{\star\}) = 1$.
- **Causality**: I is terminal, $!_Y \circ k = !_X$ holds, reflecting probability normalization.

Remark 2.3 (Interpretation)**.** In `Stoch`, objects represent the types of random outcomes (e.g., sequences, vectors, tokens). Morphisms represent stochastic processes or channels mapping inputs to probability distributions over outputs. Deterministic functions $f : X \to Y$ correspond to deterministic kernels $k_f(x, \cdot) = \delta_{f(x)}$. States $p : \mathsf{I} \to X$ correspond bijectively to probability measures $\mu_p \in \mathcal{P}(X)$ via $\mu_p(A) = p(\star, A)$. Marginalization arises from discarding information, e.g., for a joint state $p : \mathsf{I} \to X \otimes Y$, the X-marginal is $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$.

Lemma 2.4 (Standard Borel and measurability)**.** The spaces $\mathbb{V}^*$ (countable disjoint union of finite products), $\mathcal{H}_{\text{seq\_emb}}$ (countable disjoint union of Euclidean products), $\mathcal{H} \simeq \mathbb{R}^{d_{\text{model}}}$, and $\mathbb{V}$ (finite) are standard Borel. If $f_{\text{emb}}, f_{\text{bb}}$ are Borel-measurable, then the induced deterministic kernels $k_{\text{emb}}, k_{\text{bb}}$ are morphisms in `Stoch`.

Sketch. Countable disjoint unions of Polish spaces are standard Borel; products/sums preserve standard Borelness. Deterministic kernels defined by Borel maps are measurable morphisms in `Stoch`. □

2.2 Divergence Enrichment and Categorical Information Measures

The structure of `Stoch` is particularly powerful when enriched with a statistical divergence D, quantifying the dissimilarity between probability measures (states) $p, q : \mathsf{I} \to X$, written $D_X(p\|q)$ (Perrone, 2023a). Examples include KL divergence (D_{KL}), Total Variation (d_{TV}), Rényi divergences (D_α), and the broad class of f-divergences (D_f) (Amari and Nagaoka, 2000; Nowozin et al., 2016).

A fundamental property linking divergences and Markov kernels is the Data Processing Inequality (DPI), which holds for most standard divergences, including f-divergences and the usual Rényi divergences in their standard DPI ranges.

Theorem 2.5 (Data Processing Inequality (DPI))**.** Let D be a statistical divergence satisfying the DPI. For any Markov kernel $k : X \to Y$ in $\mathtt{Stoch}$ and any pair of states $p, q : \mathsf{I} \to X$:

$$D_Y(k \circ p \| k \circ q) \leq D_X(p \| q) \tag{2.1}$$

Processing through the channel k cannot increase the D-divergence between the distributions.

Remark. For Rényi divergences, standard DPI statements cover $\alpha \in (0, \infty]$ under the usual absolute-continuity conventions, with KL recovered as the limit $\alpha \to 1$ and max-divergence recovered as $\alpha \to \infty$ when finite. We use $\alpha \in (0, \infty)$ as the default and invoke endpoint cases only when the required absolute-continuity conditions are explicit.

Assumption 2.6 (Standing assumptions)**.** We work with: (i) objects that are standard Borel; (ii) Borel-measurable $f_{\text{emb}}, f_{\text{bb}}$, hence deterministic kernels in $\mathtt{Stoch}$; (iii) finite vocabulary $\mathbb{V}$; (iv) a deterministic parameterization $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ that is differentiable on a full-measure subset under p_{H_t}; (v) finite second moments $\mathbb{E}\|H_t\|^2 < \infty$; (vi) where invoked (e.g., Theorem 5.2), well-posed weak convergence of the relevant joint laws; and (vii) unless stated otherwise, analysis restricted to the *interior* of the simplex (i.e., $p_h(w) > 0$ for all w) on a full-measure set, so that Fisher–Rao and second-order KL expansions are valid. For near-boundary behavior, we work on a high-probability subset where $p_h(w) \geq \delta > 0$ and carry δ-dependent constants.

Preliminaries: NLL-KL equivalence and distance conventions. We recall the classical identity that minimizing cross-entropy is equivalent to minimizing the average KL divergence between data and model conditionals. Throughout, we use the Hellinger distance with the single global convention

$$d_H^2(p, q) \;\coloneqq\; 1 - \sum_{w \in \mathbb{V}} \sqrt{p(w)q(w)} \;=\; \tfrac{1}{2} \sum_{w \in \mathbb{V}} \left(\sqrt{p(w)} - \sqrt{q(w)}\right)^2,$$

On finite alphabets, we will use the bounds

$$d_H^2(p, q) \;\leq\; 1 - e^{-D_{\text{KL}}(p\|q)/2} \;\leq\; \tfrac{1}{2}\, D_{\text{KL}}(p\|q).$$

(Equivalently, $1 - \sum_w \sqrt{p(w)q(w)} \leq 1 - e^{-D_{\text{KL}}(p\|q)/2} \leq \frac{1}{2}\, D_{\text{KL}}(p\|q)$.) We use $\text{TV}(p, q) = \frac{1}{2}\|p - q\|_1$ when needed, with $\text{TV}^2(p, q) \leq \frac{1}{2}\, D_{\text{KL}}(p\|q)$. This convention is used consistently in all subsequent sections.

We provide a full proof of the NLL-KL equivalence in Appendix A.1.

Theorem 2.7 (NLL Minimization as Average KL Minimization)**.** It is a well-known result in information theory that minimizing the cross-entropy loss $L_{\text{CE}}(\theta)$ is equivalent to minimizing the average KL divergence between the true and model conditional distributions. We state it here in the language of our framework to ground the subsequent analysis.

$$\mathcal{L}_{\text{KL}}(\theta) \coloneqq \mathbb{E}_{\mathbf{w}_{<t} \sim p_{W_{<t}}}\Big[D_{\text{KL}}\big(k_{\text{data}}(\mathbf{w}_{<t}, \cdot) \,\|\, k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot)\big)\Big], \qquad \operatorname*{argmin}_\theta L_{\text{CE}}(\theta) = \operatorname*{argmin}_\theta \mathcal{L}_{\text{KL}}(\theta). \tag{2.2}$$

The expectation is taken over contexts $\mathbf{w}_{<t}$ drawn according to the data's marginal context distribution $p_{W_{<t}}$. The minimum value of $\mathcal{L}_{\text{KL}}(\theta)$ is non-negative. If the model class $\{k_{\text{gen},\theta} \mid \theta \in \Theta\}$ is sufficiently expressive to contain k_{data} (i.e., $k_{\text{data}} = k_{\text{gen},\theta_{\text{true}}}$ for some $\theta_{\text{true}} \in \Theta$), then the minimum value is 0, achieved if and only if $k_{\text{gen},\theta^*}(\mathbf{w}_{<t}, \cdot) = k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ for $p_{W_{<t}}$-almost every context $\mathbf{w}_{<t}$.

Perrone (Perrone, 2023a) introduced categorical definitions of entropy and mutual information intrinsically tied to the divergence D and the MC structure.

Definition 2.8 (Categorical Entropy (Perrone, 2023a))**.** Let $(\texttt{Stoch}, D)$ be enriched with a DPI-satisfying divergence D.

1. The *pointwise categorical entropy* of a kernel $k : X \to Y$ is the function $\mathcal{H}_D^{\text{cat}}(k) : X \to \mathbb{R}_{\geq 0}$ given by
$$\mathcal{H}_D^{\text{cat}}(k)(x) \coloneqq D_{Y\otimes Y}(\Delta_Y \circ k(x,\cdot) \parallel (k \otimes k) \circ \Delta_X(x,\cdot)). \tag{2.3}$$
Intuitively, it compares (1) applying k then copying the output vs. (2) copying x then applying k independently. If k is deterministic, $\mathcal{H}_D^{\text{cat}}(k)(x) = 0$.

2. Given a prior/state $p_X : \mathsf{I} \to X$, the *state-averaged categorical entropy* is the scalar
$$\overline{\mathcal{H}}_D^{\text{cat}}(k; p_X) \;\coloneqq\; \mathbb{E}_{x\sim p_X}\big[\mathcal{H}_D^{\text{cat}}(k)(x)\big] = \int_X \mathcal{H}_D^{\text{cat}}(k)(x)\, p_X(\mathrm{d}x). \tag{2.4}$$

3. The *Categorical Mutual Information* for a joint state $p : \mathsf{I} \to X \otimes Y$ is $I_D(p) \coloneqq D_{X\otimes Y}(p \| p_X \otimes p_Y)$.

Remark 2.9 (Properties and Connections)**.** When $D = D_{\text{KL}}$ and the output object Y is *finite* (in particular, $Y = \mathbb{V}$), $I_{D_{\text{KL}}}(p)$ recovers Shannon mutual information. In this discrete setting, the pointwise categorical entropy satisfies $\mathcal{H}_{D_{\text{KL}}}^{\text{cat}}(k)(x) = H\big(k(x,\cdot)\big)$, and the averaged quantity satisfies $\overline{\mathcal{H}}_{D_{\text{KL}}}^{\text{cat}}(k; p_X) = \mathbb{E}_{x\sim p_X}\big[H(k(x,\cdot))\big] = H(Y \mid X)$. For non-atomic Y, the diagonal law $\Delta_Y \circ k(x,\cdot)$ is typically singular with respect to $(k \otimes k) \circ \Delta_X(x,\cdot)$, so $\mathcal{H}_{D_{\text{KL}}}^{\text{cat}}(k)(x)$ is generally $+\infty$ unless $k(x,\cdot)$ is purely atomic; deterministic kernels are the zero-entropy special case.

3 Autoregressive Language Models as Composed Kernels

We now apply the Markov Category framework established in Section 2 to model Autoregressive language models. Specifically, we model the single-step generation mapping $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ as a composition of Markov kernels within the category $\texttt{Stoch}$.

The relevant measurable spaces (objects in $\texttt{Stoch}$) are:

- Input context space: $(\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) = (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$, where $\mathbb{V}^*$ is the set of finite sequences over the vocabulary $\mathbb{V}$, equipped with a suitable σ-algebra making it standard Borel (e.g., considering it as a disjoint union of finite products $\mathbb{V}^n$).
- Initial sequence representation space: $(\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}})) = (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$, the space of initial vector sequences (e.g., $\bigcup_n (\mathbb{R}^{d_{\text{model}}})^n$), also equipped with a standard Borel structure.
- Final hidden state space: $(\mathcal{H}, \mathscr{B}(\mathcal{H})) = (\mathcal{H}, \mathscr{B}(\mathcal{H}))$, typically $(\mathbb{R}^{d_{\text{model}}}, \mathscr{B}(\mathbb{R}^{d_{\text{model}}}))$.
- Output vocabulary space: $(\mathbb{V}, \mathcal{P}(\mathbb{V})) = (\mathbb{V}, \mathcal{P}(\mathbb{V}))$, a finite measurable space.

Standard Borel spaces are chosen because they form a well-behaved class of measurable spaces (isomorphic to Borel subsets of Polish spaces) closed under countable products, sums, and containing standard examples like $\mathbb{R}^d$ and finite sets, ensuring measure-theoretic regularity (Kallenberg and Kallenberg, 1997).

The generation process decomposes into three kernels (morphisms in `Stoch`):

1. **Embedding Layer Kernel ($k_{\text{emb}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$)**: This kernel encapsulates the initial processing of the discrete input sequence $\mathbf{w}_{<t} \in \mathbb{V}^*$. It typically involves applying a token embedding function $\mathcal{E} : \mathbb{V} \to \mathbb{R}^{d_{\text{model}}}$ to each token w_i and potentially incorporating absolute positional encodings. Let $f_{\text{emb}} : \mathbb{V}^* \to \mathcal{H}_{\text{seq\_emb}}$ denote the overall deterministic function computing the initial sequence representation $E_{<t}$. Since this mapping is deterministic, the kernel k_{emb} is defined via the Dirac measure $\delta_\cdot$:

$$k_{\text{emb}}(\mathbf{w}_{<t}, A) \coloneqq \delta_{f_{\text{emb}}(\mathbf{w}_{<t})}(A) = \mathbf{1}_A(f_{\text{emb}}(\mathbf{w}_{<t})), \quad \text{for } A \in \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}). \tag{3.1}$$

This is a valid morphism in `Stoch`.

2. **Backbone Transformation Kernel ($k_{\text{bb}} : (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}})) \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$)**: This kernel represents the core computation, usually a deep neural network like a Transformer stack. Let $f_{\text{bb}} : \mathcal{H}_{\text{seq\_emb}} \to \mathcal{H}$ be the function mapping the initial sequence representation $E_{<t}$ to the final hidden state $h_t \in \mathcal{H}$ (often the output vector at the last sequence position). This function incorporates complex operations like multi-head self-attention and feed-forward layers. Relative positional information, such as Rotary Position Embeddings (RoPE) (Su et al., 2024), is implemented within the function f_{bb} by modifying attention computations based on token positions. Assuming the backbone computation is deterministic for a given $E_{<t}$ and parameters θ, the kernel k_{bb} is also deterministic:

$$k_{\text{bb}}(E_{<t}, B) \coloneqq \delta_{f_{\text{bb}}(E_{<t})}(B) = \mathbf{1}_B(f_{\text{bb}}(E_{<t})), \quad \text{for } B \in \mathscr{B}(\mathcal{H}). \tag{3.2}$$

This is also a morphism in `Stoch`.

3. **LM Head Kernel ($k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$)**: This final step maps the summary hidden state $h_t \in \mathcal{H}$ to a probability distribution over the finite vocabulary $\mathbb{V}$. Typically, h_t is passed through a linear layer ($f_{\text{head}} : \mathcal{H} \to \mathbb{R}^{|\mathbb{V}|}$) producing logits $\mathbf{z} = f_{\text{head}}(h_t)$, followed by the softmax function: $P(w|h_t) = [\text{softmax}(\mathbf{z})]_w$. This defines a Markov kernel induced by a deterministic map into the simplex:

$$k_{\text{head}}(h, A) \coloneqq \sum_{w \in A} [\text{softmax}(f_{\text{head}}(h))]_w \quad \text{for } h \in \mathcal{H}, A \subseteq \mathbb{V}. \tag{3.3}$$

This kernel maps each point h in the representation space to a probability measure on the discrete space $\mathbb{V}$, satisfying the required measurability conditions. It is a generally non-deterministic morphism in `Stoch` that is parameterized by a deterministic map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$.

