Next-Token Prediction as Implicit Spectral Contrastive Learning

Anonymous ACL submission

Abstract

Why does simply predicting the next token create such powerful language models? This paper provides a formal answer by introducing a compositional framework that treats autoregressive generation as a flow of information through probabilistic kernels. Using this lens, we prove that the next-token prediction objective is mathematically equivalent to implicit spectral contrastive learning. This single training principle forces the model to organize its internal representations into a geometrically structured space aligned with dominant patterns of predictive similarity in the data—all without needing explicit contrastive pairs. Our framework yields two further results: it offers a formal information-theoretic justification for modern speculative decoding by quantifying the information surplus in hidden states, and it shows how models are compelled to capture the underlying uncertainty of language.

1 Introduction

Autoregressive language models (AR LMs), such as Transformers (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020), have achieved remarkable success by sequentially predicting the next token based on its context. They learn a parameterized distribution P_θ that factorizes as:

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

where $\mathbf{w}_{<t}$ is the context sequence, and parameters θ are optimized via negative log-likelihood (NLL). Despite empirical triumphs, a deep theoretical understanding of their internal mechanisms remains fragmented (Manning et al., 2020; Elhage et al., 2021). Key open questions include: why does the simple NLL objective yield such powerful representations, and what is the theoretical basis for the success of efficiency techniques like speculative decoding (Li et al., 2024)?

This paper introduces a framework using Markov Categories (MCs) (Cho and Jacobs, 2019; Fritz, 2020) to address these questions. We model the single-step generation map $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ as a formal composition of Markov kernels in the category `Stoch`:

$$k_{\text{gen},\theta} := k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}. \tag{1.2}$$

Here, k_{emb} and k_{bb} are deterministic kernels for context encoding, producing a hidden state $h_t \in \mathcal{H}$, while k_{head} is the stochastic kernel mapping h_t to a predictive distribution. Enriching this model with statistical divergences (e.g., D_{KL}) allows for a geometric and information-theoretic analysis (Perrone, 2023b). Our contributions are:

1. A Formal Compositional Model (§3). We model the AR generation step as a composite kernel, providing a formal language to reason about information flow. 2. Rationale for Speculative Decoding (§4.1). We use the Data Processing Inequality to quantify the information surplus in hidden states that methods like EAGLE (Li et al., 2024) and multi-token prediction (MTP Gloeckle et al. (2024)) exploit. 3. A Unified View of the NLL Objective (§5, §7). We show that NLL acts as an implicit spectral contrastive objective. We prove that minimizing NLL forces the model to learn a geometrically structured representation space by aligning representations with the eigenspectrum of a predictive similarity operator. This process simultaneously compels the model to learn the data's intrinsic conditional stochasticity.

This synthesis of compositional probability and information geometry provides a foundation for analyzing modern LMs. It explains how a simple training objective gives rise to complex, structured representations and offers a principled path toward understanding emergent capabilities and improving model efficiency.

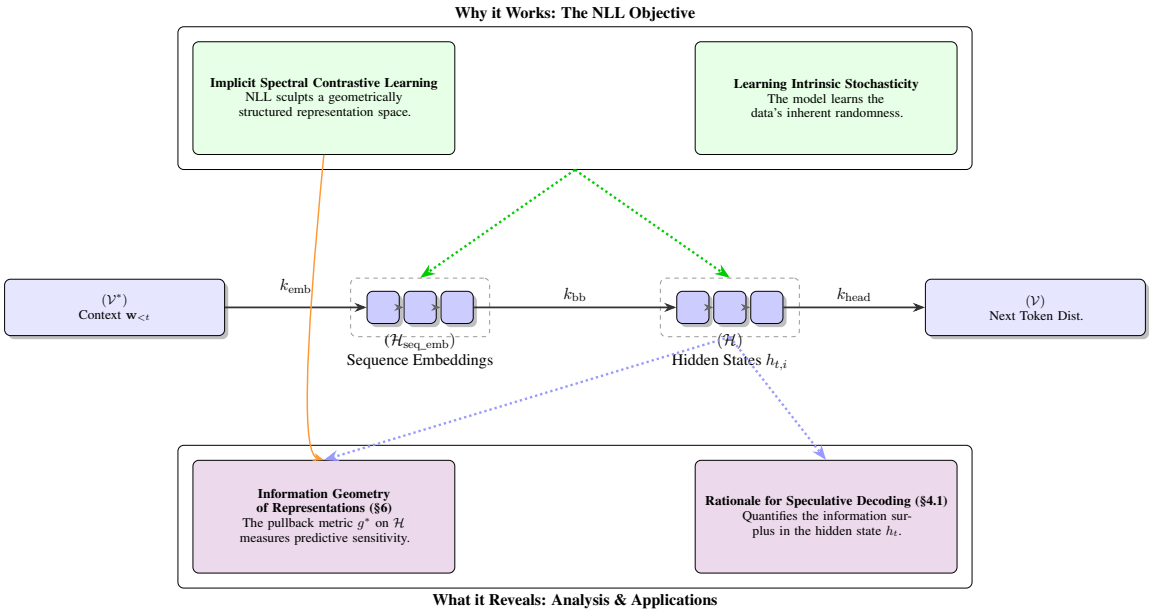

Figure 1: A conceptual overview of our framework. **Center:** The core thesis models the Autoregressive generation step as a composition of Markov kernels $k_{\text{gen}} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in the category `Stoch`. This separates the deterministic context encoding $(k_{\text{emb}}, k_{\text{bb}})$ from the stochastic prediction (k_{head}). **Top:** This compositional lens reveals the deeper mechanisms of the Negative Log-Likelihood (NLL) objective, which we re-frame as minimizing the average KL divergence between the model and true data kernels. We prove that this single objective implicitly forces the model to learn two key properties of the data: its intrinsic conditional stochasticity (measured via categorical entropy) and its underlying geometric structure, which we show is equivalent to performing spectral contrastive learning on a predictive similarity graph. **Bottom:** The framework yields new analytical tools. By pulling back the Fisher-Rao metric, we endow the representation space $\mathcal{H}$ with an information geometry that quantifies predictive sensitivity. This provides a formal basis for analyzing the information surplus in hidden states, giving a theoretical foundation for modern speculative decoding methods.

2 Background: Markov Categories and Information Geometry

Our framework models stochastic processes as morphisms in a *Markov Category* (Fritz, 2020). We focus on the category `Stoch`, whose objects are standard Borel spaces (e.g., a vocabulary $\mathbb{V}$ or representation space $\mathbb{R}^d$) and whose morphisms are Markov kernels.

Definition 2.1 (Markov Category (Fritz, 2020))**.** A Markov category $(\mathcal{C}, \otimes, I)$ is a symmetric monoidal category where each object X is equipped with a commutative comonoid structure $(\Delta_X : X \to X \otimes X, !_X : X \to I)$ that is natural in X, and the monoidal unit I is a terminal object (the *causality* axiom: $!_X$ is the unique map $X \to I$).

A kernel $k : X \to Y$ maps each input $x \in X$ to a probability distribution over Y. This provides a compositional language for probability, where composing kernels corresponds to chaining stochastic processes. When enriched with a statistical divergence D (e.g., D_{KL}), this framework provides categorical definitions of entropy and mutual information that inherently satisfy the Data Processing Inequality (DPI), making it ideal for analyzing information flow (Perrone, 2023b). A detailed introduction is provided in Appendix A.

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 `Stoch`.

The relevant measurable spaces (objects in `Stoch`) are: the input context space $(\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$, the hidden state space $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$ (typically $(\mathbb{R}^{d_{\text{model}}}, \mathscr{B}(\mathbb{R}^{d_{\text{model}}}))$), and the output vocabulary space $(\mathbb{V}, \mathcal{P}(\mathbb{V}))$. We choose standard Borel spaces as they form a well-behaved class of measurable spaces ensuring measure-theoretic regularity (Kallenberg and Kallenberg, 1997).

