Transformers as Optimal Transport: A Geometric Framework for Representation Evolution

Abstract

We prove that transformer self-attention matrices are exactly the optimal solutions to semi-relaxed entropic optimal transport problems with unit regularization ($\varepsilon = 1$). This mathematical equivalence—not an approximation or analogy—reveals that attention mechanisms inherently solve a specific optimal transport problem where each query independently redistributes unit mass across keys. The semi-relaxed formulation is fundamental in the causal setting: enforcing column-sum constraints couples all rows, which conflicts with the autoregressive factorization that forbids future positions from influencing earlier ones. From this fundamental equivalence, we derive tight bounds showing that probability distributions induced by fixed probes evolve with total variation bounded by $\|W_{\text{out}}^{\top}\|_{2\to\infty}\|h^{(\ell+1)} - h^{(\ell)}\|_2$. Through comprehensive empirical analysis of GPT-2 models (124M–1.5B parameters), we validate these theoretical predictions and discover an unexpected saturation phenomenon: when softmax confidence exceeds 0.9999, representations lock completely (TV $< 10^{-10}$) while hidden states continue evolving by 2–9%, revealing a mechanism for separating decision certainty from continued computation. Our framework provides the first exact optimal transport characterization of attention, explaining fundamental constraints on transformer representation dynamics.

1. Introduction

Understanding how representations evolve through the depth of neural networks constitutes a fundamental challenge in deep learning theory. In transformer architectures [20], which have revolutionized natural language processing and beyond, representations at each layer are shaped by attention mechanisms that redistribute information across sequence positions. These attention operations, while appearing mechanistically simple as normalized dot products, encode rich geometric structure that governs how information can flow through the network.

Prior work has explored various mathematical interpretations of attention mechanisms [19] and some recent work has considered doubly-stochastic attention matrices related to balanced optimal transport [14]. We prove that standard attention *exactly equals* the optimizer of semi-relaxed entropic OT (row constraints only) with the specific regularization $\varepsilon = 1$. This exact mathematical equivalence, the semi-relaxed formulation, and the identification of unit regularization are all novel.

We demonstrate that this geometric structure corresponds precisely to optimal transport, a mathematical framework for understanding how probability mass redistributes under cost constraints. Specifically, we prove that standard row-softmax attention exactly solves a semi-relaxed entropic optimal transport problem where each query independently determines how to redistribute unit mass across keys. This identification is not merely a mathematical curiosity but rather provides powerful tools for characterizing how representations can and cannot change as they propagate through transformer layers.

1.1. The Challenge of Representation Dynamics

Deep transformers routinely employ dozens or even hundreds of layers, raising fundamental questions about how representations evolve through such extreme depths. Prior work has observed that transformer representations appear to change gradually across layers [13], with different layers specializing in different types of linguistic phenomena [18]. However, these empirical observations lack a principled theoretical framework that explains why representations evolve as they do and what constraints govern their evolution.

The optimal transport perspective we develop addresses this gap by revealing that attention imposes strict geometric constraints on representation change. When we probe hidden states $h^{(\ell)}$ at layer ℓ to induce probability distributions over a vocabulary through $p^{(\ell)} = \text{softmax}(W_{\text{out}}^{\top} h^{(\ell)})$, the evolution from $p^{(\ell)}$ to $p^{(\ell+1)}$ cannot be arbitrary. Instead, it must respect bounds derived from the transport structure, with the total variation between consecutive layers bounded by the product of the probe's operator norm and the hidden state change.

The semi-relaxed structure we identify is not a mathematical curiosity but fundamental to transformer operation. Balanced optimal transport would require adjusting representations of previously processed tokens to satisfy column constraints, violating the causal structure essential for autoregressive generation. This explains why transformers must use row-normalization (semi-relaxed) rather than doubly-stochastic attention (balanced). The identification of this necessary asymmetry provides insight into why certain architectural choices are not merely conventional but mathematically required for the transformer's computational model.

1.2. Key Empirical Phenomena

Our theoretical framework motivates a comprehensive empirical investigation of representation evolution in GPT-2 models spanning three orders of magnitude in parameter count. This analysis reveals three striking phenomena that characterize how transformers manage representations across depth.

First, we observe that alignment residuals, which measure how coherently representations evolve across positions, grow monotonically with depth but with sublinear scaling. When model depth doubles from 12 to 24 layers, drift increases by a factor of only 2.5, and doubling again to 48 layers yields just 1.6× additional growth. This sublinear scaling suggests that architectural mechanisms actively resist the linear accumulation of drift that naive analysis would predict.

Second, we discover a remarkable saturation phenomenon where extreme softmax confidence creates what we term representation locks. In approximately 10% of GPT-2-XL samples, when the top token probability exceeds 0.9999, the probability distribution effectively freezes with total variation below 10^{-10} between consecutive layers. Crucially, this occurs while hidden states continue evolving by 2–9% in relative norm, revealing a mechanism by which transformers decouple representational stability from ongoing computation.

Third, across 4800 layer transitions spanning all model scales, we observe perfect satisfaction of our theoretical Lipschitz bounds with zero violations. The maximum observed ratio between actual and theoretical bounds is 0.126, providing an 87% safety margin. This suggests that transformers naturally operate in regimes far from worst-case theoretical limits, potentially due to implicit regularization from gradient-based training.

1.3. Contributions and Organization

This work makes four primary contributions to understanding transformer representations, establishing the first exact mathematical connection between attention and optimal transport along with its theoretical and empirical consequences.

First, we establish that attention exactly solves a semi-relaxed entropic optimal transport problem, providing a principled geometric framework for analyzing representation evolution. The row-softmax attention weights are precisely the unique optimizer of an optimal transport problem with cost matrix $C = -QK^\top/T$ and unit entropic regularization $\varepsilon = 1$. This is not an approximation or analogy but an exact mathematical equivalence. The semi-relaxed formulation, which enforces only row constraints without column constraints, is fundamental: balanced optimal transport would require modifying representations of previously processed tokens, violating the causal structure essential for autoregressive generation.

Second, we derive tight, architecture-aware bounds on how probability distributions can change between layers, with explicit dependence on probe norms and hidden state perturbations. These bounds follow directly as mathematical consequences of the optimal transport structure, establishing that $\|p^{(\ell+1)} - p^{(\ell)}\|_1 \leq \|W_{\text{out}}^\top\|_{2\to\infty}\|h^{(\ell+1)} - h^{(\ell)}\|_2$ between consecutive layers. We further prove this bound is tight and characterize how architectural components including layer normalization, residual connections, and multi-head attention modulate these constraints through their respective Lipschitz constants.

Third, we identify and characterize the saturation-induced decoupling phenomenon that enables representational stability without computational stagnation. Through empirical analysis, we discovered that when softmax confidence exceeds 0.9999, probability distributions freeze completely with total variation below 10^{-10} while hidden states continue evolving by 2-9% in relative norm. This phenomenon, observed in approximately 10% of samples in deeper models, was not predicted by our theoretical framework and reveals additional structure beyond what optimal transport alone explains.

Fourth, we validate our theoretical predictions through comprehensive experiments, revealing scale-invariant principles that govern representation dynamics from 124M to 1.5B parameters. Across 4800 layer transitions spanning three orders of magnitude in model size, we observe perfect satisfaction of our theoretical bounds with zero violations and consistent safety margins exceeding 85%. The empirical validation confirms that transformers naturally operate well within the constraints imposed by their optimal transport structure.

The remainder of this paper is organized as follows. Section 2 develops the mathematical framework connecting attention to optimal transport and derives fundamental bounds on representation evolution. Section 3 provides theoretical analysis of drift accumulation, alignment, and saturation phenomena. Section 4 presents our empirical validation across GPT-2 model scales, clearly separating validation of theoretical predictions from empirical discoveries. Section 5 discusses implications for understanding and designing deep transformers. Section 6 positions our work within the broader literature. Complete proofs, extended experimental details, and additional theoretical results appear in the appendices.

2. Representation Evolution Through Optimal Transport

2.1. Mathematical Preliminaries

We begin by establishing notation and key concepts that connect attention mechanisms to optimal transport theory [21; 15]. Throughout, vectors are column vectors, and $\mathbf{1}$ denotes

the all-ones vector of appropriate dimension. For a matrix A, we write $A_{i:}$ for its i-th row and use $\langle A, B\rangle = \sum_{ij} A_{ij}B_{ij}$ for the Frobenius inner product.

The space of probability distributions on m elements is $\Delta^{m-1} := \{p \in \mathbb{R}^m : p \geq 0,\ \mathbf{1}^\top p = 1\}$. For $p, q \in \Delta^{m-1}$, the total variation distance is $\mathrm{TV}(p,q) := \frac{1}{2}\|p - q\|_1$ (where $\|\cdot\|_1$ denotes the ℓ_1 norm). Under the Hamming ground cost $c(i,j) = \mathbf{1}_{\{i\neq j\}}$, the 1-Wasserstein distance coincides with total variation, i.e., $W_1(p,q) = \mathrm{TV}(p,q)$ (see, e.g., [12, §2.3]). We adopt the negative-entropy convention for entropic regularization, defining $\mathsf{H}_\varepsilon(\pi) := \varepsilon \sum_{i,j} \pi_{ij}(\log \pi_{ij} - 1)$ for a transport plan π with regularization strength $\varepsilon > 0$; the additive constant is inconsequential for the optimizer.

In standard attention with temperature T, the softmax operation computes $\mathrm{softmax}(QK^\top/T)$. Our analysis proves that attention solves an optimal transport problem with unit entropic regularization $\varepsilon = 1$. The temperature T and regularization ε interact through the product $\tau = T\varepsilon$, which controls the transport plan's entropy. Since we establish $\varepsilon = 1$, the effective regularization is $\tau = T$. In the common case where $T = \sqrt{d_k}$, we have $\tau = \sqrt{d_k}$.

2.2. Main Result: Attention as Semi-Relaxed Optimal Transport

We now present our fundamental result that establishes the exact equivalence between attention and optimal transport.