Remark 3.1 (Deterministic parameterization vs. stochastic kernel)**.** The head is best seen as a deterministic map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$, $h \mapsto p_h$, parameterizing a stochastic kernel via $k_{\text{head}}(h, \cdot) = p_h$. "Learned stochasticity" refers to the model learning distributions p_h that match the data's conditional uncertainty; the kernel itself is not a deterministic kernel $X \to \mathbb{V}$.

The overall single-step generation kernel $k_{\text{gen},\theta} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ is the composition $k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in the category `Stoch`. This composition precisely represents the model's learned conditional probability map $P_\theta(\cdot|\mathbf{w}_{<t})$. It is crucial to note that this formalism applies to **any** AR

model, including Transformers. The attention mechanism provides a powerful, history-dependent parameterization of this single-step kernel. The subsequent sections use this representation to analyze the model's behavior.

4 Information-Theoretic Analysis via Categorical Metrics

The MC framework allows us to define principled metrics for internal analysis and to formally reason about information flow. We focus on two key applications that are central to the paper's main arguments: quantifying the information surplus exploited by speculative decoding and measuring the intrinsic stochasticity of the prediction head.

We operate within the probabilistic setting induced by a distribution P_{ctx} over input contexts, corresponding to an initial state $p_{W_{<t}} : \mathsf{I} \to (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$. Processing this state through the composed kernels induces distributions over the hidden state H_t (state $p_{H_t} : \mathsf{I} \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$) and the next token W_t (state $p_{W_t} : \mathsf{I} \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$).

4.1 Information Flow Bounds and Rationale for Speculative Decoding

Setting. Transformers condition on the entire prefix, so the effective source is ∞-order. To avoid finite-order assumptions, we quantify multi-token predictability using conditional mutual-information (MI) tails.

Let $W_{t:t+K-1} = (W_t, \dots, W_{t+K-1})$ and let the hidden summary be $H_t = \phi(W_{<t})$. The chain rule gives

$$I(H_t; W_{t:t+K-1}) \;=\; I(H_t; W_t) \;+\; I\big(H_t; W_{t+1:t+K-1} \mid W_t\big). \tag{4.1}$$

We call the second term the *information surplus*

$$\mathsf{Surplus}_K \;\coloneqq\; I\big(H_t; W_{t+1:t+K-1} \mid W_t\big), \tag{4.2}$$

which satisfies $0 \le \mathsf{Surplus}_K \le H(W_{t+1:t+K-1} \mid W_t) \le (K-1)\log|\mathbb{V}|$. It is convenient to further decompose

$$a_k \;\coloneqq\; I\big(H_t; W_{t+k} \mid W_{t:t+k-1}\big) \;\; \ge 0, \qquad k \ge 1,$$

so that $\mathsf{Surplus}_K$ is a tail sum of nonnegative per-step contributions.

Proposition 4.1 (Tail sum and decay)**.** For any AR LM with $H_t = \phi(W_{<t})$,

$$\mathsf{Surplus}_K \;=\; \sum_{k=1}^{K-1} a_k, \qquad a_k \ge 0.$$

Hence $K \mapsto \mathsf{Surplus}_K$ is nondecreasing and $\mathsf{Surplus}_K \le (K-1)\log|\mathbb{V}|$. If there exists a nonincreasing envelope ψ with $a_k \le \psi(k)$ and $\sum_{k\ge1}\psi(k) < \infty$, then $\mathsf{Surplus}_K \to \mathsf{Surplus}_\infty$ and

$$0 \le \mathsf{Surplus}_\infty - \mathsf{Surplus}_K \;\le\; \sum_{k\ge K} \psi(k).$$

In particular, if $\psi(k) \le C\rho^k$ for some $C < \infty$ and $\rho \in (0,1)$, then $\mathsf{Surplus}_\infty - \mathsf{Surplus}_K \le \frac{C\,\rho^K}{1-\rho}$.

Corollary 4.2 (Finite memory as a special case)**.** If the source is m-th order Markov and $H_t = \phi(W_{t-m:t-1})$, then $a_k = 0$ for all $k \geq m$. Thus no new surplus is added after offset $m-1$, and $\mathsf{Surplus}_K$ is constant for all $K \geq m$.

Natural language exhibits long-tailed dependence, but the tail often decays. A concise control metric is the ε-effective surplus length

$$K_\varepsilon \;:=\; \min\{K \geq 1 :\; \mathsf{Surplus}_\infty - \mathsf{Surplus}_K \leq \varepsilon\},$$

which sets a principled draft length for speculative decoding under a tolerated missed-information budget ε. Empirically, once $\mathsf{Surplus}_K$ flattens, longer drafts yield diminishing returns unless the architecture changes.

Two-path view of speculative decoding. Drafting and verification are parallel kernels on the same hidden state:

$$\begin{array}{ccccc} \mathbb{V}^* & \xrightarrow{k_{\text{enc}} := k_{\text{bb}} \circ k_{\text{emb}}} & \mathcal{H} & \xrightarrow{k_{\text{verify}}} & \mathbb{V} \\ & & \Big\downarrow{\scriptstyle k_{\text{draft}}} & & \\ & & \mathbb{V}^K & & \end{array}$$

Here $k_{\text{verify}} \equiv k_{\text{head}}$ outputs the next-token distribution, while k_{draft} proposes K tokens in parallel. Because these kernels have different codomains, their raw categorical entropies are not directly comparable. A meaningful diagnostic should normalize the draft entropy per token or condition sequentially, and should be interpreted together with the verifier's acceptance probability. Effective drafting requires both nonzero surplus $\mathsf{Surplus}_K$ and a draft kernel that converts this surplus into high-probability proposals.

Beyond DPI, Equation (4.1) gives the conditional MI budget $\mathsf{Surplus}_K$ available to parallel proposals. This quantity bounds the information available to longer drafts and helps explain where returns can saturate. Realized speedup still depends on the draft model, verifier, and acceptance rule.

4.2 Metric: LM Head Categorical Entropy (Prediction Stochasticity)

We quantify the intrinsic stochasticity or uncertainty associated with the final prediction step, embodied by the LM head kernel $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$. This metric is crucial for understanding how NLL training forces the model to learn the data's inherent randomness (Section 5). For a given input $h \in \mathcal{H}$, the categorical entropy quantifies the stochasticity of the output distribution $k_{\text{head}}(h, \cdot)$. We obtain a single summary statistic by averaging this value over the distribution of hidden states p_{H_t}.

Definition 4.3 (Categorical Entropy of k_{head})**.** Using equation (2.3) with $X = \mathcal{H}$, $Y = \mathbb{V}$, and $k = k_{\text{head}}$, the pointwise categorical entropy is

$$\mathcal{H}_D^{\text{cat}}(k_{\text{head}})(h) \coloneqq D_{\mathbb{V}\otimes\mathbb{V}}(\Delta_{\mathbb{V}} \circ k_{\text{head}} \parallel (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}})(h). \tag{4.3}$$

The categorical entropy $\mathcal{H}_D^{\text{cat}}(k_{\text{head}})(h)$ measures the divergence between generating a correlated pair (W, W) versus an independent pair (W_1, W_2), where $W, W_1, W_2 \sim k_{\text{head}}(h, \cdot)$. This quantifies how far the output distribution for a given h is from a deterministic point mass. To obtain a single

metric for the LM head's overall stochasticity, we compute its expectation with respect to the hidden state distribution p_{H_t}:

$$\overline{\mathcal{H}}_D^{\mathrm{cat}}(k_{\mathrm{head}}; p_{H_t}) \coloneqq \mathbb{E}_{h\sim p_{H_t}}\left[D_{\mathbb{V}\otimes\mathbb{V}}\left(\sum_{w\in\mathbb{V}} k_{\mathrm{head}}(h,\{w\})\delta_{(w,w)} \quad \| \quad k_{\mathrm{head}}(h,\cdot)\otimes k_{\mathrm{head}}(h,\cdot)\right)\right]. \tag{4.4}$$

Interpretation. This metric measures the intrinsic conditional stochasticity of the LM head mapping. If k_{head} were deterministic (i.e., for each h, it mapped to a single specific w_h, so $p_h = \delta_{w_h}$), then both measures inside the divergence would be $\delta_{(w_h,w_h)}$, and the entropy would be $D(\delta_{(w_h,w_h)}\|\delta_{(w_h,w_h)}) = 0$. A higher value of $\overline{\mathcal{H}}_D^{\mathrm{cat}}(k_{\mathrm{head}}; p_{H_t})$ indicates greater average uncertainty or "spread" in the output distribution $p_h = k_{\mathrm{head}}(h,\cdot)$, meaning the kernel is inherently more stochastic. It quantifies how far the prediction process is from a deterministic assignment, measured in the geometry of $\mathbb{V}\otimes\mathbb{V}$ induced by D. In the special case $D = D_{\mathrm{KL}}$ and finite $\mathbb{V}$, $\overline{\mathcal{H}}_{D_{\mathrm{KL}}}^{\mathrm{cat}}(k_{\mathrm{head}}; p_{H_t}) \le \log|\mathbb{V}|$ is automatically bounded; this bound underwrites the continuity argument used later.

For the specific case $D = D_{\mathrm{KL}}$ and finite $\mathbb{V}$, the categorical entropy equals the Shannon (conditional) entropy:

Proposition 4.4 (KL case reduces to Shannon)**.** For the LM head kernel $k(h,\cdot) = p_h$ over a finite vocabulary,

$$\mathcal{H}_{D_{\mathrm{KL}}}^{\mathrm{cat}}(k_{\mathrm{head}})(h) = H(p_h), \qquad \overline{\mathcal{H}}_{D_{\mathrm{KL}}}^{\mathrm{cat}}(k_{\mathrm{head}}; p_{H_t}) \;=\; H(W_t \mid H_t).$$

Proof. Since $\sum_w p_h(w)\delta_{(w,w)}$ is supported on the diagonal, a direct calculation gives $D_{\mathbb{V}\otimes\mathbb{V}}\big(\sum_w p_h(w)\delta_{(w,w)} \,\big\|\, p_h\otimes p_h\big) = -\sum_w p_h(w)\log p_h(w)$, and averaging over H_t yields $H(W_t \mid H_t)$. □

5 Pretraining Objective, Compression, and Learning Intrinsic Stochasticity

A central question surrounding large language models is how the seemingly simple Autoregressive objective of next-token prediction, trained via minimizing cross-entropy loss (equivalently, negative log-likelihood or NLL), sculpts representations. The framework of Markov Categories and categorical entropy provides a lens through which to interpret this phenomenon, connecting it to compression and to matching the inherent conditional uncertainty of the data generating process.

Let $k_{\mathrm{data}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the (potentially unknown) Markov kernel representing the true data-generating process, such that $k_{\mathrm{data}}(\mathbf{w}_{<t},\cdot)$ corresponds to the true conditional probability measure $P_{\mathrm{data}}(\cdot|\mathbf{w}_{<t})$ on the vocabulary $\mathbb{V}$. Let $p_{W_{<t}}$ denote the marginal probability measure on the context space $(\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$, derived from the underlying joint distribution P_{data} over sequences observed in the training corpus.

The standard pretraining objective for an AR LM parameterized by θ is to minimize the negative log-likelihood (NLL) of the next token w_t given the preceding context $\mathbf{w}_{<t}$, averaged over the training data distribution P_{data}. This is equivalent to minimizing the average KL divergence between the data kernel and the model kernel (Theorem 2.7).

$$\mathcal{L}_{\mathrm{KL}}(\theta) = \mathbb{E}_{\mathbf{w}_{<t}\sim p_{W_{<t}}}[D_{\mathrm{KL}}(k_{\mathrm{data}}(\mathbf{w}_{<t},\cdot)\,\|\,k_{\mathrm{gen},\theta}(\mathbf{w}_{<t},\cdot))] \tag{5.1}$$

where $P_\theta(\cdot|\mathbf{w}_{<t}) = k_{\mathrm{gen},\theta}(\mathbf{w}_{<t},\cdot)$ and the expectation is taken over contexts $\mathbf{w}_{<t}$ drawn according to the data's marginal context distribution $p_{W_{<t}}$.

This theorem frames NLL training as driving the model kernel $k_{\text{gen},\theta}$ to match the data kernel k_{data}. The connection to compression arises from Shannon's source coding theorem. The minimal average code length required to losslessly encode the next token w_t, given the context $\mathbf{w}_{<t}$ and using an optimal code based on the true distribution $P_{\text{data}}(\cdot|\mathbf{w}_{<t})$, is the conditional Shannon entropy $H(W_t|W_{<t})_{\text{data}}$. The cross-entropy loss $L_{\text{CE}}(\theta)$ achieved by the model represents the average code length when using a code based on the model's distribution $P_\theta(\cdot|\mathbf{w}_{<t})$. Therefore, minimizing NLL (equation (2.2)) is equivalent to finding a model that provides the most efficient compression of the training data sequences, achieving an average code length that approaches the theoretical minimum $H(W_t|W_{<t})_{\text{data}}$. The widely discussed hypothesis that "compression implies understanding" posits that achieving high compression rates on complex data like natural language necessitates learning the underlying structure, rules, and statistical regularities, which may manifest as emergent capabilities.

Beyond matching the predictive distributions point-wise on average, successful NLL training implies that the model also learns to replicate the intrinsic stochasticity or uncertainty inherent in the data generation process at the prediction step. Within our framework, this intrinsic conditional stochasticity can be quantified using the concept of average categorical entropy (equation (4.4)). Recall the definition in Equation (4.4). This quantity quantifies the average "spread" or non-determinism of the kernel k under the input distribution p_X.

Let $k_{\text{head},\theta} : \mathcal{H} \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the LM head kernel corresponding to parameters θ. Let $p_{H_t,\theta}$ be the distribution over hidden states $h_t \in \mathcal{H}$ induced by processing contexts $\mathbf{w}_{<t} \sim p_{W_{<t}}$ through the model's encoder $k_{\text{bb}} \circ k_{\text{emb}}$ (parameterized by θ).

