Chain of Dual Structures in Transformer Attention

Masato Uchida
Waseda University
Tokyo, Japan
m.uchida@waseda.jp

Abstract

The attention mechanism in Transformers links queries, keys, and values through a sequence of computations, while the principles that constrain its form remain unclear. This paper examines attention under natural requirements that formalize basic properties of attention computation as transformations between representation spaces, and asks which forms of attention computation are admissible. We show that similarity computation, score-to-weight mapping, and value aggregation are unified by dual structures induced by convex potential functions and their Legendre transforms. In particular, the standard inner-product score is recovered as a limiting case of a potential-based score, the score-to-weight map must take a gradient-map form under structural requirements on attention normalization, and the value-aggregation rule must take a dual-coordinate barycentric form under structural requirements on representative-point aggregation. These results show that the standard components of attention are structurally constrained by the underlying dual structure, while the choice of potentials provides explicit degrees of freedom for understanding and modifying attention mechanisms.

1 Introduction

Transformer [1] is a framework that flexibly captures dependencies among elements in an input sequence by employing multiple representations referred to as queries, keys, and values: it computes similarities between elements based on inner products between queries and keys, and aggregates values through weighted summation according to the resulting similarities. Such computations are generally referred to as attention mechanisms, and they have demonstrated strong performance across a wide range of fields, including natural language processing [2]. The attention mechanism takes two forms: self-attention, in which queries, keys, and values are constructed from the same input sequence, and cross-attention, in which queries and keys/values are constructed from different input sequences. In either case, the computational structure is shared: weights are computed based on the relationship between queries and keys, and values are aggregated according to these weights. Here, queries provide reference representations that serve as the basis for comparison, keys provide representations used to measure similarity with respect to this reference, and values provide representations to be aggregated based on the resulting similarities. Thus, a fundamental characteristic of attention mechanisms lies in the fact that multiple representations with distinct roles are progressively linked through a sequence of computations.

Focusing on this characteristic, queries, keys, and values can be understood not as variables in a common representation space, but as variables in distinct representation spaces related through specific correspondences. For example, the operation of computing attention scores via inner products between queries and keys can be interpreted not as evaluating similarity between elements within the same representation space, but rather as evaluating similarity between elements belonging to different representation spaces that are linked by a correspondence. Moreover, the attention scores are transformed into another representation, referred to as attention weights, through normalization by softmax; this process can be interpreted as defining a correspondence that connects attention scores

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

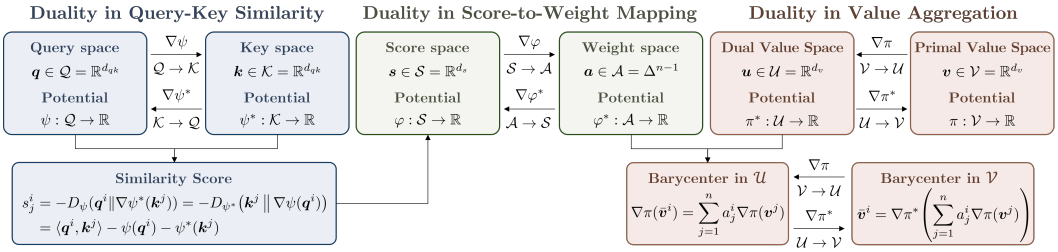

Figure 1: Dual structures underlying the attention mechanism. The potential ψ characterizes query-key similarity through a Bregman divergence, φ determines attention weights from scores through its gradient mapping, and π characterizes value aggregation by forming a weighted barycenter in dual coordinates and mapping it back to the original value space. The corresponding Legendre conjugates provide the dual representations connecting each stage of the computation.

and attention weights. Furthermore, the attention output can be understood as an operation that maps value representations into a space where aggregation by weighted summation is performed, forms a barycenter based on the attention weights, and then associates the result back to the original representation space.

From this perspective, the computational process of the attention mechanism can be organized as a structure in which transformations defined between corresponding representation spaces are progressively linked. In this paper, we decompose this structure into three stages: (1) similarity computation between queries and keys, (2) score normalization into attention weights, and (3) aggregation of value representations using the attention weights. We show that each of these stages can be described from the viewpoint of duality induced by convex potential functions, using the Legendre transform [3, 4]. Furthermore, we clarify that these descriptions correspond to information-geometric structures [5, 6, 7] via the Bregman divergence [8], which offers a unified perspective for understanding the relationships among the corresponding elements.

This paper shows that these three stages are not independent design choices, but are constrained by natural structural requirements. Specifically, the inner-product score arises as a limiting case of a potential-based score, the score-to-weight map is constrained to be a gradient map, and value aggregation is constrained to be a barycenter in the corresponding dual coordinates. Figure 1 illustrates this chain of dual structures. The identified structure further gives a local response analysis: under fixed keys and values, it characterizes the output directions locally induced by query changes, linking the dual structure to the local behavior of attention outputs. These results position standard attention within a broader structurally constrained family and show that the resulting dual structure governs both the form of attention operations and their local output behavior.

2 Related Work

2.1 Theoretical Perspectives on Transformer Attention

Motivated by the strong empirical performance of Transformers, recent theoretical work has studied their approximation power and the computational mechanisms underlying it. Yun et al. [9] showed that Transformers universally approximate broad classes of continuous sequence-to-sequence mappings on compact domains, including permutation-equivariant and general continuous mappings, while accounting for positional encodings. Kratsios et al. [10] showed the existence of probabilistic Transformers that can approximate continuous functions taking values in arbitrary convex or non-convex compact sets, while guaranteeing that the outputs always remain within the prescribed constraint sets. Furthermore, Ramapuram et al. [11] showed that Transformers employing sigmoid attention are universal function approximators.

From the perspective of memory capacity and memory mechanisms in Transformers, several studies have analyzed the role of self-attention. Kim et al. [12] quantified the memory capacity of Transformers and showed that sequence-to-sequence mappings can be memorized with a limited number of parameters. Bietti et al. [13] showed that Transformers acquire long-term knowledge stored in the weights and short-term, context-dependent memory through distinct mechanisms, and that these

forms of memory are learned at different rates during training. Mahdavi et al. [14] analyzed how multi-head attention contributes to memory capacity, clarifying the effects of the number of heads and the sequence length. Kajitsuka et al. [15] further showed that Transformers can efficiently identify and memorize input sequences in both next-token prediction and sequence-to-sequence settings.

On the other hand, several studies have investigated the computational structure of self-attention. Tsai et al. [16] interpreted the attention mechanism from a kernel perspective, showing that it can be viewed as kernel smoothing based on input similarities. Katharopoulos et al. [17] introduced specific kernel feature maps to rewrite the inner-product computation and normalization in a sequentially computable form, and proposed a method for reducing the computational complexity of autoregressive inference on long sequences. Qin et al. [18] proposed a linear Transformer that improves computational efficiency while preserving the attention formulation by replacing softmax with a reweighting mechanism based on cosine similarity. Liu et al. [19] showed that binarizing attention weights can reduce computational cost and energy consumption. Choromanski et al. [20] and Peng et al. [21] demonstrated that the kernel computation corresponding to softmax attention can be expressed as inner products based on random features, enabling attention computation in linear time while preserving the basic structure of Transformers. Chen et al. [22] modeled self-attention as an asymmetric kernel and exploited low-rank structure via singular value decomposition to improve computational efficiency. Vasylenko et al. [23] showed that α-entmax score-to-weight maps induce dense or sparse attention behaviors in long-context settings.

These studies have clarified Transformer attention from the perspectives of expressive power, memory, and computational efficiency. We complement this line of work by examining how score computation, score normalization, and value aggregation are connected through convex duality within the attention mechanism. This perspective identifies a chain of dual mappings induced by convex potentials, and further gives an operator perspective on a single attention head. It expresses the first order effect of query changes on the output as a linear operator coupling the keys and values through the local differential of score normalization. For softmax normalization with standard value aggregation, this operator reduces to the weighted covariance between keys and values under the attention weights.

2.2 Legendre Duality in Machine Learning and Information Geometry

Legendre duality is a fundamental concept in convex analysis that establishes a correspondence between a convex function and its conjugate [3, 4]. This duality provides a mathematical basis for regularized optimization, loss-function design, and the analysis of optimality conditions. In statistical models based on exponential families, the same correspondence appears as the duality between natural parameters and expectation parameters. This relation gives rise to dually flat geometric structures, which are central objects in information geometry [5, 6, 7]. From this viewpoint, inference and learning can be described as geometric operations expressed through dual coordinate systems.

The same convex-geometric structure also underlies prediction evaluation. Through the Legendre correspondence between primal and dual variables, the Bregman divergence induced by a convex potential measures discrepancies between points in the associated geometry [8]. Bregman divergences characterize conditional expectations as optimal predictors [24]. They also provide a unified perspective on risks and loss functions in binary classification [25]. Fenchel-Young losses extend this idea by defining loss functions from a convex potential and its Legendre conjugate. In this construction, the corresponding Bregman divergences appear as natural discrepancy measures [26].

Thus, Legendre duality offers a common geometric language for optimization, inference, and prediction. We use it to revisit score computation, score normalization, and value aggregation in attention. Our analysis treats these steps not as independent algebraic rules, but as transformations described through coordinates induced by convex potentials. This formulation traces how such correspondences appear across the successive stages of attention computation.

3 Preliminaries

In this section, we introduce the notation and basic concepts related to Legendre duality and Bregman divergence [3, 4, 8, 5, 6, 7].

Let $\psi : \Theta \to \mathbb{R}$ be a differentiable and strictly convex function defined on a convex domain $\Theta \subset \mathbb{R}^d$. For any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, the Bregman divergence is defined as

$$D_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}') - \langle \nabla\psi(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle. \tag{1}$$

This quantity represents the deviation of ψ from its tangent hyperplane at $\boldsymbol{\theta}'$ and is generally not symmetric. Moreover, by the convexity of ψ, it is always nonnegative, and equality holds if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}'$. In this paper, we refer to the function ψ as a *potential function*.

The Legendre conjugate of ψ, denoted by $\psi^* : \Xi \to \mathbb{R}$, is defined as

$$\psi^*(\boldsymbol{\xi}) = \sup_{\boldsymbol{\theta} \in \Theta} \{ \langle \boldsymbol{\theta}, \boldsymbol{\xi} \rangle - \psi(\boldsymbol{\theta}) \}. \tag{2}$$

Under the usual conditions for Legendre duality, the gradient mappings $\nabla\psi : \Theta \to \Xi$ and $\nabla\psi^* : \Xi \to \Theta$ are inverses of each other, satisfying

$$\boldsymbol{\xi} = \nabla\psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = \nabla\psi^*(\boldsymbol{\xi}).$$

Furthermore, based on the Legendre conjugate ψ^*, the dual Bregman divergence is defined for any $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi$ as

$$D_{\psi^*}(\boldsymbol{\xi} \| \boldsymbol{\xi}') = \psi^*(\boldsymbol{\xi}) - \psi^*(\boldsymbol{\xi}') - \langle \nabla\psi^*(\boldsymbol{\xi}'), \boldsymbol{\xi} - \boldsymbol{\xi}' \rangle. \tag{3}$$

By associating $\boldsymbol{\xi} = \nabla\psi(\boldsymbol{\theta})$ and $\boldsymbol{\xi}' = \nabla\psi(\boldsymbol{\theta}')$, we obtain the duality of Bregman divergences:

$$D_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = D_{\psi^*}(\boldsymbol{\xi}' \| \boldsymbol{\xi}). \tag{4}$$

This shows the equality of the Bregman divergences on the primal and dual spaces under the Legendre transform.

Bregman divergences can be rewritten, using the Legendre conjugate, in the following form based on the Fenchel-Young inequality (see Appendix A). Specifically, for $\boldsymbol{\theta}' \in \Theta$, let $\boldsymbol{\xi}' = \nabla\psi(\boldsymbol{\theta}')$. Then,

$$D_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) + \psi^*(\boldsymbol{\xi}') - \langle \boldsymbol{\theta}, \boldsymbol{\xi}' \rangle \tag{5}$$

holds. Similarly, the Bregman divergence on the dual space Ξ also admits a representation based on the Fenchel-Young inequality. That is, for $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Xi$, let $\boldsymbol{\theta} = \nabla\psi^*(\boldsymbol{\xi})$. Then,

$$D_{\psi^*}(\boldsymbol{\xi}' \| \boldsymbol{\xi}) = \psi^*(\boldsymbol{\xi}') + \psi(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \boldsymbol{\xi}' \rangle \tag{6}$$

holds.

In this paper, we use the dual relationships described above as a means of characterizing correspondences between variables belonging to different representation spaces. Specifically, by associating variables in one space with their counterparts in a dual space through a potential function, we describe correspondences between variables across distinct representation spaces.

4 Query, Key, and Value Spaces

In the following discussion, unless otherwise specified, we focus on the attention mechanism of a single head in self-attention, and examine how queries, keys, and values are generated from the input sequence within the Transformer and how they are related to one another. This assumption is made for simplicity and does not impose any essential restriction on the structures under consideration. Indeed, the formulation presented in this section can be applied directly to multi-head attention by treating each head separately, and to cross-attention by taking into account the differences between the input sequences. The relationship between the dual structure identified in the attention mechanism and additional mechanisms in Transformer architectures, such as positional encoding, masking, grouped-query attention, and feed-forward blocks, is discussed in Appendix B.

Within the internal processing of a Transformer, query, key, and value vectors are generated through linear transformations of the input sequence. Specifically, let d denote the dimensionality of the input sequence, d_{qk} the dimensionality of queries and keys, and d_v the dimensionality of values. Given an input sequence $\{\boldsymbol{x}^t\}_{t=1}^n$ with $\boldsymbol{x}^t = (x_1^t, \ldots, x_d^t)^\top \in \mathbb{R}^d$, we define the query, key, and value vectors using learnable matrices $W_q \in \mathbb{R}^{d_{qk} \times d}$, $W_k \in \mathbb{R}^{d_{qk} \times d}$, and $W_v \in \mathbb{R}^{d_v \times d}$ as follows:

$$\boldsymbol{q}^t = W_q \boldsymbol{x}^t, \quad \boldsymbol{k}^t = W_k \boldsymbol{x}^t, \quad \boldsymbol{v}^t = W_v \boldsymbol{x}^t. \tag{7}$$

Here, $\boldsymbol{q}^t = (q_1^t, \ldots, q_{d_{qk}}^t)^\top$, $\boldsymbol{k}^t = (k_1^t, \ldots, k_{d_{qk}}^t)^\top$, and $\boldsymbol{v}^t = (v_1^t, \ldots, v_{d_v}^t)^\top$. This operation can be interpreted as a linear mapping from the input space $\mathcal{X} \subset \mathbb{R}^d$ to three distinct representation spaces, $\mathcal{Q} \subset \mathbb{R}^{d_{qk}}$, $\mathcal{K} \subset \mathbb{R}^{d_{qk}}$, and $\mathcal{V} \subset \mathbb{R}^{d_v}$, respectively. Although all three spaces are subsets of real spaces, we denote them separately to emphasize their roles in the attention mechanism.

