A Markov Categorical Framework for Language Modeling

Yifan Zhang [1]

Abstract

Auto-regressive (AR) language models factorize sequence probabilities as $P_\theta(\mathbf{w}) = \prod_t P_\theta(w_t|\mathbf{w}_{<t})$. While empirically powerful, their internal mechanisms remain partially understood. This work introduces an analytical framework using Markov Categories (MCs), specifically the category **Stoch** of standard Borel spaces and Markov kernels. We model the AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ as a composite kernel $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$. Leveraging the enrichment of **Stoch** with statistical divergences D and associated categorical information measures (entropy $\mathcal{H}_D$, mutual information I_D), we define principled metrics: Representation Divergence $D(p_{H_t|s_1} \| p_{H_t|s_2})$, State-Prediction Information $I_D(H_t; W_t)$, Temporal Coherence $I_D(H_t; H_{t+1})$, LM Head Stochasticity $\mathcal{H}_D(k_{\text{head}})$, and Information Flow Bounds via the Data Processing Inequality (e.g., $I_D(S; H_t) \geq I_D(S; W_t)$). Beyond providing metrics, this framework analyzes the negative log-likelihood (NLL) objective itself. We argue NLL minimization equates to optimal compression and learning the data's intrinsic stochasticity ($\bar{\mathcal{H}}_D$). We employ information geometry, analyzing the pullback Fisher-Rao metric g^* on the representation space $\mathcal{H}$, to understand learned sensitivities. Furthermore, we formalize the concept that NLL acts as implicit structure learning, demonstrating how minimizing NLL forces representations of predictively dissimilar contexts apart.

1. Introduction

Autoregressive language models (AR LMs), particularly the Transformer-based architectures (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020), have become foundational in natural language processing. Their success stems from factorizing sequence probabilities $P_\theta(\mathbf{w}) = \prod_t P_\theta(w_t|\mathbf{w}_{<t})$ and optimizing parameters θ by minimizing negative log-likelihood (NLL) on large datasets. Despite their empirical power, a deep theoretical understanding of their internal information processing pathways, representation learning dynamics, and the reasons behind their emergent capabilities remains incomplete (Manning et al., 2020; Hupkes et al., 2020; Elhage et al., 2021). Such understanding is crucial for improving model interpretability, robustness, efficiency, and for guiding future development. Why the simple NLL objective yields representations capturing complex linguistic and world knowledge is a particularly central question.

This work introduces an analytical framework focused on the core AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$, rooted in the theory of Markov Categories (MCs) (Cho & Jacobs, 2019; Fritz, 2020). MCs provide an abstract algebraic setting tailored for reasoning about systems involving probability, causality, and information flow. We specifically leverage **Stoch**, the category of standard Borel spaces and Markov kernels (Kallenberg & Kallenberg, 1997; Fritz, 2020). The strength of this framework lies in its inherent compositionality, mirroring the layered structure of neural networks, its native handling of probability and stochastic transformations crucial for generative models, and its capacity for defining fundamental information-theoretic quantities. We model the generation step as a composite kernel $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in **Stoch**, separating embedding, backbone processing, and the final stochastic prediction head.

By enriching **Stoch** with a statistical divergence D (e.g., $D_{\text{KL}}, d_{\text{TV}}$) (Baez et al., 2016; Perrone, 2023a;b) and using associated categorical information measures (entropy $\mathcal{H}_D$, mutual information I_D) satisfying the Data Processing Inequality (DPI), we develop a multifaceted analysis:

1. We provide a formal MC model of the AR step, grounding the process in measure-theoretic probability and category theory.
2. We define principled metrics: Representation Divergence (RepDiv$_D$) to quantify how well hidden states distinguish context properties; Categorical Mutual Information (I_D) measuring state-prediction relevance

[1]Yifan Zhang. Correspondence to: Yifan Zhang <yifanzhangresearch@gmail.com>.

The second AI for MATH Workshop at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. 2025.

and temporal coherence; LM Head Categorical Entropy ($\mathcal{H}_D$) assessing prediction stochasticity; and DPI-based Information Flow Bounds quantifying processing bottlenecks.

3. We interpret NLL minimization as equivalent to minimizing average KL divergence, linking it directly to optimal data compression (cf. the "compression implies intelligence" hypothesis) and forcing the model to learn the data's intrinsic conditional stochasticity ($\bar{\mathcal{H}}_D$).
4. We employ information geometry, analyzing the pullback Fisher-Rao metric (g^*) on the representation space $\mathcal{H}$ induced by the LM head, to understand learned predictive sensitivities and functional anisotropy.
5. We formalize NLL as implicit structure learning, demonstrating how minimizing NLL necessarily forces representations of predictively dissimilar contexts apart along sensitive dimensions, establishing rigorous connections to spectral graph theory principles based on predictive similarity kernels.

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 $(\mathcal{C}, \otimes, I)$ 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 $X \to I$).*

Morphisms $k : X \to Y$ are interpreted as stochastic processes or channels transforming systems of type X to type Y. Composition $h \circ k$ denotes sequential processing, $k \otimes h$ parallel processing. The comonoid maps Δ_X (copy) and $!_X$ (discard) model the duplication and deletion of information. States (probability distributions) on X are morphisms $p : I \to X$.

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 $(X, \mathscr{B}(X))$. 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 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 $p : I \to X$ 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 $p : I \to X \otimes Y$, the X-marginal is $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$.*

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) $p, q : I \to X$, 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 & 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 (e.g., f-divergences, Rényi $\alpha \in [0, \infty]$).

Theorem 2.4 (Data Processing Inequality (DPI))**.** *Let D be a statistical divergence satisfying the DPI. For any Markov kernel $k : X \to Y$ in* **Stoch** *and any pair of states $p, q : I \to X$:*

$$D_Y(k \circ p \| k \circ q) \leq D_X(p\|q) \tag{1}$$

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

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

Definition 2.5 (Categorical Entropy and Mutual Information (Perrone, 2023a))**.** *Let* $(\mathbf{Stoch}, D)$ *be enriched with a DPI-satisfying divergence* D*.*

1. *The Categorical Entropy of a kernel* $k : X \to Y$ *measures its intrinsic stochasticity:*

$$\mathcal{H}_D(k) \coloneqq D_{Y\otimes Y}\left(\Delta_Y \circ k \quad \| \quad (k \otimes k) \circ \Delta_X\right) \tag{2}$$

It compares two processes producing pairs in $Y \otimes Y$*. The first* $(\Delta_Y \circ k)$ *applies* k *once* $(x \mapsto y \sim k(x, \cdot))$ *and deterministically copies the output* (y, y)*. The second* $((k \otimes k) \circ \Delta_X)$ *deterministically copies the input* (x, x) *and applies* k *independently to each component* (y_1, y_2) *where* $y_1, y_2 \sim k(x, \cdot)$ *are i.i.d. The divergence measures how different these two resulting joint distributions are, quantifying how far* k *is from being deterministic. If* k *is deterministic,* $k = k_f$*, both sides yield the same state (corresponding to* $\delta_{(f(x),f(x))}$*) and* $\mathcal{H}_D(k_f) = 0$*.*