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

1. Embedding Kernel (k_{emb}): This deterministic kernel maps the discrete input context $\mathbf{w}_{<t}$ to a sequence of initial vector representations. It is defined via a Dirac measure on the output of the embedding function.

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

2. Backbone Kernel (k_{bb}): This deterministic kernel represents the core computation (e.g., a Transformer stack) that maps the sequence of initial representations to the final hidden state $h_t \in \mathcal{H}$.

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

3. LM Head Kernel (k_{head}): This is the only stochastic kernel in the standard architecture. It maps the final hidden state h_t to a probability distribution over the vocabulary $\mathbb{V}$, typically via a linear layer followed by a softmax function.

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

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 $p_{W_{<t}} : I \to (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$. Processing this state through the composed kernels induces distributions over the hidden state H_t (state $p_{H_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$) and the next token W_t (state $p_{W_t} : I \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$).

4.1 Information Flow Bounds and Rationale for Speculative Decoding

The Data Processing Inequality (DPI) is a native property of our framework and offers a formal explanation for the success of techniques like speculative decoding (Li et al., 2024) and multi-token prediction (MTP,Gloeckle et al. (2024)). These methods hypothesize that the hidden state H_t contains information to predict a sequence of future tokens, not just the single next token W_t.

Let $W_{t:t+K-1} = (W_t, W_{t+1}, \ldots, W_{t+K-1})$ be the sequence of K tokens following the context $W_{<t}$. The generation of this sequence from the hidden state H_t forms a Markov chain: $H_t \to W_{t:t+K-1} \to W_t$. Applying the Data Processing Inequality for categorical mutual information I_D to this processing step directly yields a chain of inequalities:

$$I_D(H_t; W_{t:t+K-1}) \geq I_D(H_t; W_t). \quad (4.1)$$

The empirical success of speculative decoding is explained by the existence of a significant information surplus, which we define as:

$$\text{Information Surplus} := I_D(H_t; W_{t:t+K-1}) - I_D(H_t; W_t) \geq 0. \quad (4.2)$$

This surplus represents the information about future tokens $(W_{t+1}, \ldots, W_{t+K-1})$ that is present in the state H_t but ignored by the standard single-token prediction head k_{head}.

Our framework can model speculative decoding explicitly. A lightweight "drafting" kernel $k_{\text{draft}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}^K, \mathcal{P}(\mathbb{V}^K))$ operates in parallel with the main "verification" head $k_{\text{verify}} \equiv k_{\text{head}}$. Both take H_t as input:

$$\begin{tikzcd} \mathbb{V}^* \arrow[r, "k_{\text{enc}} := k_{\text{bb}} \circ k_{\text{emb}}"] & \mathcal{H} \arrow[r, "k_{\text{verify}}"] \arrow[d, "k_{\text{draft}}"'] & \mathbb{V} \\ & \mathbb{V}^K & \end{tikzcd}$$

The key insight is that our framework moves the understanding of speculative decoding from an empirical heuristic to a principled consequence of information theory. The information surplus is a formal, quantifiable resource that can be measured and, in principle, optimized (Gloeckle et al., 2024).

4.2 Metric: LM Head Categorical Entropy (Prediction Stochasticity)

To 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 4.1 (Categorical Entropy of k_{head})**.** The Categorical Entropy of k_{head} is defined with $X = \mathcal{H}$, $Y = \mathbb{V}$, and $k = k_{\text{head}}$:

$$\mathcal{H}_D(k_{\text{head}}) : \mathcal{H} \to \mathbb{R}_{\geq 0} := D_{\mathbb{V}\otimes\mathbb{V}}\left(\Delta_{\mathbb{V}} \circ k_{\text{head}} \quad \| \quad (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}\right). \tag{4.3}$$

The categorical entropy $\mathcal{H}_D(k_{\text{head}})(h)$ 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}:

$$\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t}) := \mathbb{E}_{h\sim p_{H_t}}\Big[D_{\mathbb{V}\otimes\mathbb{V}}\Big(\sum_{w\in\mathbb{V}} k_{\text{head}}(h, \{w\})\delta_{(w,w)} \| k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)\Big)\Big]. \tag{4.4}$$

Interpretation. This metric measures the average intrinsic conditional stochasticity of the LM head. A value of 0 indicates a deterministic mapping, while higher values indicate greater average uncertainty in the output distribution for a given hidden state.

For the specific case $D = D_{\text{KL}}$, the average categorical entropy $\bar{\mathcal{H}}_{D_{\text{KL}}}(k_{\text{head}}; p_{H_t})$ is closely related to the average conditional Shannon entropy $H(W_t|H_t)$:

$$\mathbb{E}_{h\sim p_{H_t}}[H(k_{\text{head}}(h, \cdot))] = \mathbb{E}_{h\sim p_{H_t}}\left[-\sum_{w\in\mathbb{V}} p_h(w)\log p_h(w)\right]. \tag{4.5}$$

This captures the average uncertainty in the next-token prediction given the hidden state. As we show next, Negative Log-Likelihood (NLL) minimization implicitly forces the model to match this value to the data's intrinsic stochasticity.

5 Pretraining Objective, Compression, and Learning Intrinsic Stochasticity

Let $k_{\text{data}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the (potentially unknown) Markov kernel representing the true data-generating process, such that $k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ corresponds to the true conditional probability measure $P_{\text{data}}(\cdot|\mathbf{w}_{<t})$ 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 is to minimize the negative log-likelihood (NLL), which is equivalent to minimizing the cross-entropy:

$$L_{\text{CE}}(\theta) = -\mathbb{E}_{(\mathbf{w}_{<t}, w_t)\sim P_{\text{data}}}[\log P_\theta(w_t|\mathbf{w}_{<t})] \tag{5.1}$$

where $P_\theta(w_t|\mathbf{w}_{<t})$ is the model's predicted probability. Let $k_{\text{gen},\theta} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the model's overall generation kernel, $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$, such that $P_\theta(\cdot|\mathbf{w}_{<t}) = k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot)$. The objective can be precisely stated in terms of Kullback-Leibler (KL) divergence between the data kernel and the model kernel.

Theorem 5.1 (NLL Minimization as Average KL Minimization)**.** Minimizing the cross-entropy loss $L_{\text{CE}}(\theta)$ (equation (5.1)) is equivalent to minimizing the average KL divergence between the true data kernel k_{data} and the model kernel $k_{\text{gen},\theta}$:

$$\operatorname*{argmin}_\theta L_{\text{CE}}(\theta) = \operatorname*{argmin}_\theta \mathcal{L}_{\text{KL}}(\theta) := \operatorname*{argmin}_\theta \mathbb{E}_{\mathbf{w}_{<t}\sim p_{W_{<t}}}[D_{\text{KL}}(k_{\text{data}}(\mathbf{w}_{<t}, \cdot) \| k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot))] \tag{5.2}$$

Proof Sketch. The equivalence arises from rewriting cross-entropy as $L_{\text{CE}}(\theta) = \mathcal{L}_{\text{KL}}(\theta) + H(W_t|W_{<t})_{\text{data}}$, where the conditional entropy

of the data is independent of θ. See Appendix B.1. □

This frames NLL training as driving the model kernel to match the data kernel, which, by Shannon's source coding theorem, is equivalent to finding a model that provides the most efficient compression of the training data. This connection provides a formal grounding for the "compression leads to intelligence" hypothesis.

Beyond matching predictive distributions, NLL training forces the model to learn the data's intrinsic conditional stochasticity. We formalize this in the following theorem, which shows that as the model converges, the average categorical entropy of its prediction head must converge to that of the data's true conditional distribution.

Theorem 5.2 (Convergence of Average Categorical Entropy, Informal)**.** If a sequence of models θ_n converges to an optimal model θ^* by minimizing the NLL objective, then under standard regularity conditions, the average categorical entropy of the model's prediction head $\bar{\mathcal{H}}_D(k_{\text{head},\theta_n})$ converges to that of the optimal model's head $\bar{\mathcal{H}}_D(k_{\text{head},\theta^*})$. If the model is sufficiently expressive to perfectly model the data, this value matches the intrinsic conditional entropy of the data-generating process itself.