Here, note that $\boldsymbol{q}^t$, $\boldsymbol{k}^t$, and $\boldsymbol{v}^t$ belong to distinct representation spaces. In the attention mechanism of Transformers, variables belonging to different representation spaces, namely the query, key, and value spaces, are progressively linked through a sequence of transformations, and are ultimately aggregated. However, to justify operations between variables that belong to different representation spaces, one needs to introduce a structure that associates these spaces. In this paper, we focus on three basic operations in the attention mechanism, namely, (1) similarity computation between queries and keys, (2) determination of attention weights by normalizing attention scores, and (3) aggregation of values via a weighted sum with the attention weights. For each operation, we introduce a convex potential function and its Legendre conjugate, and we describe each stage of the attention mechanism in a unified manner as a dual structure induced by the Legendre transform. Combining the three stages, the resulting attention operation takes the form

$$\mathrm{Attention}(\boldsymbol{q}) = \nabla\pi^* \left(\sum_{j=1}^{n} \left[\nabla\varphi\Big(\big(-D_\psi\big(\boldsymbol{q} \| \nabla\psi^*(\boldsymbol{k}^l)\big) \big)_{l=1}^{n} \Big) \right]_j \nabla\pi(\boldsymbol{v}^j) \right). \tag{8}$$

Here, ψ, φ, and π characterize query-key comparison, score-to-weight mapping, and value aggregation, respectively. Sections 5-7 develop these three components in this order.

5 Computation of Attention Score with Convex Potentials

5.1 Dual Coordinates for Queries and Keys

In the standard attention mechanism of Transformers, similarities are computed based on the inner products between queries $\boldsymbol{q}$ and keys $\boldsymbol{k}$, and the resulting values are converted into the attention weights. Since queries and keys are obtained through different linear mappings, they generally belong to distinct representation spaces, denoted by $\mathcal{Q}$ and $\mathcal{K}$, respectively. Therefore, we need a framework to consistently relate elements across the two spaces.

In this paper, we describe this relationship in terms of gradient mappings induced by a convex potential. Let ψ be a differentiable and strictly convex potential defined on the query space $\mathcal{Q}$. The corresponding Legendre conjugate is defined as

$$\psi^*(\boldsymbol{k}) = \sup_{\boldsymbol{q} \in \mathcal{Q}} \{ \langle \boldsymbol{q}, \boldsymbol{k} \rangle - \psi(\boldsymbol{q}) \}. \tag{9}$$

Under the usual conditions for Legendre duality, the associated gradient mappings

$$\nabla\psi : \mathcal{Q} \to \mathcal{K}, \quad \nabla\psi^* : \mathcal{K} \to \mathcal{Q} \tag{10}$$

are inverses of each other. In this setting, $\boldsymbol{q}$ and $\boldsymbol{k}$ are associated as points represented in mutually dual coordinate systems, satisfying

$$\nabla\psi(\boldsymbol{q}) = \boldsymbol{k}, \quad \nabla\psi^*(\boldsymbol{k}) = \boldsymbol{q}. \tag{11}$$

That is, the query space and the key space are dually linked through the potential pair (ψ, ψ^*).

A typical choice is the quadratic potential:

$$\psi(\boldsymbol{q}) = \frac{1}{2}\|\boldsymbol{q}\|^2, \quad \psi^*(\boldsymbol{k}) = \frac{1}{2}\|\boldsymbol{k}\|^2. \tag{12}$$

In this case, the gradient mappings reduce to the identity:

$$\nabla\psi(\boldsymbol{q}) = \boldsymbol{q}, \quad \nabla\psi^*(\boldsymbol{k}) = \boldsymbol{k}. \tag{13}$$

As another example, for $p > 1$, consider the choice:

$$\psi(\boldsymbol{q}) = \frac{1}{p} \sum_{i=1}^{d_{qk}} |q_i|^p, \quad \psi^*(\boldsymbol{k}) = \frac{1}{p'} \sum_{i=1}^{d_{qk}} |k_i|^{p'}. \tag{14}$$

where $1/p + 1/p' = 1$. In particular, when $p = 2$ and $p' = 2$, the potential reduces to a quadratic form. The associated gradient mappings are given componentwise by

$$(\nabla\psi(\boldsymbol{q}))_i = |q_i|^{p-2}q_i, \quad (\nabla\psi^*(\boldsymbol{k}))_i = |k_i|^{p'-2}k_i. \tag{15}$$

As another example, for $\beta > 0$, consider the choice:

$$\psi(\boldsymbol{q}) = \sum_{i=1}^{d_{qk}} \frac{1}{\beta^2}\left\{\cosh(\beta q_i) - 1\right\}, \quad \psi^*(\boldsymbol{k}) = \sum_{i=1}^{d_{qk}} \frac{1}{\beta^2}\left[(\beta k_i)\,\operatorname{arcsinh}(\beta k_i) - \sqrt{1+(\beta k_i)^2} + 1\right]. \tag{16}$$

In particular, this potential reduces to the quadratic potential in the limit $\beta \to 0$. The associated gradient mappings are given componentwise by

$$(\nabla\psi(\boldsymbol{q}))_i = \frac{1}{\beta}\sinh(\beta q_i), \quad (\nabla\psi^*(\boldsymbol{k}))_i = \frac{1}{\beta}\operatorname{arcsinh}(\beta k_i). \tag{17}$$

5.2 Bregman Attention Scores and the Inner-Product Limit

In the standard attention mechanism of Transformers, attention scores are defined by the inner product between a query $\boldsymbol{q}^i$ and a key $\boldsymbol{k}^j$ as

$$s^i_j = \frac{\langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle}{\sqrt{d_{qk}}}, \tag{18}$$

and the attention weights are computed by applying a softmax to these scores. For notational simplicity, we omit the conventional scaling factor $1/\sqrt{d_{qk}}$ in the potential-based score below, since it does not affect the structural discussion. In this paper, we introduce an attention score based on the Bregman divergence induced by a convex potential function $\psi(\boldsymbol{q})$ on the query space $\mathcal{Q}$ as

$$s^i_j = -D_\psi(\boldsymbol{q}^i \| \nabla\psi^*(\boldsymbol{k}^j)) = -D_{\psi^*}\left(\boldsymbol{k}^j \,\middle\|\, \nabla\psi(\boldsymbol{q}^i)\right) = \langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle - \psi(\boldsymbol{q}^i) - \psi^*(\boldsymbol{k}^j). \tag{19}$$

That is, under the definition introduced here, the attention score is not solely determined by the inner product $\langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle$, but also includes the additional terms $-\psi(\boldsymbol{q}^i) - \psi^*(\boldsymbol{k}^j)$, determined by the potential function ψ. This raises a natural question of how the potential-based score in (19) is related to the standard inner-product score in (18). We answer this question by introducing a scaling of the potential and showing that, after the subsequent softmax normalization, the standard attention mechanism is recovered as a limiting case.

We scale the potential on the query side by introducing a temperature parameter $\tau > 0$ as

$$\psi_\tau(\boldsymbol{q}) = \psi\left(\frac{\boldsymbol{q}}{\tau}\right). \tag{20}$$

Under this scaling, the Legendre conjugate of ψ_τ is given by

$$\psi^*_\tau(\boldsymbol{k}) = \psi^*(\tau\boldsymbol{k}). \tag{21}$$

With this scaling in place, the attention score defined in (19) can be characterized by the following theorem. The proof is provided in Appendix C.

Theorem 5.1. *Consider an attention mechanism in which attention weights are given by a softmax applied to the attention scores defined using the scaled potential ψ_τ in* (19)*. Then, the corresponding attention scores can be equivalently written in the following form:*

$$\tilde{s}^i_j(\tau) = \langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle - \psi^*(\tau\boldsymbol{k}^j). \tag{22}$$

Furthermore, if $\psi^*(\tau\boldsymbol{k}^j)$ converges to a constant as $\tau \to 0$, the attention weights given by the softmax reduce to the standard softmax attention based on inner products.

The role of this theorem is to show that the potential-based score formulation recovers standard inner-product softmax attention as a limiting case. From (22), we see that the temperature parameter τ controls the contribution of the key-dependent nonlinear correction term $\psi^*(\tau\boldsymbol{k}^j)$. When τ is small, this correction term is suppressed, and the attention scores exhibit behavior close to that of the standard inner product. Indeed, for all the potentials presented in (12), (14), and (16), the term

$\psi^*(\tau \boldsymbol{k}^j)$ converges to 0 as $\tau \to 0$, becoming a constant. In contrast, when τ is large, the influence of the potential is more strongly reflected in the attention scores. To gain an intuitive understanding of this behavior, we provide in Appendix D a visualization based on simple numerical examples derived from the above formulation.

Finally, when choosing a quadratic potential $\psi(\boldsymbol{q}) = \frac{1}{2}\|\boldsymbol{q}\|^2$, the corresponding Bregman divergence coincides with the squared Euclidean distance between $\boldsymbol{q}^i$ and $\boldsymbol{k}^j$. Applying softmax to the resulting score s_j^i yields a Gaussian kernel representation. While the perspective of interpreting inner-product-based attention as a kernel has also been proposed in prior work [16, 20, 21, 19, 18], these studies analyze given attention computations primarily within frameworks of kernel approximation or computational acceleration. In contrast, the essential distinction of our approach is that we take attention mechanisms defined by general potential functions as the starting point, and show that inner-product-based scores emerge naturally as a scaling limit of this more general formulation.

6 Score-to-Weight Maps from Convex Potentials

Next, we consider the process of determining the attention weights $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ from the obtained scores $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$. Here, we define the score space as $\mathcal{S} = \mathbb{R}^n$, and the space to which the attention weights belong as $\mathcal{A} \subset \mathbb{R}^n$. We introduce a differentiable convex potential function φ defined on the score space $\mathcal{S}$, and define its Legendre conjugate as

$$\varphi^*(\boldsymbol{a}) = \sup_{\boldsymbol{s} \in \mathcal{S}} \{\langle \boldsymbol{a}, \boldsymbol{s} \rangle - \varphi(\boldsymbol{s})\}. \tag{23}$$

The attention weights $\boldsymbol{a}$ corresponding to a given score vector $\boldsymbol{s}$ are determined by the gradient mapping $\nabla\varphi : \mathcal{S} \to \mathcal{A}$ as

$$\boldsymbol{a} = \nabla\varphi(\boldsymbol{s}). \tag{24}$$

The attention weights are determined by the choice of the potential function φ. A typical example is

$$\varphi(\boldsymbol{s}) = \log \sum_{j=1}^{n} \exp(s_j). \tag{25}$$

Its Legendre conjugate is given by

$$\varphi^*(\boldsymbol{a}) = \sum_{j=1}^{n} a_j \log a_j, \quad \boldsymbol{a} \in \Delta^{n-1}, \tag{26}$$

that is, the negative Shannon entropy. Here, $\Delta^{n-1} = \{\boldsymbol{a} \in \mathbb{R}^n \mid a_j \geq 0,\ \sum_{j=1}^n a_j = 1\}$ denotes the $(n-1)$-dimensional probability simplex. The dual relationships considered here, along with the basic properties of the induced gradient mappings, are part of the standard theory of convex optimization [4]. In particular, when choosing $\varphi(\boldsymbol{s}) = \log \sum_{\ell=1}^n \exp(s_\ell)$, the corresponding gradient mapping becomes

$$\nabla\varphi(\boldsymbol{s}) = \left[\frac{\exp(s_1)}{\sum_\ell \exp(s_\ell)}, \ldots, \frac{\exp(s_n)}{\sum_\ell \exp(s_\ell)}\right]^\top, \tag{27}$$

which recovers the standard softmax function [4, Example 3.25].

From the above discussion, we see that both the stage of constructing scores from queries and keys and the stage of determining attention weights from scores in the attention mechanism can be described as dual mappings induced by convex potentials. Within this framework, softmax appears as one particular example corresponding to a specific choice of potential, and the mapping from scores to weights can be understood in a more general form. An alternative but equivalent view of this score-to-weight mapping, based on entropy-regularized optimization, is presented in Appendices E and F.

The discussion above recalls the standard fact that softmax can be placed within the dual framework as an example arising from a particular choice of potential. The following theorem addresses a converse type of question: under natural requirements on attention weights, it characterizes when a

score-to-weight map is necessarily represented in a dual form induced by a convex potential. Thus, the result is not an interpretation of existing attention mechanisms through this framework, but a characterization of the form that such mappings must take under these requirements. The proof is provided in Appendix G, together with an intuitive interpretation of the structural assumptions and explicit examples that further illustrate their roles. In addition, Appendix H verifies that the standard softmax normalization used in Transformer attention satisfies the assumptions needed to apply the theorem.

Theorem 6.1. *Let $g(\boldsymbol{s}) = (g_1(\boldsymbol{s}), \dots, g_n(\boldsymbol{s}))$ be a C^1 mapping from scores $\boldsymbol{s} = (s_1, \dots, s_n) \in \mathbb{R}^n$ to the probability simplex Δ^{n-1}. Assume the following conditions.*

(i) Translation invariance: For any $\boldsymbol{s} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, $g(\boldsymbol{s} + c\mathbf{1}) = g(\boldsymbol{s})$ holds, where $\mathbf{1} = (1, \dots, 1)^\top$.

(ii) Simplex constraint: For any $\boldsymbol{s}$, $\sum_{i=1}^n g_i(\boldsymbol{s}) = 1$ and $g_i(\boldsymbol{s}) > 0$ $(i = 1, \dots, n)$ hold.

(iii) Monotonicity: For any $\boldsymbol{s}$ and $i \neq j$, $\partial g_i / \partial s_i(\boldsymbol{s}) > 0$ and $\partial g_i / \partial s_j(\boldsymbol{s}) < 0$ hold.

(iv) Symmetry: The Jacobian matrix $J(\boldsymbol{s}) = (\partial g_i / \partial s_j(\boldsymbol{s}))_{i,j=1}^n$ is symmetric for any s.

Then, there exists a C^2 function $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that

$$g(\boldsymbol{s}) = \nabla \varphi(\boldsymbol{s}). \tag{28}$$

Moreover, φ is strictly convex on the subspace orthogonal to the translation direction $\mathrm{span}\{\mathbf{1}\}$.

This theorem plays a key role in the present analysis, as it shows that the score-to-weight step is structurally constrained once the basic requirements for an attention normalization are imposed. In particular, the mapping from scores to attention weights is not an arbitrary normalization, but is forced to admit a representation as the gradient of a convex potential. This result should not be understood as a generalization of softmax; rather, it characterizes the class of normalizations that are admissible as attention weights under these requirements, through their underlying convex potentials.

This converse characterization differs from the Fenchel-Young and regularized-argmax formulations in prior work [26, 27]. Those formulations start from a convex regularizer and derive the corresponding score-to-weight map through regularized optimization, equivalently as the gradient of its convex conjugate. This theorem, in contrast, starts from a given C^1 normalization map g, without assuming a regularizer, potential, or optimization formulation, and characterizes when it admits such a convex-potential representation under assumptions (i)-(iv). It further identifies the resulting geometry: the potential is flat along $\mathrm{span}\{\mathbf{1}\}$ and strictly convex on $\mathrm{span}\{\mathbf{1}\}^\perp$.

Appendix I provides a controlled numerical experiment examining how the choice of score-to-weight potential affects prediction performance. We vary the number of label-relevant tokens and the score noise level, and compare two potential-induced mappings: softmax from the negative-entropy potential and sparsemax from a quadratic potential. The results show that neither mapping is uniformly preferable: sparsemax tends to perform better when relevant information is localized and the scores are reliable, whereas softmax tends to perform better when relevant information is distributed or the scores are noisy. This illustrates that the potential is not only a formal degree of freedom, but also controls how attention trades off selective aggregation against diffuse aggregation under different evidence structures.