NLL training aligns the model's conditional uncertainty with that of the data via two ingredients: (i) KL control of per-context discrepancies (hence Hellinger control on finite vocabularies), and (ii) uniform continuity of the head-entropy functional on the compact simplex.

Lemma 5.1 (Continuity of the average head-entropy functional)**.** Let $\mathbb{V}$ be finite and define

$$\Psi_D(p) \coloneqq D_{\mathbb{V}\otimes\mathbb{V}}\big(\sum_w p(w)\delta_{(w,w)} \,\|\, p \otimes p\big), \qquad p \in \Delta^{|\mathbb{V}|-1}.$$

Assume Ψ_D is continuous (hence bounded and uniformly continuous on $\Delta^{|\mathbb{V}|-1}$). Let $(P^{(n)})_{n\geq 1}$ and $P^\star$ be random variables in $\Delta^{|\mathbb{V}|-1}$. If

$$\mathbb{E}\big[d_H(P^{(n)}, P^\star)^2\big] \to 0,$$

then

$$\mathbb{E}\big[\Psi_D(P^{(n)})\big] \to \mathbb{E}\big[\Psi_D(P^\star)\big].$$

For $D = D_{\text{KL}}$, $\Psi_{D_{\text{KL}}}(p) = H(p)$ and $0 \leq \Psi_{D_{\text{KL}}}(p) \leq \log|\mathbb{V}|$.

Theorem 5.2 (Convergence of average categorical entropy under NLL minimization)**.** Let $X \sim p_{W_{<t}}$ be a random context and write

$$p_X(\cdot) \coloneqq k_{\text{data}}(X, \cdot), \qquad q_{X,\theta}(\cdot) \coloneqq k_{\text{gen},\theta}(X, \cdot).$$

Assume realizability: there exists $\theta^\star$ such that $q_{X,\theta^\star} = p_X$ almost surely (equivalently, $\mathcal{L}_{\text{KL}}(\theta^\star) = 0$). Let (θ_n) satisfy $\mathcal{L}_{\text{KL}}(\theta_n) \to 0$. If Ψ_D from Lemma 5.1 is continuous, then

$$\lim_{n\to\infty} \overline{\mathcal{H}}_D^{\text{cat}}(k_{\text{head},\theta_n}; p_{H_t,\theta_n}) = \mathbb{E}_X\big[\Psi_D(p_X)\big].$$

In particular, for $D = D_{\mathrm{KL}}$,

$$\lim_{n\to\infty} \overline{\mathcal{H}}^{\mathrm{cat}}_{D_{\mathrm{KL}}}(k_{\mathrm{head},\theta_n};\, p_{H_t,\theta_n}) = \mathbb{E}_X[H(p_X)] = H(W_t \mid W_{<t})_{\mathrm{data}}.$$

Since $H = f_{\mathrm{enc},\theta^\star}(W_{<t})$ is a deterministic function of $W_{<t}$, always $H(W_t \mid H) \geq H(W_t \mid W_{<t})$. At any realizable optimum $\theta^\star$, the data kernel factors through H, hence H is predictively sufficient and $H(W_t \mid H) = H(W_t \mid W_{<t})_{\mathrm{data}}$.

Theorem 5.2 provides a formal basis for the claim that NLL training *lets the model learn* not just the most likely next token, but also the degree of uncertainty or stochasticity associated with that prediction, as dictated by the data. By minimizing the average KL divergence $\mathcal{L}_{\mathrm{KL}}(\theta)$, the model $k_{\mathrm{gen},\theta}$ must align its output distributions $k_{\mathrm{gen},\theta}(\mathbf{w}_{<t}, \cdot)$ with the data distributions $k_{\mathrm{data}}(\mathbf{w}_{<t}, \cdot)$. This alignment necessarily includes matching the "shape" or "spread" of these distributions, which is precisely what is quantified by the average categorical entropy $\overline{\mathcal{H}}^{\mathrm{cat}}_D$. The parameters θ and the compositional structure $k_{\mathrm{head}} \circ k_{\mathrm{bb}} \circ k_{\mathrm{emb}}$ thus become a compressed representation capturing both the predictive dependencies and the inherent conditional randomness of the language source. This suggests that learning the correct level of stochasticity is an integral part of the compression process driven by the NLL objective, contributing to the model's ability to generate realistic and diverse text sequences.

6 Information Geometry of Representation and Prediction Spaces

The Markov Category framework, particularly $(\mathtt{Stoch}, D)$ enriched with a divergence like D_{KL}, provides a natural bridge to Information Geometry (Amari and Nagaoka, 2000; Perrone, 2023b). This allows for a geometric analysis of the spaces involved in AR language modeling, particularly the representation space $\mathcal{H}$ and the space of next-token distributions $\mathcal{P}(\mathbb{V})$.

Remark 6.1 (Gauge and invariances)**.** The pullback $g^* = g^*_{\mathrm{head}} g^{\mathrm{FR}}$ is invariant to adding a constant to logits and, in whitened coordinates with $C_{hh} = I$, to *orthogonal* reparameterizations $h = Q\widetilde{h}$ (then $g^*(h) = Q^\top g^*(\widetilde{h})Q$). General $GL(d)$ changes act by congruence and thus distort eigenvalues. When comparing spectra across checkpoints or layers, it is therefore natural to work in whitened coordinates or report coordinate-free summaries such as eigenvalue ratios and principal subspace overlaps.

The space $\mathcal{P}(\mathbb{V})$ of probability distributions over the finite vocabulary $\mathbb{V}$ forms a $(|\mathbb{V}| - 1)$-dimensional simplex $\Delta^{|\mathbb{V}|-1}$. This space possesses a well-defined Riemannian geometry induced by the Fisher-Rao information metric g^{FR}, whose components in a local coordinate system $\xi = (\xi_1, \ldots, \xi_{|\mathbb{V}|-1})$ for a distribution $p_\xi \in \mathcal{P}(\mathbb{V})$ are given by:

$$g^{\mathrm{FR}}_{ij}(\xi) = \sum_{w\in\mathbb{V}} p_\xi(w) \frac{\partial \log p_\xi(w)}{\partial \xi_i} \frac{\partial \log p_\xi(w)}{\partial \xi_j} = \mathbb{E}_{W\sim p_\xi}\left[\frac{\partial \log p_\xi(W)}{\partial \xi_i} \frac{\partial \log p_\xi(W)}{\partial \xi_j}\right]. \tag{6.1}$$

This metric quantifies the local distinguishability between nearby probability distributions, measuring the distance in terms of expected squared log-likelihood ratio gradients. The geometry of $\mathcal{P}(\mathbb{V})$ also includes dual affine connections ($\pm\alpha$-connections) related to the KL divergence, providing a richer dually flat structure (Amari and Nagaoka, 2000).

The LM Head kernel $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ corresponds to a deterministic mapping from a hidden state $h \in \mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$ to a probability distribution $p_h \coloneqq k_{\text{head}}(h, \cdot) \in \mathcal{P}(\mathbb{V})$. Let $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ denote this mapping, $p_h = g_{\text{head}}(h)$. Typically, this involves a linear layer followed by softmax: $g_{\text{head}}(h) = \text{softmax}(Wh)$ where $W \in \mathbb{R}^{|\mathbb{V}| \times d_{\text{model}}}$. This mapping g_{head} allows us to pull back the geometric structure from $\mathcal{P}(\mathbb{V})$ onto the representation space $\mathcal{H}$.

To avoid collision with the token random variables W_t, we denote the output weight matrix by $W_{\text{out}} \in \mathbb{R}^{|\mathbb{V}| \times d_{\text{model}}}$ below, so $g_{\text{head}}(h) = \text{softmax}(W_{\text{out}} h + b)$.

Specifically, the Fisher-Rao metric g^{FR} on $\mathcal{P}(\mathbb{V})$ induces a positive-semidefinite pullback tensor $g^* = g_{\text{head}}^* g^{\text{FR}}$ on $\mathcal{H}$. It is a genuine Riemannian metric only on directions where the head map has full local rank; otherwise it is a degenerate metric tensor. At a point $h \in \mathcal{H}$, its components are given by:

$$g_{ab}^*(h) = \sum_{i,j} g_{ij}^{\text{FR}}(g_{\text{head}}(h)) \frac{\partial (g_{\text{head}}(h))_i}{\partial h_a} \frac{\partial (g_{\text{head}}(h))_j}{\partial h_b}, \quad a, b \in \{1, \dots, d_{\text{model}}\}, \tag{6.2}$$

where h_a, h_b are coordinates of $h \in \mathcal{H}$, and $(g_{\text{head}}(h))_i, (g_{\text{head}}(h))_j$ represent local coordinates of the output distribution $p_h \in \mathcal{P}(\mathbb{V})$ (e.g., probabilities of specific tokens, possibly excluding one due to the sum-to-one constraint). The term $\frac{\partial (g_{\text{head}}(h))_i}{\partial h_a}$ is the Jacobian of the LM head map g_{head} evaluated at h.

Let $J(h)$ denote this Jacobian matrix ($|\mathbb{V}| - 1 \times d_{\text{model}}$ or $|\mathbb{V}| \times d_{\text{model}}$ depending on coordinates). Then $g^*(h) = J(h)^\top g^{\text{FR}}(g_{\text{head}}(h)) J(h)$. The significance of this pullback metric g^* lies in its connection to the local distinguishability of output distributions under perturbations of the input hidden state, as measured by divergences like KL divergence.

Theorem 6.2 (Pullback Metric and Local Divergence)**.** Assume $p_h \in \operatorname{int} \Delta^{|\mathbb{V}|-1}$ (i.e., $p_h(w) > 0$ for all w; see Assumption 2.6). Let $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ be the smooth map corresponding to the LM head kernel. Let $h \in \mathcal{H}$ and $v \in T_h \mathcal{H} \cong \mathcal{H}$. Consider the distributions $p_h = g_{\text{head}}(h)$ and $p_{h+\epsilon v} = g_{\text{head}}(h + \epsilon v)$ for small ϵ. The KL divergence between these output distributions, for small ϵ, is locally approximated by the quadratic form defined by the pullback metric $g^*(h)$:

$$D_{\text{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2} \epsilon^2 g^*(h)(v, v) + O(\epsilon^3) \tag{6.3}$$

where $g^*(h)(v, v) = \sum_{a,b=1}^{d_{\text{model}}} g_{ab}^*(h) v_a v_b$. A similar relationship holds for symmetric KL divergence. More generally, for a sufficiently smooth f-divergence, the leading term is $\frac{f''(1)}{2} \epsilon^2 g^*(h)(v, v)$.

Proof. Let ξ be a local coordinate system for $\mathcal{P}(\mathbb{V})$ around p_h. The KL divergence between two nearby distributions p_ξ and $p_{\xi'}$ can be expanded around p_ξ as (Amari and Nagaoka, 2000):

$$D_{\text{KL}}(p_{\xi'} \,\|\, p_\xi) = \frac{1}{2} \sum_{i,j} g_{ij}^{\text{FR}}(\xi)(\xi_i' - \xi_i)(\xi_j' - \xi_j) + O(\|\xi' - \xi\|^3).$$

Let $\xi(h)$ denote the coordinates of $p_h = g_{\text{head}}(h)$. For $p_{h+\epsilon v}$, the coordinates are $\xi(h + \epsilon v)$. By Taylor expansion in ϵ:

$$\xi_i(h + \epsilon v) = \xi_i(h) + \epsilon \sum_{a=1}^{d_{\text{model}}} \frac{\partial \xi_i}{\partial h_a}(h) v_a + O(\epsilon^2).$$

Thus, $\xi_i(h+\epsilon v) - \xi_i(h) = \epsilon J_{ia}(h)v_a + O(\epsilon^2)$, where $J(h)$ is the Jacobian matrix of the map $h \mapsto \xi(h)$ (i.e., the Jacobian of g_{head} in local coordinates ξ). Substituting this into the KL expansion:

$$\begin{aligned} D_{\mathrm{KL}}(p_{h+\epsilon v} \,\|\, p_h) &= \frac{1}{2} \sum_{i,j} g^{\mathrm{FR}}_{ij}(\xi(h)) \left(\epsilon \sum_a J_{ia}(h) v_a \right) \left(\epsilon \sum_b J_{jb}(h) v_b \right) + O(\epsilon^3) \\ &= \frac{1}{2} \epsilon^2 \sum_{a,b} \left(\sum_{i,j} J_{ia}(h) g^{\mathrm{FR}}_{ij}(\xi(h)) J_{jb}(h) \right) v_a v_b + O(\epsilon^3) \\ &= \frac{1}{2} \epsilon^2 \sum_{a,b} (J(h)^\top g^{\mathrm{FR}}(\xi(h)) J(h))_{ab} v_a v_b + O(\epsilon^3). \end{aligned}$$

The term $J(h)^\top g^{\mathrm{FR}}(\xi(h))J(h)$ is precisely the matrix representation of the pullback metric $g^*(h)$ in the standard coordinates of $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$, derived from equation (6.2). Thus, $D_{\mathrm{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2}\epsilon^2 g^*(h)(v,v) + O(\epsilon^3)$. The result for other well-behaved f-divergences follows from their similar second-order expansion involving g^{FR}. □

This theorem formally establishes that the pullback metric g^* measures how sensitive the output distribution p_h is to infinitesimal changes in the hidden state h, where sensitivity is gauged by the local divergence (specifically, KL divergence, relating to the Fisher-Rao metric) in the output space $\mathcal{P}(\mathbb{V})$.