2. *The Categorical Mutual Information of a joint state* $p : I \to X \otimes Y$ *measures the statistical dependence between* X *and* Y*:*

$$I_D(p) \coloneqq D_{X\otimes Y}\left(p \quad \| \quad p_X \otimes p_Y\right) \tag{3}$$

where $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$ *and* $p_Y = (!_X \otimes \mathrm{id}_Y) \circ p$ *are the marginal states.* $I_D(p)$ *measures how far the joint state* p *is from the product of its marginals (representing independence), according to the geometry induced by* D*.*

Remark 2.6 (Properties and Connections)**.** *When* $D = D_{\mathrm{KL}}$*,* $I_{D_{\mathrm{KL}}}(p)$ *recovers the standard Shannon mutual information* $I(X;Y)$ *for the joint distribution* p*.* $\mathcal{H}_{D_{\mathrm{KL}}}(k)$ *provides an intrinsic measure of the kernel's stochasticity, related to but distinct from average conditional Shannon entropy (Perrone, 2023a). Crucially, these categorical definitions automatically satisfy the DPI. For instance, consider a state* $p_{XY} : I \to X \otimes Y$ *and a kernel* $h : Y \to Z$*. Let* p_{XZ} *be the state obtained by applying* $\mathrm{id}_X \otimes h$ *to* p_{XY}*. The DPI for* D *applied to the states involved in the definition of* I_D *implies* $I_D(p_{XY}) \geq I_D(p_{XZ})$ *(Perrone, 2023a, Prop. 4.8). This reflects the principle that processing* $(Y \to Z)$ *cannot increase information about* X*. Furthermore, information geometry (Amari & Nagaoka, 2000) arises naturally: the Fisher-Rao metric is induced by the local quadratic approximation of the KL divergence, linking the divergence* D *to the underlying geometric structure of the space of probability measures.*

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 $\mathbf{Stoch}$.

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

- Input context space: $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) = (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, where $\mathcal{V}^*$ is the set of finite sequences over the vocabulary $\mathcal{V}$, equipped with a suitable σ-algebra making it standard Borel (e.g., considering it as a disjoint union of finite products $\mathcal{V}^n$).
- 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: $(\mathcal{V}, \mathcal{P}(\mathcal{V})) = (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, 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 & Kallenberg, 1997).

The generation process decomposes into three kernels (morphisms in $\mathbf{Stoch}$):

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

$$k_{\text{emb}}(\mathbf{w}_{<t}, A) \coloneqq \delta_{f_{\text{emb}}(\mathbf{w}_{<t})}(A) = \mathbf{1}_A(f_{\text{emb}}(\mathbf{w}_{<t})), \quad \text{for } A \in \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}). \tag{4}$$

This is a valid morphism in $\mathbf{Stoch}$.

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:

$$k_{\text{bb}}(E_{<t}, B) \coloneqq \delta_{f_{\text{bb}}(E_{<t})}(B) = \mathbf{1}_B(f_{\text{bb}}(E_{<t})), \quad \text{for } B \in \mathscr{B}(\mathcal{H}). \tag{5}$$

This is also a morphism in $\mathbf{Stoch}$.

3. **LM Head Kernel** ($k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$): This final kernel maps the summary hidden state $h_t \in \mathcal{H}$ to a probability distribution over the finite vocabulary $\mathcal{V}$. Typically, h_t is passed through a linear layer ($f_{\text{head}} : \mathcal{H} \to \mathbb{R}^{|\mathcal{V}|}$) 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 genuinely stochastic Markov kernel:

$$k_{\text{head}}(h, A) \coloneqq \sum_{w \in A} [\text{softmax}(f_{\text{head}}(h))]_w \quad \text{for } h \in \mathcal{H}, A \subseteq \mathcal{V}. \tag{6}$$

This kernel maps each point h in the representation space to a probability measure on the discrete space $\mathcal{V}$, satisfying the required measurability conditions. It is a morphism in $\mathbf{Stoch}$.

The overall single-step generation kernel $k_{\text{gen},\theta} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ is the composition $k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in the category $\mathbf{Stoch}$. This composition precisely represents the model's learned conditional probability map $P_\theta(\cdot|\mathbf{w}_{<t})$. The subsequent sections will use this representation to define and analyze information-theoretic metrics.

4. Markov Categorical Metrics

We now apply the Markov category framework $(\mathbf{Stoch}, D)$ to analyze the AR generation kernel $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$. We select a suitable statistical divergence D satisfying the Data Processing Inequality (DPI) (e.g., D_{KL}, d_{TV}, or more generally an f-divergence (Amari & Nagaoka, 2000; Nowozin et al., 2016)) and utilize the corresponding categorical information measures $\mathcal{H}_D$ and I_D (Equations (2) and (3)) to probe the information flow and transformations within the generation step. A particular focus is placed on the final hidden state $H_t \in \mathcal{H}$ and the stochastic prediction kernel k_{head}.

We operate within the probabilistic setting induced by a distribution over input contexts. Let P_{ctx} be a probability measure on the context space $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) = (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$. This corresponds to an initial *state* in the Markov category $\mathbf{Stoch}$, represented by a morphism $p_{W_{<t}} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, where I is the monoidal unit (a singleton measurable space) and $p_{W_{<t}}(\star, A) = P_{\text{ctx}}(A)$ for any $A \in \mathscr{B}(\mathcal{V}^*)$. Processing this initial state through the sequence of deterministic kernels k_{emb} and k_{bb}, and the stochastic kernel k_{head}, induces distributions (states) at subsequent stages:

- **Initial Sequence Embedding State**: Given $p_{W_{<t}} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, the distribution of the initial vector sequence representation $E_{<t} \in \mathcal{H}_{\text{seq\_emb}}$ is given by the state $p_{E_{<t}} : I \to (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$, defined as:

$$p_{E_{<t}} \coloneqq k_{\text{emb}} \circ p_{W_{<t}}. \tag{7}$$

  Since k_{emb} corresponds to the deterministic function f_{emb}, the measure associated with $p_{E_{<t}}$ is the pushforward measure $(P_{\text{ctx}}) \circ f_{\text{emb}}^{-1}$.
- **Final Hidden State**: The distribution of the final hidden state $H_t \in \mathcal{H}$ is given by the state $p_{H_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$:

$$p_{H_t} \coloneqq k_{\text{bb}} \circ p_{E_{<t}} = (k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}}. \tag{8}$$

  As k_{bb} is also deterministic (representing f_{bb}), p_{H_t} corresponds to the pushforward measure $(P_{\text{ctx}}) \circ (f_{\text{bb}} \circ f_{\text{emb}})^{-1}$.
- **Predicted Next Token State**: The marginal distribution of the predicted next token $W_t \in \mathcal{V}$, averaged over all contexts according to P_{ctx}, is given by the state $p_{W_t} : I \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$:

$$p_{W_t} \coloneqq k_{\text{head}} \circ p_{H_t} = (k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}} = k_{\text{gen},\theta} \circ p_{W_{<t}}. \tag{9}$$

  Let μ_{H_t} be the measure on $\mathcal{H}$ associated with p_{H_t}. Then the measure associated with p_{W_t} on $\mathcal{V}$ is given by $\mu_{W_t}(A) = \int_{\mathcal{H}} k_{\text{head}}(h, A)\, \mu_{H_t}(\mathrm{d}h)$ for $A \subseteq \mathcal{V}$.