7 Value Aggregation through Convex Potentials

In the standard attention mechanism, the weights specify which tokens are relevant, but they do not by themselves determine how the selected value vectors should be combined. We represent this remaining choice by the geometry of value aggregation, specified by a convex potential π. In this view, the aggregation output is a representative point of the referred values under the geometry induced by π. Requiring this representative to be insensitive to the order of the values and to the use of intermediate grouping leads to the following characterization: value aggregation must take the form of a barycenter in the coordinate system induced by π. The following theorem derives this barycentric form directly from consistency requirements on the aggregation rule. The intuition behind these requirements, the proof, and the standard variational interpretation of the resulting barycenter in terms of Bregman divergence are provided in Appendix J.

Theorem 7.1. *Let $\pi : \mathcal{V} \to \mathbb{R}$ be a differentiable strictly convex potential defined on a convex domain $\mathcal{V} \subset \mathbb{R}^{d_v}$, and assume that $\nabla\pi$ is one-to-one. Let*

$$A_\pi(\boldsymbol{a}; \boldsymbol{v}^1, \dots, \boldsymbol{v}^n) \in \mathcal{V} \tag{29}$$

be a value aggregation rule for $\boldsymbol{a} = (a_1, \dots, a_n) \in \Delta^{n-1}$ and $\boldsymbol{v}^1, \dots, \boldsymbol{v}^n \in \mathcal{V}$. Assume that, in the coordinate system induced by $\nabla\pi$, the aggregation rule is continuous, unanimous, affine-equivariant, invariant to permutations, consistent with grouping and reaggregation, and unaffected by zero-weighted inputs. Then any aggregation rule satisfying these conditions must coincide with the barycentric relation

$$\nabla\pi\left(A_\pi(\boldsymbol{a}; \boldsymbol{v}^1, \dots, \boldsymbol{v}^n)\right) = \sum_{j=1}^{n} a_j \nabla\pi(\boldsymbol{v}^j). \tag{30}$$

Consequently, if the right-hand side belongs to $\nabla\pi(\mathcal{V})$, then

$$A_\pi(\boldsymbol{a}; \boldsymbol{v}^1, \dots, \boldsymbol{v}^n) = \nabla\pi^* \left(\sum_{j=1}^{n} a_j \nabla\pi(\boldsymbol{v}^j) \right). \tag{31}$$

The usual weighted sum is recovered as the Euclidean case. Indeed, when

$$\pi(\boldsymbol{v}) = \frac{1}{2}\|\boldsymbol{v}\|^2, \tag{32}$$

the gradient map $\nabla\pi$ is the identity map, and (30) reduces to

$$A_\pi(\boldsymbol{a}; \boldsymbol{v}^1, \dots, \boldsymbol{v}^n) = \sum_{j=1}^{n} a_j \boldsymbol{v}^j. \tag{33}$$

Thus, standard Transformer value aggregation is the quadratic-potential case of Theorem 7.1. In this view, the standard weighted sum is not an independent primitive, but the aggregation rule induced by the Euclidean value-space structure. More generally, the weights specify which values are referred to, while the potential π specifies how they are averaged. Hence, changing π can change the attention output even under fixed attention weights. Appendix K gives power-type and hyperbolic-cosine-type examples, and Appendix L gives a numerical example showing that different choices of the potential π can lead to different prediction behavior even under the same attention weights.

The dual structure characterized above also yields a local analysis of the representational ability of a single attention head: how query changes are reflected in the output representation, and how much effective variation they can induce. This question is important for interpreting the representational role of attention, since changes in the attention distribution do not automatically yield rich changes in the output representation. The effective output directions depend on the arrangement of the keys and values, as well as on the value-aggregation potential.

The next theorem characterizes these directions as the image of a Hessian-based linear map. Its rank gives the effective local dimension at the attention output. The proof is provided in Appendix M, together with several consequences of this characterization, including its specialization to standard softmax attention, bounds on the local output dimension of an attention head, the extension of the rank interpretation to multi-head attention, and a rank bound on the local effect of an attention head on subsequent downstream computations. To demonstrate how this characterization can be applied to a trained Transformer, Appendix N evaluates the local output dimension characterized by the theorem in pretrained GPT-2 small [28] and compares it with the corresponding rank bound.

Theorem 7.2. *Fix the keys $\boldsymbol{k}^1, \dots, \boldsymbol{k}^n$ and values $\boldsymbol{v}^1, \dots, \boldsymbol{v}^n$, and consider the single-head attention output $\boldsymbol{z}^i$ as a function of the query $\boldsymbol{q}^i$. Assume that φ is C^2 and satisfies $\nabla\varphi(\boldsymbol{s} + c\mathbf{1}) = \nabla\varphi(\boldsymbol{s})$ for all $\boldsymbol{s}$ and $c \in \mathbb{R}$. Define*

$$\mathcal{C}_\pi^i := \sum_{\ell,m=1}^{n} \nabla\pi(\boldsymbol{v}^\ell) \left[\nabla^2\varphi(\boldsymbol{s}^i)\right]_{\ell m} (\boldsymbol{k}^m)^\top. \tag{34}$$

Then the attainable first-order directions of $\nabla\pi(\boldsymbol{z}^i)$ are $\operatorname{Im}\mathcal{C}_\pi^i = \{\mathcal{C}_\pi^i \boldsymbol{u} : \boldsymbol{u} \in \mathbb{R}^{d_{qk}}\}$. Hence, the local value-side dimension of this head is $\operatorname{rank}\mathcal{C}_\pi^i$.

8 Conclusion

This paper characterized the three basic stages of Transformer attention: score computation, score normalization, and value aggregation. The analysis was guided by a basic question: how do natural consistency requirements restrict the possible forms of these operations? Under the natural requirements expected of attention mechanisms, these three stages combine into a single chain of dual maps given in (8). This expression displays the three dual operations in succession: comparison by ψ, normalization by φ, and aggregation by π.

For score computation, we formulated query-key similarity using the Bregman divergence induced by a convex potential. The resulting score contains the usual inner-product term, but also includes potential-dependent correction terms that specify the geometry in which queries and keys are compared. Under the appropriate scaling, these correction terms vanish and the standard inner-product score emerges as the limiting case of this broader geometric formulation (Theorem 5.1).

For score normalization, we proved that any score-to-weight map satisfying translation invariance, the simplex constraint, monotonicity, and symmetry of the Jacobian must be represented as the gradient of a convex potential (Theorem 6.1). While softmax is a familiar example arising from the log-sum-exp potential, the theorem characterizes the form of admissible normalizations under these structural requirements. Different potentials therefore induce different allocation rules, such as diffuse or selective attention, within the same dual framework.

For value aggregation, we showed that once a potential fixes the value-space structure, the consistency requirements lead to a barycentric form in the associated dual coordinates (Theorem 7.1). The standard weighted sum is recovered as the quadratic case, while non-quadratic potentials define different representative values under the same attention weights. This dual formulation also supports a local response analysis, specifying which output directions can be induced by query changes through the fixed keys, fixed values, and value-aggregation potential (Theorem 7.2).

These results clarify the relation between constraints and choices in attention. The stated consistency requirements restrict the admissible forms of score normalization and value aggregation, while the choice of potentials determines the induced normalization and aggregation geometries. The controlled experiments in the appendices illustrate this relation by showing that different potentials can lead to different allocation or aggregation behavior under matched conditions. Thus, the convex-dual formulation is not only a way to describe the standard attention mechanism, but also a way to make explicit how changing the potential changes the corresponding attention operation.

Our analysis has several limitations that also suggest directions for future work. On the theoretical side, we focus on the internal structure of attention and do not characterize how the identified structures interact with other components of a Transformer or translate into performance on downstream tasks. On the empirical side, our experiments are limited to controlled studies of individual attention components and the measurement of the theoretically derived operator in a pretrained GPT-2 model; we do not perform end-to-end training or task-level evaluation of modified attention mechanisms. Extending the theoretical analysis to broader Transformer architectures and examining how the identified structures and their design choices relate to learning dynamics and downstream performance are important directions for future work.

Acknowledgments and Disclosure of Funding

The author thanks the reviewers for their insightful comments and constructive discussions. Their feedback helped clarify the contributions and further develop this work. This work was supported in part by the Japan Society for the Promotion of Science through Grants-in-Aid for Scientific Research (C) (26K14831).

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 2017.

[2] Gianni Brauwers and Flavius Frasincar. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3279–3298, 2023.

[3] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*. Springer, 1985.

[6] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.

[7] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

[8] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[9] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[10] Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[11] Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russell Webb. Theory, analysis, and best practices for sigmoid self-attention. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[12] Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[13] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[14] Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[15] Tokio Kajitsuka and Issei Sato. On the optimal memorization capacity of transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[16] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: A 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)*, 2019.

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[18] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. Cosformer: Rethinking softmax in attention. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[19] Jing Liu, Zizheng Pan, Haoyu He, Jianfei Cai, and Bohan Zhuang. Ecoformer: Energy-saving attention with linear complexity. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[21] Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[22] Yingyi Chen, Qinghua Tao, Francesco Tonin, and Johan A.K. Suykens. Primal attention: Self-attention through asymmetric kernel svd in primal representation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[23] Pavlo Vasylenko, Hugo Pitorro, Andre Martins, and Marcos Vinicius Treviso. Long-context generalization with sparse attention. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

[24] Arindam Banerjee, Xinhua Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

[25] Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

[26] Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21:1–69, 2020.

[27] Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[28] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

[29] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1/2):479–487, 1988.

[30] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

A Fenchel-Young Inequality and Bregman Divergence

Following [7], in this section we review the Fenchel-Young inequality from the definition of convex conjugates, and present the corresponding representation of Bregman divergences. Let $\psi : \Theta \subset \mathbb{R}^d \to \mathbb{R}$ be a differentiable and strictly convex function, and define its Legendre conjugate $\psi^* : \Xi \to \mathbb{R}$ by

$$\psi^*(\boldsymbol{\xi}) = \sup_{\boldsymbol{\theta}\in\Theta}\{\langle \boldsymbol{\theta}, \boldsymbol{\xi}\rangle - \psi(\boldsymbol{\theta})\}. \tag{35}$$

Fenchel-Young Inequality For any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\xi} \in \Xi$, by the definition of the convex conjugate we have

$$\psi^*(\boldsymbol{\xi}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\xi}\rangle - \psi(\boldsymbol{\theta}). \tag{36}$$

Therefore,

$$\psi(\boldsymbol{\theta}) + \psi^*(\boldsymbol{\xi}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\xi}\rangle \tag{37}$$

holds for all $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. This inequality is known as the Fenchel-Young inequality.

Furthermore, if ψ is differentiable and strictly convex, and the supremum in the definition of ψ^* is attained at an interior point, the maximizer $\boldsymbol{\theta}'$ is unique and satisfies the optimality condition

$$\boldsymbol{\xi} = \nabla\psi(\boldsymbol{\theta}'). \tag{38}$$

In this case,

$$\psi^*(\boldsymbol{\xi}) = \langle \boldsymbol{\theta}', \boldsymbol{\xi} \rangle - \psi(\boldsymbol{\theta}') \tag{39}$$

holds, and therefore

$$\psi(\boldsymbol{\theta}') + \psi^*(\boldsymbol{\xi}) = \langle \boldsymbol{\theta}', \boldsymbol{\xi} \rangle, \quad \boldsymbol{\xi} = \nabla\psi(\boldsymbol{\theta}') \tag{40}$$

is obtained. This relation is known as the Fenchel-Young equality. In addition, under the usual conditions for Legendre duality, the gradient mapping $\nabla\psi : \Theta \to \Xi$ is bijective, and its inverse coincides with $\nabla\psi^*$. Consequently,

$$\boldsymbol{\xi} = \nabla\psi(\boldsymbol{\theta}') \quad \Longleftrightarrow \quad \boldsymbol{\theta}' = \nabla\psi^*(\boldsymbol{\xi}) \tag{41}$$

holds.

Fenchel-Young Representation of the Bregman Divergence We first consider the primal Bregman divergence

$$D_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}') - \langle \nabla\psi(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle. \tag{42}$$

Letting $\boldsymbol{\xi}' = \nabla\psi(\boldsymbol{\theta}')$, we obtain

$$D_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}') - \langle \boldsymbol{\xi}', \boldsymbol{\theta} \rangle + \langle \boldsymbol{\xi}', \boldsymbol{\theta}' \rangle \tag{43}$$

$$= \psi(\boldsymbol{\theta}) + \big(\langle \boldsymbol{\theta}', \boldsymbol{\xi}' \rangle - \psi(\boldsymbol{\theta}') \big) - \langle \boldsymbol{\theta}, \boldsymbol{\xi}' \rangle. \tag{44}$$

By the Fenchel-Young equality (40), we have $\langle \boldsymbol{\theta}', \boldsymbol{\xi}' \rangle - \psi(\boldsymbol{\theta}') = \psi^*(\boldsymbol{\xi}')$, and hence

$$D_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) + \psi^*(\boldsymbol{\xi}') - \langle \boldsymbol{\theta}, \boldsymbol{\xi}' \rangle. \tag{45}$$

Similarly, for the dual Bregman divergence

$$D_{\psi^*}(\boldsymbol{\xi}' \| \boldsymbol{\xi}) = \psi^*(\boldsymbol{\xi}') - \psi^*(\boldsymbol{\xi}) - \langle \nabla\psi^*(\boldsymbol{\xi}), \boldsymbol{\xi}' - \boldsymbol{\xi} \rangle, \tag{46}$$

let $\boldsymbol{\theta} = \nabla\psi^*(\boldsymbol{\xi})$. Then, using the Fenchel–Young equality

$$\psi(\boldsymbol{\theta}) + \psi^*(\boldsymbol{\xi}) = \langle \boldsymbol{\theta}, \boldsymbol{\xi} \rangle, \tag{47}$$

we obtain

$$\psi^*(\boldsymbol{\xi}) + \langle \nabla\psi^*(\boldsymbol{\xi}), \boldsymbol{\xi}' - \boldsymbol{\xi} \rangle = \langle \boldsymbol{\theta}, \boldsymbol{\xi}' \rangle - \psi(\boldsymbol{\theta}), \tag{48}$$

which yields

$$D_{\psi^*}(\boldsymbol{\xi}' \| \boldsymbol{\xi}) = \psi^*(\boldsymbol{\xi}') + \psi(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \boldsymbol{\xi}' \rangle. \tag{49}$$

B Relation to Additional Mechanisms

Transformer architectures incorporate several mechanisms in addition to the basic attention operation. These mechanisms affect different stages of the attention computation and, consequently, interact with the dual structure considered in this paper in different ways.

Positional encoding acts on the representations from which queries and keys are formed, prior to the computation of attention scores. Thus, it modifies the queries and keys themselves, but does not alter the rule by which scores are computed from queries and keys, nor the subsequent score-to-weight map. The dual structure of this map is independent of both the representation transformation and the specific form of the score computation, and hence applies equally to the transformed representations.

Masked attention constrains the components over which attention weights are computed by excluding selected scores from the normalization. In this case, the score-to-weight map is defined in the same way, but its domain is restricted to the subset of unmasked components.

Grouped-query attention can be viewed as a variant of multi-head attention in which several query heads share the same keys and values. This sharing changes the coupling across heads, but each query head still computes its attention weights through its own score-to-weight map. Hence, the head-wise dual structure used in our analysis remains applicable.