Proof Sketch. The proof, detailed in Appendix B.2 along with the formal statement, relies on the Continuous Mapping Theorem. Convergence in the NLL objective implies weak convergence of the induced hidden state distributions and pointwise convergence of the head kernel's outputs, which in turn implies convergence of the expected categorical entropy. □

Theorem 5.2 provides a formal basis for the claim that NLL training compels the model to learn not just the most likely next token, but also the degree of uncertainty associated with that prediction. The model must learn the shape of the data's conditional distributions, a property quantified by categorical entropy.

6 Information Geometry of Representation and Prediction Spaces

The Markov Category framework, particularly $(\texttt{Stoch}, D)$ enriched with a divergence like D_{KL}, provides a natural bridge to Information Geometry (Amari and Nagaoka, 2000; Perrone, 2023a). 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 $\mathcal{P}(\mathbb{V})$. 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}, which quantifies the local distinguishability between nearby probability distributions.

The LM Head kernel corresponds to a map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$. This map allows us to use the pullback operation to induce a metric tensor $g^* = g_{\text{head}}^* g^{\text{FR}}$ on the representation space $\mathcal{H}$. This pullback metric is central to our analysis. The significance of g^* is that it endows the abstract vector space of representations $\mathcal{H}$ with a geometry that is directly tied to the model's predictive function. It measures how sensitive the output distribution is to changes in the hidden state.

Theorem 6.1 (Pullback Metric and Local Divergence)**.** 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 $g^*(h)$:

$$D_{\text{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2}\epsilon^2 g^*(h)(v, v) + O(\epsilon^3) \tag{6.1}$$

where $g^*(h)(v, v) = v^\top g^*(h) v$.

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 $\mathcal{P}(\mathbb{V})$.

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 (5.2)) 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 repre-

sentation 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).

7.1 Constraint on Output Distribution Approximation

Minimizing the NLL loss (average KL divergence) forces the model's predicted distribution $p_\theta(\cdot|x)$ to closely approximate the true data distribution $P_{\text{data}}(\cdot|x)$. This is formalized by Pinsker's inequality.

Theorem 7.1 (Output Distribution Approximation Constraint)**.** If the model fits the data well ($\mathcal{L}(\theta)$ is small), the distance between the model's output distributions for any two contexts, $d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x'))$, must approximate the distance between the true target distributions, $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$. Specifically:

$$d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \approx d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')). \tag{7.1}$$

Proof Sketch. The result follows from Pinsker's inequality and the triangle inequality. If the model prediction $p_\theta(\cdot|x)$ is close to the data $P_{\text{data}}(\cdot|x)$ for most contexts, the distance between any two model predictions must be close to the distance between the corresponding data distributions. See Appendix B.3. □

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). If two contexts x, x' have predictively dissimilar target distributions, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, then their learned representations $h_x = f_{\text{enc}}(x)$ and $h_{x'} = f_{\text{enc}}(x')$ must differ in a way that is detectable by the head mapping. Specifically, the difference vector $h_x - h_{x'}$ must have significant components along directions v of high predictive sensitivity (i.e., where the pullback metric quadratic form $g^*(h)(v, v)$ is large).

$$d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')) \text{ large } \implies \int_0^1 g^*(h_t)(h_x - h_{x'}, h_x - h_{x'})dt \text{ is large,} \tag{7.2}$$

where the integral is over the path $h_t = (1-t)h_{x'} + th_x$. Conversely, if contexts x, x' are predictively similar ($d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small), the NLL objective encourages their representations to be close along these same sensitive directions.

Proof Sketch. A large distance between the target distributions forces a large distance between the model's output distributions in $\mathcal{P}(\mathbb{V})$. This distance in the output manifold, as measured by the Fisher-Rao metric, is related by equation (6.1) to the integrated length of the path in the representation space $\mathcal{H}$, as measured by the pullback metric g^*. Therefore, a large output distance implies a large "predictive path length" between the representations, forcing them apart along directions of high predictive sensitivity. The full proof is in Appendix B.4. □

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 $p_x := P_{\text{data}}(\cdot|x)$. A predictive similarity kernel is a function $K : \mathbb{V}^* \times \mathbb{V}^* \to \mathbb{R}_{\geq 0}$ quantifying the similarity between p_x and $p_{x'}$, such as the *Bhattacharyya Coefficient Kernel* $K_{\text{BC}}(x, x') = \sum_w \sqrt{p_x(w)p_{x'}(w)}$ or the *Expected Likelihood Kernel* $K_{\text{Lin}}(x, x') = \sum_w p_x(w)p_{x'}(w)$. High values of $K(x, x')$ indicate high predictive similarity.

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 symmetric predictive similarity kernel $K(x, x')$. The quadratic form of the associated graph Laplacian corresponds to the Dirichlet energy.

Definition 7.4 (Graph Laplacian and Dirichlet Energy)**.** Let $K(x, x')$ be a predictive similarity kernel. The graph Laplacian Δ_K acts on $\phi : \mathbb{V}^* \to \mathbb{R}$. Its quadratic form is the Dirichlet energy $\mathcal{E}_K(\phi) = \frac{1}{2} \iint K(x, x')(\phi(x) - \phi(x'))^2 d\mu_{ctx} d\mu_{ctx}$, measuring smoothness over the similarity graph.

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 Corollary 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 (NLL and Implicit Dirichlet Energy Minimization)**.** Let $\phi_v(x) = \langle h_x, v\rangle$ be the projection of the representation $h_x = f_{\text{enc}}(x)$ onto a direction v of high predictive sensitivity (where $g^*(h)(v, v)$ is large). Minimizing the NLL loss $\mathcal{L}(\theta)$ implicitly encourages configurations where the Dirichlet energy $\mathcal{E}_K(\phi_v)$ is small.

Proof Sketch. The Dirichlet energy integrand is large only when predictively similar contexts ($K(x, x')$ is large) are mapped to distant representations ($\phi_v(x)$ and $\phi_v(x')$ differ significantly). Corollary 7.2 shows that NLL explicitly discourages this configuration. It encourages predictively similar contexts to have nearby representations (along sensitive directions), thereby minimizing the integrand. This aligns the NLL objective with Dirichlet energy minimization. The full proof is in Appendix B.5. □

This proposition formalizes the link: NLL pushes representations towards configurations favored by spectral clustering on the predictive similarity graph.

7.5 NLL as Spectral Objective

We now strengthen this connection, showing that NLL optimization is not merely analogous to spectral methods, but that it implicitly solves a spectral objective. This viewpoint is closer to the analysis in (Tan et al., 2024).

Definition 7.6 (Predictive Similarity Operator)**.** Let $f_{\text{enc}} : \mathbb{V}^* \to \mathcal{H}$ be a fixed encoder, inducing a distribution $\mu = p_{H_t}$ on $\mathcal{H}$ via μ_{ctx}. Let $K(x, x')$ be the predictive similarity kernel. The predictive similarity operator $M_K : L^2(\mathcal{H}, \mu) \to L^2(\mathcal{H}, \mu)$ is defined via its action on functions $\psi : \mathcal{H} \to \mathbb{R}$:

$$(M_K\psi)(h_x) \triangleq \mathbb{E}_{x'\sim\mu_{ctx}}[K(x, x')\psi(h_{x'})] = \int_{\mathbb{V}^*} K(x, x')\psi(f_{\text{enc}}(x'))\, \mu_{ctx}(\mathrm{d}x'). \tag{7.3}$$

If K is symmetric, M_K is a compact self-adjoint operator whose eigenfunctions capture dominant patterns of predictive similarity. Our key result is that NLL training implicitly aligns the representation geometry with the eigenspace of M_K.