The random variables corresponding to these stages are denoted $W_{<t}$ (context sequence), $E_{<t}$ (initial embedding sequence), H_t (final hidden state), and W_t (predicted next token). Their respective distributions (states in $\mathbf{Stoch}$) are denoted $p_{W_{<t}}$, $p_{E_{<t}}$, p_{H_t}, and p_{W_t}. Using these rigorously defined states and the categorical information measures, we propose the following metrics.

4.1. Metric 1: Representation Divergence (Context Encoding Fidelity)

To quantify how effectively the distribution of the final hidden state H_t distinguishes between different underlying properties S of the input context $\mathbf{w}_{<t}$. Consider a random variable S representing a context property. Let $p_{W_{<t}|s}$ be

the conditional input state for property value s. The conditional distribution of the hidden state H_t given $S = s$ is represented by the state:

$$p_{H_t|s} \coloneqq (k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}|s} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H})). \quad (10)$$

Let $\mu_{H_t|s}$ be the probability measure on $\mathcal{H}$ associated with $p_{H_t|s}$.

Definition 4.1 (Representation Divergence)**.** *The Representation Divergence between contexts exhibiting properties s_1 and s_2 is the statistical divergence D between the corresponding conditional hidden state measures:*

$$\begin{aligned} \textit{RepDiv}_D(s_1 \| s_2) &\coloneqq D_{\mathcal{H}}(\mu_{H_t|s_1} \| \mu_{H_t|s_2}) \\ &\equiv D_{\mathcal{H}}(p_{H_t|s_1} \| p_{H_t|s_2}). \end{aligned} \quad (11)$$

Interpretation. A large $\text{RepDiv}_D(s_1 \| s_2)$ implies H_t effectively distinguishes properties s_1, s_2. A small value suggests similar representations. The choice of D affects the notion of distinguishability.

A large value indicates that the measures $\mu_{H_t|s_1}$ and $\mu_{H_t|s_2}$ are highly distinguishable according to the chosen divergence D. This implies that the transformation $(k_{\text{bb}} \circ k_{\text{emb}})$ maps contexts with properties s_1 and s_2 to significantly different distributions in the representation space $\mathcal{H}$. The hidden state H_t thus serves as an effective statistical signature for distinguishing between properties s_1 and s_2. (Estimation details are discussed in Appendix B).

4.2. Metric 2: Categorical Mutual Information (Statistical Dependencies)

To measure the strength of statistical dependence between key random variables using the intrinsic definition I_D (Equation (3)).

Definition 4.2 (State-Prediction Dependence ($I_D(H_t; W_t)$))**.** *The joint state $p_{H_t,W_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H})) \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ representing (H_t, W_t) is obtained categorically:*

$$p_{H_t,W_t} \coloneqq (\text{id}_{\mathcal{H}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}} \circ p_{H_t}. \quad (12)$$

The categorical mutual information is:

$$\begin{aligned} I_D(H_t; W_t) &\coloneqq I_D(p_{H_t,W_t}) \\ &\equiv D_{\mathcal{H}\otimes\mathcal{V}}(p_{H_t,W_t} \quad \| \quad p_{H_t} \otimes p_{W_t}). \end{aligned} \quad (13)$$

If $D = D_{\text{KL}}$, this recovers Shannon mutual information $I(H_t; W_t)$.

Definition 4.3 (Temporal State Dependence ($I_D(H_t; H_{t+1})$))**.** *Modeling the transition $H_t \to H_{t+1}$ involves generating W_t and re-encoding. This defines an effective transition kernel $\bar{k}_{step} : \mathcal{H} \to \mathcal{H}$. The joint state $p_{H_t,H_{t+1}} : I \to \mathcal{H} \otimes \mathcal{H}$ is:*

$$p_{H_t,H_{t+1}} \coloneqq (\text{id}_{\mathcal{H}} \otimes \bar{k}_{step}) \circ \Delta_{\mathcal{H}} \circ p_{H_t}. \quad (14)$$

The temporal statistical dependence is:

$$\begin{aligned} I_D(H_t; H_{t+1}) &\coloneqq I_D(p_{H_t,H_{t+1}}) \\ &\equiv D_{\mathcal{H}\otimes\mathcal{H}}(p_{H_t,H_{t+1}} \quad \| \quad p_{H_t} \otimes p_{H_{t+1}}). \end{aligned} \quad (15)$$

Interpretation. $I_D(H_t; W_t)$ quantifies the average amount of information (relative to divergence D) that the hidden state H_t provides about the immediately following token W_t. A high value suggests H_t strongly constrains the distribution over W_t, indicating high predictive relevance. $I_D(H_t; H_{t+1})$ measures the average statistical dependency between consecutive hidden states. A high value implies that the state H_{t+1} is highly predictable from H_t, suggesting the model maintains and evolves contextual information coherently over time. Low values might indicate information loss or abrupt changes in representation between time steps. (Estimation details are discussed in Appendix B).

4.3. Metric 3: LM Head Categorical Entropy (Prediction Stochasticity)

To quantify the intrinsic stochasticity of the LM head kernel $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ using categorical entropy (Equation (2)).

Definition 4.4 (Categorical Entropy of k_{head})**.**

$$\mathcal{H}_D(k_{\text{head}}) \coloneqq D_{\mathcal{V}\otimes\mathcal{V}}\left(\Delta_{\mathcal{V}} \circ k_{\text{head}} \| (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}\right). \quad (16)$$

This compares generating (W, W) vs (W_1, W_2) where $W, W_1, W_2 \overset{i.i.d.}{\sim} k_{\text{head}}(h, \cdot)$.

A practical measure is the average categorical entropy over the input distribution p_{H_t}:

$$\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t}) \coloneqq \mathbb{E}_{h\sim p_{H_t}}\Big[\quad (17)$$

$$D_{\mathcal{V}\otimes\mathcal{V}}\Big(\sum_{w\in\mathcal{V}} k_{\text{head}}(h, \{w\})\delta_{(w,w)} \quad \| k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)\Big)\Big]. \quad (18)$$

Interpretation. Measures the average conditional stochasticity or "spread" of the output distribution $k_{\text{head}}(h, \cdot)$. A value of 0 indicates a deterministic mapping. Higher values indicate greater average uncertainty. For $D = D_{\text{KL}}$, it relates closely to the average conditional Shannon entropy $\mathbb{E}_h[H(k_{\text{head}}(h, \cdot))]$.

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 $\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t})$ indicates greater average uncertainty or spread in the output distribution $p_h = k_{\text{head}}(h, \cdot)$, meaning

the kernel is inherently more stochastic. It quantifies how far the prediction process is from a deterministic assignment, measured in the geometry of $\mathcal{V} \otimes \mathcal{V}$ induced by D. (Estimation details are discussed in Appendix B).