Theorem 1 (MAIN THEOREM - Attention equals semi-relaxed entropic OT) *The row-softmax attention weights $S = \mathrm{softmax}(QK^\top/T)$ are exactly the unique optimizer of:*

$$\min_{\pi \in \Pi_{\mathrm{row}}} \left\{ \langle \pi, C\rangle + \varepsilon \sum_{ij} \pi_{ij}(\log \pi_{ij} - 1) \right\} \tag{2.1}$$

where $C = -QK^\top/T$, $\Pi_{\mathrm{row}} = \{\pi \in \mathbb{R}^{n\times m}_{\geq 0} : \pi\mathbf{1} = \mathbf{1}\}$, and $\varepsilon = 1$.

Significance: This theorem establishes that attention is not approximately or analogously related to optimal transport—it *exactly* solves a specific OT problem. The semi-relaxed nature (only row constraints, no column constraints) is essential for causality: balanced OT would require adjusting representations of past tokens, violating autoregressive generation principles.

The proof, detailed in Appendix A.1, proceeds by applying Karush-Kuhn-Tucker conditions to the convex optimization problem. The entropic regularizer ensures strict positivity of the optimizer on the support, which through complementary slackness eliminates the nonnegativity dual variables. The row-constrained constraint structure allows the problem to decompose into n independent optimization problems, each yielding the familiar softmax form. This result connects to classical matrix scaling theory [16; 7] and computational optimal transport [3].

2.3. Probing Representations Across Layers

To study how representations evolve with depth, we use a fixed linear probe that maps hidden states to a probability distribution over a vocabulary (or output space). Given a probe matrix $W_{\mathrm{out}} \in \mathbb{R}^{d\times V}$, with d the hidden dimension and V the vocabulary size, we define for each layer ℓ:

$$z^{(\ell)} = W_{\mathrm{out}}^\top h^{(\ell)} \in \mathbb{R}^V, \qquad p^{(\ell)} = \mathrm{softmax}\big(z^{(\ell)}\big) \in \Delta^{V-1}. \tag{2.2}$$

The logits $z^{(\ell)}$ capture the raw representational content, while the distribution $p^{(\ell)}$ provides a normalized view amenable to information-theoretic analysis. Throughout, the probe uses the canonical softmax temperature 1 (no additional scaling).

2.4. Consequences of the OT Structure: Representation Bounds

The optimal transport structure proven in Theorem 1 immediately implies constraints on representation evolution. While the following result is mathematically a consequence of the main theorem, we state it as a theorem due to its fundamental importance:

Theorem 2 (Lipschitz bound from OT structure) *As a direct consequence of attention solving semi-relaxed OT, probability distributions induced by fixed probes evolve between consecutive layers with:*

$$\|p^{(\ell+1)} - p^{(\ell)}\|_1 \leq \|W_{\text{out}}^\top\|_{2\to\infty} \|h^{(\ell+1)} - h^{(\ell)}\|_2 \tag{2.3}$$

where $\|A\|_{2\to\infty} = \max_j \|A_{j:}\|_2$ is the maximum row norm. This bound is tight and achieved when hidden state changes align with the probe's maximum-norm direction.

The proof, given in Appendix A.2, combines two key observations. First, softmax is 1-Lipschitz from ℓ_∞ to ℓ_1, meaning $\|\text{softmax}(z) - \text{softmax}(w)\|_1 \leq \|z - w\|_\infty$ for any logit vectors z, w. Second, the probe induces logit changes bounded by $\|z^{(\ell+1)} - z^{(\ell)}\|_\infty \leq \|W_{\text{out}}^\top\|_{2\to\infty} \|h^{(\ell+1)} - h^{(\ell)}\|_2$ through the operator norm inequality.

3. Theoretical Analysis of Drift and Stability

3.1. Accumulation of Representational Drift

While Theorem 2 bounds single-layer transitions, understanding deep transformers requires analyzing how drift accumulates across many layers. Through telescoping, we obtain:

Corollary 3 (Cumulative drift bound) *For any depth L, the total representation change is bounded by:*

$$\|p^{(L)} - p^{(0)}\|_1 \leq \|W_{\text{out}}^\top\|_{2\to\infty} \sum_{\ell=0}^{L-1} \|h^{(\ell+1)} - h^{(\ell)}\|_2 \tag{3.1}$$

This cumulative bound would suggest linear growth with depth if per-layer changes were constant. However, our empirical analysis reveals sublinear scaling, indicating that architectural mechanisms like residual connections and layer normalization [1] provide implicit regularization against drift accumulation.

To quantify the coherence of representation evolution across sequence positions, we introduce an alignment framework. For logit changes $\Delta z^{(\ell)} = z^{(\ell+1)} - z^{(\ell)}$, we seek decomposition:

$$\Delta z_{i:}^{(\ell)} = v^{(\ell)} + \kappa_i^{(\ell)} \mathbf{1} + r_{i:}^{(\ell)} \tag{3.2}$$

where $v^{(\ell)} \in \mathbb{R}^V$ represents a common direction of change across all positions, $\kappa_i^{(\ell)} \in \mathbb{R}$ are position-specific constants that cancel under softmax, and $r_{i:}^{(\ell)}$ are residual terms capturing position-specific deviations.

The alignment quality at layer ℓ is measured by the maximum residual $R_\ell = \max_i \|r_{i:}^{(\ell)}\|_\infty$. Small residuals indicate coherent evolution where all positions change similarly, while large residuals suggest divergent representational updates across the sequence. This connects to the Hilbert metric framework [2; 9] for analyzing contraction in positive systems.

3.2. Saturation-Induced Representation Locking

In approximately 10% of GPT-2-XL samples, we observe saturation events where extreme softmax confidence ($p_{\max} > 0.9999$) creates near-zero TV distance despite continued hidden state evolution. This phenomenon is not predicted by the optimal transport framework but instead reveals a learned mechanism for managing the drift that Theorem 2 proves must accumulate.

A remarkable phenomenon emerges when softmax probabilities approach the boundary of the probability simplex. We characterize this *empirically observed* behavior mathematically:

Theorem 4 (Saturation creates representational invariance) *Let $p^{(\ell)}$ be a probability distribution with maximum element $p^{(\ell)}_{\max} = 1-\epsilon$ for small $\epsilon > 0$. For any logit perturbation Δz that preserves the argmax:*

$$\|\text{softmax}(z^{(\ell)} + \Delta z) - p^{(\ell)}\|_1 \leq 2\epsilon \tag{3.3}$$

regardless of the magnitude of Δz on non-maximal coordinates.

The proof (Appendix A.3) follows from analyzing the softmax Jacobian near saturation. When one probability approaches 1, the gradient with respect to logit changes vanishes quadratically, creating an effective invariance to perturbations.

This theoretical result guarantees that once confidence exceeds 0.9999 ($\epsilon < 10^{-4}$), the total variation change is at most $2\epsilon = 2 \times 10^{-4}$. In our measurements, the actual change is typically orders of magnitude smaller (e.g., TV $\approx 10^{-10}$), indicating additional suppression beyond this worst-case bound.

3.3. Architectural Factors Affecting Stability

Several architectural components modulate representation evolution through their impact on the bounds in Theorem 2. Layer normalization [1] with learned affine parameters (γ, β) and residual connections in the pre-normalization configuration modulate the effective Lipschitz constants that appear in our bounds. Multi-head attention introduces additional complexity through parallel transport problems, with H heads each solving its own semi-relaxed OT problem with costs $C^{(h)} = -Q^{(h)}{K^{(h)}}^\top/T$. The total drift is bounded by the sum of per-head contributions weighted by their respective projection norms. Complete mathematical derivations of the Lipschitz constants for these components appear in Appendix D.

4. Empirical Analysis of Representation Evolution

We present two distinct types of empirical results: validation of our theoretical predictions derived from the optimal transport framework, and discovery of phenomena that emerge from data but extend beyond our current theory. This separation is important for intellectual clarity—we distinguish what our theory predicts and we confirm, from what we discover empirically that requires future theoretical development.

4.1. Experimental Setup and Methodology

We conduct comprehensive experiments on three scales of GPT-2 models to validate our theoretical framework and uncover empirical phenomena. The models span three orders of magnitude in parameters: GPT-2 base with 124M parameters and 12 layers, GPT-2-medium

with 355M parameters and 24 layers, and GPT-2-XL with 1.5B parameters and 48 layers. All models share vocabulary size $V = 50,257$ and use learned positional embeddings with causal attention masking.

Our analysis examines 500 randomly sampled sequences from the WikiText-103 validation set. For each sequence and model, we extract hidden states at every layer and compute induced probability distributions using the pretrained language modeling head as our probe. Complete experimental protocol including specific measurements, numerical stability considerations, and computational details appears in Appendix E.2.

4.2. Monotonic Drift with Sublinear Scaling

Our first key empirical finding concerns how alignment residuals accumulate with depth. We normalize residuals by $\sqrt{V}$ to account for high-dimensional concentration effects, as random perturbations in V-dimensional space have expected ℓ_2 norm scaling as $\sqrt{V}$.

Table 1: Normalized alignment residuals $R_\ell/\sqrt{V}$ across model depths

Layer Range	GPT-2 (12L)	GPT-2-medium (24L)	GPT-2-XL (48L)
1-4	0.08 ± 0.02	0.12 ± 0.03	0.15 ± 0.05
5-8	0.19 ± 0.03	0.31 ± 0.06	0.42 ± 0.12
9-12	0.31 ± 0.03	0.48 ± 0.08	0.68 ± 0.19
13-24	-	0.79 ± 0.10	0.95 ± 0.28
25-48	-	-	1.30 ± 0.44

The sublinear scaling is particularly striking: doubling depth from 12 to 24 layers increases the final residual by a factor of 2.5, while doubling again to 48 layers yields only 1.6× additional growth. This contrasts sharply with the linear accumulation that naive analysis would predict and suggests that architectural mechanisms actively resist drift accumulation.

4.3. Empirical Discovery: Saturation-Induced Representation Locking

Beyond validating our theoretical predictions, we discovered an unexpected phenomenon not predicted by our optimal transport framework. In approximately 10% of GPT-2-XL samples, we observe events where extreme softmax confidence creates near-perfect representational stability despite continued computation.

When the maximum probability in $p^{(\ell)}$ exceeds 0.9999, the total variation between consecutive layers drops below 10^{-10}, effectively reaching numerical zero. Simultaneously, the underlying hidden states continue evolving with relative changes of 2-9% in ℓ_2 norm, and logit perturbations reach magnitudes up to 33.8. This dramatic decoupling reveals a mechanism by which transformers can freeze decisions while maintaining computational flexibility.