Theorem 7.7 (NLL Objective as Implicit Spectral Alignment)**.** Let the representation covariance matrix be $\Sigma_H = \text{Cov}_{x\sim\mu_{ctx}}[h_x]$. Let the head-weighted covariance be $\Sigma_{H,g^*} = \mathbb{E}_{x\sim\mu_{ctx}}[h_x h_x^\top \otimes g^*(h_x)]$, which captures variance along predictively sensitive directions. Minimizing the NLL loss $\mathcal{L}(\theta)$ implicitly encourages a solution to an objective akin to maximizing the alignment between the representations and the similarity operator, subject to constraints on predictive distortion. It favors representations where directions of high variance align with the leading eigenfunctions of M_K, particularly after filtering for predictive relevance by g^*.

Proof Sketch. Under a standard log-linear model for the prediction head, the NLL loss decomposes into an alignment term (which encourages h_x to

align with the expected prototype of its target distribution) and a uniformity term. This structure implies that the inner product similarity of representations, $\langle h_x, h_{x'} \rangle$, is driven to match the inner product similarity of the expected prototypes. This is a spectral objective: it forces the covariance matrix of the representations to align with the eigenspectrum of the predictive similarity operator M_K, thereby learning a spectral embedding of the data's predictive similarity graph. A full derivation is provided in Appendix B.6. □

8 Related Work

The theoretical understanding of how autoregressive language models learn such effective representations remains a vibrant but fragmented area of research. Progress has occurred along several parallel fronts: the development of abstract languages for probabilistic modeling (Fritz, 2020; Cho and Jacobs, 2019), the application of geometric principles to learning (HaoChen et al., 2021; Tan et al., 2024; Park et al., 2024), and the analysis of the implicit biases of optimization (Gunasekar et al., 2018).

Our framework is built at the confluence of two modern perspectives on probability theory. The first is the *synthetic view*, which uses the language of category theory to provide a high-level, algebraic, and compositional foundation for probability (Fritz, 2020; Cho and Jacobs, 2019). In this approach, stochastic processes are modeled as morphisms in a *Markov category*, which allows for reasoning about information flow and system composition via universal axioms (Perrone, 2023a). While powerful for describing structure, this perspective is not inherently metric and does not natively describe the shape or geometry of representation spaces.

The second is the *metric view* of *information geometry*, which treats a parametric family of probability distributions as a Riemannian manifold endowed with the Fisher-Rao metric (Amari and Nagaoka, 2000). This provides a powerful geometric language for analyzing the local distinguishability of distributions and the dynamics of optimization in parameter space, but it lacks a native compositional structure for describing complex systems like LMs. Our work leverages recent efforts to bridge this gap (Perrone, 2023b), applying a synthesized categorical and geometric framework not just to abstract systems, but directly to the internal representation spaces of language models.

The *Information Bottleneck (IB)* principle posits that an optimal representation should efficiently compress the input while retaining maximal predictive information about the target (Tishby et al., 2000; Shwartz-Ziv and Tishby, 2017). While insightful, IB is a high-level objective and does not specify the mechanism by which this is achieved. A more mechanistic view comes from the literature on the *implicit bias* of optimization, which has recently established that training with the NLL objective functions as a powerful form of *implicit contrastive learning* (Gunasekar et al., 2018). Seminal works have proven that NLL implicitly solves a spectral objective, structuring the representation space by aligning it with the eigenspectrum of a data similarity operator, without requiring explicit negative samples (HaoChen et al., 2021; Tan et al., 2024).

9 Conclusion

We introduced a Markov categorical framework to analyze autoregressive language models. By modeling generation as a composition of kernels, we provided a formal information-theoretic rationale for speculative decoding. Most critically, we proved that the standard NLL objective functions as an implicit spectral contrastive learning algorithm. By analyzing the information geometry of the prediction head, we showed that NLL sculpts the representation space by aligning it with the eigenspectrum of a predictive similarity operator. This result offers an explanation for how the simple next-token prediction task learns such powerfully structured representations. Our synthesis of categorical probability and information geometry provides a new, rigorous lens for understanding the foundational principles of modern language models.

Limitations

This work provides a theoretical framework, with its primary focus on the mathematical modeling of the single-step generation process. Our analysis, by necessity, isolates the core mechanisms of learning. This naturally opens up several exciting avenues for future work, such as the direct empirical validation of the predicted geometric structures in large-scale models. Extending this compositional framework to model the emergent, long-range dynamics of multi-step generation also presents a promising

direction for a more complete theory of autoregressive modeling.

References

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits.

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

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

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*.

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

Jeff Z. HaoChen, Hao Chen, Chen Wei, Adrien Gaidon, and Tengyu Ma. 2021. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 14239–14250.

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

A Detailed Background on Markov Categories **11**

B Full Proofs of Theorems and Propositions **11**

B.1 Proof of Theorem 5.1 (NLL Minimization as Average KL Minimization) 11

B.2 Proof of Theorem 5.2 (Convergence of Average Categorical Entropy) 12

B.3 Proof of Theorem 7.1 (Output Distribution Approximation Constraint) 14

B.4 Proof of Corollary 7.2 (Implicit Representation Separation) 15

B.5 Proof of Proposition 7.5 (NLL Objective and Implicit Dirichlet Energy Minimization) . . 15

B.6 Proof of Theorem 7.7 (NLL Objective as Implicit Spectral Alignment) 17

C Full Related Work Discussion **18**

C.1 The Languages of Probability 18

C.2 The Geometry of Learning and Representation 19

A Detailed Background on Markov Categories

Definition A.1 (Category `Stoch` (Fritz, 2020; Perrone, 2023b))**.** 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 $(h \circ k)(x, C) := \int_Y h(y, C)\, k(x, \mathrm{d}y)$ (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 $(k \otimes h)((x, y), \cdot) := k(x, \cdot) \otimes h(y, \cdot)$ (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 $!_X : X \to I$ 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 A.2 (Interpretation)**.** In `Stoch`, objects represent types of random outcomes (sequences, vectors, tokens). Morphisms represent stochastic processes mapping inputs to probability distributions over outputs. Deterministic functions correspond to Dirac kernels. States $p : I \to X$ correspond to probability measures on X. Marginalization arises from discarding information via the map $!_Y$.

Definition A.3 (Categorical Entropy of a Kernel)**.** Let $k : X \to Y$ be a kernel in `Stoch`. The Categorical Entropy of k is a map $\mathcal{H}_D(k) : X \to \mathbb{R}_{\geq 0}$ defined as:

$$\mathcal{H}_D(k)(x) := D_{Y \otimes Y}\left(\Delta_Y \circ k(x, \cdot) \quad \| \quad (k(x, \cdot) \otimes k(x, \cdot)) \circ \Delta_X(x, \cdot)\right). \tag{A.1}$$

For the LM head kernel $k = k_{\text{head}}$, this gives a value for each hidden state $h \in \mathcal{H}$:

$$\mathcal{H}_D(k_{\text{head}})(h) := D_{\mathbb{V} \otimes \mathbb{V}}\left(\sum_{w \in \mathbb{V}} k_{\text{head}}(h, \{w\}) \delta_{(w,w)} \quad \| \quad k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)\right). \tag{A.2}$$

The average categorical entropy is then $\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t}) := \mathbb{E}_{h \sim p_{H_t}}[\mathcal{H}_D(k_{\text{head}})(h)]$.

B Full Proofs of Theorems and Propositions

B.1 Proof of Theorem 5.1 (NLL Minimization as Average KL Minimization)

Let $p_x(\cdot) := k_{\text{data}}(x, \cdot)$ 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:

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

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:

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

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 $L_{\mathrm{CE}}(\theta)$ with respect to θ is equivalent to minimizing $\mathcal{L}_{\mathrm{KL}}(\theta)$.

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

The minimum value $\mathcal{L}_{\mathrm{KL}}(\theta) = 0$ is achieved if and only if the integrand is zero $p_{W_{<t}}$-almost everywhere. That is, $D_{\mathrm{KL}}(p_x(\cdot)\,\|\,q_{x,\theta^*}(\cdot)) = 0$ for $p_{W_{<t}}$-almost every x. This occurs if and only if $p_x(\cdot) = q_{x,\theta^*}(\cdot)$ for $p_{W_{<t}}$-almost every x. In terms of kernels, this means $k_{\text{data}}(x,\cdot) = k_{\text{gen},\theta^*}(x,\cdot)$ for $p_{W_{<t}}$-almost every x.