4.4. Metric 4: Information Flow Bounds via Data Processing Inequality

To leverage the DPI to bound information flow about a context property S to the output token W_t. The sequence $S \to \mathbf{w}_{<t} \to E_{<t} \to H_t \to W_t$ forms a Markov chain $S \to H_t \to W_t$. The joint state p_{S,W_t} is obtained from p_{S,H_t} via the kernel $\mathrm{id}_S \otimes k_{\mathrm{head}}$: $p_{S,W_t} = (\mathrm{id}_S \otimes k_{\mathrm{head}}) \circ p_{S,H_t}$.

Theorem 4.5 (Categorical Information Flow Bound)**.** *Let $I_D(S;X) \coloneqq D_{S\otimes X}(p_{S,X} \| p_S \otimes p_X)$. The DPI applied to the definition of I_D implies:*

$$I_D(S;H_t) \geq I_D(S;W_t). \tag{19}$$

Proof. We want to show $I_D(S;H_t) \geq I_D(S;W_t)$. Recall $I_D(S;X) \coloneqq D_{S\otimes X}(p_{S,X} \| p_S \otimes p_X)$. The kernel relating the joint states is $k = \mathrm{id}_S \otimes k_{\mathrm{head}} : S \otimes \mathcal{H} \to S \otimes \mathcal{V}$. We have $p_{S,W_t} = k \circ p_{S,H_t}$. Also, $p_S \otimes p_{W_t} = p_S \otimes (k_{\mathrm{head}} \circ p_{H_t}) = (\mathrm{id}_S \otimes k_{\mathrm{head}}) \circ (p_S \otimes p_{H_t}) = k \circ (p_S \otimes p_{H_t})$. The DPI states that for any kernel $k : A \to B$ and states $p, q : I \to A$, we have $D_B(k \circ p \| k \circ q) \leq D_A(p\|q)$. Apply this with $A = S \otimes \mathcal{H}$, $B = S \otimes \mathcal{V}$, $k = \mathrm{id}_S \otimes k_{\mathrm{head}}$, $p = p_{S,H_t}$, and $q = p_S \otimes p_{H_t}$. Then $k \circ p = p_{S,W_t}$ and $k \circ q = p_S \otimes p_{W_t}$. The inequality becomes: $D_{S\otimes\mathcal{V}}(p_{S,W_t} \| p_S \otimes p_{W_t}) \leq D_{S\otimes\mathcal{H}}(p_{S,H_t} \| p_S \otimes p_{H_t})$. By definition, this is $I_D(S;W_t) \leq I_D(S;H_t)$. □

Interpretation. Information about S present in the representation H_t cannot be increased by the final prediction step k_{head}. The difference $I_D(S;H_t) - I_D(S;W_t) \geq 0$ quantifies information loss about S at this stage.

This fundamental inequality asserts that the amount of statistical information (measured by I_D) that the next token W_t carries about the context property S cannot exceed the amount of information about S that is already encoded in the intermediate hidden representation H_t. The final stochastic step $k_{\mathrm{head}} : H_t \to W_t$ can only preserve or lose information about S; it cannot create it. The difference $I_D(S;H_t) - I_D(S;W_t) \geq 0$ quantifies the information about S that is present in the representation H_t but is "lost" or not utilized in the immediate prediction of W_t. This loss could be due to the inherent stochasticity of k_{head} (as measured by $\mathcal{H}_D(k_{\mathrm{head}})$) or because the mapping discards aspects of H_t relevant to S but not relevant for predicting W_t. This unused information might still be crucial for predicting subsequent tokens $(W_{t+1}, \dots)$. (Estimation details are discussed in Appendix B).

5. Pretraining Objective, Compression, and Learning Intrinsic Stochasticity

The NLL objective $\min L_{\mathrm{CE}}(\theta) = -\mathbb{E}_{(\mathbf{w}_{<t},w_t)\sim P_{\text{data}}}[\log P_\theta(w_t|\mathbf{w}_{<t})]$ drives AR LM training. Let k_{data} be the true data kernel and $k_{\mathrm{gen},\theta}$ the model kernel.

Theorem 5.1 (NLL Minimization as Average KL Minimization)**.** *Minimizing $L_{\mathrm{CE}}(\theta)$ is equivalent to minimizing the average KL divergence:*

$$\begin{aligned}\arg\min_\theta L_{\mathrm{CE}}(\theta) &= \arg\min_\theta \mathcal{L}_{\mathrm{KL}}(\theta) \\ &\coloneqq \arg\min_\theta \mathbb{E}_{\mathbf{w}_{<t}\sim p_{W_{<t}}}[D_{\mathrm{KL}}(k_{\mathrm{data}}(\mathbf{w}_{<t},\cdot) \\ &\qquad \| \, k_{\mathrm{gen},\theta}(\mathbf{w}_{<t},\cdot))] \end{aligned} \tag{20}$$

The minimum $\mathcal{L}_{\mathrm{KL}}(\theta) \geq 0$ is achieved iff $k_{\mathrm{gen},\theta}(\mathbf{w}_{<t},\cdot) = k_{data}(\mathbf{w}_{<t},\cdot)$ almost everywhere.

Proof. 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 $q_{x,\theta}(\cdot) = k_{\mathrm{gen},\theta}(x,\cdot)$ 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:

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

The average KL divergence is defined as:

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

where $H(p_x(\cdot)) = -\sum_w p_x(w)\log p_x(w)$ 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 $L_{\text{CE}}(\theta)$ 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 $\mathcal{L}_{\text{KL}}(\theta) = 0$ is achieved if and only if the integrand is zero $p_{W_{<t}}$-almost everywhere. That is, $D_{\text{KL}}(p_x(\cdot) \,\|\, q_{x,\theta^*}(\cdot)) = 0$ 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 $\mathcal{L}_{\text{KL}}(\theta^*) = 0$, which is the minimum possible value. □

This frames training as density estimation, forcing the model kernel to match the data kernel. Via source coding theory, minimizing $L_{\text{CE}}(\theta)$ corresponds to finding a model that provides efficient data compression, approaching the conditional entropy $H(W_t|W_{<t})_{\text{data}}$. Success implies the model must also replicate the intrinsic stochasticity of k_{data}, quantifiable via average categorical entropy (Equation (18)).

Theorem 5.2 (Convergence of Average Categorical Entropy via NLL Minimization)**.** *Under suitable convergence conditions (pointwise kernel convergence, weak state convergence, divergence continuity), if $\mathcal{L}_{\text{KL}}(\theta_n) \to \inf_\theta \mathcal{L}_{\text{KL}}(\theta)$, then the average categorical entropy of the learned LM head converges to that of the optimal head: $\lim_{n\to\infty} \bar{\mathcal{H}}_D(k_{head,n}; p_{H_t,\theta_n}) = \bar{\mathcal{H}}_D(k_{head,\theta^*}; p_{H_t,\theta^*})$. If $\mathcal{L}_{\text{KL}}(\theta^*) = 0$, the learned entropy approximates the true data's intrinsic conditional stochasticity. (Proof in Appendix A.1).*

Theorem 5.2 formalizes that NLL training forces the model to learn the correct degree of uncertainty dictated by the data, measured by $\bar{\mathcal{H}}_D$. This is integral to the compression process.