Lemma 6.3 (Softmax–linear head: closed form of g^*)**.** Suppose $p_h = \mathrm{softmax}(W_{\text{out}}h + b)$ with $W_{\text{out}} \in \mathbb{R}^{|\mathbb{V}| \times d}$. Then

$$g^*(h) \;=\; W_{\text{out}}^\top \Big[\mathrm{Diag}(p_h) - p_h p_h^\top \Big] W_{\text{out}}.$$

Proof. For softmax logits $z_w = \langle W_{\text{out},w}, h \rangle + b_w$, $\nabla_h \log p_h(w) = W_{\text{out}}^\top (e_w - p_h)$. Therefore $g^*(h) = \mathbb{E}_{W \sim p_h}[\nabla_h \log p_h(W) \nabla_h \log p_h(W)^\top] = W_{\text{out}}^\top (\mathrm{Diag}(p_h) - p_h p_h^\top) W_{\text{out}}$. □

Lemma 6.4 (FR–Lipschitzness and Hellinger control)**.** Let $p_h = \mathrm{softmax}(W_{\text{out}}h + b)$ with $W_{\text{out}} \in \mathbb{R}^{|\mathbb{V}| \times d}$. Then for all h,

$$g^*(h) = W_{\text{out}}^\top \big(\mathrm{Diag}(p_h) - p_h p_h^\top \big) W_{\text{out}} \;\preceq\; \frac{\|W_{\text{out}}\|_2^2}{2} I_d.$$

Consequently, the head map $g_{\text{head}} : h \mapsto p_h$ is globally Lipschitz from Euclidean to Fisher–Rao:

$$d_{\mathrm{FR}}\big(p_h, p_{h'}\big) \;\le\; \frac{\|W_{\text{out}}\|_2}{\sqrt{2}} \, \|h - h'\|_2, \qquad \forall\, h, h'.$$

Moreover, using $d_H(p,q) \le d_{\mathrm{FR}}(p,q)/(2\sqrt{2})$ on the simplex,

$$d_H\big(p_h, p_{h'}\big) \;\le\; \frac{\|W_{\text{out}}\|_2}{4} \, \|h - h'\|_2, \qquad \forall\, h, h'.$$

Remark 6.5 (FR–Hellinger relation)**.** On the probability simplex, $d_{\mathrm{FR}}(p,q) = 2 \arccos \sum_w \sqrt{p(w)q(w)}$ and $d_H(p,q) = \sqrt{1 - \sum_w \sqrt{p(w)q(w)}} = \sqrt{2}\, \sin\big(d_{\mathrm{FR}}(p,q)/4\big)$. Thus $d_H(p,q) \le d_{\mathrm{FR}}(p,q)/(2\sqrt{2})$ by $\sin x \le x$, and also $d_{\mathrm{FR}}(p,q) \le \pi\sqrt{2}\, d_H(p,q)$ by $\sin x \ge 2x/\pi$ on $x \in [0, \pi/2]$.

Proof sketch. By Lemma 6.3, $g^*(h) = W_{\text{out}}^\top(\text{Diag}(p_h) - p_h p_h^\top)W_{\text{out}}$. The covariance factor $\text{Diag}(p_h) - p_h p_h^\top$ has spectral norm at most $1/2$, hence $g^*(h) \preceq \frac{\|W_{\text{out}}\|_2^2}{2} I_d$. For any h, h', consider the straight path $h_t = (1-t)h + th'$. The Fisher–Rao distance is upper bounded by the length of the image path: $d_{\text{FR}}(p_h, p_{h'}) \le \int_0^1 \sqrt{(h'-h)^\top g^*(h_t)(h'-h)}\,\mathrm{d}t \le \frac{\|W_{\text{out}}\|_2}{\sqrt{2}}\|h - h'\|_2$. The Hellinger bound follows from $d_H \le d_{\text{FR}}/(2\sqrt{2})$. □

Remark 6.6 (Weight tying)**.** If the head ties weights with the input embedding ($W_{\text{out}} = E^\top$), bounds and spectra involving W_{out} inherit the input-embedding geometry. In particular, in whitened coordinates (Theorem 7.13), the principal directions of g^* align with the principal directions of E's covariance weighted by p_h.

Remark 6.7 (Local vs. global distances)**.** For small displacements, the FR geodesic distance agrees to second order with the quadratic approximation in Equation (6.3). For larger moves, one must integrate along a curve; the straight-line path in $\mathcal{H}$ yields an upper bound on the FR distance between $g_{\text{head}}(h)$ and $g_{\text{head}}(h')$. At *interior* points of the simplex, FR geodesic distance and (symmetric) KL are locally equivalent (second order). Near the boundary, restrict attention to a high-probability subset where $p_h(w) \ge \delta > 0$ so that constants in the local equivalence remain controlled; equivalently, operate with tempered/clipped logits to maintain interiority.

Remark 6.8 (Pullback Metric as Expected Score Outer Product)**.** The Fisher-Rao metric g^{FR} is the expected outer product of the score function $\nabla_\xi \log p_\xi(W)$. This property pulls back to $\mathcal{H}$. Let $p_h(w) = k_{\text{head}}(h, \{w\})$. The score vector for token w with respect to the representation is $\nabla_h \log p_h(w) \in \mathcal{H}$. The pullback metric tensor is precisely the expected outer product of this score:

$$g^*(h) = \mathbb{E}_{W \sim p_h}[(\nabla_h \log p_h(W))(\nabla_h \log p_h(W))^\top]. \tag{6.4}$$

This directly connects the information geometry of $\mathcal{H}$ to the sensitivity of log-probabilities to changes in the representation h. This score vector $\nabla_h \log p_h(W)$ is analogous to that used in score-based generative models, but here taken with respect to the conditioning variable h.

Connection to optimization. Under standard regularity, $g^*(h)$ coincides with the Fisher information (and Gauss–Newton) matrix of the per-step NLL with respect to h, furnishing a direct link between second-order training dynamics and the pullback geometry on $\mathcal{H}$.

The rank of the pullback metric depends on the dimensions of the spaces involved.

Proposition 6.9 (Rank of the Pullback Metric)**.** The rank of the pullback Fisher-Rao metric $g^*(h)$ at a point $h \in \mathcal{H}$ is bounded by the minimum of the representation dimension and the dimension of the probability simplex:

$$\text{rank}(g^*(h)) \le \min(d_{\text{model}}, |\mathbb{V}| - 1). \tag{6.5}$$

Proof. The pullback metric $g^*(h)$ is defined as $g^*(h) = J(h)^\top g^{\text{FR}}(g_{\text{head}}(h))J(h)$, where $J(h)$ is the Jacobian of the map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ (represented in appropriate local coordinates). The dimension of $\mathcal{H}$ is d_{model}, and the dimension of $\mathcal{P}(\mathbb{V})$ is $d_{\text{prob}} = |\mathbb{V}| - 1$. The Jacobian $J(h)$ is a $d_{\text{prob}} \times d_{\text{model}}$ matrix. The Fisher-Rao metric g^{FR} at $g_{\text{head}}(h)$ is a $d_{\text{prob}} \times d_{\text{prob}}$ positive definite matrix (and thus has rank d_{prob}). Using the property that $\text{rank}(A^\top B A) = \text{rank}(A)$ if B is positive definite, we have $\text{rank}(g^*(h)) = \text{rank}(J(h))$. The rank of a matrix is bounded by its dimensions, so

$$\text{rank}(g^*(h)) \le \min(d_{\text{model}}, d_{\text{prob}}) = \min(d_{\text{model}}, |\mathbb{V}| - 1).$$

On the interior of the simplex (Assumption 2.6), g^{FR} is positive definite; near the boundary the argument applies on a high-probability subset with $p_h(w) \ge \delta > 0$. □

6.1 Interpretation and Implications

This geometric perspective provides several insights:

- The quadratic form $g^*(h)(v, v)$ quantifies the local distinguishability (via KL divergence, equation (6.3)) between the output distributions p_h and $p_{h+\epsilon v}$. It measures how sensitive the model's prediction is to perturbations of the hidden state h in a direction v. Directions v with large $g^*(h)(v, v)$ correspond to changes in h that significantly alter the output distribution.
- In modern LMs, the representation dimension is usually much smaller than the vocabulary size ($d_{\text{model}} \ll |\mathbb{V}|$). By Prop. 6.9, $\text{rank}(g^*(h)) \leq d_{\text{model}}$. This upper bound does *not* imply degeneracy on $\mathcal{H}$: if the head Jacobian has full column rank modulo the all-ones logit direction, $g^*(h)$ can be full rank on $\mathbb{R}^{d_{\text{model}}}$. The important phenomenon is instead *anisotropy*: some directions have much larger predictive sensitivity than others.
- The eigenvalues and eigenvectors of the matrix for $g^*(h)$ reveal the local principal directions of predictive sensitivity in $\mathcal{H}$. Directions with large eigenvalues are those where small changes in h induce large changes, geometrically measured by g^{FR}, in the predicted distribution p_h. These directions are important for the head's current prediction map; whether they coincide with globally important semantic directions depends on the encoder distribution and the region of $\mathcal{H}$ being analyzed.

Thus separation should be understood through the head map. Contexts with different predictive futures need not be far in every Euclidean direction of $\mathcal{H}$; they must differ along directions visible to g_{head}. The geometry induced by g^* characterizes this local separation capability. Training shapes the encoder ($k_{\text{bb}} \circ k_{\text{emb}}$) and the LM head k_{head} so that contexts with different true conditionals are mapped to hidden states whose images under g_{head} are appropriately separated in $\mathcal{P}(\mathbb{V})$.

7 NLL as Implicit Spectral Contrastive Learning

A central thesis of this work is that the simple objective of minimizing the negative log-likelihood (NLL) of the next token (equation (2.2)) implicitly functions as a powerful form of contrastive learning. While lacking the explicit positive/negative pairs of standard contrastive methods, we prove that NLL optimization inherently structures the learned representation space $\mathcal{H}$ according to predictive similarity. It achieves this by implicitly solving a spectral objective that aligns the geometry of representations with the underlying predictive structure of the data $P_{\text{data}}(\cdot|x)$, a principle we formalize by connecting NLL to the eigenspectrum of a predictive similarity operator (HaoChen et al., 2021; Tan et al., 2024).

Let $f_{\text{enc}} : \mathbb{V}^* \to \mathcal{H}$ denote the deterministic encoder mapping a context sequence $x = \mathbf{w}_{<t}$ to its hidden representation $h_x = f_{\text{enc}}(x)$, implemented by the composition $k_{\text{bb}} \circ k_{\text{emb}}$. Let $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ be the deterministic mapping from the hidden state to the next-token distribution, corresponding to the LM head kernel k_{head}, such that $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$. The training objective is to minimize the expected KL divergence over the context distribution $\mu_{ctx} = p_{W_{<t}}$:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x) \,\|\, g_{\text{head}}(f_{\text{enc}}(x)))] \tag{7.1}$$

where $P_{\text{data}}(\cdot|x)$ represents the true conditional distribution of the next token given context x, assumed to be derived from the data-generating process.

Successful optimization of $\mathcal{L}(\theta)$ drives the model's output distribution $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$ towards the target distribution $P_{\text{data}}(\cdot|x)$ in the sense of minimizing average KL divergence. As we argue below, this fundamental requirement indirectly imposes geometric constraints on the distribution of representations $h_x = f_{\text{enc}}(x)$ in $\mathcal{H}$.

7.1 Constraint on Output Distribution Approximation

Minimizing the NLL loss (equation (7.1)) directly forces the model's predicted distribution $p_\theta(\cdot|x)$ to closely approximate the target distribution $P_{\text{data}}(\cdot|x)$. This closeness can be measured not only by KL divergence but also by other standard metrics on probability distributions, due to well-known inequalities relating them.

Theorem 7.1 (Output Distribution Approximation Constraint)**.** Assume the model parameters θ yield a small average KL divergence $\mathcal{L}_{\text{KL}}(\theta) \coloneqq \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))]$, where $p_\theta(\cdot|x) = g_{\text{head}}(f_{\text{enc}}(x))$. With our global convention for Hellinger, $d_H^2(p,q) \le \frac{1}{2}D_{\text{KL}}(p\|q)$ on finite alphabets. Therefore,

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_H(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^2] \le \tfrac{1}{2}\,\mathcal{L}_{\text{KL}}(\theta). \tag{7.2}$$

Consequently, if the model fits the data well ($\mathcal{L}_{\text{KL}}(\theta)$ is small), then for any pair (x, x'),

$$\left|d_H(p_\theta(\cdot|x), p_\theta(\cdot|x')) - d_H(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))\right| \le \epsilon_x + \epsilon_{x'}, \qquad \epsilon_x := d_H(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x)), \tag{7.3}$$

and $\mathbb{P}(\epsilon_X \ge \delta) \le \frac{1}{2}\,\mathcal{L}_{\text{KL}}(\theta)/\delta^2$ by Markov. Hence, for independent $X, X' \sim \mu_{\text{ctx}}$,

$$\mathbb{P}\Big(\left|d_H(p_\theta^X, p_\theta^{X'}) - d_H(p_{\text{data}}^X, p_{\text{data}}^{X'})\right| > 2\delta\Big) \le \frac{\mathcal{L}_{\text{KL}}(\theta)}{\delta^2},$$

where $p_\theta^X = p_\theta(\cdot|X)$ and $p_{\text{data}}^X = P_{\text{data}}(\cdot|X)$. Thus pairwise predictive distances are approximated on typical random pairs whenever the average KL is small.

Proof Sketch. Apply $d_H^2 \le \frac{1}{2}\,\text{KL}$ pointwise with $p = P_{\text{data}}(\cdot|x)$ and $q = p_\theta(\cdot|x)$ and take expectations to obtain Equation (7.2). The pairwise inequality follows from the triangle inequality for d_H. Markov's inequality gives the one-point tail bound, and a union bound over independent X, X' gives the displayed random-pair bound. □

Different normalizations of d_{TV} and d_H only change constants; we use the convention fixed in the preliminaries throughout.

This theorem formalizes the intuition that minimizing the NLL objective forces the model's predictions to mirror the structure of the true predictive distributions, specifically in terms of their pairwise distances.

7.2 Consequences for Representation Geometry

Theorem 7.1 establishes that predictively dissimilar contexts x, x' must lead to distinct model output distributions $p_\theta(\cdot|x), p_\theta(\cdot|x')$. Since $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$ and $p_\theta(\cdot|x') = g_{\text{head}}(h_{x'})$, this requirement imposes constraints on the corresponding representations $h_x = f_{\text{enc}}(x)$ and $h_{x'} = f_{\text{enc}}(x')$. Specifically, h_x and $h_{x'}$ must differ in ways that are discernible by the head mapping g_{head}. The information geometry of the head mapping, captured by the pullback metric $g^*(h)$ (Section 6), determines which differences in representation space are discernible.