Analysis of the layer distribution of saturation events reveals they concentrate in layers 12-41 of the 48-layer model, notably absent from both early and final layers. This suggests a computational strategy where early layers build representations, middle layers lock confident predictions, and late layers refine remaining uncertain positions. The pattern is consistent across different sequence positions and appears to correlate with linguistic boundaries such as punctuation and sentence endings.

This saturation phenomenon was discovered empirically and is not explained by our current theoretical framework, suggesting rich dynamics beyond the optimal transport structure.

4.4. Validation of Theoretical Bounds

Across all 4800 layer transitions examined (500 samples $\times$ varying layer counts per model), we observe perfect satisfaction of the Lipschitz bound from Theorem 2. Not a single violation occurs, and the distribution of tightness ratios reveals substantial safety margins.

The empirical tightness ratio $r = \|\Delta p\|_1 / (\|W_{\text{out}}^\top\|_{2\to\infty} \|\Delta h\|_2)$ has maximum value 0.126 across all observations, mean 0.043 ± 0.021, median 0.038, and 95th percentile 0.080. This 87% safety margin between worst-case observed and theoretical bounds suggests that gradient-based training naturally avoids adversarial configurations that would approach theoretical limits.

For computational efficiency validation, we implement Sinkhorn iterations for computing entropic optimal transport distances using the rank-1 kernel structure available for Hamming ground cost. With entropic regularization $\varepsilon = 1.0$, convergence to tolerance 10^{-6} requires 42.3 ± 8.1 iterations on average, consistent with geometric convergence. The rank-1 kernel structure reduces computational complexity from $O(V^2)$ to $O(V)$ per iteration, with the explicit derivation provided in Appendix F.2.

5. Implications and Discussion

5.1. Mechanisms Enabling Deep Transformer Stability

Our analysis reveals three complementary mechanisms that enable transformers to maintain coherent representations across extreme depths. First, the optimal transport structure imposes geometric constraints that inherently limit representation drift, preventing chaotic evolution that could otherwise emerge in deep networks. Second, architectural components including residual connections and layer normalization create implicit regularization that resists linear drift accumulation, as evidenced by the sublinear scaling we observe empirically. Third, the saturation-based locking mechanism provides a way to freeze confident predictions while allowing continued refinement of uncertain positions, effectively implementing a form of progressive decision-making.

5.2. Design Principles for Representation Control

Our framework suggests several actionable principles for controlling representation evolution in transformer architectures. The probe norm $\|W_{\text{out}}^\top\|_{2\to\infty}$ directly appears in drift bounds, offering a concrete parameter for tuning stability. Practitioners can modulate this through weight regularization or normalization schemes applied to the output head. Temperature management in attention offers particularly fine-grained control. Lower temperatures increase the propensity for saturation, enabling more aggressive representation locking, while higher temperatures maintain flexibility at the cost of increased drift.

6. Related Work

Our work connects to several research threads in deep learning and transformer analysis. Studies of representation dynamics in deep networks [13; 11] have characterized how representations evolve empirically but lack the tight theoretical bounds we provide through the

optimal transport framework. Work on neural network optimal transport [12] has primarily focused on using OT for distribution matching and generative modeling, while we reveal OT as the inherent structure of attention mechanisms themselves.

Theoretical analysis of transformers has examined expressiveness [22], optimization properties [10], and approximation capabilities [5]. Our contribution is orthogonal, characterizing the geometric constraints that govern representation evolution independent of these other properties. Recent work on mechanistic interpretability [6] seeks to understand transformer computations, and our framework provides a mathematical foundation for understanding how information flows through attention layers.

Efficient attention mechanisms including FlashAttention [4], linear-time kernelized approaches [8], and sparse attention variants [17] modify the computational structure while preserving the basic softmax operation. Our analysis applies to all variants that maintain row-constrained normalization. Work on alternative normalization schemes like Sinkformers [14] enforces doubly-stochastic constraints, creating balanced rather than semi-relaxed transport problems.

7. Conclusion

7.1. Summary and Contributions

We have shown that the optimal transport structure inherent in transformer attention fundamentally constrains how representations evolve through network depth. By proving tight bounds on representation drift and discovering the saturation-induced locking phenomenon, we explained how deep transformers maintain stable representations while preserving computational flexibility. Our framework provides both theoretical understanding and practical tools for analyzing and controlling representation dynamics in large-scale models.

7.2. Limitations and Future Work

Our framework has several important limitations. We characterize only self-attention, not cross-attention or the full transformer architecture including feed-forward network layers. The saturation phenomenon, while empirically robust across all model scales examined, remains theoretically unexplained within our optimal transport framework. Our analysis assumes fixed pretrained models and does not characterize training dynamics. While we prove that attention solves semi-relaxed optimal transport, this mathematical equivalence does not explain why this particular optimization problem leads to effective language modeling.

The principles we uncovered—geometric constraints from optimal transport, architectural modulation of drift, and saturation-based stability—appear to be fundamental to transformer operation rather than artifacts of particular implementations. Future work should investigate whether these principles extend to other attention variants, whether the saturation phenomenon can be leveraged for improved efficiency through selective computation, and how these insights might inform the design of next-generation architectures that better balance stability and expressiveness.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26 (NeurIPS 2013)*, pages 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf. Preprint available at arXiv:1306.0895.

[4] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, 2022.

[5] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. *International Conference on Machine Learning*, pages 5793–5831, 2022.

[6] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[7] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[8] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165, 2020.

[9] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, UK, 2012.

[10] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5763. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.463.

[11] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems 31*, pages 5727–5736. Curran Associates, Inc., 2018.

[12] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[13] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International*

Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854. PMLR, 2017.

[14] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

[15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

[16] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[17] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2020.

[18] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

[19] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4344–4353. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1443.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[21] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, 2008.

[22] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *International Conference on Learning Representations*, 2020.

Supplementary Material

Appendix A. Complete Proofs for Main Theoretical Results

This appendix provides complete proofs and technical background for all results stated in the main text. We maintain the notation established in Sections 2-3, where vectors are column vectors, $\mathbf{1}$ denotes the all-ones vector of appropriate dimension, and $\langle A, B\rangle = \sum_{ij} A_{ij}B_{ij}$ represents the Frobenius inner product. Throughout this appendix, all logarithms are natural unless otherwise specified.

A.1. Proof of Theorem 1: Attention as Semi-Relaxed Entropic Optimal Transport

We now provide the complete proof that standard row-softmax attention exactly equals the unique primal optimizer of the row-constrained, semi-relaxed entropic optimal transport program. This proof makes explicit the complete Karush-Kuhn-Tucker (KKT) system including complementary slackness, demonstrates how masking enters through feasibility constraints, and establishes why the row dual variables are unique up to row-additive shifts of the cost matrix.

Proof Consider the semi-relaxed entropic optimal transport problem with unmasked support $\Omega \subseteq [n] \times [m]$. The optimization program is:

$$\min_{\pi \in \Pi_{\text{row}}(\mu)} \left\{ \langle \pi, C \rangle + \varepsilon \sum_{(i,j)\in\Omega} \pi_{ij}(\log \pi_{ij} - 1) \right\} \tag{A.1}$$

where the feasible set is defined as $\Pi_{\text{row}}(\mu) = \{\pi \geq 0 : \sum_{j:(i,j)\in\Omega} \pi_{ij} = \mu_i \ \forall i\}$.

Masked entries $(i, j) \notin \Omega$ are excluded from the feasible set, which is equivalent to setting $C_{ij} = +\infty$ for these entries and enforcing $\pi_{ij} = 0$. We assume feasibility, meaning each row i has at least one unmasked column. Since $\mu_i > 0$ for all i, a strictly positive feasible point exists on Ω by distributing μ_i across the allowed columns of row i.

To solve this convex optimization problem, we apply the Karush-Kuhn-Tucker conditions. We introduce Lagrange multipliers $\alpha_i \in \mathbb{R}$ for the row-sum constraints and multipliers $\beta_{ij} \geq 0$ for the nonnegativity constraints $\pi_{ij} \geq 0$ on $(i, j) \in \Omega$. The Lagrangian is:

$$\mathcal{L}(\pi, \alpha, \beta) = \sum_{(i,j)\in\Omega} (C_{ij}\pi_{ij} + \varepsilon\pi_{ij}(\log \pi_{ij} - 1) - \beta_{ij}\pi_{ij}) + \sum_i \alpha_i \left(\mu_i - \sum_{j:(i,j)\in\Omega} \pi_{ij} \right) \tag{A.2}$$

The KKT conditions for optimality are:

$$\text{(Primal feasibility)} \quad \pi_{ij} \geq 0 \text{ for } (i,j) \in \Omega, \quad \sum_{j:(i,j)\in\Omega} \pi_{ij} = \mu_i \ \forall i \tag{A.3}$$

$$\text{(Dual feasibility)} \quad \beta_{ij} \geq 0 \text{ for } (i,j) \in \Omega \tag{A.4}$$

$$\text{(Stationarity)} \quad \frac{\partial \mathcal{L}}{\partial \pi_{ij}} = C_{ij} + \varepsilon \log \pi_{ij} - \alpha_i - \beta_{ij} = 0 \text{ for } (i,j) \in \Omega \tag{A.5}$$

$$\text{(Complementary slackness)} \quad \beta_{ij}\pi_{ij} = 0 \text{ for } (i,j) \in \Omega \tag{A.6}$$

Note that no KKT equations are written for masked entries $(i, j) \notin \Omega$ because they are excluded from the feasible set by construction.

To establish strict positivity on the unmasked support, consider the rowwise contribution to the objective for fixed row i:

$$f_i(\pi_{i:}) = \sum_{j:(i,j)\in\Omega} (C_{ij}\pi_{ij} + \varepsilon\pi_{ij}(\log \pi_{ij} - 1)) \tag{A.7}$$

subject to the constraints $\sum_{j:(i,j)\in\Omega} \pi_{ij} = \mu_i$ and $\pi_{ij} \geq 0$.

The function $g(t) = \varepsilon t(\log t - 1) + C_{ij}t$ is strictly convex on $(0, \infty)$ with derivative:

$$\frac{dg}{dt} = \varepsilon \log t + C_{ij} \tag{A.8}$$

As $t \downarrow 0$, we have $\log t \to -\infty$, so the derivative approaches $-\infty$. This implies that the minimizer on the probability simplex must be strictly in the interior, giving us $\pi^*_{ij} > 0$ for all $(i, j) \in \Omega$.