6. Information Geometry of Representation and Prediction Spaces

The Markov Category $(\mathbf{Stoch}, D_{\text{KL}})$ connects naturally to Information Geometry (Amari & Nagaoka, 2000; Perrone, 2023b). The space of next-token distributions $\mathcal{P}(\mathcal{V})$ is a simplex $\Delta^{|\mathcal{V}|-1}$ with Fisher-Rao metric g^{FR}. The LM Head kernel corresponds to a map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$, typically $g_{\text{head}}(h) = \text{softmax}(Wh)$. This map pulls back the Fisher-Rao metric g^{FR} from $\mathcal{P}(\mathcal{V})$ to a (possibly degenerate) metric $g^* = g_{\text{head}}^* g^{\text{FR}}$ on $\mathcal{H}$. Let $J(h)$ be the Jacobian of g_{head}. Then $g^*(h) = J(h)^\top g^{\text{FR}}(g_{\text{head}}(h))J(h)$.

Theorem 6.1 (Pullback Metric and Local Divergence)**.** *For $h \in \mathcal{H}$ and $v \in T_h\mathcal{H}$, the local KL divergence is related to the pullback metric:*

$$D_{\text{KL}}(g_{head}(h + \epsilon v) \,\|\, g_{head}(h)) = \frac{1}{2}\epsilon^2 g^*(h)(v, v) + O(\epsilon^3) \tag{21}$$

where $g^*(h)(v, v) = v^\top g^*(h)v$. (Proof follows standard pullback arguments, see Appendix A.2 in Appendix).

This shows $g^*(h)$ measures the local sensitivity of the output distribution p_h to changes in h, quantified by KL divergence.

Remark 6.2 (Connection to Score Function)**.** *The pullback metric can be expressed via the score function $\nabla_h \log p_h(W)$: $g^*(h) = \mathbb{E}_{W\sim p_h}[(\nabla_h \log p_h(W))(\nabla_h \log p_h(W))^\top]$.*

Proposition 6.3 (Rank of the Pullback Metric)**.** $\text{rank}(g^*(h)) \leq \min(d_{\text{model}}, |\mathcal{V}| - 1)$.

Remark 6.4 (Typical Rank and Degeneracy)**.** *In typical LLMs, $d_{\text{model}} \ll |\mathcal{V}|$. If the Jacobian $J(h)$ has rank d_{model}, then $g^*(h)$ is usually non-degenerate (full rank d_{model}). The geometry is anisotropic, meaning sensitivity varies with direction v, but there isn't a significant null space of directions irrelevant to prediction.*

Implications. The pullback metric g^* characterizes the functional geometry of $\mathcal{H}$. Its spectrum reveals principal directions of predictive sensitivity. Directions v with large $g^*(h)(v, v)$ are those where changes in h strongly affect the output distribution p_h. Training shapes the encoder and head to structure this manifold $(\mathcal{H}, g^*)$, separating representations based on their predictive consequences.

7. Implicit Spectral Structuring via NLL Optimization

Why does minimizing NLL (Equation (20)) yield structured representations $h_t = f_{\text{enc}}(x)$? We argue NLL implicitly imposes geometric constraints related to spectral methods (HaoChen et al., 2021; Tan et al., 2024). Let $p_\theta(\cdot|x) = g_{\text{head}}(f_{\text{enc}}(x))$ and $P_{\text{data}}(\cdot|x)$ be the target.

Theorem 7.1 (Output Distribution Approximation Constraint)**.** *If $\mathcal{L}(\theta) = \mathbb{E}_x[D_{\text{KL}}(P_{data}(\cdot|x)\|p_\theta(\cdot|x))]$ is small, then for metrics d_{out} satisfying $d_{out}(p, q)^k \leq C \cdot D_{\text{KL}}(p\|q)$ (e.g., Hellinger, TV), the average output distance is small:*

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{out}(P_{data}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathcal{L}(\theta). \tag{22}$$

By the triangle inequality, this implies

$$\begin{aligned} d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x')) \leq\ & d_{out}(P_{data}(\cdot|x), p_\theta(\cdot|x)) \\ & + d_{out}(p_\theta(\cdot|x), p_\theta(\cdot|x')) + d_{out}(p_\theta(\cdot|x'), P_{data}(\cdot|x')), \end{aligned} \tag{23}$$

and $d_{out}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \approx d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x'))$. *(Proof in Appendix A.3).*

This theorem shows NLL forces model outputs to mirror the pairwise distance structure of the true data distributions.

Definition 7.2 (Sensitive Directions and Head Sensitivity)**.** *A direction* $v \in \mathcal{H}$ *is sensitive for* g_{head} *at* h *if the pullback metric* $g^*(h)(v, v) > 0$. g_{head} *is sufficiently sensitive if* $d_{out}(g_{head}(h_1), g_{head}(h_2)) > \delta$ *implies* $(h_1 - h_2)$ *has significant components along sensitive directions.*

Corollary 7.3 (Implicit Representation Separation)**.** *Assume conditions of Thm. 7.1 and Def. 7.2 hold. If contexts* x, x' *are predictively dissimilar (*$d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x'))$ *large), their representations* $h_x, h_{x'}$ *must differ along sensitive directions for* g_{head}*. Conversely, NLL does not strongly constrain the distance between* $h_x, h_{x'}$ *for predictively similar contexts. (Proof in Appendix A.4).*

This corollary shows NLL implicitly separates representations based on predictive dissimilarity along head-sensitive dimensions.

Definition 7.4 (Predictive Similarity Kernel)**.** *A kernel* $K(x, x') \geq 0$ *measures similarity between* $P_{data}(\cdot|x)$ *and* $P_{data}(\cdot|x')$ *(e.g., based on Bhattacharyya coefficient* K_{BC}*, Hellinger distance* K_H*, or Expected Likelihood* K_{Lin}*). High* $K(x, x')$ *means high predictive similarity.*

Definition 7.5 (Graph Laplacian and Dirichlet Energy)**.** *Let* $K(x, x')$ *be a predictive similarity kernel. The graph Laplacian* Δ_K *acts on* $\phi : \mathcal{V}^* \to \mathbb{R}$*. 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.*

Proposition 7.6 (NLL Objective and Implicit Dirichlet Energy Minimization)**.** *Let* $\phi_v(x) = \langle f_{\text{enc}}(x), v \rangle$ *be the projection onto a sensitive direction* v*. Minimizing NLL* $\mathcal{L}(\theta)$ *implicitly encourages representations such that* $\mathcal{E}_K(\phi_v)$ *tends to be small for sensitive directions* v*, by pushing representations closer when* $K(x, x')$ *is high. (Proof in Appendix A.4.1).*

This links NLL to minimizing Dirichlet energy on the predictive similarity graph along head-sensitive dimensions, akin to spectral clustering.

Definition 7.7 (Predictive Similarity Operator)**.** *The operator* M_K *averages functions* $\psi(h_{x'})$ *over contexts* x'*, weighted by* $K(x, x')$*:*

$$(M_K\psi)(h_x) \triangleq \mathbb{E}_{x' \sim \mu_{ctx}}[K(x, x')\psi(h_{x'})] = \int_{\mathcal{V}^*} K(x, x')\psi(f_{\text{enc}}(x'))\, \mu_{ctx}(\mathrm{d}x'). \quad (24)$$

Its eigenspace captures patterns of predictive similarity.