If the model class $\{k_{\text{gen},\theta}\}$ contains k_{data}, say $k_{\text{data}} = k_{\text{gen},\theta_{\text{true}}}$, then choosing $\theta^* = \theta_{\text{true}}$ achieves $\mathcal{L}_{\mathrm{KL}}(\theta^*) = 0$, which is the minimum possible value. □

B.2 Proof of Theorem 5.2 (Convergence of Average Categorical Entropy)

We want to show that $\lim_{n\to\infty} \bar{\mathcal{H}}_D(k_{\text{head},n}; p_{H_t,\theta_n}) = \bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*})$.

Formal Statement of Theorem 5.2. *Let D be a statistical divergence (e.g., an f-divergence with f differentiable at 1, $f(1) = 0$, or d_{TV}) defined on $\mathcal{P}(\mathbb{V}\times\mathbb{V})$. Assume a sequence of model parameters θ_n achieves convergence in the training objective, such that the average KL divergence $\mathcal{L}_{\mathrm{KL}}(\theta_n) \to \inf_\theta \mathcal{L}_{\mathrm{KL}}(\theta)$ as $n \to \infty$. Let $k_n = k_{\text{gen},\theta_n}$ be the corresponding sequence of model kernels, and $k_{head,n} = k_{head,\theta_n}$ be the LM head kernels. Assume this convergence implies:*

(i) Pointwise convergence of the head kernel's output distributions: $k_{head,n}(h,\cdot) \to k_{head,\theta^*}(h,\cdot)$ in a suitable topology on $\mathcal{P}(\mathbb{V})$ (e.g., in total variation, which is strong for finite $\mathbb{V}$) for p_{H_t,θ^*}-almost every h, where θ^* minimizes $\mathcal{L}_{\mathrm{KL}}(\theta)$.

(ii) Weak convergence of the induced hidden state distributions: $p_{H_t,\theta_n} \Rightarrow p_{H_t,\theta^*}$.

(iii) The function $\Psi_D(h,p) := D_{\mathbb{V}\otimes\mathbb{V}}(\sum_w p(w)\delta_{(w,w)}\|p\otimes p)$ is continuous and bounded as a function of $p = k_{head}(h,\cdot)$ for h in the support of p_{H_t,θ^*}, with respect to the convergence topology in (i). (This holds for standard divergences like KL and TV on finite $\mathbb{V}$).

Then, the average categorical entropy of the learned LM head converges to that of the optimal LM head:

$$\lim_{n\to\infty} \bar{\mathcal{H}}_D(k_{\text{head},n}; p_{H_t,\theta_n}) = \bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*}). \tag{B.1}$$

Furthermore, if the model class is sufficiently expressive such that the optimal model kernel k_{gen,θ^*} matches the data kernel k_{data} (i.e., $\mathcal{L}_{\text{KL}}(\theta^*) = 0$), then the learned average categorical entropy approximates that of the implicit final stochastic step of the true data generating process. Assuming k_{data} admits a similar factorization with a final stochastic kernel $k_{head,\ data}$ acting on some "true" state h_{data}, then:

$$\bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*}) \approx \bar{\mathcal{H}}_D(k_{\text{head, data}}; p_{H_t,\text{data}}). \tag{B.2}$$

Recall the definition:

$$\bar{\mathcal{H}}_D(k_{\text{head},\theta}; p_{H_t,\theta}) = \mathbb{E}_{h\sim p_{H_t,\theta}}[\Psi_D(h, k_{\text{head},\theta}(h,\cdot))],$$

where $\Psi_D(h,p) := D_{\mathbb{V}\otimes\mathbb{V}}(\sum_{w\in\mathbb{V}} p(w)\delta_{(w,w)} \,\|\, p\otimes p)$, and $p = k_{\text{head},\theta}(h,\cdot)$.

Let X_n be the random variable $\Psi_D(H_n, k_{\text{head},n}(H_n,\cdot))$ where $H_n \sim p_{H_t,\theta_n}$. We want to show $\lim_{n\to\infty}\mathbb{E}[X_n] = \mathbb{E}[X^*]$, where $X^* = \Psi_D(H^*, k_{\text{head},\theta^*}(H^*,\cdot))$ with $H^* \sim p_{H_t,\theta^*}$.

We are given: (i) $k_{\text{head},n}(h,\cdot) \to k_{\text{head},\theta^*}(h,\cdot)$ in a suitable topology (e.g., total variation) for p_{H_t,θ^*}-almost every h. Let's denote this $p_n(h) \to p^*(h)$.

(ii) $p_{H_t,\theta_n} \Rightarrow p_{H_t,\theta^*}$ (weak convergence). This means $\int g(h)p_{H_t,\theta_n}(dh) \to \int g(h)p_{H_t,\theta^*}(dh)$ for all bounded continuous functions $g : \mathcal{H} \to \mathbb{R}$.

(iii) The function $\Psi_D(h,p)$ is continuous and bounded in p (with respect to the topology in (i)) for relevant h. Since $\mathbb{V}$ is finite, standard divergences like KL and TV are continuous functions of the probability vectors $p \in \Delta^{|\mathbb{V}|-1}$. The map $p \mapsto \sum p(w)\delta_{(w,w)}$ and $p \mapsto p\otimes p$ are also continuous. Thus, $p \mapsto \Psi_D(h,p)$ is continuous for fixed h. Boundedness also holds for typical divergences on finite spaces. Let M be an upper bound: $|\Psi_D(h,p)| \le M$.

Let $\Phi_n(h) = \Psi_D(h, k_{\text{head},n}(h,\cdot))$ and $\Phi^*(h) = \Psi_D(h, k_{\text{head},\theta^*}(h,\cdot))$. From (i) and the continuity part of (iii), we have $\Phi_n(h) \to \Phi^*(h)$ for p_{H_t,θ^*}-almost every h.

We want to show $\lim_{n\to\infty}\int \Phi_n(h)p_{H_t,\theta_n}(dh) = \int \Phi^*(h)p_{H_t,\theta^*}(dh)$.

Consider the difference:

$$\begin{aligned}
|\mathbb{E}[X_n] - \mathbb{E}[X^*]| &= \left|\int \Phi_n(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta^*}(dh)\right| \\
&\le \left|\int \Phi_n(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta_n}(dh)\right| \\
&+ \left|\int \Phi^*(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta^*}(dh)\right| \\
&= \left|\int (\Phi_n(h) - \Phi^*(h))p_{H_t,\theta_n}(dh)\right| + \left|\int \Phi^*(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta^*}(dh)\right|
\end{aligned}$$

The second term converges to 0 as $n \to \infty$ due to the weak convergence (ii), provided $\Phi^*(h)$ is bounded and continuous. While $\Phi^*(h)$ might not be continuous in h, if it is bounded and continuous p_{H_t,θ^*}-almost everywhere, weak convergence is often sufficient. Let's assume $\Phi^*(h)$ behaves well enough (e.g., is bounded and continuous almost everywhere w.r.t. the limiting measure p_{H_t,θ^*}) for $\int \Phi^*(h)p_{H_t,\theta_n}(dh) \to \int \Phi^*(h)p_{H_t,\theta^*}(dh)$. (This is sometimes known as the Generalized Continuous Mapping Theorem or Portmanteau Theorem).

For the first term, we have $\Phi_n(h) \to \Phi^*(h)$ for p_{H_t,θ^*}-almost every h. We also have the bound $|\Phi_n(h) - \Phi^*(h)| \le |\Phi_n(h)| + |\Phi^*(h)| \le 2M$ from the boundedness assumption (iii). We can use a variant of the Dominated Convergence Theorem adapted for converging measures. Since $p_{H_t,\theta_n} \Rightarrow p_{H_t,\theta^*}$ and $\Phi_n \to \Phi^*$ pointwise a.e. (w.r.t. p_{H_t,θ^*}), and the sequence Φ_n is uniformly bounded, we can conclude that $\int(\Phi_n(h) - \Phi^*(h))p_{H_t,\theta_n}(dh) \to 0$. Combining these, we get $\lim_{n\to\infty}\mathbb{E}[X_n] = \mathbb{E}[X^*]$.