Since $\pi^*_{ij} > 0$ on Ω, the complementary slackness condition (A.6) forces $\beta_{ij} = 0$ for all $(i, j) \in \Omega$. The stationarity equations (A.5) then simplify to:

$$C_{ij} + \varepsilon \log \pi^*_{ij} - \alpha_i = 0 \tag{A.9}$$

Solving for π^*_{ij}:

$$\pi^*_{ij} = \exp\left(\frac{\alpha_i - C_{ij}}{\varepsilon}\right) = u_i \exp\left(-\frac{C_{ij}}{\varepsilon}\right) \tag{A.10}$$

where we define $u_i := \exp(\alpha_i/\varepsilon) > 0$.

To determine the row scaling factors u_i, we enforce the row constraint from (A.3):

$$\sum_{j:(i,j)\in\Omega} \pi^*_{ij} = \mu_i \tag{A.11}$$

Substituting our expression for π^*_{ij}:

$$u_i \sum_{j:(i,j)\in\Omega} \exp\left(-\frac{C_{ij}}{\varepsilon}\right) = \mu_i \tag{A.12}$$

Therefore:

$$u_i = \frac{\mu_i}{\sum_{j:(i,j)\in\Omega} \exp(-C_{ij}/\varepsilon)} \tag{A.13}$$

This gives us the final form for each row:

$$\pi^*_{i:} = \mu_i \cdot \frac{\exp(-C_{i:}/\varepsilon)}{\mathbf{1}^\top \exp(-C_{i:}/\varepsilon)} = \mu_i \cdot \mathrm{softmax}(-C_{i:}/\varepsilon) \quad \text{on } \Omega \tag{A.14}$$

with $\pi^*_{ij} = 0$ for $(i, j) \notin \Omega$.

For the specific case of attention with $\mu = \mathbf{1}$ and $C = -QK^\top/T$, the normalized plan $S := \mathrm{diag}(\mu)^{-1}\pi^*$ equals the usual row-softmax attention map.

To establish uniqueness, note that the objective $\langle \pi, C\rangle + \mathsf{H}_\varepsilon(\pi)$ is strictly convex on the affine feasible set (each row constraint defines an affine subspace). Therefore, the primal optimizer π^* is unique.

For the dual variables, the multipliers α_i are uniquely determined by the row sums through the relation $u_i = \exp(\alpha_i/\varepsilon)$ and the equations above, given a fixed cost matrix C.

Finally, we verify row-additive invariance. If we replace C_{ij} by $C_{ij} + c_i$ for some constants $c_i \in \mathbb{R}$, this multiplies $\exp(-C_{i:}/\varepsilon)$ by $e^{-c_i/\varepsilon}$, which cancels in the row normalization, leaving π^* unchanged. Under this transformation, the dual variables transform as $\alpha_i \mapsto \alpha_i + c_i$ to maintain the stationarity equations. Thus, the dual variables are unique modulo the row-additive invariance of the primal. ■

A.2. Proof of Theorem 2: Lipschitz Bound on Representation Evolution

We provide two complete proofs that softmax is 1-Lipschitz from ℓ_∞ to ℓ_1 and demonstrate that this constant is tight. We then use this result to establish the bound on representation evolution.

Proof [Proof via Jacobian operator norm] Let $p = \text{softmax}(z)$ for some $z \in \mathbb{R}^m$. The Jacobian of the softmax function at z is:

$$J(z) = \text{diag}(p) - pp^\top \tag{A.15}$$

We need to establish that $\|J(z)\|_{\infty\to 1} \leq 1$. For any vector $v \in \mathbb{R}^m$ with $\|v\|_\infty \leq 1$, the i-th component of $J(z)v$ is:

$$(J(z)v)_i = p_i v_i - p_i \sum_{j=1}^{m} p_j v_j = p_i(v_i - \mu) \tag{A.16}$$

where $\mu = \sum_{j=1}^{m} p_j v_j$ represents the weighted average of v with respect to the probability distribution p.

Since p is a probability vector and $\|v\|_\infty \leq 1$, we have $\mu \in [-1, 1]$. Therefore:

$$\|J(z)v\|_1 = \sum_{i=1}^{m} p_i |v_i - \mu| = \mathbb{E}_p[|V - \mu|] \tag{A.17}$$

where V is a random variable taking value v_i with probability p_i. This represents the expected absolute deviation from the mean.

The functional $v \mapsto \sum_i p_i |v_i - \mu|$ is convex in each coordinate v_i. Therefore, its maximum over the hypercube $v \in [-1, 1]^m$ is attained at the vertices, meaning $v \in \{-1, +1\}^m$.

For such a vertex vector, let $p^+ = \sum_{i:v_i=+1} p_i$ and $p^- = 1 - p^+$. Then $\mu = p^+ - p^- = 2p^+ - 1$, and:

$$\sum_i p_i |v_i - \mu| = p^+ |1 - (2p^+ - 1)| + p^- | - 1 - (2p^+ - 1)| \tag{A.18}$$

$$= p^+ |2 - 2p^+| + p^- |2p^+| \tag{A.19}$$

$$= 2p^+(1 - p^+) + 2p^- p^+ \tag{A.20}$$

$$= 2p^+(1 - p^+) + 2(1 - p^+)p^+ = 4p^+(1 - p^+) \leq 1 \tag{A.21}$$

with equality when $p^+ = 1/2$. Hence $\|J(z)\|_{\infty\to 1} \leq 1$. ■

Proof [Proof via interpolation path] Let $s, w \in \mathbb{R}^m$ and define the path $\phi(\tau) = \text{softmax}(s + \tau(w - s))$ for $\tau \in [0, 1]$. By the fundamental theorem of calculus:

$$\phi(1) - \phi(0) = \int_0^1 \phi'(\tau) d\tau = \int_0^1 J(s + \tau(w - s))(w - s) d\tau \tag{A.22}$$

Taking ℓ_1 norms and using the result from the first proof:

$$\|\text{softmax}(w) - \text{softmax}(s)\|_1 \leq \int_0^1 \|J(s + \tau(w - s))\|_{\infty\to 1} \|w - s\|_\infty d\tau \leq \|w - s\|_\infty \tag{A.23}$$

■

Tightness. Let $m = 2$, $s = (0, 0)$, and $v = (1, -1)$. For $0 < \varepsilon \ll 1$ set $w := s + \varepsilon v$. Then $p = \mathrm{softmax}(s) = (\frac{1}{2}, \frac{1}{2})$ and

$$J(s) = \mathrm{diag}(p) - pp^\top = \frac{1}{4}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Hence $J(s)v = \frac{1}{2}v$, so $\|J(s)v\|_1 = 1$ and $\|v\|_\infty = 1$. By a first-order expansion,

$$\mathrm{softmax}(w) - \mathrm{softmax}(s) = J(s)(w - s) + o(\varepsilon) = \varepsilon\, J(s)v + o(\varepsilon),$$

and therefore

$$\frac{\|\mathrm{softmax}(w) - \mathrm{softmax}(s)\|_1}{\|w - s\|_\infty} = \frac{\varepsilon\|J(s)v\|_1 + o(\varepsilon)}{\varepsilon\|v\|_\infty} \xrightarrow[\varepsilon\to 0]{} 1.$$

Thus the $\ell_\infty \to \ell_1$ Lipschitz constant 1 is tight. For temperature $T > 0$, the same construction for $\mathrm{softmax}(\cdot/T)$ yields the tight constant $1/T$.

Now we apply this result to prove the representation evolution bound. For consecutive layers ℓ and $\ell + 1$:

$$\|p^{(\ell+1)} - p^{(\ell)}\|_1 = \|\mathrm{softmax}(z^{(\ell+1)}) - \mathrm{softmax}(z^{(\ell)})\|_1 \le \|z^{(\ell+1)} - z^{(\ell)}\|_\infty \tag{A.24}$$

Since $z^{(\ell)} = W_{\text{out}}^\top h^{(\ell)}$, we have:

$$z^{(\ell+1)} - z^{(\ell)} = W_{\text{out}}^\top (h^{(\ell+1)} - h^{(\ell)}) \tag{A.25}$$

Therefore:

$$\|z^{(\ell+1)} - z^{(\ell)}\|_\infty = \max_j |(W_{\text{out}}^\top (h^{(\ell+1)} - h^{(\ell)}))_j| \tag{A.26}$$

$$= \max_j |(W_{\text{out}}^\top)_{j:} \cdot (h^{(\ell+1)} - h^{(\ell)})| \tag{A.27}$$

$$\le \max_j \|(W_{\text{out}}^\top)_{j:}\|_2 \|h^{(\ell+1)} - h^{(\ell)}\|_2 \tag{A.28}$$

$$= \|W_{\text{out}}^\top\|_{2\to\infty} \|h^{(\ell+1)} - h^{(\ell)}\|_2 \tag{A.29}$$

Combining these inequalities yields the desired bound.

A.3. Proof of Theorem 4: Saturation Creates Representational Invariance

We analyze the behavior of softmax near the boundary of the probability simplex to establish the saturation-induced invariance result.

Proof Let $p^{(\ell)} = \mathrm{softmax}(z^{(\ell)})$ be a probability distribution with maximum element $p_{\max}^{(\ell)} = 1 - \epsilon$ for small $\epsilon > 0$. Without loss of generality, assume the maximum is achieved at index 1, so $p_1^{(\ell)} = 1 - \epsilon$.