Hypothesis 7.8 (NLL Objective and Alignment with Operator Eigenspace)**.** *Minimizing NLL likely aligns representations with the eigenspace of* M_K*. It may compress representations along directions associated with large eigenvalues (high similarity) but primarily along components less sensitive to the head (*$g^*(h)(v, v)$ *small), while preserving variance along head-sensitive components needed to distinguish dissimilar contexts. (Argument in Appendix A.5).*

In summary, NLL optimization implicitly enforces geometric constraints mirroring spectral methods on the predictive similarity graph, pushing representations of predictively similar contexts together along relevant dimensions, thus providing a basis for understanding how NLL induces structured representations.

8. Conclusion

We introduced a Markov Categorical framework for analyzing AR language model generation steps, modeling the process as a kernel composition $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in $\mathbf{Stoch}$. Using divergence enrichment (D) and categorical information measures ($\mathcal{H}_D, I_D$), we defined metrics for Representation Divergence, State-Prediction/Temporal Information, Head Stochasticity, and Information Flow bounds via DPI. This framework interprets NLL minimization as equivalent to average KL divergence minimization (Section 5), linking it to compression and learning the data's intrinsic stochasticity ($\bar{\mathcal{H}}_D$). Information geometry (Section 6) analyzes the functional geometry of $\mathcal{H}$ via the pullback metric g^*, revealing predictive sensitivities. We formalized NLL as implicit structure learning (Section 7), demonstrating it imposes geometric constraints separating representations based on predictive dissimilarity, connecting it formally to spectral methods on predictive similarity graphs.

Impact Statement

This work is primarily theoretical, aiming to establish a mathematical framework for analyzing the internal mechanisms of autoregressive language models. The principal positive impact of this research is the advancement of fundamental AI understanding, which is crucial for building more reliable and transparent systems. Our framework and the proposed metrics offer a principled approach to: 1) Providing tools to dissect information flow and representation geometry, allowing researchers to better understand how models make predictions; 2) A deeper theoretical grasp can inform the development of more robust, efficient, and capable future architectures; 3) Metrics like Representation Divergence could be used to formally audit models for biases or to quantify how they represent sensitive concepts. We do not foresee direct negative societal consequences from this theoretical work.

References

Amari, S.-i. and Nagaoka, H. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Baez, J. C., Fong, B., and Pollard, B. S. A compositional framework for markov processes. *Journal of Mathematical Physics*, 57(3), 2016.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 2018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pp. 1877–1901, 2020.

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

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits, 2021.

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

HaoChen, J. Z., Chen, H., Wei, C., Gaidon, A., and Ma, T. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 14239–14250, 2021.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

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

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

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

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pérez-Cruz, F. Estimation of information theoretic measures for continuous random variables. *Advances in neural information processing systems*, 21, 2008.

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

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

Wang, Q., Kulkarni, S. R., and Verdú, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

A. Proofs and Arguments

A.1. Proof of Theorem 5.2 (Convergence of Average Categorical Entropy)

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

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 $\Psi_D(h,p) := D_{\mathcal{V}\otimes\mathcal{V}}(\sum_{w\in\mathcal{V}} p(w)\delta_{(w,w)} \,\|\, p\otimes p)$, 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 $\mathcal{V}$ is finite, standard divergences like KL and TV are continuous functions of the probability vectors $p \in \Delta^{|\mathcal{V}|-1}$. 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:

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

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 (sometimes related to uniform integrability or Pratt's Lemma). 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 $\lim_{n\to\infty}\int |\Phi_n(h) - \Phi^*(h)|p_{H_t,\theta_n}(dh) = 0$. (A rigorous justification might need Skorokhod's representation theorem or related results, but under these conditions, this convergence generally holds).

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

A.2. Proof for Theorem 6.1 (Pullback Metric and Local Divergence)

Let ξ be local coordinates for $\mathcal{P}(\mathcal{V})$ around $p_h = g_{\text{head}}(h)$. The KL divergence between nearby $p_{\xi'}$ and p_ξ is $D_{\text{KL}}(p_{\xi'} \,\|\, p_\xi) = \frac{1}{2}\sum_{i,j} g^{\text{FR}}_{ij}(\xi)(\xi'_i - \xi_i)(\xi'_j - \xi_j) + O(\|\xi' - \xi\|^3)$, where g^{FR} is the Fisher-Rao metric in these coordinates (Amari & Nagaoka, 2000). Let $\xi(h)$ be the coordinates of $p_h = g_{\text{head}}(h)$. For $p_{h+\epsilon v} = g_{\text{head}}(h + \epsilon v)$, the coordinates are $\xi(h + \epsilon v)$. By Taylor expansion:

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

Let $J(h)$ be the Jacobian matrix of the map $h \mapsto \xi(h)$, with entries $J_{ia}(h) = \frac{\partial \xi_i}{\partial h_a}(h)$. Then $\xi_i(h + \epsilon v) - \xi_i(h) = \epsilon \sum_a J_{ia}(h) v_a + O(\epsilon^2) = \epsilon (J(h)v)_i + O(\epsilon^2)$. Substituting into the KL expansion:

$$\begin{aligned}
D_{\text{KL}}(p_{h+\epsilon v} \,\|\, p_h) &\approx \frac{1}{2} \sum_{i,j} g^{\text{FR}}_{ij}(\xi(h))(\epsilon (J(h)v)_i)(\epsilon (J(h)v)_j) \\
&= \frac{1}{2}\epsilon^2 \sum_{i,j} (J(h)v)_i g^{\text{FR}}_{ij}(\xi(h))(J(h)v)_j + O(\epsilon^3) \\
&= \frac{1}{2}\epsilon^2 (J(h)v)^\top g^{\text{FR}}(\xi(h))(J(h)v) \\
&= \frac{1}{2}\epsilon^2 v^\top J(h)^\top g^{\text{FR}}(\xi(h)) J(h) v
\end{aligned}$$

The term $J(h)^\top g^{\text{FR}}(\xi(h)) J(h)$ is the definition of the pullback metric $g^*(h)$ in the standard basis of $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$. Thus, $D_{\text{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2}\epsilon^2 v^\top g^*(h) v + O(\epsilon^3) = \frac{1}{2}\epsilon^2 g^*(h)(v, v) + O(\epsilon^3)$. □

A.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 $\mathcal{P}(\mathcal{V})$ 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, Equation (22).

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:

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

This is Equation (23) in the main text. 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. From Equation (22), $\mathbb{E}_{x\sim\mu_{ctx}}[\epsilon_x^k] \leq C\mathcal{L}(\theta)$, 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:

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

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. □

A.4. Proof of Corollary 7.3 (Implicit Representation Separation)

We assume the two conditions hold: (i) $\mathcal{L}(\theta)$ is sufficiently small such that for typical x, x', the errors $\epsilon_x = d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))$ and $\epsilon_{x'} = d_{\text{out}}(p_\theta(\cdot|x'), P_{\text{data}}(\cdot|x'))$ are negligible compared to $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$. (ii) The head mapping $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ is sufficiently sensitive: if $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) > \delta > 0$ for $h_x, h_{x'}$ in the populated region, then h_x and $h_{x'}$ must differ along directions sensitive to g_{head}. These sensitive directions span the subspace orthogonal to the null space of the Jacobian $J_{g_{\text{head}}}(h)$, which corresponds to the support of the pullback metric $g^*(h)$ (Section 6).