For the final part: if the model class is expressive such that $k_{\text{gen},\theta^*} = k_{\text{data}}$ (meaning $\mathcal{L}_{\text{KL}}(\theta^*) = 0$), then the model perfectly matches the data generating process almost everywhere. If we assume the data process can be similarly factorized $k_{\text{data}} = k_{\text{head, data}} \circ k_{\text{enc, data}}$, then matching $k_{\text{gen},\theta^*} = k_{\text{data}}$ implies that the components must match (up to potential identifiability issues, e.g., transformations between the encoder output and head input that cancel out). Under reasonable assumptions (e.g., the factorization is unique in the relevant sense), we would have $k_{\text{head},\theta^*} \approx k_{\text{head, data}}$ and the distribution induced by the encoder $k_{\text{bb}} \circ k_{\text{emb}}$ would approximate the distribution of the "true" internal state feeding into $k_{\text{head, data}}$, i.e., $p_{H_t,\theta^*} \approx p_{H_t,\text{data}}$. Therefore, $\bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*}) \approx \bar{\mathcal{H}}_D(k_{\text{head, data}}; p_{H_t,\text{data}})$. □

B.3 Proof of Theorem 7.1 (Output Distribution Approximation Constraint)

We are given the average KL divergence loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x) \,\|\, p_\theta(\cdot|x))]$$

where $p_\theta(\cdot|x) = g_{\text{head}}(f_{\text{enc}}(x))$ and μ_{ctx} is the distribution over contexts. We are also given a metric d_{out} on $\mathcal{P}(\mathbb{V})$ satisfying a Pinsker-type inequality:

$$d_{\text{out}}(p, q)^k \leq C \cdot D_{\text{KL}}(p\|q)$$

for some constants $k, C > 0$. Examples include Hellinger distance d_H ($k = 2, C = 1/2$) and Total Variation distance d_{TV} ($k = 2, C = 1$).

Let $p = P_{\text{data}}(\cdot|x)$ and $q = p_\theta(\cdot|x)$ for a specific context x. Applying the inequality yields:

$$d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k \leq C \cdot D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x)).$$

Now, we take the expectation of both sides with respect to the context distribution $x \sim \mu_{ctx}$. Since expectation is linear and the inequality holds pointwise for each x, we get:

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq \mathbb{E}_{x\sim\mu_{ctx}}[C \cdot D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))]$$

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))]$$

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathcal{L}(\theta).$$

This establishes the first part of the theorem.

For the second part, consider any two contexts x, x'. Let $p_x^{\text{data}} = P_{\text{data}}(\cdot|x)$, $p_{x'}^{\text{data}} = P_{\text{data}}(\cdot|x')$, $p_x^\theta = p_\theta(\cdot|x)$, and $p_{x'}^\theta = p_\theta(\cdot|x')$. The triangle inequality for the metric d_{out} states:

$$d_{\text{out}}(A, C) \leq d_{\text{out}}(A, B) + d_{\text{out}}(B, C)$$

Applying this twice:

$$\begin{aligned} d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) &\leq d_{\text{out}}(p_x^{\text{data}}, p_x^\theta) + d_{\text{out}}(p_x^\theta, p_{x'}^{\text{data}}) \\ &\leq d_{\text{out}}(p_x^{\text{data}}, p_x^\theta) + (d_{\text{out}}(p_x^\theta, p_{x'}^\theta) + d_{\text{out}}(p_{x'}^\theta, p_{x'}^{\text{data}})) \\ &= d_{\text{out}}(p_x^{\text{data}}, p_x^\theta) + d_{\text{out}}(p_x^\theta, p_{x'}^\theta) + d_{\text{out}}(p_{x'}^\theta, p_{x'}^{\text{data}}). \end{aligned}$$

Let $\epsilon_x = d_{\text{out}}(p_x^{\text{data}}, p_x^\theta)$ and $\epsilon_{x'} = d_{\text{out}}(p_{x'}^\theta, p_{x'}^{\text{data}})$. If the model fits the data well, $\mathcal{L}(\theta)$ is small. Notice that $\mathbb{E}_{x\sim\mu_{ctx}}[\epsilon_x^k] \leq C\mathcal{L}(\theta)$, meaning the expected error (to the power k) is small. By Markov's inequality, for any $\delta > 0$,

$$\mathbb{P}(\epsilon_x^k \geq \delta^k) \leq \frac{\mathbb{E}[\epsilon_x^k]}{\delta^k} \leq \frac{C\mathcal{L}(\theta)}{\delta^k}.$$

Thus, $\mathbb{P}(\epsilon_x \geq \delta)$ is small if $\mathcal{L}(\theta)$ is small, implying that for a vast majority of contexts x drawn from μ_{ctx}, the individual error ϵ_x is small. Therefore, for typical pairs (x, x'), both ϵ_x and $\epsilon_{x'}$ are small.

Rearranging the triangle inequality gives:

$$d_{\text{out}}(p_x^\theta, p_{x'}^\theta) \geq d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) - (\epsilon_x + \epsilon_{x'})$$

$$d_{\text{out}}(p_x^\theta, p_{x'}^\theta) \leq d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) + (\epsilon_x + \epsilon_{x'})$$

When ϵ_x and $\epsilon_{x'}$ are small, these inequalities show that the distance between the model's output distributions, $d_{\text{out}}(p_x^\theta, p_{x'}^\theta)$, must be close to the distance between the true data distributions, $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})$. In particular, if $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})$ is large (predictively dissimilar contexts), then $d_{\text{out}}(p_x^\theta, p_{x'}^\theta)$ must also be large, as the difference is bounded by small error terms. □

B.4 Proof of Corollary 7.2 (Implicit Representation Separation)

From Theorem 7.1, if $\mathcal{L}(\theta)$ is small, then for typical contexts x, x':

$$d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) \approx d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')).$$

If contexts x, x' are predictively dissimilar, the RHS is large, implying the LHS must also be large. Let $p_x = g_{\text{head}}(h_x)$ and $p_{x'} = g_{\text{head}}(h_{x'})$ be two points in the manifold $(\mathcal{P}(\mathbb{V}), g^{\text{FR}})$. The distance between them, under the geometry induced by the Fisher-Rao metric, is the infimum of the lengths of all paths connecting them. The length of any such path $\gamma(t)$ from $p_{x'}$ to p_x is given by $L(\gamma) = \int_0^1 \sqrt{g^{\text{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t))}\, dt$.

Consider the specific path induced by the straight line in representation space, $\gamma(t) = g_{\text{head}}((1-t)h_{x'} + th_x)$. Let $h(t) = (1-t)h_{x'} + th_x$. The tangent vector to this path in $\mathcal{P}(\mathbb{V})$ is $\gamma'(t) = J_{g_{\text{head}}}(h(t)) \cdot (h_x - h_{x'})$, where $J_{g_{\text{head}}}$ is the Jacobian of g_{head}. The squared length of this tangent vector is:

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

The total squared length of this particular path is $(\int_0^1 \sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})} dt)^2$. Since the geodesic distance $d_{\text{FR}}(p_x, p_{x'})$ is the shortest path length, it is bounded above by this length.

A large distance $d_{\text{out}}(p_x, p_{x'})$ (which is closely related to $d_{\text{FR}}(p_x, p_{x'})$) implies that the length of any path connecting them must be large. This in turn requires that the integrand $\sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})}$ must be large over a substantial portion of the path. This can only happen if the difference vector $h_x - h_{x'}$ has significant components along directions v where the pullback metric g^* is large (i.e., directions of high predictive sensitivity).