MLP blocks are nonlinear transformations placed outside the attention operation. They may change the representations supplied to or produced by attention, but they are not part of the score computation or the normalization step. They therefore lie outside the dual structure analyzed here.

Thus, these additional mechanisms may transform representations, restrict the domain of normalization, or replicate the attention computation across heads, but they do not modify the dual form of the attention computation identified in this paper.

C Proof of Theorem 5.1

In this section, we derive Theorem 5.1 by expanding the Bregman divergence associated with a scaled potential function.

Let ψ be a differentiable and strictly convex potential function, and define the scaled potential by

$$\psi_\tau(\boldsymbol{q}) = \psi(\boldsymbol{q}/\tau). \tag{50}$$

Then, by definition, its Legendre conjugate is given by

$$\begin{aligned}\psi_\tau^*(\boldsymbol{k}) &= \sup_{\boldsymbol{q}}\{\langle \boldsymbol{k}, \boldsymbol{q}\rangle - \psi(\boldsymbol{q}/\tau)\} &(51)\\ &= \sup_{\boldsymbol{u}}\{\langle \tau\boldsymbol{k}, \boldsymbol{u}\rangle - \psi(\boldsymbol{u})\} &(52)\\ &= \psi^*(\tau\boldsymbol{k}). &(53)\end{aligned}$$

By definition, the Bregman divergence induced by ψ_τ is

$$D_{\psi_\tau}\left(\boldsymbol{q}^i \middle\| \nabla\psi_\tau^*(\boldsymbol{k}^j)\right) = \psi_\tau(\boldsymbol{q}^i) - \psi_\tau\left(\nabla\psi_\tau^*(\boldsymbol{k}^j)\right) - \left\langle \nabla\psi_\tau\left(\nabla\psi_\tau^*(\boldsymbol{k}^j)\right), \boldsymbol{q}^i - \nabla\psi_\tau^*(\boldsymbol{k}^j)\right\rangle. \tag{54}$$

Using the identities $\psi_\tau(\boldsymbol{x}) = \psi(\boldsymbol{x}/\tau)$ and $\psi_\tau^*(\boldsymbol{k}) = \psi^*(\tau\boldsymbol{k})$, together with $\nabla\psi_\tau(\boldsymbol{x}) = \frac{1}{\tau}\nabla\psi(\boldsymbol{x}/\tau)$ and $\nabla\psi_\tau^*(\boldsymbol{k}) = \tau\,\nabla\psi^*(\tau\boldsymbol{k})$, we obtain

$$D_{\psi_\tau}\left(\boldsymbol{q}^i \,\middle\|\, \nabla\psi_\tau^*(\boldsymbol{k}^j)\right) = \psi\left(\frac{\boldsymbol{q}^i}{\tau}\right) - \psi\left(\nabla\psi^*(\tau\boldsymbol{k}^j)\right) - \left\langle \boldsymbol{k}^j, \boldsymbol{q}^i - \tau\,\nabla\psi^*(\tau\boldsymbol{k}^j)\right\rangle. \tag{55}$$

Applying the Fenchel-Young equality $\psi(\boldsymbol{x}) + \psi^*(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$ with $\boldsymbol{y} = \nabla\psi(\boldsymbol{x})$, we obtain

$$D_{\psi_\tau}\left(\boldsymbol{q}^i \,\middle\|\, \nabla\psi_\tau^*(\boldsymbol{k}^j)\right) = \psi\left(\frac{\boldsymbol{q}^i}{\tau}\right) + \psi^*(\tau\boldsymbol{k}^j) - \langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle. \tag{56}$$

Therefore, the attention score $s_j^i(\tau)$ can be written as

$$s_j^i(\tau) = -D_{\psi_\tau}\left(\boldsymbol{q}^i \,\middle\|\, \nabla\psi_\tau^*(\boldsymbol{k}^j)\right) = -\psi\left(\frac{\boldsymbol{q}^i}{\tau}\right) - \psi^*(\tau\boldsymbol{k}^j) + \langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle. \tag{57}$$

Among these terms, the first term $-\psi(\boldsymbol{q}^i/\tau)$ is a constant that does not depend on the index j. Writing the softmax normalization explicitly, we have

$$a_j^i(\tau) = \frac{\exp\left(\langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle - \psi^*(\tau\boldsymbol{k}^j) - \psi(\boldsymbol{q}^i/\tau)\right)}{\sum_{j'}\exp\left(\langle \boldsymbol{q}^i, \boldsymbol{k}^{j'}\rangle - \psi^*(\tau\boldsymbol{k}^{j'}) - \psi(\boldsymbol{q}^i/\tau)\right)} = \frac{\exp\left(\langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle - \psi^*(\tau\boldsymbol{k}^j)\right)}{\sum_{j'}\exp\left(\langle \boldsymbol{q}^i, \boldsymbol{k}^{j'}\rangle - \psi^*(\tau\boldsymbol{k}^{j'})\right)}. \tag{58}$$

Therefore, the term $-\psi(\boldsymbol{q}^i/\tau)$ cancels out in the softmax normalization, and the effective score reduces to

$$\tilde{s}_j^i(\tau) = \langle \boldsymbol{q}^i, \boldsymbol{k}^j\rangle - \psi^*(\tau\boldsymbol{k}^j). \tag{59}$$

This coincides with (22).

Furthermore, if $\psi^*(\tau\boldsymbol{k}^j)$ converges to a constant independent of j in the limit $\tau \to 0$, this term can also be eliminated by the softmax normalization. As a result, the attention weights converge to the standard softmax attention based solely on inner products.

D A Numerical Illustration of Temperature-Dependent Attention Behavior

In this section, we present numerical experiments illustrating the behavior of the attention scores involving the temperature parameter τ, as well as the resulting attention weights determined from them.

First, the attention score is defined by (22) as

$$\tilde{s}_j(\tau) = \langle \boldsymbol{q}, \boldsymbol{k}^j\rangle - \psi^*(\tau\boldsymbol{k}^j). \tag{22}$$

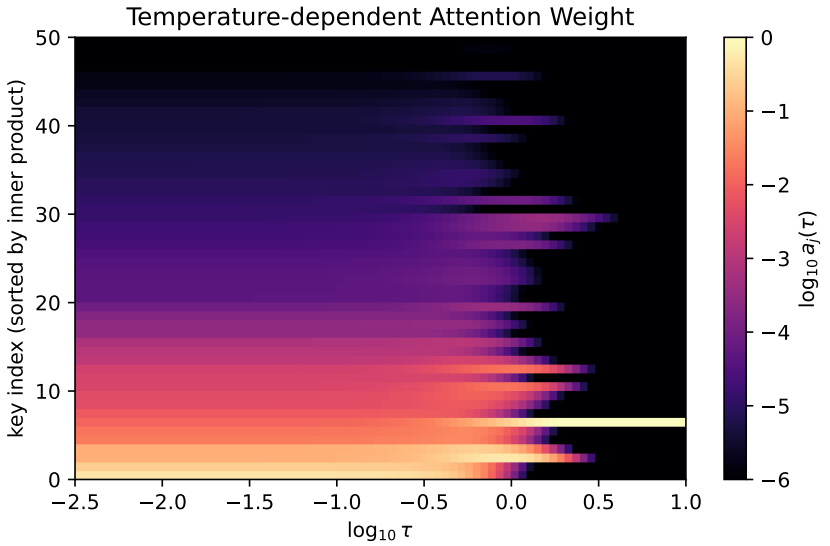

Figure 2: Heatmap showing how attention weights vary with the temperature τ. The horizontal axis represents the temperature τ (on a logarithmic scale), and the vertical axis represents the indices of the keys, ordered by the magnitude of the inner product $\langle \boldsymbol{q}, \boldsymbol{k}^j \rangle$.

Here, the first term represents the similarity given by the inner product between the query and the key, while the second term is a potential term which depends on the magnitude of the key.

The attention weights $a_j(\tau)$ obtained from these scores via the softmax are then defined as

$$a_j(\tau) = \frac{\exp\left(\tilde{s}_j(\tau)\right)}{\sum_{l=1}^{n} \exp\left(\tilde{s}_l(\tau)\right)}. \tag{60}$$

In this experiment, we generate a single query $\boldsymbol{q}$ of dimension $d = 16$ and use $n = 50$ keys $\{\boldsymbol{k}^j\}$. The query $\boldsymbol{q}$ and each key $\boldsymbol{k}^j$ are independently drawn from a normal distribution with mean 0 and variance 1. We then vary the temperature parameter τ and compute the corresponding attention weights $\{a_j(\tau)\}_{j=1}^n$ for each value of τ. Figure 2 visualizes the resulting attention weights as a heatmap, where the horizontal axis represents the temperature τ, and the vertical axis represents the indices of the keys, ordered by decreasing values of the inner product $\langle \boldsymbol{q}, \boldsymbol{k}^j \rangle$, which corresponds to the ordering at $\tau = 0$.

The horizontal axis and the color scale are shown on a logarithmic scale. As can be seen from the figure, as τ decreases, the relative contribution of the correction term $\psi^*(\tau \boldsymbol{k}^j)$ in (22) becomes weaker, and the ranking induced by the inner-product term becomes dominant. As a result, the relative ordering of attention scores changes, and rank reversals can occur across different values of τ. This experiment was conducted on a standard laptop computer and completed within a short execution time.

E Score-Attention Correspondence as Entropy-Regularized Optimization

For the potential function φ^* defined in (23), its Legendre conjugate φ is given by

$$\varphi(\boldsymbol{s}) = \sup_{\boldsymbol{a} \in \Delta^{n-1}} \left\{ \langle \boldsymbol{a}, \boldsymbol{s} \rangle - \varphi^*(\boldsymbol{a}) \right\}. \tag{61}$$

This optimization problem defines a mapping from the score $\boldsymbol{s}$ to the weight vector $\boldsymbol{a}$ under the dual structure induced by the potential φ^* on the simplex Δ^{n-1}. In particular, this correspondence is uniquely characterized by the gradient mapping

$$\boldsymbol{a} = \nabla \varphi(\boldsymbol{s}). \tag{62}$$

In (61), the first term $\langle \boldsymbol{a}, \boldsymbol{s} \rangle$ represents the similarity between the score and the weight vector, and the second term $\varphi^*(\boldsymbol{a})$ is interpreted as a regularization term that specifies the shape of admissible weight distributions on the probability simplex. For example, when $\varphi^*(\boldsymbol{a})$ is chosen as the negative Shannon entropy, (61) becomes the conjugate formulation corresponding to the softmax function, resulting in a regularized optimization problem that enhances consistency with the score while suppressing excessive concentration of the distribution.

We can generalize this formulation by considering Tsallis entropy [29] as the regularization term. For $q > 1$, the Tsallis entropy is defined as

$$H_q(\boldsymbol{a}) = \frac{1}{q-1}\Big(1 - \sum_{j=1}^{n} a_j^q\Big), \quad \boldsymbol{a} \in \Delta^{n-1}, \tag{63}$$

which reduces to the Shannon entropy in the limit $q \to 1$. In this case, by defining

$$\varphi^*(\boldsymbol{a}) = -H_q(\boldsymbol{a}) = \frac{1}{q-1}\sum_{j=1}^{n} a_j^q + \text{const.}, \tag{64}$$

the expression in (61) can be written as

$$\varphi(\boldsymbol{s}) = \sup_{\boldsymbol{a} \in \Delta^{n-1}} \left\{ \langle \boldsymbol{a}, \boldsymbol{s} \rangle - \frac{1}{q-1}\sum_{j=1}^{n} a_j^q \right\}. \tag{65}$$

We note that an optimization problem isomorphic to (65) has also been studied in [27] as a Tsallis-entropy-based generalization of attention mechanisms. From the perspective of Theorem 6.1, this formulation can be viewed as a concrete instance of a score-to-weight mapping that shares the potential-based structure of the theorem without satisfying all its assumptions globally. A discussion of this relation, together with the derivation of the resulting attention weights and their basic properties, such as the fact that only components whose scores exceed a certain threshold receive positive weights, while the remaining components are assigned exactly zero weights, is provided in Appendix F. Within the general framework developed here, Tsallis-entropy-based attention is therefore understood not as an isolated modification, but as a natural instance arising from a particular choice of convex potential function.

F Derivation and Properties of the Optimal Solution to (65)

In this section, we derive the optimal solution to the Tsallis-entropy-regularized constrained optimization problem corresponding to (65), and discuss its fundamental properties.

Problem setting The optimization problem under consideration is

$$\max_{\boldsymbol{a} \in \mathbb{R}^n} \quad \sum_{i=1}^{n} a_i s_i - \frac{1}{q-1}\sum_{i=1}^{n} a_i^q \tag{66}$$

$$\text{s.t.} \quad a_i \geq 0 \quad (i = 1, \ldots, n), \tag{67}$$

$$\sum_{i=1}^{n} a_i = 1, \tag{68}$$

where $q > 1$.

Since the function $t \mapsto -t^q$ is strictly concave on $t \geq 0$, the objective function (66) is strictly concave with respect to $\boldsymbol{a}$. Moreover, the constraint set

$$\Delta^{n-1} = \{\boldsymbol{a} \in \mathbb{R}^n : a_i \geq 0,\ \sum_i a_i = 1\} \tag{69}$$

is compact and convex. Therefore, the optimal solution to this problem exists and is unique.

Lagrangian and KKT conditions We rewrite the inequality constraints (67) as $-a_i \le 0$ and introduce the corresponding Lagrange multipliers $\mu_i \ge 0$, and denote by $\lambda \in \mathbb{R}$ the Lagrange multiplier associated with the equality constraint (68). The Lagrangian is then given by

$$\mathcal{L}(\boldsymbol{a}, \lambda, \boldsymbol{\mu}) = \sum_{i=1}^{n} a_i s_i - \frac{1}{q-1} \sum_{i=1}^{n} a_i^q + \lambda\Big(1 - \sum_{i=1}^{n} a_i\Big) + \sum_{i=1}^{n} \mu_i a_i. \tag{70}$$

By the KKT conditions, the optimal solution $(\boldsymbol{a}^*, \lambda^*, \boldsymbol{\mu}^*)$ satisfies the following:

1. (Stationarity)
$$\frac{\partial \mathcal{L}}{\partial a_i} = s_i - \frac{q}{q-1}(a_i^*)^{q-1} - \lambda^* + \mu_i^* = 0, \quad i = 1, \dots, n; \tag{71}$$
2. (Feasibility and complementary slackness)
$$a_i^* \ge 0, \quad \mu_i^* \ge 0, \quad \mu_i^* a_i^* = 0, \quad i = 1, \dots, n; \tag{72}$$
3. (Normalization constraint)
$$\sum_{i=1}^{n} a_i^* = 1. \tag{73}$$

Closed-form of the optimal solution When $a_i^* > 0$, the complementary slackness condition (72) implies $\mu_i^* = 0$, and from (71) we obtain

$$(a_i^*)^{q-1} = \frac{q-1}{q}(s_i - \lambda^*). \tag{74}$$

On the other hand, when $a_i^* = 0$, the complementary slackness condition implies no restriction on μ_i^* other than $\mu_i^* \ge 0$. Substituting $a_i^* = 0$ into (71), we obtain $s_i - \lambda^* = -\mu_i^*$, and hence

$$s_i - \lambda^* \le 0. \tag{75}$$

Combining these cases, the optimal solution can be written as

$$a_i^* = \left[\frac{q-1}{q}(s_i - \lambda^*)\right]_+^{\frac{1}{q-1}}, \quad [x]_+ = \max\{x, 0\}, \tag{76}$$

which implies

$$a_i^* > 0 \quad \Longleftrightarrow \quad s_i > \lambda^*. \tag{77}$$

Therefore, the support of the attention distribution, $\{i : s_i > \lambda^*\}$, is finite, and components satisfying $s_i \le \lambda^*$ are assigned exactly zero weight. In this respect, the attention distribution in (76) differs from the Shannon-entropy–regularized case, where all components receive strictly positive weights. However, in the limit $q \to 1$, (76) converges to the standard softmax, and the corresponding potential function $\varphi(\boldsymbol{s})$ coincides with $\log \sum_{j=1}^{n} \exp(s_j)$. Thus, Tsallis-entropy regularization can be understood as a framework that continuously interpolates between diffuse, softmax-based weighting and threshold-based, selective weighting.