From the proof of Theorem 7.1, under assumption (i), we have:

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

Substitute $p_\theta(\cdot|y) = g_{\text{head}}(h_y)$ where $h_y = f_{\text{enc}}(y)$:

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

Now, consider the case where contexts x, x' are predictively dissimilar, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large. Specifically, assume it is significantly larger than the approximation error $(\epsilon_x + \epsilon_{x'})$ and also larger than the sensitivity threshold δ from assumption (ii). Then, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be large, and in particular, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) > \delta$.

By assumption (ii), if the distance between the outputs $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ exceeds the threshold δ, then the inputs h_x and $h_{x'}$ must differ along directions to which g_{head} is sensitive. Therefore, we conclude that if $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, then h_x and $h_{x'}$ must differ along the sensitive directions for g_{head} (as defined in Definition 7.2).

Conversely, consider the case where x, x' are predictively similar, i.e., $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small (e.g., close to zero). Then, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be small. This condition ($g_{\text{head}}(h_x)$ close to $g_{\text{head}}(h_{x'})$) can potentially be satisfied even if h_x and $h_{x'}$ differ significantly, provided their difference lies primarily within the null space of the Jacobian of g_{head} (directions insensitive to the head). However, the condition does not require h_x and $h_{x'}$ to be far apart along sensitive directions. In fact, mapping them closely together along sensitive dimensions is consistent with achieving a small $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ and thus satisfying the NLL objective constraint in this case. Therefore, NLL optimization does not strongly constrain the distance between h_x and $h_{x'}$ along relevant dimensions when contexts are predictively similar. □

A.4.1. Proof of Proposition 7.6 (NLL Objective and Implicit Dirichlet Energy Minimization)

We are given a sensitive direction v (in the support of $g^*(h)$) and the projection $\phi_v(x) = \langle h_x, v\rangle$, where $h_x = f_{\text{enc}}(x)$. The Dirichlet energy with respect to a predictive similarity kernel $K(x, x')$ is:

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

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

We assume $\mathcal{L}(\theta)$ is small, and the conditions of Corollary 7.3 hold.

Consider a pair of contexts (x, x') where the predictive similarity $K(x, x')$ is high. By Definition 7.4, high $K(x, x')$ implies that the true conditional distributions $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$ are similar, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small. From the converse part of Corollary 7.3, when $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small, the NLL objective does not force h_x and $h_{x'}$ apart along sensitive directions v. In fact, to ensure that $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$ is close to $p_\theta(\cdot|x') = g_{\text{head}}(h_{x'})$, which is required to approximate the small distance between $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$, the representations h_x and $h_{x'}$ are encouraged to be close along these sensitive directions v. That is, if $K(x, x')$ is large, minimizing NLL encourages $\langle h_x - h_{x'}, v\rangle$ to be small for sensitive v.

Now examine the integral defining $\mathcal{E}_K(\phi_v)$. The integrand is $K(x, x')(\langle h_x - h_{x'}, v\rangle)^2$. This term makes a significant contribution only when $K(x, x')$ is large (otherwise the factor $K(x, x')$ makes it small) and $(\langle h_x - h_{x'}, v\rangle)^2$ is large. However, we just argued that minimizing NLL exerts pressure such that when $K(x, x')$ is large, the term $(\langle h_x - h_{x'}, v\rangle)^2$ tends to be small for sensitive directions v.

Therefore, NLL minimization actively discourages configurations where the integrand is large for the pairs (x, x') that contribute most due to high $K(x, x')$. This means the optimization process implicitly favors representations h_x such that the overall integral $\mathcal{E}_K(\phi_v)$ is small for projections ϕ_v onto directions v sensitive to the prediction head. While this is not a direct minimization of $\mathcal{E}_K(\phi_v)$, the pressure exerted by NLL aligns with reducing the terms that dominate the Dirichlet energy integral, thus implicitly favoring lower energy configurations along predictively relevant dimensions. □

A.5. Argument of Hypothesis 7.8 (NLL Objective and Alignment with Operator Eigenspace)

This argument remains more interpretative, formalizing the sketch. We assume the setup: encoder f_{enc}, head g_{head}, predictive similarity kernel K, similarity operator M_K (Equation (24)) with eigenfunctions $\{\phi_i\}$ and eigenvalues $\{\lambda_i\}$. Assume Corollary 7.3 holds.

The operator M_K acts on functions ψ defined on the representation space $\mathcal{H}$. Its eigenfunctions ϕ_i represent directions or patterns in $\mathcal{H}$ that are stable under averaging weighted by predictive similarity. A large eigenvalue λ_i signifies that the corresponding eigenfunction ϕ_i captures a dominant structure of predictive similarity: contexts x whose representations h_x have high values of $\phi_i(h_x)$ tend to be predictively similar to other contexts x' whose representations $h_{x'}$ also have high values of $\phi_i(h_{x'})$.

From Corollary 7.3, minimizing NLL requires:

- If $K(x, x')$ is low (dissimilar predictions), then h_x and $h_{x'}$ must differ along sensitive directions for g_{head}.
- If $K(x, x')$ is high (similar predictions), then h_x and $h_{x'}$ are allowed (and encouraged) to be close along sensitive directions for g_{head}.

Consider directions u in $\mathcal{H}$ that are strongly correlated with eigenfunctions ϕ_i having large eigenvalues λ_i. These directions capture clusters or variations associated with high predictive similarity. For contexts x, x' within such a cluster (high $K(x, x')$), NLL allows their representations $h_x, h_{x'}$ to be close along the sensitive components of u.

Now, invoke a principle of representational efficiency or compression (Section 5). A model minimizing prediction error (NLL) might also implicitly seek compact representations, discarding information not necessary for the immediate task. Dimensions in $\mathcal{H}$ that are insensitive to the head g_{head} (i.e., directions w where $g^*(h)(w, w) \approx 0$) carry information not used for the next-token prediction $p_\theta(\cdot|x)$. These directions are precisely those that are not sensitive according to Definition 7.2. As g^* is likely non-degenerate (Remark 6.4), the insensitive subspace is trivial. The relevant distinction is between directions v where $g^*(h)(v, v)$ is large (highly sensitive) vs small (weakly sensitive).

Consider the variation of representations $\{h_x\}$ along a direction u associated with a large eigenvalue λ_i. This variation reflects differences among contexts that are generally predictively similar. NLL requires $g_{\text{head}}(h_x)$ to be close for these contexts. This allows h_x to be close along highly sensitive directions (large $g^*(h)(v,v)$). An efficient model might compress representations by reducing variance along directions where variation is less critical for NLL. This could preferentially target directions v where $g^*(h)(v,v)$ is small (weakly sensitive directions), as variation along these directions has less impact on the output distribution $p_\theta(\cdot|x)$ and thus the NLL loss.