Conversely, if contexts are predictively similar, the RHS is small, implying the LHS must be small. This means the representations h_x and $h_{x'}$ must be mapped to nearby points in $\mathcal{P}(\mathbb{V})$. This is achieved if $h_x, h_{x'}$ are close along the directions of high predictive sensitivity, as these are the directions that would otherwise cause a large separation in the output space. Hence, NLL encourages representations of predictively similar contexts to be close along sensitive directions. □

B.5 Proof of Proposition 7.5 (NLL Objective and Implicit Dirichlet Energy Minimization)

1. Formal Setup. The proposition states that minimizing the NLL objective implicitly encourages the minimization of the Dirichlet energy of the learned representations when projected onto predictively sensitive directions. Let us first formalize the components.

The Dirichlet energy $\mathcal{E}_K(\phi_v)$ of a function $\phi_v : \mathbb{V}^* \to \mathbb{R}$ with respect to a predictive similarity kernel $K(x, x')$ is given by:

$$\mathcal{E}_K(\phi_v) = \frac{1}{2} \iint_{\mathbb{V}^* \times \mathbb{V}^*} K(x, x')(\phi_v(x) - \phi_v(x'))^2 \, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x'). \tag{B.3}$$

We consider the function $\phi_v(x) = \langle h_x, v \rangle$, which is the projection of the learned representation $h_x = f_{\text{enc}}(x)$ onto a vector $v \in \mathcal{H}$. The vector v is assumed to be a direction of high predictive sensitivity,

meaning it is a direction where the pullback metric $g^*(h)$ from Section 6 has a large magnitude (i.e., $g^*(h)(v, v)$ is large). This signifies that changes to a representation h along direction v cause large changes in the model's output distribution.

The predictive similarity kernel $K(x, x') \geq 0$ quantifies the similarity between the true target distributions $p_x := P_{\text{data}}(\cdot|x)$ and $p_{x'} := P_{\text{data}}(\cdot|x')$. A high value of $K(x, x')$ implies that p_x and $p_{x'}$ are similar.

Our core argument relies on Corollary 7.2, which establishes that a successful NLL optimization (i.e., achieving a small loss) forces the model's predictive distributions $p_\theta(\cdot|x)$ to approximate the true data distributions p_x. Consequently, the geometric distance between two model predictions, $d(p_\theta(\cdot|x), p_\theta(\cdot|x'))$, must approximate the distance between the corresponding true distributions, $d(p_x, p_{x'})$.

2. Analysis of the Dirichlet Energy Integrand. We analyze the integrand of the Dirichlet energy, $I(x, x') = K(x, x')(\phi_v(x) - \phi_v(x'))^2$, by considering two complementary cases based on the similarity of the context pair (x, x').

Case 1: High Predictive Similarity ($K(x, x')$ is large). If two contexts x and x' are predictively similar, $K(x, x')$ is large, which implies that their true target distributions are close, i.e., $d(p_x, p_{x'})$ is small.

1. NLL minimization ensures that the model's predictions approximate the true distributions. By the triangle inequality, if $p_\theta(\cdot|x) \approx p_x$ and $p_\theta(\cdot|x') \approx p_{x'}$, then we must have $d(p_\theta(\cdot|x), p_\theta(\cdot|x')) \approx d(p_x, p_{x'})$, which is small.
2. As established by the properties of the pullback metric g^*, a small distance between the output distributions $p_\theta(\cdot|x)$ and $p_\theta(\cdot|x')$ necessitates a small "predictive path length" between their corresponding latent representations h_x and $h_{x'}$.
3. For this path length to be small, the difference vector $h_x - h_{x'}$ must have a small magnitude when measured by the metric g^*. Specifically, for a direction v of high predictive sensitivity, this forces the projected difference $|\phi_v(x) - \phi_v(x')| = |\langle h_x - h_{x'}, v\rangle|$ to be small.
4. Therefore, for this pair (x, x'), the integrand $I(x, x')$ is a product of a *large* term ($K(x, x')$) and a term $(\phi_v(x) - \phi_v(x'))^2$ that is actively driven to be *small* by the NLL objective. The resulting contribution to the total energy $\mathcal{E}_K(\phi_v)$ is thus minimized.

Case 2: Low Predictive Similarity ($K(x, x')$ is small). If two contexts are predictively dissimilar, $K(x, x')$ is small. This implies that their true target distributions are far apart, i.e., $d(p_x, p_{x'})$ is large.

1. NLL minimization requires $p_\theta(\cdot|x) \approx p_x$ and $p_\theta(\cdot|x') \approx p_{x'}$, which in turn forces the model's output distributions to be far apart: $d(p_\theta(\cdot|x), p_\theta(\cdot|x'))$ must be large.
2. A large distance between output distributions requires a large predictive path length between h_x and $h_{x'}$. Consequently, the projected difference $|\phi_v(x) - \phi_v(x')|$ can be, and often is encouraged to be, large.
3. In this case, the integrand $I(x, x')$ is a product of a term $(\phi_v(x) - \phi_v(x'))^2$ that can be large, but it is multiplied by the term $K(x, x')$, which is by definition *small*.
4. The small similarity weight $K(x, x')$ ensures that the contribution of such predictively dissimilar pairs to the total energy $\mathcal{E}_K(\phi_v)$ is suppressed, regardless of how far apart their representations are pushed.

In conclusion, the NLL optimization process systematically reduces the Dirichlet energy. It actively minimizes the terms in the integral that could be large (i.e., those where $K(x, x')$ is large) by forcing the corresponding representations to be close. It is permissive of large representation differences only when pairs are predictively dissimilar, but the contribution of these pairs to the Dirichlet energy is already negligible due to the small kernel weight. Thus, the learning dynamics of NLL are inherently biased towards finding a representation function f_{enc} that produces features that are smooth over the predictive similarity graph, fulfilling the core principle of spectral methods and Dirichlet energy minimization. □

B.6 Proof of Theorem 7.7 (NLL Objective as Implicit Spectral Alignment)

This theorem states that minimizing the Negative Log-Likelihood (NLL) objective implicitly performs a spectral decomposition of the data's predictive similarity structure, forcing the learned representations to align with the dominant eigenfunctions of a predictive similarity operator M_K. We provide a formal proof by first recasting the NLL objective as a multi-positive contrastive loss and then leveraging established results from the theory of self-supervised learning that connect such losses to spectral methods (HaoChen et al., 2021; Tan et al., 2024).

1. Formal Setup: NLL as a Contrastive Objective. The NLL training objective seeks to minimize the KL divergence between the true data distribution and the model's predictive distribution, averaged over all contexts:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}}\left[D_{\mathrm{KL}}\left(P_{\text{data}}(\cdot|x)\,\|\,p_\theta(\cdot|x)\right)\right], \tag{B.4}$$

where μ_{ctx} is the distribution over contexts, $p_x := P_{\text{data}}(\cdot|x)$ is the true conditional distribution, and $p_\theta(\cdot|x)$ is the model's prediction. Minimizing this is equivalent to minimizing the cross-entropy loss. For a single context x, the loss (ignoring the constant data entropy $H(p_x)$) is:

$$J_x(\theta) = -\mathbb{E}_{w\sim p_x}[\log p_\theta(w|x)]. \tag{B.5}$$

We adopt a standard log-linear model for the prediction head, where $p_\theta(w|x) = \mathrm{softmax}(f_{\text{head}}(h_x))_w$. Let $h_x = f_{\text{enc}}(x)$ be the representation and $g(w)$ be the output logit vector (prototype) for token $w \in \mathcal{V}$. The probability is:

$$p_\theta(w|x) = \frac{\exp(\langle g(w), h_x\rangle/\tau)}{\sum_{w'\in\mathcal{V}}\exp(\langle g(w'), h_x\rangle/\tau)}, \tag{B.6}$$

where τ is a temperature parameter. The loss for context x becomes:

$$J_x(\theta) = \underbrace{-\frac{1}{\tau}\mathbb{E}_{w\sim p_x}[\langle g(w), h_x\rangle]}_{\mathcal{L}_{\text{align}}} + \underbrace{\log\sum_{w'\in\mathcal{V}}\exp\left(\frac{1}{\tau}\langle g(w'), h_x\rangle\right)}_{\mathcal{L}_{\text{unif}}}. \tag{B.7}$$

This is precisely a **multi-positive contrastive loss**.