Since the probabilities sum to 1, we have:

$$\sum_{j=2}^{m} p_j^{(\ell)} = \epsilon \tag{A.30}$$

For any logit perturbation Δz that preserves the argmax (meaning $z_1^{(\ell)} + \Delta z_1 \geq z_j^{(\ell)} + \Delta z_j$ for all j), we need to bound:

$$\|\text{softmax}(z^{(\ell)} + \Delta z) - p^{(\ell)}\|_1 \tag{A.31}$$

Let $\tilde{p} = \text{softmax}(z^{(\ell)} + \Delta z)$. Since the argmax is preserved, $\tilde{p}_1 \geq \tilde{p}_j$ for all j. We can write:

$$\tilde{p}_j = \frac{\exp(z_j^{(\ell)} + \Delta z_j)}{\sum_{k=1}^m \exp(z_k^{(\ell)} + \Delta z_k)} \tag{A.32}$$

Since $p_1^{(\ell)} = 1 - \epsilon$, we have $z_1^{(\ell)} - z_j^{(\ell)} = \log((1-\epsilon)/p_j^{(\ell)})$ for $j \geq 2$. For small ϵ and $p_j^{(\ell)} \leq \epsilon$:

$$z_1^{(\ell)} - z_j^{(\ell)} \geq \log\left(\frac{1-\epsilon}{\epsilon}\right) \approx \log\left(\frac{1}{\epsilon}\right) \tag{A.33}$$

This large separation means that even with bounded perturbations Δz, the denominator of $\tilde{p}_j$ is dominated by the $\exp(z_1^{(\ell)} + \Delta z_1)$ term:

$$\sum_{k=1}^m \exp(z_k^{(\ell)} + \Delta z_k) \approx \exp(z_1^{(\ell)} + \Delta z_1)\,(1 + O(\epsilon)) \tag{A.34}$$

Therefore:

$$\tilde{p}_1 \approx \frac{\exp(z_1^{(\ell)} + \Delta z_1)}{\exp(z_1^{(\ell)} + \Delta z_1)(1 + O(\epsilon))} = \frac{1}{1 + O(\epsilon)} = 1 - O(\epsilon) \tag{A.35}$$

More precisely, we can show that $|\tilde{p}_1 - p_1^{(\ell)}| \leq \epsilon$ and $\sum_{j=2}^m |\tilde{p}_j - p_j^{(\ell)}| \leq \epsilon$, giving:

$$\|\tilde{p} - p^{(\ell)}\|_1 = |\tilde{p}_1 - p_1^{(\ell)}| + \sum_{j=2}^m |\tilde{p}_j - p_j^{(\ell)}| \leq 2\epsilon \tag{A.36}$$

This bound is independent of the magnitude of Δz on non-maximal coordinates, as long as the argmax is preserved. ■

Appendix B. Extended Mathematical Background

B.1. Notation and Entropy Conventions

Throughout this appendix, we maintain consistent notation with the main text. Vectors are column vectors unless otherwise specified, and $\mathbf{1}$ denotes the all-ones vector of appropriate size. For a matrix A, we write $A_{i:}$ for row i and $A^{:j}$ for column j. The Frobenius inner product is defined as $\langle A, B \rangle = \sum_{ij} A_{ij} B_{ij}$.

We adopt the negative-entropy convention for entropic regularization:

$$\mathsf{H}_\varepsilon(\pi) = \varepsilon \sum_{(i,j)\in\Omega} \pi_{ij}(\log \pi_{ij} - 1) \tag{A.1}$$

The additive constant $-\varepsilon \sum_{ij} \pi_{ij}$ does not affect optimizers under fixed-mass constraints but ensures convenient cancellation properties in the KKT conditions. This convention is standard in the computational optimal transport literature [12].

B.2. Balanced versus Semi-Relaxed Entropic Optimal Transport

For completeness and to highlight the distinction with our semi-relaxed setting, we briefly review the balanced entropic optimal transport problem. Given marginal distributions $\mu \in \mathbb{R}^n_{>0}$ and $\nu \in \mathbb{R}^m_{>0}$, the balanced entropic OT problem is:

$$\mathrm{OT}_\varepsilon(\mu, \nu; C) = \min_{\pi \in \Pi(\mu,\nu)} \{\langle \pi, C\rangle + \mathsf{H}_\varepsilon(\pi)\} \tag{A.2}$$

where the feasible set enforces both row and column marginal constraints:

$$\Pi(\mu, \nu) = \{\pi \in \mathbb{R}^{n\times m}_{\geq 0} : \pi \mathbf{1} = \mu, \pi^\top \mathbf{1} = \nu\} \tag{A.3}$$

The key difference from our semi-relaxed setting is the additional column-sum constraint $\pi^\top \mathbf{1} = \nu$. This leads to a different optimizer that requires iterative Sinkhorn-Knopp scaling to satisfy both marginal constraints, rather than the closed-form row-constrained softmax we obtain in the semi-relaxed case.

B.3. The Sinkhorn-Knopp Algorithm and Matrix Scaling

For the balanced entropic OT problem, the optimizer has the form:

$$\pi^*_{ij} = u_i \mathcal{K}_{ij} v_j \tag{A.4}$$

where $\mathcal{K} = \exp(-C/\varepsilon)$ is the Gibbs kernel and (u, v) are positive scaling vectors determined by the marginal constraints.

The Sinkhorn-Knopp algorithm [16] alternately projects onto the row and column constraint sets:

$$u^{(k+1)} = \frac{\mu}{\mathcal{K} v^{(k)}} \quad \text{(entrywise division)} \tag{A.5}$$

$$v^{(k+1)} = \frac{\nu}{\mathcal{K}^\top u^{(k+1)}} \tag{A.6}$$

This iterative procedure converges to the unique scaling vectors that satisfy both marginal constraints. The convergence rate depends on the condition number of the kernel $\mathcal{K}$, which we analyze in detail in Section B.4.

B.4. Hilbert Metric and Birkhoff Contraction Theory

We provide a complete treatment of the Hilbert (projective) metric and its application to analyzing Sinkhorn iteration convergence, following [2; 9].

Definition A.1 (Hilbert metric) *For vectors $x, y \in \mathbb{R}^V_{>0}$, the Hilbert metric is defined as:*

$$d_H(x, y) = \log\left(\max_i \frac{x_i}{y_i}\right) - \log\left(\min_i \frac{x_i}{y_i}\right) = \log\left(\frac{\max_i(x_i/y_i)}{\min_i(x_i/y_i)}\right) \tag{A.7}$$

Key properties of the Hilbert metric include:

1. **Scale invariance:** $d_H(ax, by) = d_H(x, y)$ for all $a, b > 0$
2. **Triangle inequality:** $d_H(x, z) \leq d_H(x, y) + d_H(y, z)$

3. **Multiplicative error characterization:** $d_H(x, y) \leq \eta$ if and only if $e^{-\eta} \leq x_i/y_i \leq e^{\eta}$ for all i

Theorem A.2 (Birkhoff's contraction theorem) *Let A be a matrix with strictly positive entries. Define the projective diameter:*

$$\Delta(A) = \sup_{x,y>0} d_H(Ax, Ay) \tag{A.8}$$

Then for all $x, y > 0$:

$$d_H(Ax, Ay) \leq \tanh\left(\frac{\Delta(A)}{4}\right) d_H(x, y) \tag{A.9}$$

The contraction factor $\rho = \tanh(\Delta(A)/4) < 1$ whenever $\Delta(A) < \infty$.

For our Gibbs kernel $\mathcal{K} = \exp(-C/\varepsilon)$, the projective diameter can be computed using the cross-ratio formula:

$$\Delta(\mathcal{K}) = \log\left(\max_{i,j,k,\ell} \frac{\mathcal{K}_{ik}\mathcal{K}_{j\ell}}{\mathcal{K}_{i\ell}\mathcal{K}_{jk}}\right) \tag{A.10}$$

When the kernel entries lie in the range $[\kappa_{\min}, \kappa_{\max}]$, we have:

$$\Delta(\mathcal{K}) \leq 2\log\left(\frac{\kappa_{\max}}{\kappa_{\min}}\right) \tag{A.11}$$

This yields the contraction factor:

$$\rho = \tanh\left(\frac{\log(\kappa_{\max}/\kappa_{\min})}{2}\right) \tag{A.12}$$

The Sinkhorn algorithm alternates between multiplication by $\mathcal{K}$ or $\mathcal{K}^\top$ (which contracts by factor ρ) and entrywise division (which preserves distances due to scale invariance). Therefore, each half-iteration contracts multiplicative errors by at most ρ, and a full iteration contracts by at most ρ^2.

B.5. Temperature-Regularization Scaling Equivalence

An important technical point concerns the interaction between temperature T and entropic regularization parameter ε. With cost matrix $C = -QK^\top/T$, the Gibbs kernel is:

$$\mathcal{K} = \exp(-C/\varepsilon) = \exp\left(\frac{QK^\top}{T\varepsilon}\right) \tag{A.13}$$

Define the effective regularization $\tau = T\varepsilon$. Then:

$$\mathcal{K} = \exp(QK^\top/\tau) \tag{A.14}$$

This shows that only the product τ governs the kernel's dynamic range and entropy. For the semi-relaxed optimizer, we have:

$$\pi^*_{i:} = \mu_i \cdot \mathrm{softmax}(-C_{i:}/\varepsilon) = \mu_i \cdot \mathrm{softmax}((QK^\top)_{i:}/\tau) \tag{A.15}$$

The rescaling $(T, \varepsilon) \mapsto (\alpha T, \alpha^{-1}\varepsilon)$ leaves τ unchanged and hence preserves the optimizer.

However, this equivalence does not extend to all derived quantities. Lipschitz bounds for the map $z \mapsto \mathrm{softmax}(z/T)$ depend on T alone:

$$\|\mathrm{softmax}(z/T) - \mathrm{softmax}(w/T)\|_1 \leq \frac{1}{T}\|z - w\|_\infty \tag{A.16}$$

This distinction is important when comparing optimal transport-theoretic statements (which depend on τ) with Lipschitz-based drift bounds (which depend on T).

Appendix C. Detailed Analysis of Probability Evolution

C.1. Softmax Range and Concentration Analysis

We provide detailed analysis of how the range of logit values controls the concentration of softmax probabilities.

Lemma A.3 (Extremal softmax under fixed range) *Let $p = \mathrm{softmax}(z/\varepsilon) \in \Delta^{m-1}$ and define the range $R = \mathrm{range}(z) = \max_j z_j - \min_j z_j \geq 0$. Then:*

$$\max_j p_j \geq \frac{1}{1 + (m-1)e^{-R/\varepsilon}}, \quad \min_j p_j \geq \frac{1}{1 + (m-1)e^{R/\varepsilon}} \tag{A.1}$$

with equality achieved at specific extremal configurations.

Proof Without loss of generality (by translation invariance of softmax), assume $\min_j z_j = 0$ and $\max_j z_j = R$.