Corollary 7.2 (Implicit Representation Separation)**.** Assume the model fits the data well ($\mathcal{L}(\theta)$ is small). Let $p_h \coloneqq g_{\text{head}}(h)$ and $h_x = f_{\text{enc}}(x)$. For two contexts x, x', set $v \coloneqq h_x - h_{x'}$ and consider the straight path $h_t = (1-t)h_{x'} + th_x$. The image path $\gamma(t) = g_{\text{head}}(h_t)$ in the output simplex has length

$$L(\gamma) = \int_0^1 \sqrt{g^*(h_t)(v,v)}\,\mathrm{d}t,$$

so the Fisher–Rao geodesic distance satisfies

$$d_{\text{FR}}(p_{h_x}, p_{h_{x'}}) \;\le\; \int_0^1 \sqrt{g^*(h_t)(v,v)}\,\mathrm{d}t \;\le\; \sqrt{\int_0^1 g^*(h_t)(v,v)\,\mathrm{d}t}. \tag{7.4}$$

Thus $d_{\text{FR}}(p_{h_x}, p_{h_{x'}})^2 \le \int_0^1 g^*(h_t)(v,v)\,\mathrm{d}t$. Combined with Theorem 7.1 and the monotone relation between Hellinger and Fisher–Rao distance on the simplex, predictively dissimilar contexts force a large right-hand side in Eq. (7.4). Equivalently, v must have substantial components along directions where g^* is large, unless the head is locally insensitive along the path.

Proof Sketch. From Theorem 7.1, if $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, then $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be large. The squared distance between two points in a Riemannian manifold is related to the integrated metric along a geodesic. For small distances in output space, we have $d_{\text{out}}(p,q)^2 \approx D_{\text{KL}}(p\|q)$, which from Eq. (6.3) is related to the pullback metric g^*. A large distance between $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ implies a large integrated path length according to the pullback geometry, forcing h_x and $h_{x'}$ to differ along directions where g^* is large. □

This corollary establishes that NLL minimization implicitly acts like a contrastive learning objective: it pushes representations $h_x, h_{x'}$ apart if their corresponding contexts are predictively dissimilar. This differential pressure based on predictive similarity forms the basis for our connection to spectral methods.

7.3 Predictive Similarity Kernels

The preceding analysis suggests that NLL shapes the representation geometry based on the *dissimilarity* between the true next-token distributions $P_{\text{data}}(\cdot|x)$. To connect this to spectral methods, which operate on similarity structures, we formalize the complementary notion of *predictive similarity*.

Definition 7.3 (Predictive Similarity Kernel)**.** Let $p_x \coloneqq P_{\text{data}}(\cdot|x)$ denote the true conditional distribution for context x. A predictive similarity kernel is a bounded symmetric real-valued function $K : \mathbb{V}^* \times \mathbb{V}^* \to \mathbb{R}$ quantifying the similarity between p_x and $p_{x'}$. For graph-Laplacian and Dirichlet-energy statements below, we additionally require $K(x,x') \ge 0$ so that it can be interpreted as an edge weight. For operator and CCA statements, the important condition is positive semidefiniteness of the quadratic form; after centering or whitening, such kernels may take negative pairwise values. Examples include:

- **Bhattacharyya Coefficient Kernel**: $K_{\text{BC}}(x,x') \coloneqq \text{BC}(p_x, p_{x'}) = \sum_{w\in\mathbb{V}} \sqrt{p_x(w)p_{x'}(w)}$. This measures the cosine similarity between the square-root vectors $(\sqrt{p_x(w)})_w$. Under our global convention $d_H^2(p,q) = 1 - \sum_w \sqrt{p(w)q(w)}$, we have $d_H^2(p_x, p_{x'}) = 1 - K_{\text{BC}}(x,x')$, and K_{BC} is positive semidefinite. High K_{BC} corresponds to low d_H.

- **Hellinger-based Kernel (Gaussian Kernel on $\sqrt{p}$)**: $K_{\mathrm{H}}(x, x') \coloneqq \exp(-\beta d_H^2(p_x, p_{x'}))$ for some scale $\beta > 0$. Since $d_H^2(p, q) = 1 - \langle \sqrt{p}, \sqrt{q} \rangle$, this kernel equals $e^{-\beta} \exp(\beta \langle \sqrt{p_x}, \sqrt{p_{x'}} \rangle)$ and is positive semidefinite.
- **Expected Likelihood Kernel (Linear Kernel)**: $K_{\mathrm{Lin}}(x, x') \coloneqq \langle p_x, p_{x'} \rangle = \sum_{w \in \mathbb{V}} p_x(w) p_{x'}(w)$. This is the standard linear kernel (inner product) between probability vectors $p_x, p_{x'}$ and is positive semidefinite. High values indicate significant overlap between the distributions. It can be interpreted as the expected likelihood $p_{x'}(W)$ under $W \sim p_x$.
- **Divergence-based Kernels**: $K(x, x') = \exp(-\beta S(p_x, p_{x'}))$ for a symmetric divergence S, such as Jensen–Shannon divergence. Positive semidefiniteness is not automatic for arbitrary choices of S and must be verified separately.

In general, larger values of $K(x, x')$ indicate higher predictive similarity. To act on representations, we disintegrate K through the encoder f_{enc} by defining the induced kernel on $\mathcal{H}$:

$$\widetilde{K}(h, h') \triangleq \mathbb{E}\big[K(X, X') \;\big|\; f_{\mathrm{enc}}(X) = h, f_{\mathrm{enc}}(X') = h'\big], \tag{7.5}$$

whenever the conditional expectation exists. Since all spaces here are standard Borel, regular conditional probabilities exist; hence, the disintegration above is well-defined. This produces a bounded symmetric measurable kernel on $(\mathcal{H}, \mu)$, where $\mu = (f_{\mathrm{enc}})_{\#}\mu_{ctx}$, suitable for operator-theoretic analysis. For the operator results below we assume K is *positive semidefinite* (PSD), meaning that its integral quadratic form is nonnegative for all bounded measurable test functions. The examples K_{BC}, K_{H}, and K_{Lin} are PSD. For graph-energy results, we also assume pointwise nonnegativity.

Lemma 7.4 (PSD preserved under disintegration)**.** If K is PSD on $\mathbb{V}^*$ in the quadratic-form sense, then $\widetilde{K}$ defined in (7.5) is PSD on $(\mathcal{H}, \mu)$. Consequently, for all $\psi \in L^2(\mathcal{H}, \mu)$,

$$\iint \psi(h)\, \widetilde{K}(h, h')\, \psi(h')\, \mu(\mathrm{d}h) \mu(\mathrm{d}h') \;\geq\; 0.$$

Proof sketch. Let $(X, X') \sim \mu_{ctx} \otimes \mu_{ctx}$ and $H = f_{\mathrm{enc}}(X)$, $H' = f_{\mathrm{enc}}(X')$. For any bounded measurable ψ, by the tower property,

$$\mathbb{E}[\psi(H)\psi(H')K(X, X')] \;=\; \mathbb{E}\big[\,\mathbb{E}[K(X, X') \mid H, H']\, \psi(H)\psi(H')\big] = \iint \psi(h)\widetilde{K}(h, h')\psi(h')\, \mathrm{d}\mu(h)\mathrm{d}\mu(h').$$

Since K is PSD, the left-hand side is ≥ 0, hence so is the right-hand side. □

7.4 Connection to Graph Laplacian and Dirichlet Energy Minimization

Consider an undirected graph where contexts $x \in \mathbb{V}^*$ are nodes distributed according to μ_{ctx}, and edge weights are given by a bounded nonnegative symmetric predictive similarity kernel $K(x, x')$. The quadratic form of the associated graph Laplacian corresponds to the Dirichlet energy, which measures how "smooth" a function ϕ (e.g., a 1D projection of the representations) is over the graph.

$$\mathcal{E}_K(\phi) \coloneqq \frac{1}{2} \iint K(x, x')(\phi(x) - \phi(x'))^2\, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x') = \langle \phi, \Delta_K \phi \rangle_{L^2(\mu_{ctx})}. \tag{7.6}$$

Spectral clustering aims to find embeddings (represented by functions ϕ) that minimize this energy subject to constraints, effectively mapping similar contexts close together. The NLL objective, through Theorem 7.2, exerts a related pressure.

Recent work (Park et al., 2024) connects in-context learning to Dirichlet energy minimization on a task-similarity graph. Our work shows this is a foundational principle of NLL pre-training itself, where similarity is defined by the intrinsic next-token distributions $P_{\text{data}}(\cdot|x)$.

Proposition 7.5 (Dirichlet energy bound under local bi-Lipschitzness)**.** Fix a bounded nonnegative symmetric predictive-similarity kernel K and define, for $v \in \mathcal{H}$, the projection $\phi_v(x) = \langle h_x, v\rangle$ with $h_x = f_{\text{enc}}(x)$. Assume moreover that g_{head} is locally bi-Lipschitz on a high-probability compact set: there exist $0 < m \le L < \infty$ such that for all h, h' in this set,

$$m\,\|h - h'\|_2 \;\le\; d_{\text{FR}}(p_h, p_{h'}) \;\le\; L\,\|h - h'\|_2,$$

and use $d_{\text{FR}}(p, q) \le \pi\sqrt{2}\, d_H(p, q)$, so that

$$\|h - h'\|_2 \;\le\; \frac{\pi\sqrt{2}}{m}\, d_H(p_h, p_{h'}).$$

Suppose

$$\varepsilon \;\coloneqq\; \mathbb{E}_{x\sim\mu_{ctx}}\Big[d_H\big(P_{\text{data}}(\cdot\,|\,x),\, p_\theta(\cdot\,|\,x)\big)^2\Big]$$

is small and let $\|K\|_\infty < \infty$. Then

$$\begin{aligned}\mathcal{E}_K(\phi_v) \;&\le\; \frac{\pi^2\|v\|^2}{m^2} \iint K(x,x')\, d_H\big(p_\theta(\cdot\,|\,x), p_\theta(\cdot\,|\,x')\big)^2\, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x') \\ &\le\; C_{\text{geom}} \iint K(x,x')\, d_H\big(P_{\text{data}}(\cdot\,|\,x), P_{\text{data}}(\cdot\,|\,x')\big)^2\, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x') \;+\; \frac{12\pi^2\|v\|^2}{m^2}\|K\|_\infty\,\varepsilon,\end{aligned}$$

where $C_{\text{geom}} = 3\pi^2\|v\|^2/m^2$.

Proof. The local lower Lipschitz condition and the FR–Hellinger relation imply

$$\|h_x - h_{x'}\|_2^2 \le \frac{2\pi^2}{m^2}\, d_H\big(p_\theta(\cdot|x), p_\theta(\cdot|x')\big)^2.$$

Therefore

$$\mathcal{E}_K(\phi_v) \le \frac{1}{2}\|v\|^2 \iint K(x,x')\|h_x - h_{x'}\|^2\, \mathrm{d}\mu(x)\mathrm{d}\mu(x'),$$

which gives the first bound. Let $e_x = d_H(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))$. The triangle inequality gives

$$d_H(p_\theta(\cdot|x), p_\theta(\cdot|x')) \le d_H(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')) + e_x + e_{x'}.$$

Using $(a+b+c)^2 \le 3a^2 + 6b^2 + 6c^2$ and boundedness of K gives the stated second inequality. □

This proposition is a conditional control statement: if NLL makes the predicted conditionals close to the data conditionals and if the head is locally injective on the relevant region, then projections of the representation have small Dirichlet energy whenever the nonnegative predictive-similarity kernel assigns large weight mainly to close predictive conditionals.

7.5 NLL as Spectral Objective

We now state a conditional spectral result. The statement concerns a calibrated quadratic surrogate to the NLL objective and the variance-normalized alignment problem induced by that surrogate; it should not be read as an exact equivalence for the original nonconvex likelihood. This viewpoint is closest to spectral analyses of contrastive learning such as (Tan et al., 2024).

Definition 7.6 (Predictive Similarity Operator)**.** Let $\mu = (f_{\mathrm{enc}})_{\#}\mu_{ctx}$ on $\mathcal{H}$ and $\widetilde{K}$ be as in (7.5). Define the integral operator $M_{\widetilde{K}} : L^2(\mathcal{H}, \mu) \to L^2(\mathcal{H}, \mu)$ by

$$(M_{\widetilde{K}}\psi)(h) := \int_{\mathcal{H}} \widetilde{K}(h, h')\, \psi(h')\, \mu(\mathrm{d}h'). \tag{7.7}$$

Under boundedness and symmetry, $M_{\widetilde{K}}$ is compact and self-adjoint with real spectrum. If, in addition, K is PSD (hence $\widetilde{K}$ is PSD by Lemma 7.4), then $M_{\widetilde{K}}$ is positive semidefinite and its spectrum is nonnegative.

Remark 7.7 (Estimating $\widetilde{K}$ in practice)**.** While $K(x, x')$ is defined using $P_{\mathrm{data}}(\cdot \mid x)$, in practice one may approximate it via: (i) a stronger teacher model to produce $\widehat{p}(\cdot \mid x)$; (ii) the learner's own predictions $p_\theta(\cdot \mid x)$ (late-training snapshot); or (iii) surrogate "predictive prototypes" $\bar{g}_x = \mathbb{E}[g(W) \mid x]$ and the linear kernel $K(x, x') = \langle \bar{g}_x, \bar{g}_{x'} \rangle$. Each choice induces a corresponding $\widetilde{K}$ via disintegration (7.5), and leads to the same operator-eigenfunction pipeline under whitening. For the CCA/spectral form in Theorem 7.10, it is natural to use *whitened* prototypes

$$\widetilde{g}_x := C_{\bar{g}\bar{g}}^{-1/2}\, \bar{g}_x,$$

so that $K(x, x') = \langle \widetilde{g}_x, \widetilde{g}_{x'} \rangle$ matches the $C_{\bar{g}\bar{g}}^{-1}$ factor. This whitened linear kernel is PSD as a kernel, but after centering it need not be pointwise nonnegative; it is therefore appropriate for operator/CCA analysis, not necessarily as a graph edge-weight kernel without further modification.