Finally, for any constant $c \in \mathbb{R}$, letting $s_i' = s_i + c$ yields the corresponding dual variable $\lambda'^* = \lambda^* + c$, and the optimal solution satisfies $\boldsymbol{a}^*(\boldsymbol{s}') = \boldsymbol{a}^*(\boldsymbol{s})$. Hence, the attention distribution is invariant under translations of the scores.

Uniqueness of the dual variable λ The parameter λ is determined by the normalization condition (73) so as to satisfy

$$\sum_{i=1}^{n} \left[\frac{q-1}{q}(s_i - \lambda)\right]_+^{\frac{1}{q-1}} = 1. \tag{78}$$

Define the left-hand side as

$$\phi(\lambda) = \sum_{i=1}^{n} \left[\frac{q-1}{q}(s_i - \lambda)\right]_+^{\frac{1}{q-1}}. \tag{79}$$

Then $\phi(\lambda)$ is continuous and monotonically decreasing with respect to λ. Moreover,

$$\lambda \ge \max_i s_i \;\Rightarrow\; \phi(\lambda) = 0, \quad \lambda \to -\infty \;\Rightarrow\; \phi(\lambda) \to \infty. \tag{80}$$

Therefore, by the intermediate value theorem together with monotonicity, there exists a unique λ satisfying (78).

Relation to Theorem 6.1 We briefly relate the Tsallis-entropy-based formulation discussed above to Theorem 6.1. As shown in this section, the mapping $\boldsymbol{s} \mapsto \boldsymbol{a}^*(\boldsymbol{s})$ defined by (76) maps scores to the probability simplex by construction. Moreover, the invariance $\boldsymbol{a}^*(\boldsymbol{s} + c\mathbf{1}) = \boldsymbol{a}^*(\boldsymbol{s})$ follows from the fact that shifting the scores as $s_i \mapsto s_i + c$ induces the corresponding shift $\lambda^* \mapsto \lambda^* + c$, so that $(s_i + c) - (\lambda^* + c) = s_i - \lambda^*$ remains unchanged. The closed-form solution in (76) also shows that, on regions where the set of positive components remains fixed, each positive component is a nondecreasing function of its own score and a nonincreasing function of the other scores. Since the attention weights are obtained from the Tsallis-Legendre potential φ in (65), the corresponding Jacobian is symmetric wherever φ is twice differentiable.

The sparse cases, however, are not globally covered by Theorem 6.1. Sparsemax and Tsallis-based entmax can assign exactly zero weight to some components and are not differentiable at points where the set of positive components changes. They therefore do not satisfy the strict positivity and smoothness assumptions of the theorem over the entire score space. On each region where the set of positive components remains unchanged, the active components are smooth functions of the scores and the corresponding Jacobian block is symmetric. Thus, the sparse constructions share the same local potential structure on such regions, while Theorem 6.1 characterizes the globally smooth and strictly positive case.

Translation invariance at the level of the potential The translation invariance of the resulting attention weights can also be understood directly at the level of the potential function. For any $c \in \mathbb{R}$, we have

$$\varphi(\boldsymbol{s} + c\mathbf{1}) = \sup_{\boldsymbol{a} \in \Delta^{n-1}} \left\{ \langle \boldsymbol{a}, \boldsymbol{s} + c\mathbf{1} \rangle - \frac{1}{q-1} \sum_{j=1}^{n} a_j^q \right\} \tag{81}$$

$$= \sup_{\boldsymbol{a} \in \Delta^{n-1}} \left\{ \langle \boldsymbol{a}, \boldsymbol{s} \rangle + c\langle \boldsymbol{a}, \mathbf{1} \rangle - \frac{1}{q-1} \sum_{j=1}^{n} a_j^q \right\}. \tag{82}$$

Since $\langle \boldsymbol{a}, \mathbf{1} \rangle = 1$ for any $\boldsymbol{a} \in \Delta^{n-1}$, this yields

$$\varphi(\boldsymbol{s} + c\mathbf{1}) = \varphi(\boldsymbol{s}) + c. \tag{83}$$

Consequently, taking gradients with respect to $\boldsymbol{s}$ gives

$$\nabla\varphi(\boldsymbol{s} + c\mathbf{1}) = \nabla\varphi(\boldsymbol{s}), \tag{84}$$

showing that the induced score-to-weight mapping is invariant under translations of the score vector. This argument has the same structure as in the log-sum-exp case.

We emphasize that this argument does not rely on the specific form of the Tsallis-type regularizer. Rather, the translation invariance $\nabla\varphi(\boldsymbol{s} + c\mathbf{1}) = \nabla\varphi(\boldsymbol{s})$ follows directly from the structural form of the Legendre-type representation, in which the score vector enters only through the linear term $\langle \boldsymbol{a}, \boldsymbol{s} \rangle$ under the simplex constraint $\langle \boldsymbol{a}, \mathbf{1} \rangle = 1$. In this sense, translation invariance is a generic consequence of expressing attention weights via a Legendre transform, and not an incidental property specific to a particular choice of potential.

G Proof of Theorem 6.1

First, by the simplex constraint, we have

$$\sum_{i=1}^{n} g_i(\boldsymbol{s}) = 1, \quad (\forall \boldsymbol{s}). \tag{85}$$

Differentiating both sides with respect to s_j yields

$$\sum_{i=1}^{n} \frac{\partial g_i}{\partial s_j}(\boldsymbol{s}) = 0, \quad (j = 1, \ldots, n). \tag{86}$$

This implies that the sum of each column of the Jacobian matrix $J(\boldsymbol{s}) = \big(J_{ij}(\boldsymbol{s})\big)_{i,j}$ is zero, where the row index i corresponds to the output component g_i and the column index j corresponds to the input coordinate s_j, as a direct consequence of the normalization condition $\sum_i g_i(\boldsymbol{s}) = 1$.

Next, by translation invariance, the mapping g satisfies

$$g(\boldsymbol{s}) = g(\boldsymbol{s} + c\mathbf{1}), \quad (\forall c \in \mathbb{R}), \tag{87}$$

where $\mathbf{1} = (1, \ldots, 1)^\top$. For each fixed s, the right-hand side defines a composite function $c \mapsto g(\boldsymbol{s} + c\mathbf{1})$. Differentiating this composite function with respect to c and applying the chain rule, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}c} g_i(\boldsymbol{s} + c\mathbf{1}) = \sum_{j=1}^{n} \frac{\partial g_i}{\partial s_j}(\boldsymbol{s} + c\mathbf{1}) \frac{\partial (s_j + c)}{\partial c} = \sum_{j=1}^{n} \frac{\partial g_i}{\partial s_j}(\boldsymbol{s} + c\mathbf{1}), \tag{88}$$

for each $i = 1, \ldots, n$. Since $g(s + c\mathbf{1})$ is independent of c, the left-hand side vanishes. Evaluating the above expression at $c = 0$ yields

$$\sum_{j=1}^{n} \frac{\partial g_i}{\partial s_j}(\boldsymbol{s}) = 0, \quad (i = 1, \ldots, n). \tag{89}$$

Hence, the sum of each row of the Jacobian matrix $J(\boldsymbol{s})$ is zero.

By the monotonicity assumption, we have $J_{ii}(\boldsymbol{s}) > 0$ and $J_{ij}(\boldsymbol{s}) < 0$ for $i \neq j$. Moreover, from (89),

$$J_{ii}(\boldsymbol{s}) = -\sum_{j \neq i} J_{ij}(\boldsymbol{s}). \tag{90}$$

Therefore, for any $x \in \mathbb{R}^n$,

$$x^\top J(\boldsymbol{s}) x = \sum_i J_{ii}(\boldsymbol{s}) x_i^2 + \sum_{i \neq j} J_{ij}(\boldsymbol{s}) x_i x_j \tag{91}$$

$$= \frac{1}{2} \sum_{i \neq j} \big(-J_{ij}(\boldsymbol{s})\big)(x_i - x_j)^2. \tag{92}$$

Since $-J_{ij}(\boldsymbol{s}) > 0$, it follows that

$$x^\top J(\boldsymbol{s}) x = 0 \quad \Longleftrightarrow \quad x_i = x_j, \ (\forall i, j). \tag{93}$$

At this point, it is important to clarify the scope in which positive definiteness can be meaningfully discussed. From (93), $x^\top J(s) x$ vanishes if and only if $x_i = x_j$ for all i, j, that is, when x is proportional to the vector $\mathbf{1}$. This is consistent with the assumption of translation invariance. Indeed, translation invariance yields (89), which implies

$$J(\boldsymbol{s})\mathbf{1} = \mathbf{0}. \tag{94}$$

As a consequence, the vector $\mathbf{1}$ necessarily belongs to the kernel of the Jacobian, and $J(\boldsymbol{s})$ cannot be positive definite on the whole space $\mathbb{R}^n$.

To exclude this trivial direction and examine all remaining variations, we restrict attention to vectors that are not constant. A convenient way to eliminate the constant direction is to impose

$$\sum_{i=1}^{n} x_i = 0. \tag{95}$$

Indeed, any nonzero constant vector $x = c\mathbf{1}$ satisfies $\sum_{i=1}^{n} x_i = nc \neq 0$, whereas the only constant vector satisfying $\sum_{i=1}^{n} x_i = 0$ is the zero vector. Therefore, the subspace

$$\left\{ \boldsymbol{x} \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i = 0 \right\} \tag{96}$$

excludes the constant direction $\mathrm{span}\{\mathbf{1}\}$. Note that, since $\sum_{i=1}^{n} x_i = \langle \boldsymbol{x}, \mathbf{1} \rangle$ and, by definition,

$$\mathrm{span}\{\mathbf{1}\}^\perp = \{\boldsymbol{x} \in \mathbb{R}^n \mid \langle \boldsymbol{x}, \mathbf{1} \rangle = 0\}, \tag{97}$$

we obtain

$$\left\{ \boldsymbol{x} \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i = 0 \right\} = \mathrm{span}\{\mathbf{1}\}^{\perp}. \tag{98}$$

For any nonzero $x \in \mathrm{span}\{\mathbf{1}\}^{\perp}$, at least one difference $x_i - x_j$ is nonzero, so that the quadratic form in (92) is strictly positive. Since $-J_{ij}(\boldsymbol{s}) > 0$ for all $i \neq j$, we conclude that

$$\boldsymbol{x}^{\top} J(\boldsymbol{s}) \boldsymbol{x} > 0 \quad \text{for all } \boldsymbol{x} \in \mathrm{span}\{\mathbf{1}\}^{\perp} \setminus \{0\}. \tag{99}$$

To show the existence of a potential function explicitly, we construct it directly. Define $\varphi : \mathbb{R}^n \to \mathbb{R}$ as

$$\varphi(\boldsymbol{s}) = \int_0^1 \langle g(t\boldsymbol{s}),\, \boldsymbol{s} \rangle \, \mathrm{d}t, \tag{100}$$

where $t \in [0, 1]$ parametrizes the straight line from the origin to $\boldsymbol{s}$.

We show that $\nabla\varphi(\boldsymbol{s}) = g(\boldsymbol{s})$. Since g is C^1, we may differentiate under the integral sign. For each $i = 1, \ldots, n$, we compute

$$\frac{\partial \varphi}{\partial s_i}(\boldsymbol{s}) = \int_0^1 \frac{\partial}{\partial s_i} \left(\sum_{k=1}^{n} g_k(t\boldsymbol{s}) s_k \right) \mathrm{d}t \tag{101}$$

$$= \int_0^1 g_i(t\boldsymbol{s}) \mathrm{d}t + \int_0^1 t \sum_{k=1}^{n} s_k \frac{\partial g_k}{\partial s_i}(t\boldsymbol{s}) \mathrm{d}t. \tag{102}$$

By the symmetry of the Jacobian, $\partial g_k / \partial s_i = \partial g_i / \partial s_k$, and hence

$$\sum_{k=1}^{n} s_k \frac{\partial g_k}{\partial s_i}(t\boldsymbol{s}) = \sum_{k=1}^{n} s_k \frac{\partial g_i}{\partial s_k}(t\boldsymbol{s}). \tag{103}$$

By the chain rule, the right-hand side satisfies

$$\sum_{k=1}^{n} s_k \frac{\partial g_i}{\partial s_k}(t\boldsymbol{s}) = \frac{\mathrm{d}}{\mathrm{d}t} g_i(t\boldsymbol{s}). \tag{104}$$

Hence,

$$\frac{\partial \varphi}{\partial s_i}(\boldsymbol{s}) = \int_0^1 g_i(t\boldsymbol{s}) \mathrm{d}t + \int_0^1 t \frac{\mathrm{d}}{\mathrm{d}t} g_i(t\boldsymbol{s}) \mathrm{d}t. \tag{105}$$

The second term is evaluated by integration by parts:

$$\int_0^1 t \frac{\mathrm{d}}{\mathrm{d}t} g_i(t\boldsymbol{s}) \mathrm{d}t = \Big[t g_i(t\boldsymbol{s}) \Big]_0^1 - \int_0^1 g_i(t\boldsymbol{s}) \mathrm{d}t \tag{106}$$

$$= g_i(\boldsymbol{s}) - \int_0^1 g_i(t\boldsymbol{s}) \mathrm{d}t. \tag{107}$$

Substituting this into the previous expression, the integral terms cancel, and we obtain

$$\frac{\partial \varphi}{\partial s_i}(\boldsymbol{s}) = g_i(\boldsymbol{s}). \tag{108}$$

Since this holds for all $i = 1, \ldots, n$, it follows that

$$\nabla\varphi(\boldsymbol{s}) = g(\boldsymbol{s}). \tag{109}$$

In this case, $\nabla^2\varphi(\boldsymbol{s}) = J(\boldsymbol{s})$ holds, and from the positive definiteness of $J(\boldsymbol{s})$ shown above, it follows that φ is strictly convex along directions satisfying $\sum_{i=1}^{n} x_i = 0$.

Remark: Intuition behind the structural requirements The assumptions imposed in Theorem 6.1 are not technical artifacts introduced for mathematical convenience, but reflect fundamental structural requirements that a mapping from scores to attention weights is naturally expected to satisfy. We briefly discuss the role and interpretation of each condition.

(i) Translation invariance. The translation invariance condition $g(\boldsymbol{s} + c\mathbf{1}) = g(\boldsymbol{s})$ expresses the fact that attention weights should depend only on *relative* scores, not on their absolute offset. In the context of attention mechanisms, scores are typically computed from inner products or similarity measures, and adding a constant offset to all scores does not alter the relative preference among candidates. This invariance is therefore intrinsic to the notion of attention as a comparative mechanism, rather than an absolute evaluation. From a geometric viewpoint, this condition implies that the mapping g is insensitive to variations along the direction $\mathbf{1}$, which later manifests as the flat direction of the potential function.