To maximize $\max_j p_j$ under fixed range R, consider configurations where k coordinates equal R and $m - k$ coordinates equal 0. The softmax probabilities are:

$$p_{\max} = \frac{e^{R/\varepsilon}}{ke^{R/\varepsilon} + (m-k)}, \quad p_{\min} = \frac{1}{ke^{R/\varepsilon} + (m-k)} \tag{A.2}$$

The maximum entry $p_{\max}$ is maximized when $k = 1$:

$$\max_j p_j \geq \frac{e^{R/\varepsilon}}{e^{R/\varepsilon} + (m-1)} = \frac{1}{1 + (m-1)e^{-R/\varepsilon}} \tag{A.3}$$

Similarly, the minimum entry is minimized when $k = m - 1$:

$$\min_j p_j \geq \frac{1}{(m-1)e^{R/\varepsilon} + 1} = \frac{1}{1 + (m-1)e^{R/\varepsilon}} \tag{A.4}$$

Equality is achieved when one coordinate of z equals $\min z$ and the remaining $m - 1$ coordinates equal $\max z$ (or vice versa). ■

Corollary A.4 (Entropy and total variation bounds from range) *Let $p = \mathrm{softmax}(z/\varepsilon)$ with range R and let u denote the uniform distribution on m elements.*

1. ***Entropy bound:*** *With $p^* = \max_j p_j$:*

$$H(p) \le h_b(p^*) + (1 - p^*) \log(m - 1) \tag{A.5}$$

where $h_b(t) = -t \log t - (1 - t) \log(1 - t)$ is the binary entropy function.

2. ***Total variation to uniform:*** *With $p_{\min} = \min_j p_j$:*

$$\mathrm{TV}(p, u) = \frac{1}{2}\|p - u\|_1 \le 1 - m \cdot p_{\min} \le 1 - \frac{m}{1 + (m - 1)e^{R/\varepsilon}} \tag{A.6}$$

C.2. Alignment Analysis and Diagnostic Framework

We develop a comprehensive framework for analyzing the coherence of representation evolution across sequence positions.

Definition A.5 (Alignment decomposition) *For logit changes $\Delta z^{(\ell)} = z^{(\ell+1)} - z^{(\ell)} \in \mathbb{R}^{n \times V}$ where n is the sequence length and V is the vocabulary size, we seek the decomposition:*

$$\Delta z_{i:}^{(\ell)} = v^{(\ell)} + \kappa_i^{(\ell)} \mathbf{1} + r_{i:}^{(\ell)} \tag{A.7}$$

where:

- $v^{(\ell)} \in \mathbb{R}^V$ *is a common direction of change*
- $\kappa_i^{(\ell)} \in \mathbb{R}$ *are position-specific constants (cancelled by softmax)*
- $r_{i:}^{(\ell)} \in \mathbb{R}^V$ *are residual terms*

The optimal decomposition minimizing the maximum residual norm can be computed as:

$$v^{(\ell)} = \arg\min_{v} \max_{i} \min_{\kappa_i} \|\Delta z_{i:}^{(\ell)} - v - \kappa_i \mathbf{1}\|_\infty \tag{A.8}$$

This optimization problem can be solved efficiently using linear programming techniques.

Proposition A.6 (Diagnostic bound for approximate alignment) *Under the alignment decomposition above, for each position i and depth $L \ge 1$:*

$$\left\| p_i^{(L)} - \mathrm{softmax}\left(z_i^{(0)} + \sum_{\ell=0}^{L-1} v^{(\ell)} \right) \right\|_1 \le \sum_{\ell=0}^{L-1} \|r_{i:}^{(\ell)}\|_\infty \tag{A.9}$$

Consequently, with $R_\ell = \max_i \|r_{i:}^{(\ell)}\|_\infty$:

$$\max_i \left\| p_i^{(L)} - \mathrm{softmax}\left(z_i^{(0)} + \sum_{\ell=0}^{L-1} v^{(\ell)} \right) \right\|_1 \le \sum_{\ell=0}^{L-1} R_\ell \tag{A.10}$$

C.3. Rank Obstruction for Global Transport Representation

We prove that composing multiple attention layers generally cannot be represented as a single attention layer with the same query-key dimension.

Proposition A.7 (Rank obstruction for single-layer collapse) *Let $S^{(1)}, S^{(2)}$ be positive row-stochastic matrices of appropriate dimensions, and let $S = S^{(2)}S^{(1)}$ be their composition. Define logit-like quantities:*

$$\widetilde{Z}_{ij} = \log S_{ij} + c_i \tag{A.11}$$

for arbitrary row offsets (c_i).

If there exist matrices $Q \in \mathbb{R}^{n\times d_k}$ and $K \in \mathbb{R}^{m\times d_k}$ such that $\widetilde{Z} = QK^\top$, then all column-difference vectors:

$$\Delta^{(j)} = \widetilde{Z}_{:j} - \widetilde{Z}_{:1} \in \mathbb{R}^n \tag{A.12}$$

must lie in $\mathrm{span}(Q)$, which has dimension at most d_k.

Proof If $\widetilde{Z} = QK^\top$, then:

$$\Delta^{(j)} = \widetilde{Z}_{:j} - \widetilde{Z}_{:1} = Q(K_{j:} - K_{1:})^\top \in \mathrm{span}(Q) \tag{A.13}$$

For $d_k = 1$, Q is a column vector, so all $\Delta^{(j)}$ must be scalar multiples of Q, meaning they are collinear.

To show this generically fails, consider the concrete example with $n = 2$, $m = 3$:

$$S^{(1)} = \begin{pmatrix} 0.40 & 0.35 & 0.25 \\ 0.25 & 0.40 & 0.35 \end{pmatrix}, \quad S^{(2)} = \begin{pmatrix} 0.50 & 0.30 & 0.20 \\ 0.20 & 0.50 & 0.30 \\ 0.30 & 0.20 & 0.50 \end{pmatrix} \tag{A.14}$$

Computing $S = S^{(2)}S^{(1)}$:

$$S = \begin{pmatrix} 0.335 & 0.355 & 0.310 \\ 0.268 & 0.330 & 0.402 \end{pmatrix} \tag{A.15}$$

The difference vectors are:

$$\Delta^{(2)} = \begin{pmatrix} \log(0.355/0.335) \\ \log(0.330/0.268) \end{pmatrix} \approx \begin{pmatrix} 0.058 \\ 0.208 \end{pmatrix} \tag{A.16}$$

$$\Delta^{(3)} = \begin{pmatrix} \log(0.310/0.335) \\ \log(0.402/0.268) \end{pmatrix} \approx \begin{pmatrix} -0.078 \\ 0.405 \end{pmatrix} \tag{A.17}$$

The determinant $\det[\Delta^{(2)}|\Delta^{(3)}] \approx 0.058 \times 0.405 - (-0.078) \times 0.208 \approx 0.040 \neq 0$ shows these vectors are not collinear, so no rank-1 factorization exists.

The collinearity condition requires the vanishing of a nontrivial polynomial (the determinant), which defines a measure-zero subset of the space of matrix pairs. Therefore, the obstruction holds for almost all $(S^{(1)}, S^{(2)})$ with respect to Lebesgue measure. ■

C.4. Worked Example: 3×4 Transport Analysis

We provide a complete worked example illustrating the semi-relaxed optimizer computation.

Consider $n = 3$ queries, $m = 4$ keys, temperature $T = 1$, no masking, and $\varepsilon = 1$. Let:

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad K = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{A.18}$$

The score matrix and cost are:

$$QK^\top = \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & -1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}, \quad C = -QK^\top = \begin{pmatrix} -1 & 0 & -1 & 1 \\ 0 & -1 & 1 & -1 \\ -1 & -1 & 0 & 0 \end{pmatrix} \tag{A.19}$$

For $\mu = \mathbf{1}$, the semi-relaxed optimizer is $S_{i:} = \text{softmax}(-C_{i:}) = \text{softmax}((QK^\top)_{i:})$.
Row 1: $S_{1:} = \text{softmax}(1, 0, 1, -1)$
The exponentials are $(e^1, e^0, e^1, e^{-1}) = (e, 1, e, 1/e)$.
The normalization constant is $Z_1 = e + 1 + e + 1/e = 2e + 1 + 1/e \approx 6.873$.
Therefore: $S_{1:} \approx (0.395, 0.145, 0.395, 0.053)$.
Row 2: $S_{2:} = \text{softmax}(0, 1, -1, 1)$
The exponentials are $(1, e, 1/e, e)$.
The normalization constant is $Z_2 = 1 + e + 1/e + e = 1 + 2e + 1/e \approx 6.873$.
Therefore: $S_{2:} \approx (0.145, 0.395, 0.053, 0.395)$.
Row 3: $S_{3:} = \text{softmax}(1, 1, 0, 0)$
The exponentials are $(e, e, 1, 1)$.
The normalization constant is $Z_3 = 2e + 2 \approx 7.437$.
Therefore: $S_{3:} \approx (0.365, 0.365, 0.134, 0.134)$.
The normalized effective target (average of attention rows) is:

$$\bar{\nu}_{\text{eff}} = \frac{1}{3}(S_{1:} + S_{2:} + S_{3:}) \approx (0.302, 0.302, 0.194, 0.194) \tag{A.20}$$

To verify row-additive invariance, adding constant c to row 1 of C multiplies $\exp(-C_{1:})$ by e^{-c}, which cancels in the softmax normalization, leaving $S_{1:}$ unchanged.

Appendix D. Architectural Components: Complete Analysis

D.1. Causal and Structured Masks

We provide comprehensive analysis of how different masking patterns affect the semi-relaxed transport structure.

Let $M \in \{0, +\infty\}^{n \times m}$ encode a mask where $M_{ij} = +\infty$ if position i cannot attend to position j. The masked cost becomes:

$$C_{ij}^{\text{mask}} = \begin{cases} C_{ij} & \text{if } M_{ij} = 0 \text{ (unmasked)} \\ +\infty & \text{if } M_{ij} = +\infty \text{ (masked)} \end{cases} \tag{A.1}$$

The semi-relaxed optimizer becomes:

$$\pi^*_{ij} = \begin{cases} \mu_i \cdot \frac{\exp(-C_{ij}/\varepsilon)}{\sum_{k:M_{ik}=0}\exp(-C_{ik}/\varepsilon)} & \text{if } M_{ij} = 0 \\ 0 & \text{if } M_{ij} = +\infty \end{cases} \tag{A.2}$$

Common masking patterns include:

Causal masking: For autoregressive models, $M_{ij} = +\infty$ if $j > i$, preventing attention to future positions.