Since all spaces are standard Borel and f_{enc} is Borel-measurable, regular conditional probabilities exist; hence the conditional expectation in Eq. (7.5) is well defined a.s. and $\widetilde{K}$ is measurable and bounded. If $\widetilde{K}$ is symmetric and bounded, the induced operator is Hilbert–Schmidt, compact, and self-adjoint; its eigenfunctions capture dominant patterns of predictive similarity. Our key result is that, under a linear-softmax LM head, NLL training aligns the representation geometry with these eigenspaces in a generalized spectral/Canonical Correlation Analysis (CCA) sense.

Theorem 7.8 (Calibrated Fenchel-Young surrogate)**.** Assume a linear-softmax LM head $p_\theta(w \mid x) \propto \exp\{\langle g(w), h_x \rangle + b_w\}$ with $h_x = f_{\mathrm{enc}}(x)$ and $\max_w \|g(w)\| \le R < \infty$. For any $\tau > 0$ and all $h \in \mathcal{H}$,

$$\log \sum_w \exp\{\langle g(w), h \rangle + b_w\} \;\le\; C_\tau + \tfrac{\tau}{2}\|h\|^2, \qquad C_\tau = \log \sum_w \exp\big\{b_w + \tfrac{1}{2\tau}\|g(w)\|^2\big\}.$$

Equality in the per-term Young bound holds iff $g(w) = \tau h$; thus, equality in the sum requires this for every contributing w, which is possible only in degenerate cases. Consequently,

$$\mathbb{E}\big[-\log p_\theta(W \mid X)\big] \;\le\; (C_\tau - \mathbb{E}[b_W]) \;-\; \mathbb{E}\langle \bar{g}_X, h_X \rangle \;+\; \tfrac{\tau}{2}\,\mathbb{E}\|h_X\|^2, \quad \bar{g}_X := \mathbb{E}[g(W) \mid X].$$

The RHS equals $\frac{\tau}{2}\,\mathbb{E}\big\|h_X - \frac{1}{\tau}\bar{g}_X\big\|^2 - \frac{1}{2\tau}\,\mathbb{E}\|\bar{g}_X\|^2 + (C_\tau - \mathbb{E}[b_W])$, i.e., a *calibrated regression* objective for h_X onto $\bar{g}_X/\tau$.

Practical calibration of τ. Minimizing the bound w.r.t. τ balances C_τ (decreasing in τ) against $\frac{\tau}{2}\mathbb{E}\|h_X\|^2$ (increasing). A practical scheme is to (i) estimate $R \approx \max_w \|g(w)\|$, (ii) track $\widehat{m} = \mathbb{E}\|h_X\|^2$ on a held-out set, and (iii) choose τ by a 1D line search minimizing $C_\tau + \frac{\tau}{2}\widehat{m}$ (cost dominated by evaluating C_τ once per candidate). For fixed representations and prototypes, τ only rescales the prototype side of the regression surrogate and therefore cancels from the CCA directions in Theorem 7.10. During actual training, changing τ changes the surrogate objective and should be viewed as a calibration choice rather than an invariance of the original NLL.

Proof. Young's inequality with parameter τ gives

$$\langle g(w), h\rangle \le \frac{1}{2\tau}\|g(w)\|^2 + \frac{\tau}{2}\|h\|^2.$$

Substituting this bound into each exponential term in the log-partition function yields the stated bound with constant C_τ. Taking expectation over $W \mid X$ gives the linear term $-\mathbb{E}\langle \bar{g}_X, h_X\rangle$ and the bias term $-\mathbb{E}[b_W]$. Completing the square gives the calibrated regression form. Equality in the Young bound occurs exactly when $g(w) = \tau h$ for the corresponding term, so equality for the whole log-sum requires this simultaneously for all tokens with nonzero contribution, which is degenerate except in special cases. □

Proposition 7.9 (Closed-form in a linearized encoder)**.** Let $h_X = A\,\phi(X)$ with fixed feature map ϕ and $C_{\phi\phi} = \mathbb{E}[\phi\phi^\top] \succ 0$. Minimizing the surrogate in Theorem 7.8 yields the normal equations $A^\star C_{\phi\phi} = \frac{1}{\tau}C_{\phi\bar{g}}^\top$, i.e., $A^\star = \frac{1}{\tau}C_{\phi\bar{g}}^\top C_{\phi\phi}^{-1}$ (ridge modifications if desired).

Theorem 7.10 (CCA form of the NLL surrogate)**.** Let $h_X \in \mathbb{R}^d$ and $\bar{g}_X \in \mathbb{R}^{d_g}$ be (centered) square-integrable random vectors with $C_{hh} \succ 0$, $C_{\bar{g}\bar{g}} \succ 0$, and cross-covariance $C_{h\bar{g}}$. Consider the canonical-correlation objective

$$\max_{u,v\neq 0} \frac{u^\top C_{h\bar{g}} v}{\sqrt{u^\top C_{hh} u}\,\sqrt{v^\top C_{\bar{g}\bar{g}} v}}.$$

Optimizing out v yields the equivalent squared Rayleigh quotient

$$\max_{u\neq 0} \frac{u^\top C_{h\bar{g}}\, C_{\bar{g}\bar{g}}^{-1}\, C_{h\bar{g}}^\top u}{u^\top C_{hh}\, u}.$$

The maximizers are generalized eigenvectors of

$$C_{h\bar{g}} C_{\bar{g}\bar{g}}^{-1} C_{h\bar{g}}^\top u = \lambda C_{hh} u.$$

Equivalently, they are right eigenvectors of

$$M = C_{hh}^{-1} C_{h\bar{g}} C_{\bar{g}\bar{g}}^{-1} C_{h\bar{g}}^\top,$$

which is generally not symmetric in the Euclidean inner product but is self-adjoint with respect to the C_{hh} inner product. Multiple directions with constraints $u_i^\top C_{hh} u_j = \delta_{ij}$ are given by the top generalized eigenvectors. In whitened coordinates ($C_{hh} = I$), this reduces to an ordinary symmetric eigenproblem.

Corollary 7.11 (Operator view (functional version))**.** When optimizing over linear functionals of $h \in L^2(\mathcal{H}, \mu)$ with whitening, the variational problem in Theorem 7.10 corresponds to an eigenfunction problem for the Hilbert–Schmidt operator with kernel

$$\widetilde{K}(h, h') \;=\; \mathbb{E}\big[\langle \widetilde{g}_X, \widetilde{g}_{X'}\rangle \;\big|\; f_{\mathrm{enc}}(X) = h,\; f_{\mathrm{enc}}(X') = h'\big], \qquad \widetilde{g}_X \coloneqq C_{\bar{g}\bar{g}}^{-1/2}\bar{g}_X.$$

Corollary 7.12 (Finite-sample analogue)**.** Let $\widehat{C}_{hh}, \widehat{C}_{\bar{g}\bar{g}}, \widehat{C}_{h\bar{g}}$ be centered empirical second moments computed from n i.i.d. contexts and their targets, and let $\lambda > 0$ be a ridge parameter. Under standard boundedness or sub-Gaussian moment assumptions and a positive eigengap γ for the corresponding ridge-regularized population problem, maximizing

$$u^\top \widehat{C}_{h\bar{g}} (\widehat{C}_{\bar{g}\bar{g}} + \lambda I)^{-1} \widehat{C}_{h\bar{g}}^\top u \quad \text{subject to} \quad u^\top (\widehat{C}_{hh} + \lambda I) u = 1$$

yields a subspace whose principal-angle error is $O_p(n^{-1/2}/\gamma)$ by covariance concentration and Davis–Kahan perturbation theory, with constants depending on the moment bounds and ridge level.

Corollary 7.13 (Whitened special case)**.** If $C_{hh} = I$ (representation whitening) and we restrict to variance-one directions u, the problem reduces to an ordinary eigenproblem for $C_{h\bar{g}}\, C_{\bar{g}\bar{g}}^{-1}\, C_{h\bar{g}}^\top$. Equivalently, defining whitened prototypes $\widetilde{g}_X = C_{\bar{g}\bar{g}}^{-1/2} \bar{g}_X$, we have

$$C_{h\bar{g}}\, C_{\bar{g}\bar{g}}^{-1}\, C_{h\bar{g}}^\top \;=\; C_{h\widetilde{g}}\, C_{h\widetilde{g}}^\top,$$

so the leading eigendirections depend only on the cross-covariance between h_X and $\widetilde{g}_X$.

Proof Sketch. The surrogate in Theorem 7.8 yields a regression view with target $\bar{g}_X/\tau$. The regression objective alone does not identify a unique spectral basis; the spectral form appears after imposing whitening or variance/orthogonality constraints on linear functionals. Under these constraints, maximizing alignment becomes the CCA Rayleigh quotient above. The operator form follows from the Hilbert–Schmidt correspondence under whitening. The scalar calibration parameter τ rescales the prototype side and cancels in the CCA eigendirections. □

In summary, for linear-softmax LM heads, a calibrated quadratic surrogate to NLL gives a regression-to-prototypes view. Once one imposes whitening or variance-normalized linear-function constraints, the induced alignment problem becomes a generalized spectral/CCA problem. This provides a principled bridge from next-token likelihood to spectral organization of $\mathcal{H}$ under explicit assumptions.

8 Related Work

The theoretical understanding of representation learning in deep neural networks has been advanced along several parallel, yet largely disconnected, fronts. A significant challenge lies in the absence of a unified mathematical language capable of connecting a model's compositional architecture and its training dynamics to the emergent geometric structure of its learned representations. This work is situated at the confluence of these disparate research programs, aiming to synthesize the algebraic, compositional perspective of categorical probability with the metric, differential-geometric view of information geometry. We structure our review of the related literature into two parts. First, we introduce the foundational languages of probability that our framework unifies: the synthetic view of probability rooted in category theory and the metric view rooted in information geometry. Second, we survey the tools and concepts used to analyze the geometry of learned representations, focusing on the training objectives that guide learning, the spectral methods used to measure the resulting geometry, and the optimization mechanisms that shape it.

8.1 The Languages of Probability

Probability as a Category. A burgeoning field of research seeks to reformulate probability theory on a more abstract, algebraic foundation using the language of category theory (Baez et al., 2016; Fong and Spivak, 2018). This synthetic approach, in contrast to the classical analytic approach built upon measure theory, aims to derive probabilistic concepts from a small set of powerful axioms (Fritz, 2020; Perrone, 2023b). The central object of study in this domain is the Markov category (Cho and Jacobs, 2019; Fritz, 2020). Formally, a Markov category is a symmetric monoidal category where each object is equipped with a commutative comonoid structure, consisting of morphisms that represent the abstract operations of copying and discarding information (Cho and Jacobs, 2019; Fritz, 2020; Perrone, 2023b). The morphisms in such a category are interpreted as stochastic maps, or Markov kernels, which are probabilistic mappings between objects (Baez et al., 2016; Pardo-Guerra et al., 2025).

The pioneering work of Fritz (2020) has established Markov categories as a robust framework for synthetic probability and statistics. A key advantage of this formalism is its generality; it provides a uniform treatment of vastly different probabilistic settings. For instance, the category `FinStoch`, with finite sets as objects and stochastic matrices as morphisms, and the category `BorelStoch`, with standard Borel spaces as objects and their corresponding Markov kernels as morphisms, are both canonical examples of Markov categories (Fritz, 2020). This high level of abstraction allows for the proof of fundamental statistical theorems, such as the Fisher-Neyman factorization theorem and Kolmogorov's zero-one law, in a purely diagrammatic and synthetic manner, avoiding the low-level complexities of measure theory (Fritz, 2020; Fritz and Rischel, 2020). As Fritz (2020) argues, relying on measure theory is akin to programming in machine code, whereas the categorical approach provides a higher-level language that facilitates reasoning about complex, compositional systems.

This line of inquiry is not merely a formal exercise; it is directly motivated by the challenges of understanding modern machine learning systems (Yuan, 2023). The compositional structure of deep neural networks, finds a natural description in the language of category theory (Fong and Spivak, 2018; Yuan, 2023; Pardo-Guerra et al., 2025).

Probability as a Manifold. In parallel to the algebraic developments in category theory, the field of information geometry (IG) has provided a powerful differential geometric lens for studying machine learning (Amari and Nagaoka, 2000). Foundational work by Amari and Nagaoka (2000) demonstrated that a parametric family of probability distributions can be viewed as a smooth manifold endowed with a canonical Riemannian metric, the Fisher-Rao metric, and a pair of dually-coupled affine connections. This geometric structure is not arbitrary; it can be intrinsically derived from statistical divergence functions, such as the Kullback-Leibler (KL) divergence, which serves as a measure of dissimilarity between distributions.

When applying these geometric tools to deep learning, it is crucial to draw a distinction between IG and the related field of geometric deep learning (GDL). GDL is primarily concerned with generalizing neural network architectures to operate on data that resides in non-Euclidean domains, such as graphs or manifolds; its focus is the geometry of the input data space. In contrast, IG has traditionally been used to analyze the geometry of the parameter space of a model. By viewing the set of all possible model parameters as a manifold, IG provides sophisticated tools for understanding the dynamics of training, offering a more nuanced perspective on optimization than that afforded by standard ℓ_1 or ℓ_2 regularization (Amari and Nagaoka, 2000).

Towards Categorical Information Geometry. While the algebraic and geometric approaches have largely evolved independently, a new frontier is emerging at their intersection. Recent work has begun to explicitly forge a synthesis, aiming to create a categorical information geometry (Perrone, 2023b). This research program, led by researchers such as Perrone (2023b), seeks to enrich the abstract, compositional structures of Markov categories with the metric and quantitative notions central to information theory and geometry, such as entropy and divergence (Perrone, 2023a,b).

This emerging synthesis recognizes that a complete theoretical picture requires both the compositional language of categories and the metric language of geometry. However, to date, the applications of this nascent field have focused primarily on reformulating abstract probability theory. The critical connection to the analysis of learned representations in practical, large-scale deep learning models remains largely unexplored. This work aims to bridge that gap, demonstrating that a categorical information geometry provides the ideal framework for analyzing the structure of the representation spaces sculpted by the learning process.