This leads to an implicit alignment: directions u associated with high predictive similarity (large λ_i) may exhibit reduced variance along components v where $g^*(h)(v,v)$ is small. Conversely, directions needed to distinguish predictively dissimilar contexts (low $K(x,x')$) must maintain variance along components v where $g^*(h)(v,v)$ is large, as this is necessary to separate their output distributions $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ as required by NLL.

This behavior mirrors the outcome of spectral contrastive learning, where representations are collapsed along directions of assumed similarity (analogous to large λ_i) while preserving discriminative information. While NLL optimization doesn't explicitly target the spectrum of M_K, the combined pressure of accurate prediction and potential representational compression leads to a structure where the geometry of $\mathcal{H}$ implicitly reflects the eigenspectrum of the predictive similarity operator, particularly in how variance is distributed between head-sensitive and head-insensitive dimensions. A fully rigorous proof connecting NLL optimization dynamics directly to the spectrum of M_K would require stronger assumptions about the optimization process and the model's implicit biases towards compression.

B. Estimation Details for Metrics

This section provides more details on the practical estimation of the metrics defined in Section 4. Estimation generally relies on Monte Carlo methods, drawing samples from the relevant distributions and using statistical estimators.

Metric 1: Representation Divergence RepDiv $_D(s_1\|s_2) = D_{\mathcal{H}}(\mu_{H_t|s_1}\|\mu_{H_t|s_2})$

- *Sampling*: Draw N_1 contexts $\{\mathbf{w}_{<t}^{(i)}\}_{i=1}^{N_1} \sim P_{\text{ctx}}(\cdot|s_1)$ and N_2 contexts $\{\mathbf{w}_{<t}^{(j)}\}_{j=1}^{N_2} \sim P_{\text{ctx}}(\cdot|s_2)$. Compute the corresponding hidden states $\{h_t^{(i)}\}_{i=1}^{N_1}$ and $\{h_t^{(j)}\}_{j=1}^{N_2}$ using $f_{\text{enc}} = f_{\text{bb}} \circ f_{\text{emb}}$.
- *Estimators for D*:
  - *kNN-based*: For KL or Rényi divergences, estimators based on distances to k-nearest neighbors in the pooled sample can be used (Wang et al., 2009; Pérez-Cruz, 2008). These are non-parametric but require large N_1, N_2 especially for high d_{model} (curse of dimensionality).
  - *Variational (f-GAN / MINE)*: Use neural networks to estimate density ratios or bounds based on variational principles (Nowozin et al., 2016; Belghazi et al., 2018). For KL divergence using the Donsker-Varadhan representation ($D_{\text{KL}}(\mu\|\nu) = \sup_T(\mathbb{E}_\mu[T] - \log \mathbb{E}_\nu[e^T])$), a neural network $T_\phi : \mathcal{H} \to \mathbb{R}$ is trained to maximize a sample-based lower bound: $\sup_\phi \left(\frac{1}{N_1}\sum_i T_\phi(h_t^{(i)}) - \log\left(\frac{1}{N_2}\sum_j e^{T_\phi(h_t^{(j)})}\right)\right)$. These can potentially handle higher dimensions better but introduce optimization challenges and approximation errors based on the capacity of T_ϕ.

Metric 2: Mutual Information $I_D(H_t; W_t)$ **and** $I_D(H_t; H_{t+1})$

- *Sampling*: Generate sequences by iteratively sampling contexts, computing states, sampling next tokens, and computing next states. Collect sample pairs $(h_t^{(i)}, w_t^{(i)})$ for $I_D(H_t; W_t)$ and $(h_t^{(i)}, h_{t+1}^{(i)})$ for $I_D(H_t; H_{t+1})$.
- *Estimators for I_D (often focused on $D = D_{\text{KL}}$)*:
  - *kNN-based (KSG)*: The Kraskov-Stögbauer-Grassberger estimator (Kraskov et al., 2004) estimates Shannon MI $I(X;Y)$ based on nearest neighbor distances in the joint space $X \times Y$ compared to marginal spaces X and Y. It requires careful choice of k and metric on the spaces. For $I_D(H_t; W_t)$, the space $\mathcal{H} \times \mathcal{V}$ has mixed continuous-discrete nature. For $I_D(H_t; H_{t+1})$, the space is high-dimensional continuous $\mathcal{H} \times \mathcal{H}$. These estimators also suffer dimensionality issues.
  - *Variational (MINE/InfoNCE)*: Maximize lower bounds on MI using neural networks (Belghazi et al., 2018; Oord et al., 2018). MINE uses the Donsker-Varadhan representation of KL divergence applied to $I(X;Y) = D_{\text{KL}}(P_{XY}\|P_X P_Y)$. InfoNCE provides another lower bound often optimized via noise contrastive estimation. These methods generally handle higher dimensions better but depend on the expressiveness and training of the

auxiliary neural networks.

Metric 3: Average Categorical Entropy $\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t})$

- *Sampling*: Draw N contexts $\{\mathbf{w}_{<t}^{(i)}\}_{i=1}^N \sim P_{\text{ctx}}$, compute hidden states $\{h_t^{(i)}\}_{i=1}^N$.
- *Computation*: For each $h_t^{(i)}$: 1. Compute the output probability vector $p_{h_t^{(i)}} = k_{\text{head}}(h_t^{(i)}, \cdot)$ (via softmax). 2. Construct the two measures on $\mathcal{V} \times \mathcal{V}$: $\mu_1^{(i)} = \sum_w p_{h_t^{(i)}}(w)\delta_{(w,w)}$ and $\mu_2^{(i)} = p_{h_t^{(i)}} \otimes p_{h_t^{(i)}}$. 3. Compute the divergence $d^{(i)} = D_{\mathcal{V}\otimes\mathcal{V}}(\mu_1^{(i)} \| \mu_2^{(i)})$. Since $\mathcal{V} \times \mathcal{V}$ is a finite space, this computation is generally straightforward (e.g., a summation for KL divergence).
- *Averaging*: Estimate the average entropy as $\frac{1}{N}\sum_{i=1}^N d^{(i)}$. This Monte Carlo estimate converges as $N \to \infty$. Estimation is generally feasible as the high-dimensional part only involves sampling h_t, and the divergence is computed on the low-dimensional space $\mathcal{V} \times \mathcal{V}$.

Metric 4: Information Flow Bound $I_D(S; H_t) \geq I_D(S; W_t)$

- *Sampling*: Requires samples $(s^{(i)}, h_t^{(i)})$ and $(s^{(i)}, w_t^{(i)})$. Obtain these by sampling contexts conditioned on s, computing h_t, and sampling w_t.
- *Estimation*:
  - $I_D(S; W_t)$: If S is discrete, this is MI between two discrete variables, estimable from contingency tables (if sample size is sufficient) or standard discrete MI estimators. If S is continuous, it's a mixed-type MI estimation problem, potentially simpler than involving H_t.
  - $I_D(S; H_t)$: This involves MI between S and the high-dimensional continuous hidden state H_t. Estimation requires robust methods suited for high dimensions (kNN or variational), considering whether S is discrete or continuous.

In all cases involving high-dimensional spaces ($\mathcal{H}$), the choice of estimator, its hyperparameters (like k in kNN or the architecture of variational networks), and the number of samples used are important considerations for obtaining reliable estimates