Local window: $M_{ij} = +\infty$ if $|i - j| > w$ for window size w, limiting attention to nearby positions.

Strided patterns: $M_{ij} = +\infty$ if $(i - j) \bmod s \neq 0$ for stride s, creating regular sparse patterns.

Block-diagonal: Partitioning sequences into blocks with attention only within blocks.

Each masking pattern modifies the feasible support Ω but preserves the row-constrained decomposition structure of the semi-relaxed problem.

D.2. Positional Encodings as Cost Modifications

Different positional encoding schemes modify the cost structure in distinct ways.

Absolute positional embeddings: Adding position vectors p_i to inputs before projection:

$$Q_i = (x_i + p_i)W_Q, \quad K_j = (x_j + p_j)W_K \tag{A.3}$$

This yields cost:

$$C_{ij} = -\frac{1}{T}[(x_i + p_i)W_Q W_K^\top (x_j + p_j)^\top] \tag{A.4}$$

Relative positional biases: Adding learned biases b_{i-j} based on relative position:

$$C^{\text{rel}}_{ij} = -\frac{Q_i K_j^\top + b_{i-j}}{T} \tag{A.5}$$

Rotary positional embeddings (RoPE): Applying position-dependent rotations $R(t) \in \mathbb{R}^{d_k \times d_k}$:

$$Q'_i = R(t_i)Q_i, \quad K'_j = R(t_j)K_j \tag{A.6}$$

The rotation matrices are typically block-diagonal with 2×2 rotation blocks:

$$R(t) = \begin{pmatrix} \cos(t\theta_1) & -\sin(t\theta_1) & & & \\ \sin(t\theta_1) & \cos(t\theta_1) & & & \\ & & \ddots & & \\ & & & \cos(t\theta_{d_k/2}) & -\sin(t\theta_{d_k/2}) \\ & & & \sin(t\theta_{d_k/2}) & \cos(t\theta_{d_k/2}) \end{pmatrix} \tag{A.7}$$

where $\theta_i = 10000^{-2(i-1)/d_k}$ are the frequency parameters.

The resulting cost is:

$$C^{\text{rope}}_{ij} = -\frac{Q_i^\top R(t_i)^\top R(t_j) K_j}{T} = -\frac{Q_i^\top R(t_j - t_i) K_j}{T} \tag{A.8}$$

D.3. Multi-Head Attention Decomposition

Multi-head attention creates multiple parallel transport problems that are combined to form the final representation update.

With H heads and head dimension $d_h = d_k/H$, the parameters are:

- Query projections: $W_Q^{(h)} \in \mathbb{R}^{d \times d_h}$
- Key projections: $W_K^{(h)} \in \mathbb{R}^{d \times d_h}$
- Value projections: $W_V^{(h)} \in \mathbb{R}^{d \times d_h}$
- Output projections: $W_O^{(h)} \in \mathbb{R}^{d_h \times d}$

For each head h:

$$Q^{(h)} = XW_Q^{(h)} \in \mathbb{R}^{n \times d_h} \tag{A.9}$$

$$K^{(h)} = XW_K^{(h)} \in \mathbb{R}^{m \times d_h} \tag{A.10}$$

$$V^{(h)} = XW_V^{(h)} \in \mathbb{R}^{m \times d_h} \tag{A.11}$$

$$C^{(h)} = -\frac{Q^{(h)}{K^{(h)}}^\top}{T} \in \mathbb{R}^{n \times m} \tag{A.12}$$

Each head solves its own semi-relaxed OT problem:

$$S_{i:}^{(h)} = \text{softmax}(Q_i^{(h)}{K^{(h)}}^\top/T) \tag{A.13}$$

The outputs are combined:

$$Y = \sum_{h=1}^{H} S^{(h)} V^{(h)} W_O^{(h)} \tag{A.14}$$

Proposition A.8 (Multi-head attention stability) *For a single row i, the output satisfies:*

$$\|y_i - y_i'\|_2 \leq \sum_{h=1}^{H} \|W_O^{(h)}\|_{2\to 2} \left(\|S_{i:}^{(h)} - S_{i:}^{\prime(h)}\|_1 \|V^{(h)}\|_{2\to 2} + \|V^{(h)} - V^{\prime(h)}\|_{2\to 2} \right) \tag{A.15}$$

where the attention weight changes are bounded by:

$$\|S_{i:}^{(h)} - S_{i:}^{\prime(h)}\|_1 \leq \frac{1}{T} \left(\|K^{(h)} - K^{\prime(h)}\|_{2\to\infty} \|Q_i^{(h)}\|_2 + \|K^{\prime(h)}\|_{2\to\infty} \|Q_i^{(h)} - Q_i^{\prime(h)}\|_2 \right) \tag{A.16}$$

D.4. LayerNorm: Complete Eigenvalue Analysis

We provide the complete eigenvalue decomposition of the LayerNorm Jacobian.

Lemma A.9 (LayerNorm Jacobian eigendecomposition) *Let d be the feature dimension, and define:*

$$m(u) = \frac{1}{d}\mathbf{1}^\top u \quad \textit{(mean)} \tag{A.17}$$

$$c(u) = u - m(u)\mathbf{1} \quad \textit{(centered vector)} \tag{A.18}$$

$$\sigma(u) = \sqrt{\frac{1}{d}\|c(u)\|_2^2 + \epsilon} \quad \textit{(standard deviation)} \tag{A.19}$$

The normalized map without affine parameters is $g(u) = c(u)/\sigma(u)$. Its Jacobian is:

$$J_g(u) = \frac{1}{\sigma(u)}P - \frac{1}{\sigma(u)^3 d}cc^\top \tag{A.20}$$

where $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ is the centering projection and $c = c(u)$.

The eigenvalues and eigenvectors are:

1. *Along* $\mathrm{span}\{c\}$*: eigenvalue* $\lambda_\parallel = \frac{\epsilon}{\sigma(u)^3}$
2. *In the subspace* $\mathbf{1}^\perp \cap c^\perp$ *(dimension* $d-2$*): eigenvalue* $\lambda_\perp = \frac{1}{\sigma(u)}$
3. *Along* $\mathrm{span}\{\mathbf{1}\}$*: eigenvalue* $\lambda_{\mathbf{1}} = 0$

Proof For $v = c$:

$$J_g(u)c = \frac{1}{\sigma}Pc - \frac{\|c\|^2}{\sigma^3 d}c = \frac{1}{\sigma}c - \frac{\|c\|^2}{\sigma^3 d}c \tag{A.21}$$

Since $\sigma^2 = \frac{\|c\|^2}{d} + \epsilon$, we have $\|c\|^2 = d(\sigma^2 - \epsilon)$. Therefore:

$$\frac{\|c\|^2}{\sigma^3 d} = \frac{d(\sigma^2-\epsilon)}{\sigma^3 d} = \frac{\sigma^2-\epsilon}{\sigma^3} = \frac{1}{\sigma} - \frac{\epsilon}{\sigma^3} \tag{A.22}$$

Thus $J_g(u)c = \frac{\epsilon}{\sigma^3}c$.

For $v \perp \mathbf{1}$ with $c^\top v = 0$:

$$J_g(u)v = \frac{1}{\sigma}Pv - \frac{1}{\sigma^3 d}c(c^\top v) = \frac{1}{\sigma}v \tag{A.23}$$

For $v = \mathbf{1}$:

$$J_g(u)\mathbf{1} = \frac{1}{\sigma}P\mathbf{1} - \frac{1}{\sigma^3 d}c(c^\top \mathbf{1}) = 0 \tag{A.24}$$

since $P\mathbf{1} = 0$ and $c^\top \mathbf{1} = 0$. ■

With affine parameters (γ, β), the full LayerNorm is:

$$\mathrm{LN}_{\gamma,\beta}(u) = \Gamma g(u) + \beta \tag{A.25}$$

where $\Gamma = \mathrm{diag}(\gamma)$. The Lipschitz constant is:

$$\mathrm{Lip}(\mathrm{LN}_{\gamma,\beta}) = \|\Gamma\|_{2\to 2}\|J_g\|_{2\to 2} = \frac{\|\gamma\|_\infty}{\sigma_{\min}} \tag{A.26}$$

where $\sigma_{\min}$ is the minimum standard deviation over the domain.

Table 2: Detailed GPT-2 model configurations used in experiments

Model	Parameters	Layers	Hidden	Heads	Head Dim	FFN Dim
GPT-2	124M	12	768	12	64	3072
GPT-2-medium	355M	24	1024	16	64	4096
GPT-2-large	774M	36	1280	20	64	5120
GPT-2-XL	1.5B	48	1600	25	64	6400

Appendix E. Extended Empirical Analysis

E.1. Complete Model Configurations

All models use:

- Vocabulary size: 50,257 (byte-pair encoding)
- Context length: 1024 tokens
- Positional embeddings: Learned absolute positions
- Activation function: GELU
- LayerNorm epsilon: 10^{-5}
- Dropout: 0.1 (disabled during evaluation)
- Weight initialization: Modified scaled initialization

E.2. Detailed Experimental Protocol

Data preparation:

1. Sample 500 sequences from WikiText-103 validation set
2. Tokenize using GPT-2 tokenizer (byte-pair encoding)
3. Truncate or pad sequences to exactly 512 tokens
4. Create attention masks for padded positions

Feature extraction:

1. Load pretrained models from HuggingFace
2. Set models to evaluation mode (disable dropout)
3. Extract hidden states at every layer using forward hooks
4. Store activations in half-precision to manage memory

Metric computation:

1. Convert hidden states to float32 for analysis

2. Compute logits $z^{(\ell)} = W_{\text{out}}^\top h^{(\ell)}$
3. Apply softmax to obtain probability distributions
4. Calculate all metrics with numerical stability checks
5. Use float64 for extreme probability values near 0 or 1

E.3. Statistical Analysis of Drift Patterns

We provide comprehensive statistical analysis of drift accumulation patterns.