(ii) Simplex constraint. The simplex constraint encodes the interpretation of attention weights as a probability distribution over indices. The normalization condition $\sum_{i=1}^{n} g_i(\boldsymbol{s}) = 1$ ensures that the weights form a convex combination, while strict positivity $g_i(\boldsymbol{s}) > 0$ excludes degenerate assignments and guarantees smooth dependence on the scores. This condition is standard in probabilistic formulations of attention, and is independent of any specific choice of normalization such as softmax.

(iii) Monotonicity. The monotonicity condition formalizes the minimal consistency requirement between scores and weights: increasing the score s_i of an element should increase its own weight g_i, while decreasing the weights assigned to competing elements. This captures the intuitive idea that attention reflects preference ordering induced by the scores.

(iv) Symmetry of the Jacobian. The symmetry assumption concerns the local cross-component effects of the score-to-weight map. Let $J(s)$ denote the Jacobian of g. The entry $J_{ij}(s) = \partial g_i / \partial s_j(s)$ measures the local effect of the j-th score on the i-th attention weight. The condition $J_{ij}(s) = J_{ji}(s)$ requires the corresponding cross-effects between two components to be symmetric. Intuitively, this prevents the score-to-weight map itself from introducing directional asymmetries between components that are not induced by the score values. In particular, when the scores do not distinguish between two components, an asymmetric local interaction between them would have to arise from the normalization map itself.

Taken together, these conditions characterize attention-weight mappings that are translation-invariant, normalized, monotone with respect to the scores, and do not introduce such additional directional asymmetries. Theorem 6.1 shows that a mapping satisfying these conditions admits a representation as the gradient of a convex potential.

Mathematically, assumption (iv) also corresponds to the integrability condition for the score-to-weight map on the domain considered in Theorem 6.1: symmetry of the Jacobian ensures that g can be represented as the gradient of a potential function. The following example makes the role of this assumption more explicit by introducing an antisymmetric component into the normalization map and showing that assumptions (i)-(iii) can still hold while assumption (iv) fails.

Role of the Jacobian symmetry assumption. Assumption (iv) has a direct mathematical relation to the potential representation in Theorem 6.1. On the simply connected domain considered here, symmetry of the Jacobian is the integrability condition for the existence of a potential φ satisfying $g = \nabla\varphi$. Thus, assumption (iv) is not independent of the existence of the potential representation. The remaining assumptions, together with (iv), further determine the convexity properties of the resulting potential.

At the same time, assumption (iv) is not implied by assumptions (i)-(iii). The following example gives a normalization map that satisfies translation invariance, the simplex constraint, and monotonicity, while having asymmetric cross-component effects.

Let $n = 3$, let

$$p(\boldsymbol{s}) = \mathrm{softmax}(\boldsymbol{s}), \tag{110}$$

and define

$$A = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}. \tag{111}$$

For $0 < |\varepsilon| < 1/3$, define

$$g_i(\boldsymbol{s}) = p_i(\boldsymbol{s}) \left(1 + \varepsilon (Ap(\boldsymbol{s}))_i\right), \qquad i = 1, 2, 3. \tag{112}$$

Equivalently,

$$g(\boldsymbol{s}) = \mathrm{Diag}(p(\boldsymbol{s})) \left(\mathbf{1} + \varepsilon Ap(\boldsymbol{s})\right) = \begin{pmatrix} p_1(\boldsymbol{s}) & 0 & 0 \\ 0 & p_2(\boldsymbol{s}) & 0 \\ 0 & 0 & p_3(\boldsymbol{s}) \end{pmatrix} \begin{pmatrix} 1 + \varepsilon\big(p_2(\boldsymbol{s}) - p_3(\boldsymbol{s})\big) \\ 1 + \varepsilon\big(p_3(\boldsymbol{s}) - p_1(\boldsymbol{s})\big) \\ 1 + \varepsilon\big(p_1(\boldsymbol{s}) - p_2(\boldsymbol{s})\big) \end{pmatrix}. \tag{113}$$

Since $p(\boldsymbol{s})$ is translation invariant, $g(\boldsymbol{s})$ is also translation invariant. Moreover, because A is skew-symmetric,

$$\sum_{i=1}^{3} g_i(\boldsymbol{s}) = 1 + \varepsilon p(\boldsymbol{s})^\top Ap(\boldsymbol{s}) = 1. \tag{114}$$

Since $|(Ap(\boldsymbol{s}))_i| \le 1$,

$$1 + \varepsilon (Ap(\boldsymbol{s}))_i \ge 1 - |\varepsilon| > 0. \tag{115}$$

Together with $p_i(\boldsymbol{s}) > 0$, this gives $g_i(\boldsymbol{s}) > 0$ for all i. Thus, assumptions (i) and (ii) are satisfied.

For the softmax probabilities, we have

$$\frac{\partial p_k(\boldsymbol{s})}{\partial s_j} = \frac{\partial}{\partial s_j} \frac{\exp(s_k)}{\sum_{\ell=1}^{3} \exp(s_\ell)} \tag{116}$$

$$= \frac{\delta_{kj} \exp(s_k) \sum_{\ell=1}^{3} \exp(s_\ell) - \exp(s_k) \exp(s_j)}{\left(\sum_{\ell=1}^{3} \exp(s_\ell)\right)^2} \tag{117}$$

$$= p_k(\boldsymbol{s})\big(\delta_{kj} - p_j(\boldsymbol{s})\big). \tag{118}$$

Therefore,

$$\frac{\partial (Ap(\boldsymbol{s}))_i}{\partial s_j} = \sum_{k=1}^{3} A_{ik} \frac{\partial p_k(\boldsymbol{s})}{\partial s_j} \tag{119}$$

$$= \sum_{k=1}^{3} A_{ik} p_k(\boldsymbol{s})\big(\delta_{kj} - p_j(\boldsymbol{s})\big) \tag{120}$$

$$= A_{ij} p_j(\boldsymbol{s}) - p_j(\boldsymbol{s}) \sum_{k=1}^{3} A_{ik} p_k(\boldsymbol{s}) \tag{121}$$

$$= p_j(\boldsymbol{s}) \left(A_{ij} - (Ap(\boldsymbol{s}))_i\right). \tag{122}$$

For $i \neq j$,

$$J_{ij}(\boldsymbol{s}) = \frac{\partial p_i(\boldsymbol{s})}{\partial s_j} \left(1 + \varepsilon (Ap(\boldsymbol{s}))_i\right) + \varepsilon p_i(\boldsymbol{s}) \frac{\partial (Ap(\boldsymbol{s}))_i}{\partial s_j} \tag{123}$$

$$= -p_i(\boldsymbol{s}) p_j(\boldsymbol{s}) \left(1 + \varepsilon (Ap(\boldsymbol{s}))_i\right) + \varepsilon p_i(\boldsymbol{s}) p_j(\boldsymbol{s}) \left(A_{ij} - (Ap(\boldsymbol{s}))_i\right) \tag{124}$$

$$= p_i(\boldsymbol{s}) p_j(\boldsymbol{s}) \left(-1 + \varepsilon \left[A_{ij} - 2(Ap(\boldsymbol{s}))_i\right]\right). \tag{125}$$

Since

$$|A_{ij} - 2(Ap(\boldsymbol{s}))_i| \le 3, \tag{126}$$

we have $J_{ij}(\boldsymbol{s}) < 0$ for $i \neq j$. Translation invariance implies

$$\sum_{j=1}^{3} J_{ij}(\boldsymbol{s}) = 0, \tag{127}$$

and hence $J_{ii}(\boldsymbol{s}) > 0$. Therefore, assumptions (i)-(iii) are satisfied.

However, at $\boldsymbol{s} = \mathbf{0}$,

$$p_1(\mathbf{0}) = p_2(\mathbf{0}) = p_3(\mathbf{0}) = \frac{1}{3}, \quad Ap(\mathbf{0}) = \mathbf{0}. \tag{128}$$

Consequently,

$$J_{12}(\mathbf{0}) = \frac{-1+\varepsilon}{9}, \quad J_{21}(\mathbf{0}) = \frac{-1-\varepsilon}{9}, \tag{129}$$

so that, for $\varepsilon \neq 0$,

$$J_{12}(\mathbf{0}) \neq J_{21}(\mathbf{0}). \tag{130}$$

Thus, assumption (iv) does not follow from assumptions (i)-(iii). In particular, because all scores are equal at $\boldsymbol{s} = \mathbf{0}$, the asymmetric cross-component effects in this example arise from the normalization map rather than from differences among the scores. Accordingly, Theorem 6.1 should be interpreted as a characterization of normalization maps within the class defined by assumptions (i)-(iv), rather than as a claim that every attention normalization must satisfy assumption (iv).

H Verification of the Structural Conditions for Softmax

In this section, we verify that the standard softmax normalization used in Transformer attention satisfies all of the structural conditions assumed in Theorem 6.1.

Let $g : \mathbb{R}^n \to \Delta^{n-1}$ be defined by the softmax mapping

$$g_i(\boldsymbol{s}) = \frac{\exp(s_i)}{\sum_{j=1}^n \exp(s_j)}, \qquad i = 1, \dots, n, \tag{131}$$

where $\boldsymbol{s} = (s_1, \dots, s_n) \in \mathbb{R}^n$. It is immediate from the definition that g is a C^∞ function on $\mathbb{R}^n$. We verify each assumption of Theorem 6.1 in turn.

Translation invariance. For any $c \in \mathbb{R}$,

$$g_i(\boldsymbol{s} + c\mathbf{1}) = \frac{\exp(s_i + c)}{\sum_{j=1}^n \exp(s_j + c)} = \frac{\exp(s_i)}{\sum_{j=1}^n \exp(s_j)} = g_i(s), \tag{132}$$

where $\mathbf{1} = (1, \dots, 1)^\top$. Hence, $g(s + c\mathbf{1}) = g(s)$ holds.

Simplex constraint. By definition,

$$\sum_{i=1}^n g_i(\boldsymbol{s}) = 1, \qquad g_i(\boldsymbol{s}) > 0 \quad (i = 1, \dots, n), \tag{133}$$

for all $s \in \mathbb{R}^n$. Thus, $g(s)$ takes values in the interior of the probability simplex Δ^{n-1}.

Monotonicity. The partial derivatives of g_i are given by

$$\frac{\partial g_i}{\partial s_i}(\boldsymbol{s}) = g_i(\boldsymbol{s})\big(1 - g_i(\boldsymbol{s})\big) > 0, \tag{134}$$

$$\frac{\partial g_i}{\partial s_j}(\boldsymbol{s}) = -g_i(\boldsymbol{s}) g_j(\boldsymbol{s}) < 0, \qquad (i \neq j). \tag{135}$$

Hence, the monotonicity condition is satisfied.

Symmetry of the Jacobian. The Jacobian matrix $J(\boldsymbol{s}) = (\partial g_i / \partial s_j)_{i,j=1}^n$ has entries

$$J_{ij}(s) = \begin{cases} g_i(\boldsymbol{s})\big(1 - g_i(\boldsymbol{s})\big), & i = j, \\ -g_i(\boldsymbol{s}) g_j(\boldsymbol{s}), & i \neq j. \end{cases} \tag{136}$$

Since $g_i(\boldsymbol{s}) g_j(\boldsymbol{s}) = g_j(\boldsymbol{s}) g_i(\boldsymbol{s})$, it follows immediately that

$$\frac{\partial g_i}{\partial s_j}(\boldsymbol{s}) = \frac{\partial g_j}{\partial s_i}(\boldsymbol{s}) \tag{137}$$

for all i, j, and therefore the Jacobian matrix is symmetric.

I Numerical Example of Sparse and Dense Attention

I.1 Purpose of the Experiment

This section presents a numerical experiment for examining how the choice of the score-to-weight mapping affects final prediction performance. We compare the standard softmax mapping with sparsemax, which produces sparse attention weights. This experiment is not an end-to-end training experiment for Transformers. In particular, we do not train the matrices that generate queries, keys, and values, nor do we train a classifier. Instead, we synthetically generate token representations and compute attention scores from them. We then apply softmax and sparsemax to the same score vector and examine how the resulting difference in attention weights is reflected in the aggregated representation and prediction performance.

The reason for this design is to isolate the effect of the score-to-weight mapping. If a Transformer is trained end-to-end, differences in prediction performance depend on many factors, including representation learning, query-key score learning, optimization, and downstream classifiers. It is therefore difficult to identify the effect of replacing softmax with sparsemax alone. In this experiment, we remove these factors and address the following question. Given the same token representations and the same scores, how does the final prediction change depending on whether the scores are converted into weights by softmax or sparsemax?

In particular, we control two factors. First, we compare the case where label-relevant information is localized in a small number of tokens with the case where it is distributed across many tokens. Second, we compare the case where attention scores are reliable with the case where they are noisy and less reliable. This allows us to examine when sparse attention is advantageous and when dense attention is advantageous.

I.2 Experimental Setup

I.2.1 Synthetic Data Generation

Each data point is given as a sequence of tokens of length n. Each token is a d-dimensional vector, written as $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^n)$, where $\boldsymbol{x}^j \in \mathbb{R}^d$. In this experiment, we set $n = 64$ and $d = 24$. Each data point has a binary label $y \in \{-1, +1\}$, which is sampled uniformly at random. Among the tokens, only a subset contains information useful for label prediction. We call these tokens *relevant tokens* and denote their index set by $R \subset \{1, \dots, n\}$. The remaining tokens are called *distractor tokens*. The number of relevant tokens is denoted by $m = |R|$. In this experiment, we vary $m \in \{1, 2, 4, 8, 16, 24, 32\}$. When m is small, label-relevant information is localized in a small number of tokens. When m is large, label-relevant information is distributed across many tokens.

To construct the token representations, we use two unit vectors. The first is a direction related to label prediction, $\boldsymbol{u} \in \mathbb{R}^d$. The second is a direction used for computing attention scores, $\boldsymbol{r} \in \mathbb{R}^d$. They are chosen to satisfy $\|\boldsymbol{u}\| = \|\boldsymbol{r}\| = 1$ and $\langle \boldsymbol{u}, \boldsymbol{r} \rangle = 0$.

The vector $\boldsymbol{u}$ corresponds to the weight direction of a classification head in a standard supervised learning setting. In a usual classification model, after attention produces an aggregated representation $\boldsymbol{z}$, a classification head computes a logit of the form $\ell = \langle \boldsymbol{w}, \boldsymbol{z} \rangle + b$. The weight direction $\boldsymbol{w}$ in this expression corresponds to $\boldsymbol{u}$ in our experiment. However, we do not learn the classification head $\boldsymbol{w}$. Since the label direction $\boldsymbol{u}$ is known by construction in the synthetic data generation process, we use $\boldsymbol{u}$ as a fixed classifier at prediction time. This can be understood as an idealized setting in which the classification head has already learned the correct label direction. This removes the effect of classifier training and allows us to isolate the influence of the score-to-weight mapping on prediction.

On the other hand, $\boldsymbol{r}$ is the direction used for computing attention scores. As described below, we regard $\boldsymbol{r}$ as a fixed query and compute scores by taking inner products between $\boldsymbol{r}$ and the token representations. The reason for choosing $\boldsymbol{u}$ and $\boldsymbol{r}$ to be orthogonal is to separate the information used for label prediction from the information used for score computation. This orthogonal decomposition is not intended to claim that a trained Transformer representation space actually decomposes in this way. Rather, it is an idealized assumption introduced to construct a controlled synthetic experiment. Its purpose is to remove the effects of representation learning and classifier learning, making it easier to observe the effect of the score-to-weight mapping.