8.2 The Geometry of Learning and Representation

Objectives of Learning Representations. A guiding principle for understanding the purpose of representation learning is the Information Bottleneck (IB) theory, introduced by Tishby et al. (2000). The IB principle posits that an optimal representation T of some input data X should be a bottleneck that is maximally informative about a relevant target variable Y while being maximally compressed with respect to the input X (Tishby et al., 2000; Shwartz-Ziv and Tishby, 2017). This trade-off between predictive accuracy and compressional complexity provides a powerful, high-level objective for representation learning.

The IB framework has been particularly influential in the theoretical analysis of deep learning. It led to the hypothesis that the training of deep neural networks proceeds in two distinct phases: an initial fitting phase, where the mutual information $I(T;Y)$ between the representation and the target increases, followed by a "compression" phase, where the mutual information $I(X;T)$ between the input and the representation decreases (Shwartz-Ziv and Tishby, 2017). While the universality of this two-phase dynamic is a subject of ongoing debate, the core intuition—that effective training involves not just memorization but also a form of structured compression—provides a compelling motivation for investigating the geometry of the learned representations.

Measurements of Representation Geometry. To move from high-level principles to concrete analysis, we require quantitative tools to probe the geometric properties of the high-dimensional activation spaces within a neural network. A particularly effective set of tools for this purpose comes from spectral graph theory. Given a graph with adjacency matrix A and degree matrix D, the Graph Laplacian is defined as $L = D - A$ (Berahmand et al., 2025). For any function f defined on the nodes of the graph (e.g., a feature activation), the quadratic form $f^\top Lf$ defines the graph's Dirichlet energy. This quantity measures the smoothness of the function with respect to the graph structure; a low Dirichlet energy indicates that connected nodes have similar function values (Park et al., 2024).

This concept has a rich history in machine learning, forming the basis of spectral clustering algorithms (Berahmand et al., 2025) and, more recently, being used to analyze and mitigate the over-smoothing problem in Graph Neural Networks (GNNs). Over-smoothing occurs when stacking many GNN layers causes the representations of all nodes to converge to an indistinguishable point, a phenomenon characterized by the Dirichlet energy of the representations collapsing to zero.

Most critically for the present work, this classical geometric tool is deeply relevant to the internal dynamics of the most advanced models. A recent study demonstrates that during in-context learning, Large Language Models (LLMs) dynamically reorganize their internal concept representations in a manner that explicitly minimizes the Dirichlet energy with respect to an implicit graph structure defined by the context (Park et al., 2024). This groundbreaking result elevates Dirichlet energy from a tool for analyzing explicit graphs to a general principle governing the emergent geometry of learned representations. This trend towards spectral analysis is further evidenced by other recent work using methods like Centered Kernel Alignment (CKA) to track representation dynamics and spectral editing of activations (SEA) to control model behavior (Qiu et al., 2024).

Mechanisms of Learning Representations. The geometric structures observed in learned representations are not accidental; they are a direct consequence of the implicit biases of the training algorithm. For modern, highly overparameterized models, the optimization process itself, typically driven by variants of gradient descent, imparts an implicit bias or implicit regularization on the final solution (Vardi, 2023). Even when multiple parameter settings can achieve zero training error, the optimization algorithm preferentially converges to a "simple" solution that generalizes well. For linear models trained on classification tasks, this bias often corresponds to finding the maximum-margin separator, a classic geometric concept (Gunasekar et al., 2018; Soudry et al., 2018; Chizat and Bach, 2020).

A useful modern perspective is that training with the standard Negative Log-Likelihood (NLL) objective has a contrastive-like component. The NLL loss for a sample (x, y_{true}), given by $-\log P(y_{\text{true}}|x) = -\log \frac{\exp(z_{\text{true}})}{\sum_j \exp(z_j)}$, is minimized by increasing the logit z_{true} for the observed class while reducing its relative competition with other logits. This resembles a contrastive loss that pulls the representation of x toward a positive target while pushing it away from competing targets, although the equivalence is exact only under specific modeling choices. This perspective links likelihood training to the broader literature on self-supervised contrastive learning, where objectives explicitly sculpt representation geometry.

9 Conclusion

In this work, we introduced a mathematical framework for analyzing the Autoregressive generation step in language models, leveraging the expressive power of Markov Categories. By modeling the process as a composition of Markov kernels, $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$, we established a foundation for a compositional, information-theoretic analysis. This allowed us to formally quantify the information surplus in hidden states, providing a clear theoretical rationale for the success of modern speculative decoding techniques like EAGLE.

More importantly, our framework clarifies why the negative log-likelihood (NLL) objective can shape useful internal structure. Under realizability and vanishing average KL, NLL matches the data's intrinsic conditional stochasticity, a process we measure with categorical entropy. Under a linear-softmax head, a calibrated quadratic surrogate to NLL yields a regression-to-predictive-prototypes objective whose whitened alignment form is a CCA/eigenproblem. By analyzing the information geometry of the prediction head via the pullback Fisher–Rao metric, we showed how training can emphasize directions of high predictive sensitivity and how, under explicit assumptions, these directions connect to a predictive-similarity operator. This provides a mathematically grounded

explanation for how likelihood training can produce organized representations without explicit contrastive pairs.

This compositional, probabilistic, and information-geometric perspective offers a principled alternative to purely empirical or heuristic analysis, unifying concepts from information theory, geometry, and spectral methods to study the mechanisms behind large language models.

Acknowledgments

We thank anonymous reviewers for their constructive and helpful feedback. We used LLMs for word-editing as well as figure plots in this work.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

John C Baez, Brendan Fong, and Blake S Pollard. A compositional framework for markov processes. *Journal of Mathematical Physics*, 57(3):033301, 2016.

Kamal Berahmand, Farid Saberi-Movahed, Razieh Sheikhpour, Yuefeng Li, and Mahdi Jalili. A comprehensive survey on spectral clustering with graph structure learning. *arXiv preprint arXiv:2501.13597*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020.

Kenta Cho and Bart Jacobs. Disintegration and bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits, 2021.

Brendan Fong and David I Spivak. Seven sketches in compositionality: An invitation to applied category theory. *arXiv preprint arXiv:1803.05316*, 2018.

Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

Tobias Fritz and Eigil Fjeldgren Rischel. Infinite products and zero-one laws in categorical probability. *Compositionality*, 2:3, 2020. doi: 10.32408/compositionality-2-3.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

Olav Kallenberg and Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, volume 29, 2016.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Sebastian Pardo-Guerra, Johnny Jingze Li, Kalyan Basu, and Gabriel A Silva. Neural networks and markov categories. *AppliedMath*, 5(3):93, 2025.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. *arXiv preprint arXiv:2501.00070*, 2024.

Paolo Perrone. Markov categories and entropy. *IEEE Transactions on Information Theory*, 70(3): 1671–1692, 2023a.

Paolo Perrone. Categorical information geometry. In *International Conference on Geometric Science of Information*, pages 268–277. Springer, 2023b.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay Cohen. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. Contrastive Learning Is Spectral Clustering On Similarity Graph. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

Yang Yuan. On the power of foundation models. In *International Conference on Machine Learning*, pages 40519–40530. PMLR, 2023.

Appendix

A Full Proofs of Theorems

A.1 Proof of Theorem 2.7 (NLL Minimization as Average KL Minimization)

Let $p_x(\cdot) \coloneqq k_{\text{data}}(x, \cdot)$ denote the true conditional probability distribution $P_{\text{data}}(\cdot|x)$ for context $x = \mathbf{w}_{<t}$. Let $q_{x,\theta}(\cdot) = k_{\text{gen},\theta}(x, \cdot)$ denote the model's conditional probability distribution $P_\theta(\cdot|x)$. The context distribution is $p_{W_{<t}}$.

The cross-entropy loss is defined as:

$$\begin{aligned} L_{\text{CE}}(\theta) &= -\mathbb{E}_{(x,w)\sim P_{\text{data}}}[\log q_{x,\theta}(w)] \\ &= -\mathbb{E}_{x\sim p_{W_{<t}}}\left[\mathbb{E}_{W\sim p_x(\cdot)}[\log q_{x,\theta}(W)]\right] \\ &= -\mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log q_{x,\theta}(w)\right] \end{aligned}$$

The average KL divergence is defined as:

$$\begin{aligned} \mathcal{L}_{\text{KL}}(\theta) &= \mathbb{E}_{x\sim p_{W_{<t}}}[D_{\text{KL}}(p_x(\cdot)\,\|\,q_{x,\theta}(\cdot))] \\ &= \mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log\frac{p_x(w)}{q_{x,\theta}(w)}\right] \\ &= \mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log p_x(w) - \sum_{w\in\mathbb{V}} p_x(w)\log q_{x,\theta}(w)\right] \\ &= \mathbb{E}_{x\sim p_{W_{<t}}}[-H(p_x(\cdot))] - \mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log q_{x,\theta}(w)\right] \\ &= -H(W_t|W_{<t})_{\text{data}} + L_{\text{CE}}(\theta) \end{aligned}$$

where $H(p_x(\cdot))$ is the Shannon entropy of the distribution $p_x(\cdot)$, and $H(W_t|W_{<t})_{\text{data}} = \mathbb{E}_{x\sim p_{W_{<t}}}[H(p_x(\cdot))]$ is the average conditional Shannon entropy of the data generating process.

Rearranging gives:

$$L_{\text{CE}}(\theta) = \mathcal{L}_{\text{KL}}(\theta) + H(W_t|W_{<t})_{\text{data}}$$

Since $H(W_t|W_{<t})_{\text{data}}$ is a property of the data distribution and does not depend on the model parameters θ, minimizing $L_{\text{CE}}(\theta)$ with respect to θ is equivalent to minimizing $\mathcal{L}_{\text{KL}}(\theta)$.

The KL divergence $D_{\text{KL}}(p\|q) \geq 0$ for any probability distributions p, q, with equality if and only if $p = q$. Therefore, the average KL divergence $\mathcal{L}_{\text{KL}}(\theta) = \mathbb{E}_{x\sim p_{W_{<t}}}[D_{\text{KL}}(p_x(\cdot)\,\|\,q_{x,\theta}(\cdot))]$ is also non-negative, as it is an expectation of non-negative values. The minimum value is achieved if and only if $k_{\text{data}}(x, \cdot) = k_{\text{gen},\theta^*}(x, \cdot)$ for $p_{W_{<t}}$-almost every x. □

A.2 Proof of Proposition 4.4 (Categorical Entropy equals Shannon Entropy)

The pointwise categorical entropy for a single input h and $D = D_{\text{KL}}$ is:

$$\mathcal{H}^{\text{cat}}_{D_{\text{KL}}}(k_{\text{head}})(h) = D_{\mathbb{V}\otimes\mathbb{V}}(\Delta_{\mathbb{V}} \circ k_{\text{head}}(h, \cdot) \parallel (k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)))\,.$$

Let $p_h(\cdot) = k_{\text{head}}(h, \cdot)$. The first distribution is $\sum_{w \in \mathbb{V}} p_h(w)\delta_{(w,w)}$, which is supported on the diagonal of $\mathbb{V} \times \mathbb{V}$. The second is the product distribution $p_h \otimes p_h$. The KL divergence is:

$$\begin{aligned}
&\sum_{(w,w') \in \mathbb{V}\times\mathbb{V}} \left(\sum_{v \in \mathbb{V}} p_h(v)\delta_{(v,v)}(w, w') \right) \log \left(\frac{\sum_{v \in \mathbb{V}} p_h(v)\delta_{(v,v)}(w, w')}{p_h(w)p_h(w')} \right) \\
&= \sum_{w \in \mathbb{V}} p_h(w) \log \left(\frac{p_h(w)}{p_h(w)p_h(w)} \right) \\
&= \sum_{w \in \mathbb{V}} p_h(w) \log \left(\frac{1}{p_h(w)} \right) = -\sum_{w \in \mathbb{V}} p_h(w) \log p_h(w) = H(p_h).
\end{aligned}$$

The averaged categorical entropy is then $\overline{\mathcal{H}}^{\text{cat}}_{D_{\text{KL}}}(k_{\text{head}}; p_{H_t}) = \mathbb{E}_{h \sim p_{H_t}}[H(p_h)]$, which is precisely the definition of the conditional Shannon entropy $H(W_t \mid H_t)$. □

A.3 Information Surplus Identity (Equation 4.1)

The chain rule for mutual information states that for random variables A, B, C:

$$I(A; B, C) = I(A; B) + I(A; C \mid B).$$

Let $A = H_t$, $B = W_t$, and $C = W_{t+1:t+K-1}$. Substituting these into the chain rule gives:

$$I(H_t; W_t, W_{t+1:t+K-1}) = I(H_t; W_t) + I(H_t; W_{t+1:t+K-1} \mid W_t).$$

Recognizing that $(W_t, W_{t+1:t+K-1})$ is the sequence $W_{t:t+K-1}$, we have:

$$I(H_t; W_{t:t+K-1}) = I(H_t; W_t) + I(H_t; W_{t+1:t+K-1} \mid W_t).$$

This directly yields the identity in Equation 4.1, where the information surplus is identified as the conditional mutual information term. □

A.4 Proof of Lemma 5.1 and Theorem 5.2

We give short self-contained proofs. Throughout, $\mathbb{V}$ is finite so $\Delta^{|\mathbb{V}|-1}$ is compact and d_H is a metric on the simplex.

Proof of Lemma 5.1. Since Ψ_D is continuous on the compact simplex, it is uniformly continuous and bounded; write $\|\Psi_D\|_\infty < \infty$. Fix $\varepsilon > 0$ and choose $\delta > 0$ such that $d_H(p, q) \le \delta$ implies $|\Psi_D(p) - \Psi_D(q)| \le \varepsilon$. Then for any n,

$$\mathbb{E}\big|\Psi_D(P^{(n)}) - \Psi_D(P^\star)\big| \le \varepsilon + 2\|\Psi_D\|_\infty \, \mathbb{P}\big(d_H(P^{(n)}, P^\star) > \delta\big) \le \varepsilon + \frac{2\|\Psi_D\|_\infty}{\delta^2} \, \mathbb{E}\big[d_H(P^{(n)}, P^\star)^2\big],$$

where the last step is Markov. Taking $n \to \infty$ and using $\mathbb{E}[d_H^2] \to 0$ yields the claim.