- **Alignment Term ($\mathcal{L}_{\text{align}}$):** This term encourages the representation h_x to align with the expected prototype vector of its true next-token distribution, $\bar{g}_x := \mathbb{E}_{w\sim p_x}[g(w)]$. It pulls the representation h_x towards every "positive" token w with a force proportional to its true probability $p_x(w)$.
- **Uniformity Term ($\mathcal{L}_{\text{unif}}$):** The log-partition function acts as a repulsion term. It pushes h_x away from all token prototypes $g(w')$ in the vocabulary, preventing a trivial collapse where all representations and prototypes map to the same point.

2. The Induced Predictive Similarity Kernel. The structure of the alignment term naturally induces a notion of similarity between two contexts, x and x'. If their target distributions p_x and $p_{x'}$ are similar, their expected prototypes $\bar{g}_x$ and $\bar{g}_{x'}$ will be close. We can formalize this by defining the **Predictive Similarity Kernel** using these prototypes:

$$K(x, x') := \langle \bar{g}_x, \bar{g}_{x'}\rangle = \left\langle \mathbb{E}_{w\sim p_x}[g(w)], \mathbb{E}_{w'\sim p_{x'}}[g(w')]\right\rangle. \tag{B.8}$$

This kernel measures the similarity of the target distributions as mediated by the geometry of the output vocabulary embeddings $\{g(w)\}$. If the output embeddings $\{g(w)\}$ are encouraged by the uniformity term to be roughly orthogonal, then $K(x, x') \approx \sum_w p_x(w)p_{x'}(w)\|g(w)\|^2$, which is closely related to the Expected Likelihood Kernel.

3. Alignment of Covariance and Kernel Matrices. The optimization pushes the learned representations to mirror the structure of this kernel. The alignment objective encourages $h_x \propto \bar{g}_x$. If this alignment were perfect, the inner product of representations would be directly proportional to the kernel:

$$\langle h_x, h_{x'} \rangle \propto \langle \bar{g}_x, \bar{g}_{x'} \rangle = K(x, x'). \tag{B.9}$$

This implies that the Gram matrix of representations, $\mathbf{H}^\top \mathbf{H}$ (where $\mathbf{H}$ is the matrix of representations $[h_{x_1}, h_{x_2}, \dots]$), is driven to align with the kernel matrix $\mathbf{K} = [K(x_i, x_j)]_{ij}$. The goal of representation learning under this objective can be seen as finding an encoder f_{enc} and prototypes $\{g(w)\}$ that jointly solve:

$$\min_{f_{\text{enc}}, \{g(w)\}} \left\| \mathbf{H}^\top \mathbf{H} - \mathbf{K} \right\|_F^2, \tag{B.10}$$

subject to regularizing constraints imposed by the uniformity term, which ensures that the representations do not collapse and the prototypes $\{g(w)\}$ spread out. This is a classic spectral alignment problem.

4. Connection to the Spectral Operator M_K. The matrix alignment problem has a direct correspondence in the continuous operator setting. The Gram matrix $\mathbf{H}^\top \mathbf{H}$ is a sample-based version of the representation covariance operator $\Sigma_H = \mathbb{E}_x[h_x \otimes h_x]$. The kernel matrix $\mathbf{K}$ is the sample-based version of the **Predictive Similarity Operator** M_K defined in Definition 7.6.

The theory of spectral methods shows that the principal components of the representations (the leading eigenvectors of Σ_H) must align with the leading eigenfunctions of the operator M_K. The NLL optimization process, via stochastic gradient descent, effectively acts as a stochastic power iteration method to find this alignment. The gradients derived from Eq. B.7 continuously update the representations h_x and prototypes $g(w)$ to better align the representation geometry with the predictive similarity structure.

In summary, the Negative Log-Likelihood (NLL) objective is not merely a mechanism for maximizing predictive accuracy. Its mathematical structure is equivalent to a multi-positive contrastive loss that implicitly defines a predictive similarity graph over the data. Minimizing this loss forces the learned representations to become a spectral embedding of this graph. This means the principal directions of variance in the representation space $\mathcal{H}$ are forced to align with the dominant eigenfunctions of the predictive similarity operator M_K. The information geometry induced by the head's pullback metric g^* (as described in Section 6) provides the crucial link, translating the abstract goal of matching probability distributions into concrete gradients that sculpt the geometry of the representation space $\mathcal{H}$ to achieve this spectral alignment. □

C Full Related Work Discussion

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.

C.1 The Languages of Probability

The Synthetic View: 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

Table 1: A Comparative Taxonomy of Theoretical Frameworks for Representation Learning. This table situates our proposed framework relative to existing theoretical paradigms, highlighting the unique synthesis of algebraic and geometric perspectives it offers to analyze the representation space directly.

Framework	**Core Mathematical Object**	**Primary Domain of Analysis**	**Key Insight/Contribution**	**Limitation Addressed by Our Work**	**Key Citations**
Information Geometry (IG)	Riemannian Manifold (Fisher-Rao Metric)	Model Parameter Space	Provides a metric for the "distance" between models; explains optimization via natural gradients.	Lacks a compositional, algebraic structure; difficult to apply to the dynamics of representations themselves.	(Amari and Nagaoka, 2000)
Categorical Probability	Markov Category (Symmetric Monoidal Category + Comonoids)	Abstract Probabilistic Systems	Provides a high-level, compositional language for probability, unifying disparate theories via algebraic axioms.	Inherently algebraic and non-metric; does not natively describe the shape or geometry of representation spaces.	(Fritz, 2020), (Perrone, 2023a), (Cho and Jacobs, 2019)
Information Bottleneck (IB)	Mutual Information	Information-Theoretic Channels	Defines representation learning as an optimal trade-off between compression and predictive relevance.	A high-level principle, not a constructive framework; does not specify the geometric mechanism for achieving the bottleneck.	(Tishby et al., 2000)
Spectral Graph Theory	Graph Laplacian / Dirichlet Energy	Graph-structured Data / Activations	Quantifies the "smoothness" of functions on a graph; explains oversmoothing in GNNs and representation reorganization in LLMs.	A measurement tool, not a full-fledged theoretical framework; describes *what* geometry emerges but not *why*.	(Yuan, 2023; Park et al., 2024)
Implicit Bias Theory	Optimization Dynamics (Gradient Flow)	Model Weights & Activations	Explains generalization in overparameterized models; frames NLL as an implicit contrastive, geometry-sculpting process.	Provides a mechanism but lacks a unified language to describe the resulting geometric and probabilistic structures.	(Gunasekar et al., 2018)
Our Framework (Proposed)	**Geometric Markov Category**	**Representation Space**	**Unifies the algebraic structure of probability with the metric geometry of representations.**	-	**This Paper**

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, 2023a). The central object of study in this domain is the *Markov category* (Cho and Jacobs, 2019; Fritz, 2020).

The Metric View: 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.

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, 2023a). This research program 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, 2023b,a).

C.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).

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 $L = D - A$. For any function f defined on the nodes of the graph (e.g., a feature activation), the quadratic form $f^\top L f$ defines the graph's *Dirichlet energy*. This quantity measures the smoothness of the function with respect to the graph structure.

Most critically for the present work, this classical geometric tool has been shown to be 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).

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).

A crucial modern insight is that training with the standard Negative Log-Likelihood (NLL) objective can be understood as a form of *implicit contrastive learning*. The NLL loss for a sample (x, y_{true}) is functionally equivalent to a contrastive loss that pulls the representation of x towards a "positive" target while pushing it away from all negative targets. This connection is profound, as it links the vast body of work on self-supervised contrastive learning—which explicitly aims to sculpt a geometrically structured representation space—to standard supervised training.