Table 3: Statistical properties of alignment residuals $R_\ell/\sqrt{V}$

Statistic	GPT-2	GPT-2-medium	GPT-2-XL
Mean (final layer)	0.31	0.79	1.30
Std Dev (final layer)	0.03	0.10	0.44
Coefficient of Variation	0.10	0.13	0.34
95th Percentile	0.36	0.95	2.01
99th Percentile	0.39	1.08	2.74
Skewness	0.42	0.58	1.23
Kurtosis	2.89	3.21	4.56

The increasing skewness and kurtosis with model size indicate heavier tails in the drift distribution for larger models, suggesting occasional positions undergo dramatically larger representation updates.

E.4. Layer-wise Diagnostic Measurements

For each layer transition, we compute the following comprehensive set of diagnostics:

1. **Total variation:** $\text{TV}(p^{(\ell+1)}, p^{(\ell)}) = \frac{1}{2}\|p^{(\ell+1)} - p^{(\ell)}\|_1$
2. **Wasserstein distance:** $W_1(p^{(\ell+1)}, p^{(\ell)}) = \frac{1}{2}\|p^{(\ell+1)} - p^{(\ell)}\|_1$ (under Hamming cost)
3. **KL divergence:** $\text{KL}(p^{(\ell+1)}\|p^{(\ell)}) = \sum_j p_j^{(\ell+1)} \log \frac{p_j^{(\ell+1)}}{p_j^{(\ell)}}$
4. **JS divergence:** $\text{JS}(p^{(\ell+1)}, p^{(\ell)}) = \frac{1}{2}\text{KL}(p^{(\ell+1)}\|m) + \frac{1}{2}\text{KL}(p^{(\ell)}\|m)$ where $m = \frac{1}{2}(p^{(\ell+1)} + p^{(\ell)})$
5. **Entropy change:** $\Delta H^{(\ell)} = H(p^{(\ell+1)}) - H(p^{(\ell)})$
6. **Top-k overlap:** Fraction of top-k tokens shared between $p^{(\ell+1)}$ and $p^{(\ell)}$ for $k \in \{1, 5, 10, 100\}$
7. **Rank correlation:** Spearman correlation between probability rankings
8. **Hidden state metrics:**
   - Relative ℓ_2 change: $\|h^{(\ell+1)} - h^{(\ell)}\|_2 / \|h^{(\ell)}\|_2$

- Cosine similarity: $\cos(h^{(\ell+1)}, h^{(\ell)})$
- Maximum coordinate change: $\|h^{(\ell+1)} - h^{(\ell)}\|_\infty$

9. **Logit metrics:**

- Maximum logit change: $\|z^{(\ell+1)} - z^{(\ell)}\|_\infty$
- Logit correlation: $\text{corr}(z^{(\ell+1)}, z^{(\ell)})$
- Effective temperature: $1/\text{std}(z^{(\ell)})$

E.5. Saturation Phenomenon: Detailed Analysis

Table 4: Properties of saturation events in GPT-2-XL

Property	Value
Frequency of occurrence	10.2%
Layer range	12-41 (of 48)
Peak occurrence layer	28
Mean $p_{\max}$ during saturation	0.99994
Mean TV between layers	8.3×10^{-11}
Mean hidden state change	5.2%
Mean logit perturbation	18.7
Correlation with punctuation	0.67
Correlation with sentence boundaries	0.54

The saturation events show strong statistical regularities:

- They never occur in the first 10% or last 10% of layers
- They cluster in runs of 2-5 consecutive layers
- They are more frequent for common words and punctuation
- They are preserved across different random seeds (deterministic given input)

E.6. Convergence Analysis for Sinkhorn Iterations

The iteration counts confirm the theoretical predictions:

- For $\varepsilon = 1.0$: Geometric (linear) convergence with empirical rate ≈ 0.85 per iteration.
- For $\varepsilon = 0.1$: Transition regime with slower geometric convergence (occasional plateaus from moderate dynamic range).
- For $\varepsilon = 0.01$: Markedly slower convergence consistent with an ill-conditioned kernel (large dynamic range).

Table 5: Sinkhorn iteration counts for different configurations

ε	Tolerance δ	Mean Iterations	Std Dev	Max Iterations
1.0	10^{-6}	42.3	8.1	67
1.0	10^{-8}	58.7	11.2	94
1.0	10^{-10}	75.1	14.8	125
0.1	10^{-6}	187.4	31.2	298
0.1	10^{-8}	251.8	42.7	412
0.01	10^{-6}	823.5	127.4	1342
0.01	10^{-8}	1156.2	189.3	1987

Appendix F. Computational Complexity: Complete Analysis

F.1. General Dense Costs

For a general cost matrix $C \in \mathbb{R}^{V \times V}$ without special structure, each Sinkhorn iteration requires:

1. **Matrix-vector multiplication:** $\mathcal{K}v$ requires V^2 multiplications and $V(V-1)$ additions
2. **Entrywise division:** $\mu/(\mathcal{K}v)$ requires V divisions
3. **Total per iteration:** $O(V^2)$ operations

Space complexity: $O(V^2)$ to store the kernel $\mathcal{K}$.
For the semi-relaxed problem, we need only row normalization:

- Compute $\mathcal{K}_{i:}$ for each row: $O(V)$ per row
- Normalize to obtain softmax: $O(V)$ per row
- Total for all n rows: $O(nV)$

F.2. Hamming Ground Cost: Rank-1 Structure

For the Hamming ground cost $C_{ij} = \mathbf{1}_{i \neq j}$, the Gibbs kernel has special structure:

$$\mathcal{K}_{ij} = \begin{cases} 1 & \text{if } i = j \\ e^{-1/\varepsilon} & \text{if } i \neq j \end{cases} \tag{A.1}$$

Setting $\alpha = e^{-1/\varepsilon} \in (0, 1)$:

$$\mathcal{K} = (1-\alpha)I + \alpha \mathbf{1}\mathbf{1}^\top \tag{A.2}$$

This rank-1 perturbation of the identity enables fast matrix-vector multiplication:
Complexity analysis:

- Computing sum s: $O(V)$ operations
- Computing each y_i: $O(1)$ operations
- Total: $O(V)$ operations
- Memory: $O(V)$ (no need to store full kernel)

Algorithm 1 Fast Hamming kernel multiplication

Input: Vector $x \in \mathbb{R}^V$, parameter $\alpha = e^{-1/\varepsilon}$
Output: $y = \mathcal{K}x$
$s \leftarrow \sum_{i=1}^{V} x_i$ ▷ Sum all components: $O(V)$
for $i = 1$ to V **do**
    $y_i \leftarrow (1-\alpha)x_i + \alpha s$ ▷ $O(1)$ per component
end for
return y

F.3. Convergence Rates: Detailed Analysis

Theorem A.10 (Conservative convergence guarantee) *For any entropic OT problem with regularization $\varepsilon > 0$, there exists a problem-dependent constant C such that Sinkhorn iterations achieve tolerance δ in marginal constraints within:*

$$t_\delta \leq C \cdot \delta^{-2} \tag{A.3}$$

iterations, where C depends on the kernel entries and marginal distributions.

Theorem A.11 (Geometric convergence under bounded dynamic range) *If the kernel entries satisfy $\kappa_{\min} \leq \mathcal{K}_{ij} \leq \kappa_{\max}$ with $0 < \kappa_{\min} \leq \kappa_{\max} < \infty$, then:*

$$t_\delta^{\text{full}} \lesssim \frac{\log(C_0/\delta)}{2\log(1/\rho)} \tag{A.4}$$

where $\rho = \tanh(\log(\kappa_{\max}/\kappa_{\min})/2) < 1$ and C_0 depends on initial conditions.

The transition between regimes occurs when $\kappa_{\max}/\kappa_{\min} \approx 1/\varepsilon^2$. For smaller ε, the kernel becomes increasingly ill-conditioned, leading to slower convergence.

F.4. Practical Implementation Considerations

Numerical stability enhancements:

1. **Log-domain computation:** Work with log-scalings $\tilde{u} = \log u$, $\tilde{v} = \log v$ to avoid numerical overflow/underflow
2. **Centered costs:** Subtract row means from cost matrix to improve conditioning
3. **Adaptive precision:** Use float64 when $\varepsilon < 0.1$ or $\delta < 10^{-8}$
4. **Early stopping:** Monitor both marginal violations and scaling changes

Parallelization opportunities:

- Row-constrained operations in semi-relaxed case are embarrassingly parallel
- Batch matrix-vector products for multiple distributions
- GPU acceleration for large vocabulary sizes

Appendix G. Extended Notation and Technical Definitions

Table 6: Complete notation reference

Symbol	Meaning
Core mathematical objects	
Q, K, V	Query, key, value matrices
C	Cost matrix $C = -QK^\top/T$
T	Temperature parameter
ε	Entropic regularization strength
τ	Effective regularization $\tau = T\varepsilon$
$\mathcal{K}$	Gibbs kernel $\mathcal{K} = \exp(-C/\varepsilon)$
π	Transport plan
S	Row-stochastic attention weights
Probability and information theory	
Δ^{m-1}	Probability simplex in $\mathbb{R}^m$
$H(p)$	Shannon entropy of distribution p
$\mathsf{H}_\varepsilon(\pi)$	Negative entropy of transport plan
$\mathrm{KL}(p\|q)$	Kullback-Leibler divergence
$\mathrm{TV}(p, q)$	Total variation distance
$W_1(p, q)$	1-Wasserstein distance
$S_\varepsilon(p, q; C)$	Debiased Sinkhorn divergence
Linear algebra and norms	
$\|A\|_{2\to\infty}$	Maximum row norm of matrix A
$\|A\|_{2\to 2}$	Spectral norm of matrix A
$\langle A, B\rangle$	Frobenius inner product
$\mathrm{diag}(v)$	Diagonal matrix with entries from v
P	Centering projection $I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$
Network components	
$h^{(\ell)}$	Hidden state at layer ℓ
$z^{(\ell)}$	Logits at layer ℓ
$p^{(\ell)}$	Probability distribution at layer ℓ
W_{out}	Output projection matrix (probe)
$\mathrm{LN}_{\gamma,\beta}$	LayerNorm with parameters (γ, β)
Optimization and convergence	
$\Pi_{\text{row}}(\mu)$	Row-constrained feasible set
$\Pi(\mu, \nu)$	Balanced feasible set
$d_H(x, y)$	Hilbert (projective) metric
$\Delta(\mathcal{K})$	Projective diameter of kernel
ρ	Contraction factor
t_δ	Iterations to tolerance δ