Proof of Theorem 5.2. Let $X \sim p_{W_{<t}}$, $p_X = k_{\text{data}}(X, \cdot)$, and $q_X^{(n)} = k_{\text{gen},\theta_n}(X, \cdot)$. On a finite alphabet, $d_H^2(p, q) \le \frac{1}{2} D_{\text{KL}}(p \| q)$, so

$$\mathbb{E}\big[d_H(p_X, q_X^{(n)})^2\big] \le \tfrac{1}{2}\,\mathbb{E}\big[D_{\text{KL}}(p_X \| q_X^{(n)})\big] = \tfrac{1}{2}\,\mathcal{L}_{\text{KL}}(\theta_n) \xrightarrow[n\to\infty]{} 0.$$

Apply Lemma 5.1 with $P^{(n)} = q_X^{(n)}$ and $P^\star = p_X$ to get $\mathbb{E}[\Psi_D(q_X^{(n)})] \to \mathbb{E}[\Psi_D(p_X)]$. Finally, since $H_t^{(\theta_n)} = f_{\text{enc},\theta_n}(X)$ and $k_{\text{head},\theta_n}(H_t^{(\theta_n)}, \cdot) = k_{\text{gen},\theta_n}(X, \cdot)$,

$$\mathbb{E}[\Psi_D(q_X^{(n)})] = \overline{\mathcal{H}}_D^{\text{cat}}(k_{\text{head},\theta_n};\, p_{H_t,\theta_n}),$$

which proves the theorem. For $D = D_{\text{KL}}$, $\Psi_{D_{\text{KL}}}(p) = H(p)$ (Proposition 4.4), yielding the stated identification with $H(W_t \mid W_{<t})_{\text{data}}$. At a realizable optimum, $k_{\text{data}}(x, \cdot) = k_{\text{head},\theta^\star}(f_{\text{enc},\theta^\star}(x), \cdot)$ depends on x only via $H = f_{\text{enc},\theta^\star}(x)$, i.e., the data kernel factors through H, so H is predictively sufficient. □

A.5 Proof of Theorem 7.1 (Output Distribution Approximation Constraint)

Let $p_x^{\text{data}} = P_{\text{data}}(\cdot|x)$ and $p_x^\theta = p_\theta(\cdot|x)$. By the Hellinger–KL inequality under our convention,

$$d_H(p_x^{\text{data}}, p_x^\theta)^2 \le \frac{1}{2} D_{\text{KL}}(p_x^{\text{data}} \| p_x^\theta).$$

Taking expectation over $x \sim \mu_{\text{ctx}}$ gives

$$\mathbb{E}_x d_H(p_x^{\text{data}}, p_x^\theta)^2 \le \frac{1}{2} \mathcal{L}_{\text{KL}}(\theta).$$

For any pair x, x', the triangle inequality gives

$$\big|d_H(p_x^\theta, p_{x'}^\theta) - d_H(p_x^{\text{data}}, p_{x'}^{\text{data}})\big| \le d_H(p_x^\theta, p_x^{\text{data}}) + d_H(p_{x'}^\theta, p_{x'}^{\text{data}}).$$

Writing $\epsilon_x = d_H(p_x^\theta, p_x^{\text{data}})$, Markov's inequality gives

$$\mathbb{P}(\epsilon_X \ge \delta) \le \frac{\mathbb{E}[\epsilon_X^2]}{\delta^2} \le \frac{\mathcal{L}_{\text{KL}}(\theta)}{2\delta^2}.$$

For independent X, X', a union bound gives

$$\mathbb{P}(\epsilon_X + \epsilon_{X'} > 2\delta) \le \frac{\mathcal{L}_{\text{KL}}(\theta)}{\delta^2}.$$

Combining the last two displays proves the random-pair statement in the theorem. □

A.6 Proof of Corollary 7.2 (Implicit Representation Separation)

From Theorem 7.1, if the model fits the data well, then for predictively dissimilar contexts x, x', the distance between their model output distributions, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$, must be large.

The interior of the output simplex $\mathcal{P}(\mathbb{V})$ is a Riemannian manifold endowed with the Fisher-Rao metric g^{FR}. The distance between two interior points, p_1 and p_2, is the infimum of the lengths of all smooth paths connecting them. The length of a path $\gamma : [0, 1] \to \mathcal{P}(\mathbb{V})$ is given by the integral:

$$L(\gamma) = \int_0^1 \sqrt{g^{\text{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t))}\, dt.$$

The mapping $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ allows us to map paths from the representation space to the output space. Consider the straight-line path in representation space connecting $h_{x'}$ to h_x:

$$h(t) = (1-t)h_{x'} + th_x, \quad \text{for } t \in [0,1].$$

The tangent vector to this path is constant: $h'(t) = h_x - h_{x'}$. This path in $\mathcal{H}$ induces a corresponding path in $\mathcal{P}(\mathbb{V})$ given by $\gamma(t) = g_{\text{head}}(h(t))$. The tangent vector to this induced path is found using the chain rule:

$$\gamma'(t) = J_{g_{\text{head}}}(h(t)) \cdot h'(t) = J_{g_{\text{head}}}(h(t)) \cdot (h_x - h_{x'}),$$

where $J_{g_{\text{head}}}(h)$ is the Jacobian of the map g_{head} evaluated at h.

The squared length of this tangent vector at point $\gamma(t)$ is given by the quadratic form of the metric g^{FR}:

$$\begin{aligned} g^{\text{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t)) &= g^{\text{FR}}_{g_{\text{head}}(h(t))}(J_{g_{\text{head}}}(h(t))(h_x - h_{x'}), J_{g_{\text{head}}}(h(t))(h_x - h_{x'})) \\ &= (h_x - h_{x'})^\top \left[J_{g_{\text{head}}}(h(t))^\top g^{\text{FR}}_{g_{\text{head}}(h(t))} J_{g_{\text{head}}}(h(t)) \right] (h_x - h_{x'}). \end{aligned}$$

The term in the square brackets is precisely the definition of the pullback metric tensor g^* evaluated at the point $h(t) \in \mathcal{H}$. Thus, the squared length of the tangent vector is:

$$g^{\text{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t)) = g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'}).$$

The total length of this specific path in $\mathcal{P}(\mathbb{V})$ is therefore:

$$L(\gamma) = \int_0^1 \sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})}\, dt.$$

The Riemannian distance $d_{\text{FR}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ is the infimum of path lengths, so it is bounded above by the length of our chosen path:

$$d_{\text{FR}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) \le \int_0^1 \sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})}\, dt.$$

By Cauchy–Schwarz,

$$d_{\text{FR}}\big(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})\big)^2 \le \int_0^1 g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})\, dt.$$

On the simplex, d_H and d_{FR} are related exactly by

$$d_H(p,q) = \sqrt{2}\sin\big(d_{\text{FR}}(p,q)/4\big),$$

so large Hellinger separation of the model outputs implies large Fisher–Rao separation. Together with Theorem 7.1, predictively dissimilar data conditionals therefore force the integrated pullback quadratic form above to be large on typical well-fit pairs. Equivalently, the difference vector $v = h_x - h_{x'}$ must be visible along directions where the head has nontrivial predictive sensitivity along the path.

Conversely, if contexts are predictively similar and the model fits them well, then the output separation is small. This permits representations to differ in directions invisible to the head, but it discourages unnecessary separation along directions where g^* is large. □

A.7 Proof of Proposition 7.5 (NLL Objective and Implicit Dirichlet Energy Minimization)

The Dirichlet energy is

$$\mathcal{E}_K(\phi_v) = \frac{1}{2}\iint K(x,x')\langle h_x - h_{x'}, v\rangle^2 \,\mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x').$$

By Cauchy–Schwarz,

$$\mathcal{E}_K(\phi_v) \le \frac{1}{2}\|v\|^2 \iint K(x,x')\|h_x - h_{x'}\|^2 \,\mathrm{d}\mu(x)\mathrm{d}\mu(x').$$

The local lower Lipschitz condition and $d_{\mathrm{FR}}(p,q) \le \pi\sqrt{2}\, d_H(p,q)$ imply

$$\|h_x - h_{x'}\|^2 \le \frac{2\pi^2}{m^2}\, d_H(p_\theta(\cdot|x), p_\theta(\cdot|x'))^2.$$

This proves the first inequality in Proposition 7.5. For the second, write

$$e_x = d_H(P_{\mathrm{data}}(\cdot|x), p_\theta(\cdot|x)).$$

The triangle inequality gives

$$d_H(p_\theta(\cdot|x), p_\theta(\cdot|x')) \le d_H(P_{\mathrm{data}}(\cdot|x), P_{\mathrm{data}}(\cdot|x')) + e_x + e_{x'}.$$

Squaring and using $(a+b+c)^2 \le 3a^2 + 6b^2 + 6c^2$, then integrating against bounded K, yields

$$\iint K\, d_H(p_\theta^x, p_\theta^{x'})^2 \le 3\iint K\, d_H(P_{\mathrm{data}}^x, P_{\mathrm{data}}^{x'})^2 + 12\|K\|_\infty\, \mathbb{E}[e_X^2].$$

Since $\mathbb{E}[e_X^2] = \varepsilon$, the stated bound follows. □

A.8 Well-posedness of the Predictive Similarity Operator $M_{\widetilde{K}}$

The operator $M_{\widetilde{K}} : L^2(\mathcal{H},\mu) \to L^2(\mathcal{H},\mu)$ is defined by the integral $(M_{\widetilde{K}}\psi)(h) = \int_{\mathcal{H}} \widetilde{K}(h,h')\psi(h')\mu(\mathrm{d}h')$. For $M_{\widetilde{K}}$ to be a compact self-adjoint operator, its kernel $\widetilde{K}(h,h')$ must satisfy certain conditions.

1. **Symmetry:** Since the original kernel $K(x,x')$ is assumed to be symmetric, the disintegrated kernel $\widetilde{K}(h,h') = \mathbb{E}[K(X,X') \mid f(X) = h, f(X') = h']$ is also symmetric.
2. **Square-Integrability:** The space $(\mathcal{H},\mu)$ is a probability space. A sufficient condition for compactness on a probability space is that the kernel is square-integrable, i.e., $\iint |\widetilde{K}(h,h')|^2\mu(\mathrm{d}h)\mu(\mathrm{d}h') < \infty$. As we assumed K is a bounded function, its conditional expectation $\widetilde{K}$ is also bounded. A bounded measurable function on a finite measure space is always square-integrable.

Since $\widetilde{K}$ is symmetric and square-integrable with respect to the probability measure μ, it is a Hilbert-Schmidt kernel. Every Hilbert-Schmidt integral operator is compact. Because the kernel is also real and symmetric, the operator is self-adjoint. By the spectral theorem for compact self-adjoint operators, $M_{\widetilde{K}}$ has a real point spectrum away from zero with possible accumulation only at zero; including the zero-eigenspace, one obtains an orthonormal eigenbasis for $L^2(\mathcal{H},\mu)$. □

If, moreover, K is PSD in the quadratic-form sense, then $\widetilde{K}$ is PSD by Lemma 7.4. Hence for every $\psi \in L^2(\mathcal{H},\mu)$ we have $\langle \psi, M_{\widetilde{K}}\psi\rangle \ge 0$. Thus $M_{\widetilde{K}}$ is positive semidefinite and its spectrum lies in $[0,\infty)$.

A.9 Proof of Theorems 7.8 and 7.10

The proof separates the analytic surrogate from the CCA algebra.

Proof. For Theorem 7.8, the linear-softmax likelihood has

$$-\log p_\theta(W|X) = -\langle g(W), h_X\rangle - b_W + \log \sum_w \exp\{\langle g(w), h_X\rangle + b_w\}.$$

Young's inequality with parameter τ gives

$$\langle g(w), h_X\rangle \le \frac{1}{2\tau}\|g(w)\|^2 + \frac{\tau}{2}\|h_X\|^2.$$

Substitution into the log-partition function gives

$$\log \sum_w e^{\langle g(w), h_X\rangle + b_w} \le C_\tau + \frac{\tau}{2}\|h_X\|^2, \quad C_\tau = \log \sum_w e^{b_w + \|g(w)\|^2/(2\tau)}.$$

Taking conditional expectation over $W|X$ yields

$$\mathbb{E}[-\log p_\theta(W|X)] \le C_\tau - \mathbb{E}[b_W] - \mathbb{E}\langle \bar{g}_X, h_X\rangle + \frac{\tau}{2}\mathbb{E}\|h_X\|^2.$$

Completing the square gives

$$\frac{\tau}{2}\mathbb{E}\left\|h_X - \frac{1}{\tau}\bar{g}_X\right\|^2 - \frac{1}{2\tau}\mathbb{E}\|\bar{g}_X\|^2 + C_\tau - \mathbb{E}[b_W].$$

This proves the surrogate statement.

For Theorem 7.10, assume h_X and $\bar{g}_X$ are centered. The regression surrogate above by itself does not identify a spectral basis; the CCA basis appears after imposing variance normalization on linear functionals of h_X and $\bar{g}_X$. For fixed u, maximizing

$$\frac{u^\top C_{h\bar{g}} v}{\sqrt{u^\top C_{hh} u}\sqrt{v^\top C_{\bar{g}\bar{g}} v}}$$

over v gives $v \propto C_{\bar{g}\bar{g}}^{-1} C_{h\bar{g}}^\top u$ and value squared

$$\frac{u^\top C_{h\bar{g}} C_{\bar{g}\bar{g}}^{-1} C_{h\bar{g}}^\top u}{u^\top C_{hh} u}.$$

The stationary points of this Rayleigh quotient satisfy

$$C_{h\bar{g}} C_{\bar{g}\bar{g}}^{-1} C_{h\bar{g}}^\top u = \lambda C_{hh} u,$$

and the top constrained directions are the top generalized eigenvectors by the Courant–Fischer variational principle. In whitened coordinates $C_{hh} = I$, this is the ordinary symmetric eigenproblem stated in the main text. The scalar τ from the surrogate rescales the prototype side and cancels from these normalized CCA directions. □