In standard attention, scores are computed by inner products between queries and keys. For example, given a query $\boldsymbol{q}$ and a key $\boldsymbol{k}^j$ corresponding to token j, one may define $s_j = \langle \boldsymbol{q}, \boldsymbol{k}^j \rangle$. In self-attention, both queries and keys are produced from input tokens through learned linear transformations, but this experiment does not model that full process. Instead, to isolate the effect of the score-to-weight mapping, we consider a simpler single-query attention setting. That is, the same fixed query is used for all tokens. Specifically, we regard the unit vector $\boldsymbol{r}$ as a fixed query and set $\boldsymbol{q} = \boldsymbol{r}$. We also regard the token representation itself as the key, so that $\boldsymbol{k}^j = \boldsymbol{x}^j$. Then the attention score is given by $s_j = \langle \boldsymbol{q}, \boldsymbol{k}^j \rangle + \eta_j = \langle \boldsymbol{r}, \boldsymbol{x}^j \rangle + \eta_j$, where η_j is score noise. Thus, the score computation in this experiment, $s_j = \langle \boldsymbol{x}^j, r \rangle + \eta_j$, can be viewed as a simplified inner-product attention score using the fixed query r and token key $\boldsymbol{x}^j$. The query is fixed because our focus is not on learning query-key scores, but on the subsequent score-to-weight mapping.

We next generate token representations synthetically so that this score computation and the prediction rule are meaningful. Each token $\boldsymbol{x}^j$ is generated as the sum of four components:

$$\boldsymbol{x}^j = \boldsymbol{\xi}_j + \epsilon_j \boldsymbol{u} + \mathbb{1}[j \in R] \frac{C}{\sqrt{m}} y \boldsymbol{u} + \mathbb{1}[j \in R] \alpha \boldsymbol{r}, \quad j = 1, \ldots, n. \tag{138}$$

Here, $\boldsymbol{\xi}_j \sim \mathcal{N}(0, \sigma_x^2 I_d)$ and $\epsilon_j \sim \mathcal{N}(0, \sigma_d^2)$. In the experiment, we set $\sigma_x = 0.35$, $\sigma_d = 1.0$, $C = 3.0$, and $\alpha = 1.0$. The first term, $\boldsymbol{\xi}_j$, is background noise. It represents generic variation in each token that is not directly related to the label. The second term, $\epsilon_j \boldsymbol{u}$, is random noise along the label direction. This component is added to all tokens, so distractor tokens may also contain components in the label direction. Since this component may point either in the correct or incorrect label direction, it acts as noise that can interfere with prediction. The third term, $\mathbb{1}[j \in R] \frac{C}{\sqrt{m}} y \boldsymbol{u}$, is the label signal added only to relevant tokens. This term is added only when $j \in R$. Since it points in the direction $y\boldsymbol{u}$, it provides useful information for predicting the label y. The scaling factor $C/\sqrt{m}$ is used to prevent the total signal strength from trivially increasing as the number of relevant tokens increases. This makes it easier to interpret changes in m as changes in whether useful information is localized in a small number of tokens or distributed across many tokens, rather than simply as changes in the total amount of useful signal. The fourth term, $\mathbb{1}[j \in R] \alpha \boldsymbol{r}$, is an additional component added only to relevant tokens. This component is used so that the attention score can reflect whether a token is relevant. Since we regard $\boldsymbol{r}$ as a fixed query and compute scores using $\langle \boldsymbol{x}^j, \boldsymbol{r} \rangle$, relevant tokens containing an $\boldsymbol{r}$-direction component tend to receive higher scores on average.

I.2.2 Attention Score Construction

Each token score is computed as the inner product between the fixed query r and the token representation $\boldsymbol{x}^j$: $s_j = \langle \boldsymbol{x}^j, \boldsymbol{r} \rangle + \eta_j$, where $\eta_j \sim \mathcal{N}(0, \sigma_s^2)$. Substituting the token generation formula yields

$$s_j = \left\langle \boldsymbol{\xi}_j + \epsilon_j \boldsymbol{u} + \mathbb{1}[j \in R] \frac{C}{\sqrt{m}} y \boldsymbol{u} + \mathbb{1}[j \in R] \alpha \boldsymbol{r}, \boldsymbol{r} \right\rangle + \eta_j \tag{139}$$

$$= \langle \boldsymbol{\xi}_j, \boldsymbol{r} \rangle + \epsilon_j \langle \boldsymbol{u}, \boldsymbol{r} \rangle + \mathbb{1}[j \in R] \frac{C}{\sqrt{m}} y \langle \boldsymbol{u}, \boldsymbol{r} \rangle + \mathbb{1}[j \in R] \alpha \langle \boldsymbol{r}, \boldsymbol{r} \rangle + \eta_j. \tag{140}$$

Since $\langle \boldsymbol{u}, \boldsymbol{r} \rangle = 0$ and $\langle \boldsymbol{r}, \boldsymbol{r} \rangle = 1$, we obtain

$$s_j = \langle \boldsymbol{\xi}_j, \boldsymbol{r} \rangle + \mathbb{1}[j \in R] \alpha + \eta_j. \tag{141}$$

This expression clarifies the meaning of the score. A relevant token, namely a token with $j \in R$, contains the $\alpha \boldsymbol{r}$ component and therefore receives a higher score on average. A distractor token does not contain this component and therefore has a lower score on average. However, because of $\langle \boldsymbol{\xi}_j, \boldsymbol{r} \rangle$ and η_j, the score is imperfect: a distractor token can accidentally receive a high score, and a relevant token can receive a low score. In the experiment, the score noise level σ_s is varied over $\sigma_s \in \{0.05, 0.10, 0.20, 0.35, 0.50, 0.75, 1.00\}$. When σ_s is small, the scores identify relevant tokens relatively accurately. When σ_s is large, the scores become unreliable, and distractor tokens are more likely to receive large weights by mistake.

The orthogonality between $\boldsymbol{u}$ and $\boldsymbol{r}$ is important here. If $\boldsymbol{r}$ were not orthogonal to $\boldsymbol{u}$, the score would contain the additional term $\mathbb{1}[j \in R] \frac{C}{\sqrt{m}} y \langle \boldsymbol{u}, \boldsymbol{r} \rangle$. In that case, the attention score could directly reflect the label y, rather than only whether the token is relevant. This would be undesirable for an

experiment intended to compare score-to-weight mappings. We therefore choose $\boldsymbol{r}$ to be orthogonal to $\boldsymbol{u}$, so that the score reflects information for finding relevant tokens but does not directly reveal the label itself. In other words, the information used for score computation and the information used for label prediction are artificially separated.

I.2.3 Score-to-Weight Mappings

Given the score vector $\boldsymbol{s} = (s_1, \ldots, s_n)$, we compare two score-to-weight mappings. The first is softmax:

$$a_j^{\mathrm{soft}} = \frac{\exp(s_j)}{\sum_{\ell=1}^{n} \exp(s_\ell)}. \tag{142}$$

Softmax assigns a positive weight to every token and therefore produces a dense attention distribution. Even tokens with low scores are not completely ignored. The second is sparsemax, defined as the Euclidean projection of the score vector onto the probability simplex:

$$a^{\mathrm{sparse}} = \arg\min_{\boldsymbol{p} \in \Delta^{n-1}} \|\boldsymbol{p} - \boldsymbol{s}\|^2. \tag{143}$$

Sparsemax can assign exactly zero weight to low-scoring tokens. It therefore produces a sparse attention distribution and behaves like a mechanism that selects a smaller number of tokens. In this experiment, softmax and sparsemax receive the same score vector. Therefore, the only difference between the two methods is the mapping from scores to attention weights.

I.2.4 Prediction Rule and Signal-Noise Trade-off

After obtaining the attention weights $\boldsymbol{a} = (a_1, \ldots, a_n)$, we aggregate the token representations by a weighted sum $\boldsymbol{z} = \sum_{j=1}^{n} a_j \boldsymbol{x}^j$. We then project the aggregated vector $\boldsymbol{z}$ onto the fixed classification direction $\boldsymbol{u}$ and define the prediction logit as $\ell = \langle \boldsymbol{z}, \boldsymbol{u} \rangle$. The predicted label is

$$\hat{y} = \begin{cases} +1, & \ell \geq 0, \\ -1, & \ell < 0. \end{cases} \tag{144}$$

In a standard supervised learning setting, one would learn a classification head w and compute a logit of the form $\ell = \langle w, z \rangle + b$. In this experiment, however, we do not train a classifier. Instead, we use the label direction u used in data generation as a fixed classifier. This idealizes the situation in which the classification head has already learned the correct label direction. As a result, we can observe how the attention weights obtained from the same scores affect prediction, without confounding the comparison with the quality of classifier learning.

The role of the attention weights in prediction can be understood by decomposing the logit:

$$\ell = \left\langle \sum_{j=1}^{n} a_j \boldsymbol{x}^j, \boldsymbol{u} \right\rangle \tag{145}$$

$$= \sum_{j=1}^{n} a_j \langle \boldsymbol{\xi}_j, \boldsymbol{u} \rangle + \sum_{j=1}^{n} a_j \epsilon_j + \frac{C}{\sqrt{m}} y \sum_{j \in R} a_j + \alpha \sum_{j \in R} a_j \langle \boldsymbol{r}, \boldsymbol{u} \rangle \tag{146}$$

$$= \sum_{j=1}^{n} a_j \langle \boldsymbol{\xi}_j, \boldsymbol{u} \rangle + \sum_{j=1}^{n} a_j \epsilon_j + \frac{C}{\sqrt{m}} y \sum_{j \in R} a_j, \tag{147}$$

where the last equality follows from $\langle \boldsymbol{r}, \boldsymbol{u} \rangle = 0$.

This decomposition shows two ways in which attention weights affect prediction. First, the last term is the useful label signal. It is proportional to $\sum_{j \in R} a_j$. Thus, assigning more attention mass to relevant tokens strengthens the signal in the correct label direction. Second, the first two terms are noise terms. Their magnitude depends on how the weights are distributed across tokens. Importantly, the noise magnitude is not determined only by the total mass assigned to relevant tokens. If attention weights are spread over many tokens, independent noise components tend to be averaged out. If the weights are concentrated on a small number of tokens, the prediction becomes more sensitive to the noise contained in those tokens. This can also be seen from a simple variance calculation. Let

the noise term be written abstractly as $N = \sum_{j=1}^{n} a_j \zeta_j$, where ζ_j are independent zero-mean noise variables with variance σ_ζ^2. Then $\mathrm{Var}(N) = \sigma_\zeta^2 \sum_{j=1}^{n} a_j^2$. Thus, the variance of the noise depends on $\sum_{j=1}^{n} a_j^2$. This quantity is small when the weights are distributed over many tokens and large when they are concentrated on a small number of tokens.

Therefore, prediction performance is not determined only by the attention mass assigned to relevant tokens, $\sum_{j \in R} a_j$. The useful signal becomes stronger as this relevant mass increases, but the magnitude of the noise also depends on the concentration of the attention distribution, $\sum_{j=1}^{n} a_j^2$. Sparsemax can discard unnecessary tokens when it selects relevant tokens correctly, but because it concentrates mass on a smaller number of tokens, it can also become sensitive to the noise in those selected tokens. Softmax may leave some weight on unnecessary tokens, but by spreading the weights more smoothly, it can average out noise in some settings. This signal-noise trade-off is important for understanding the performance difference between softmax and sparsemax.

For each pair (m, σ_s), we generate 5000 samples and repeat the experiment over 10 random seeds. We report the difference in accuracy, $\Delta_{\mathrm{acc}} = \mathrm{Acc}_{\mathrm{sparsemax}} - \mathrm{Acc}_{\mathrm{softmax}}$. A positive value of Δ_{acc} means that sparsemax achieves higher accuracy than softmax, whereas a negative value means that softmax achieves higher accuracy.

I.3 Experimental Results

In the experiment, we vary the number of relevant tokens m and the score noise level σ_s. Figure 3 summarizes the results for the score-to-weight comparison. The top panel shows the accuracy of the standard setting, which combines softmax attention weights with the usual linear value aggregation. The bottom panel shows the accuracy difference between sparsemax with linear aggregation and the standard setting. That is, each value in the bottom panel is computed as the accuracy of sparsemax with linear aggregation minus the accuracy of the standard setting.

As shown in the figure, neither softmax nor sparsemax is uniformly better. Rather, the preferable mapping depends on how label-relevant information is distributed and how reliable the scores are. The experiments in this section and Appendix L were conducted on the free version of Google Colab, and all runs completed within a few minutes without using specialized compute resources.

When m is small and σ_s is also small, sparsemax tends to perform better. In this regime, label-relevant information is localized in a small number of relevant tokens, and the scores can identify these tokens relatively accurately. Sparsemax can assign zero weight to low-scoring distractor tokens, and therefore can more easily discard unnecessary tokens. As a result, the influence of distractors mixed into the aggregated representation z is reduced, improving prediction accuracy.

When m is large, softmax tends to perform better. In this regime, label-relevant information is distributed across many relevant tokens. Since sparsemax concentrates weight on a smaller number of tokens, it may discard some tokens that contain useful information. In contrast, softmax distributes weight over a wider range of tokens and can therefore aggregate distributed information more effectively.

Softmax also tends to perform better when σ_s is large. When the scores are unreliable, distractor tokens may accidentally receive high scores. Sparsemax concentrates weight on high-scoring tokens and is therefore strongly affected by such erroneous high scores. Softmax, by spreading the weights more smoothly, is less likely to depend excessively on a small number of spuriously high-scoring distractors.

These results can be interpreted through the signal-noise trade-off described above. When the scores are reliable, sparsemax can concentrate attention on relevant tokens and discard unnecessary tokens. It is therefore advantageous when relevant information is localized. However, when the scores are unreliable or when relevant information is distributed across many tokens, concentrating weight on a small number of tokens can be disadvantageous. Softmax spreads the weights more broadly. Although this means that it also mixes in some unnecessary tokens, it can capture distributed useful information and may average out noise more effectively.

These observations show that the choice of the score-to-weight potential is not merely a formal degree of freedom, but a factor that determines the operational behavior of attention. A sparse potential is useful when the task requires selecting a small number of reliable tokens. A dense

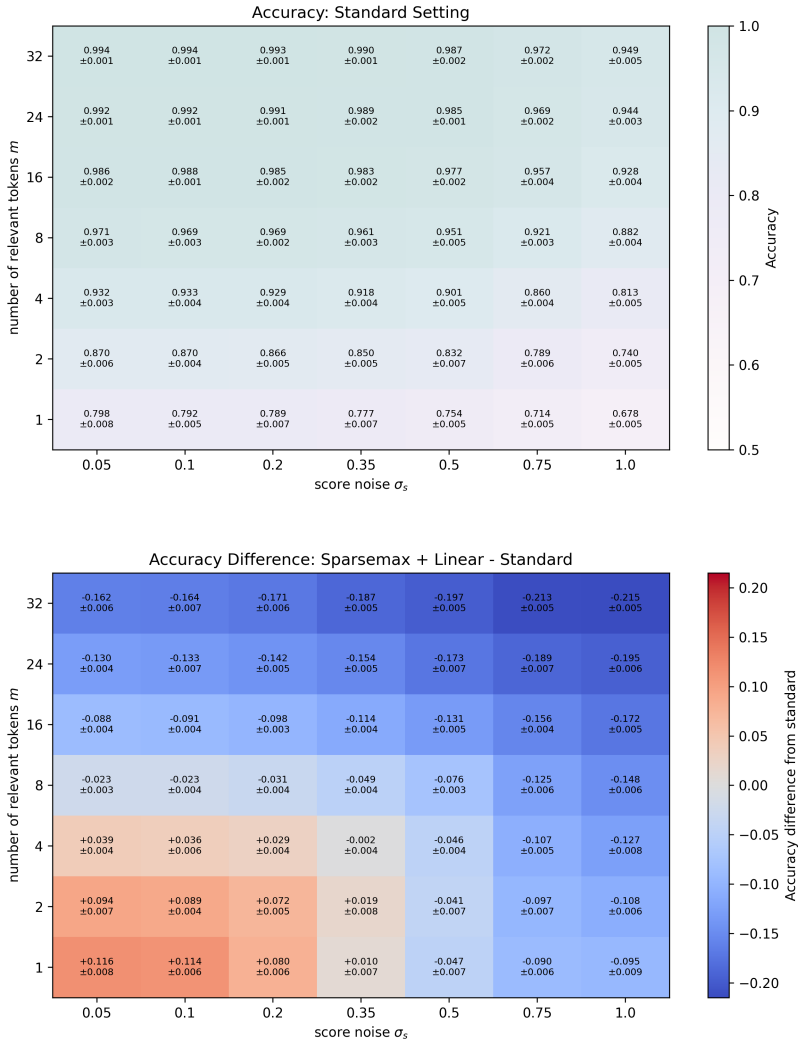

Figure 3: Accuracy of the standard setting and the accuracy difference from it as m and σ_s are varied. The standard setting uses softmax attention weights with linear value aggregation. The top panel shows the accuracy of the standard setting, while the bottom panel shows $\Delta_{\text{acc}} = \text{Acc}_{\text{sparsemax}} - \text{Acc}_{\text{softmax}}$.

potential is useful when information is distributed across many tokens or when the scores contain substantial noise. Therefore, this experiment does not show that sparsemax is generally superior to softmax. Nor does it claim that the same boundary will appear unchanged in an end-to-end trained Transformer. Rather, the experiment shows that the choice of score-to-weight potential changes how attention mass is allocated, and that this difference appears as a difference in prediction performance depending on the localization of relevant information and the reliability of the scores.

J Proof of Theorem 7.1

We first make precise what is meant by consistency of aggregation under the geometry specified by the potential π. The gradient map

$$\nabla\pi : \mathcal{V} \to \nabla\pi(\mathcal{V}) \tag{148}$$

gives the coordinate representation associated with this geometry. Let

$$\mathcal{U} = \nabla\pi(\mathcal{V}) \tag{149}$$

be the corresponding coordinate space. We refer to $\mathcal{V}$ as the primal value space and to $\mathcal{U} = \nabla\pi(\mathcal{V})$ as the dual value space induced by π.

For an aggregation rule A_π on $\mathcal{V}$, define the induced aggregation rule B on $\mathcal{U}$ by

$$B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n) = \nabla\pi\left(A_\pi(\boldsymbol{a}; \boldsymbol{v}^1, \dots, \boldsymbol{v}^n)\right), \quad \boldsymbol{u}^j = \nabla\pi(\boldsymbol{v}^j). \tag{150}$$

Thus, B describes the aggregation rule after value representations are expressed in the coordinates determined by the geometry specified by π.

We assume that B satisfies the following consistency requirements.

First, B satisfies unanimity:

$$B(\boldsymbol{a}; \boldsymbol{u}, \dots, \boldsymbol{u}) = \boldsymbol{u}. \tag{151}$$

This means that if all input values coincide, the representative is that common value.

Second, B is invariant under simultaneous permutations of weights and values. That is, for any permutation σ of $\{1, \dots, n\}$,

$$B(\boldsymbol{a}_\sigma; \boldsymbol{u}^{\sigma(1)}, \dots, \boldsymbol{u}^{\sigma(n)}) = B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n), \tag{152}$$

where

$$\boldsymbol{a}_\sigma = (a_{\sigma(1)}, \dots, a_{\sigma(n)}). \tag{153}$$

Thus, the output depends on the weighted collection itself, not on the order in which the values are listed.

Third, B is consistent with grouping and reaggregation. Let $G_1, \dots, G_m$ be a partition of $\{1, \dots, n\}$, and define

$$\alpha_r = \sum_{j \in G_r} a_j, \qquad a_{j|r} = \frac{a_j}{\alpha_r} \quad (j \in G_r,\ \alpha_r > 0). \tag{154}$$

Then aggregation is consistent with first aggregating inside each group and then aggregating the resulting representatives:

$$B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n) = B\left(\alpha; B(\boldsymbol{a}_{\cdot|1}; \{\boldsymbol{u}^j\}_{j \in G_1}), \dots, B(\boldsymbol{a}_{\cdot|m}; \{\boldsymbol{u}^j\}_{j \in G_m})\right). \tag{155}$$

Fourth, B is equivariant with respect to affine transformations in the coordinate space. For every affine map T whose relevant images remain in $\mathcal{U}$,

$$B(\boldsymbol{a}; T\boldsymbol{u}^1, \dots, T\boldsymbol{u}^n) = TB(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n). \tag{156}$$

This condition means that aggregation is compatible with affine changes of the coordinate system in which the geometry is represented.

We also assume that B is continuous with respect to the weights and that zero-weighted values do not affect the aggregation. In particular,

$$B((1, 0); \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}, \quad B((0, 1); \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}. \tag{157}$$

We now show that these conditions imply that aggregation in the coordinate space $\mathcal{U}$ must be the weighted barycenter:

$$B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n) = \sum_{j=1}^{n} a_j \boldsymbol{u}^j. \tag{158}$$

First consider the binary case. Define

$$b_t(\boldsymbol{x}, \boldsymbol{y}) = B((t, 1-t); \boldsymbol{x}, \boldsymbol{y}), \quad 0 \le t \le 1. \tag{159}$$

We first show that $b_t(\boldsymbol{x}, \boldsymbol{y})$ lies on the affine line through $\boldsymbol{x}$ and $\boldsymbol{y}$. Let P be an affine projection onto this line such that

$$P\boldsymbol{x} = \boldsymbol{x}, \quad P\boldsymbol{y} = \boldsymbol{y}. \tag{160}$$

By affine equivariance,

$$B((t, 1-t); P\boldsymbol{x}, P\boldsymbol{y}) = PB((t, 1-t); \boldsymbol{x}, \boldsymbol{y}). \tag{161}$$

Since $P\boldsymbol{x} = \boldsymbol{x}$ and $P\boldsymbol{y} = \boldsymbol{y}$, the left-hand side is

$$B((t, 1-t); \boldsymbol{x}, \boldsymbol{y}) = b_t(\boldsymbol{x}, \boldsymbol{y}). \tag{162}$$

Hence

$$b_t(\boldsymbol{x}, \boldsymbol{y}) = Pb_t(\boldsymbol{x}, \boldsymbol{y}). \tag{163}$$

Therefore $b_t(\boldsymbol{x}, \boldsymbol{y})$ is fixed by P, and so it lies on the affine line through $\boldsymbol{x}$ and $\boldsymbol{y}$. Thus, for distinct $\boldsymbol{x}$ and $\boldsymbol{y}$, there exists a scalar $\lambda(t, \boldsymbol{x}, \boldsymbol{y})$ such that

$$b_t(\boldsymbol{x}, \boldsymbol{y}) = \lambda(t, \boldsymbol{x}, \boldsymbol{y})\boldsymbol{x} + \{1 - \lambda(t, \boldsymbol{x}, \boldsymbol{y})\}\boldsymbol{y}. \tag{164}$$

We next show that this coefficient depends only on t. Let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}', \boldsymbol{y}'$ be points such that $\boldsymbol{x} \neq \boldsymbol{y}$ and $\boldsymbol{x}' \neq \boldsymbol{y}'$. Choose an affine map T satisfying

$$T\boldsymbol{x} = \boldsymbol{x}', \quad T\boldsymbol{y} = \boldsymbol{y}'. \tag{165}$$

By affine equivariance,

$$b_t(\boldsymbol{x}', \boldsymbol{y}') = b_t(T\boldsymbol{x}, T\boldsymbol{y}) = Tb_t(\boldsymbol{x}, \boldsymbol{y}). \tag{166}$$

Using the representation above, we obtain

$$Tb_t(\boldsymbol{x}, \boldsymbol{y}) = \lambda(t, \boldsymbol{x}, \boldsymbol{y})\boldsymbol{x}' + \{1 - \lambda(t, \boldsymbol{x}, \boldsymbol{y})\}\boldsymbol{y}'. \tag{167}$$

Since the affine representation on the line through distinct $\boldsymbol{x}'$ and $\boldsymbol{y}'$ is unique, we have

$$\lambda(t, \boldsymbol{x}', \boldsymbol{y}') = \lambda(t, \boldsymbol{x}, \boldsymbol{y}). \tag{168}$$

Hence the coefficient depends only on t. We write it as $\lambda(t)$:

$$b_t(\boldsymbol{x}, \boldsymbol{y}) = \lambda(t)\boldsymbol{x} + \{1 - \lambda(t)\}\boldsymbol{y}. \tag{169}$$

Permutation invariance gives

$$b_t(\boldsymbol{x}, \boldsymbol{y}) = b_{1-t}(\boldsymbol{y}, \boldsymbol{x}). \tag{170}$$

Substituting the above representation into both sides gives

$$\lambda(1 - t) = 1 - \lambda(t). \tag{171}$$

Next, grouping consistency gives a multiplicative relation. Aggregating $\boldsymbol{x}$ and $\boldsymbol{y}$ with weights t and $1 - t$, and then aggregating the result with $\boldsymbol{y}$ using weights s and $1 - s$, is the same as aggregating $\boldsymbol{x}$ and $\boldsymbol{y}$ directly with weights st and $1 - st$. This follows by applying grouping consistency to $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y})$ with weights $(st, s(1 - t), 1 - s)$, using unanimity for the repeated $\boldsymbol{y}$. Thus,

$$b_s(b_t(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y}) = b_{st}(\boldsymbol{x}, \boldsymbol{y}). \tag{172}$$

The coefficient of $\boldsymbol{x}$ on the left-hand side is

$$\lambda(s)\lambda(t), \tag{173}$$

whereas the coefficient of $\boldsymbol{x}$ on the right-hand side is

$$\lambda(st). \tag{174}$$

Therefore,

$$\lambda(st) = \lambda(s)\lambda(t). \tag{175}$$

By the zero-weight condition,

$$\lambda(1) = 1. \tag{176}$$

Together with continuity, this implies that $\lambda(t) > 0$ for $0 < t < 1$. Indeed, for any $t \in (0,1)$, choose m large enough so that $t^{1/m}$ is close to 1. Since λ is continuous and $\lambda(1) = 1$, we have $\lambda(t^{1/m}) > 0$. Using multiplicativity,

$$\lambda(t) = \lambda\Big((t^{1/m})^m\Big) = \lambda(t^{1/m})^m > 0. \tag{177}$$

Hence $\log \lambda(t)$ is well-defined. Let

$$g(r) = \log \lambda(e^r), \qquad r < 0. \tag{178}$$

The multiplicative relation for λ gives

$$g(r+s) = g(r) + g(s) \tag{179}$$

whenever $r < 0$, $s < 0$, and $r + s < 0$. By the standard result for Cauchy's functional equation, continuity of g implies that g is linear. Hence,

$$g(r) = cr \tag{180}$$

for some constant c. Therefore,

$$\lambda(t) = t^c. \tag{181}$$

Using the symmetry relation in (171) together with $\lambda(t) = t^c$, we obtain

$$(1-t)^c = 1 - t^c \tag{182}$$

for all $t \in (0,1)$. This identity holds only when $c = 1$. Hence

$$\lambda(t) = t. \tag{183}$$

Thus,

$$b_t(\boldsymbol{x}, \boldsymbol{y}) = t\boldsymbol{x} + (1-t)\boldsymbol{y}. \tag{184}$$

We now prove the general n-point case by induction. Let

$$\beta = 1 - a_1, \qquad \tilde{a}_j = \frac{a_j}{\beta} \quad (j = 2, \dots, n), \tag{185}$$

and assume first that $\beta > 0$. By grouping consistency,

$$B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n) = B\left((a_1, \beta); \boldsymbol{u}^1, B(\tilde{\boldsymbol{a}}; \boldsymbol{u}^2, \dots, \boldsymbol{u}^n)\right). \tag{186}$$

Using the binary aggregation formula in (184) and the induction hypothesis, we obtain

$$B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n) = a_1\boldsymbol{u}^1 + \beta B(\tilde{\boldsymbol{a}}; \boldsymbol{u}^2, \dots, \boldsymbol{u}^n) \tag{187}$$

$$= a_1\boldsymbol{u}^1 + \beta \sum_{j=2}^{n} \tilde{a}_j \boldsymbol{u}^j \tag{188}$$

$$= a_1\boldsymbol{u}^1 + \sum_{j=2}^{n} a_j \boldsymbol{u}^j \tag{189}$$

$$= \sum_{j=1}^{n} a_j \boldsymbol{u}^j. \tag{190}$$

The cases with zero weights follow by continuity, or equivalently by removing zero-weighted values. This completes the induction and proves the weighted barycenter form for all n.

We now return to the original value space $\mathcal{V}$. By definition,

$$\boldsymbol{u}^j = \nabla\pi(\boldsymbol{v}^j), \tag{191}$$

and

$$B(\boldsymbol{a}; \boldsymbol{u}^1, \dots, \boldsymbol{u}^n) = \nabla\pi\left(A_\pi(\boldsymbol{a}; \boldsymbol{v}^1, \dots, \boldsymbol{v}^n)\right). \tag{192}$$

Therefore,

$$\nabla\pi\left(A_\pi(\boldsymbol{a};\boldsymbol{v}^1,\dots,\boldsymbol{v}^n)\right)=\sum_{j=1}^n a_j\nabla\pi(\boldsymbol{v}^j). \tag{193}$$

This proves the barycentric relation in the coordinate system induced by π.

Assume now that

$$\sum_{j=1}^n a_j\nabla\pi(\boldsymbol{v}^j)\in\nabla\pi(\mathcal{V}). \tag{194}$$

Since $\nabla\pi$ is one-to-one, the representative is uniquely determined. Moreover, on the range of $\nabla\pi$, the inverse of $\nabla\pi$ is $\nabla\pi^*$, where π^* is the Legendre conjugate of π. Thus,

$$A_\pi(\boldsymbol{a};\boldsymbol{v}^1,\dots,\boldsymbol{v}^n)=\nabla\pi^*\left(\sum_{j=1}^n a_j\nabla\pi(\boldsymbol{v}^j)\right). \tag{195}$$

This completes the proof.

Variational interpretation. The dual barycentric relation in (30) also admits the standard variational interpretation in Bregman geometry. Define

$$F(\boldsymbol{v})=\sum_{j=1}^n a_j D_\pi(\boldsymbol{v}\|\boldsymbol{v}^j). \tag{196}$$

By the definition of the Bregman divergence,

$$D_\pi(\boldsymbol{v}\|\boldsymbol{v}^j)=\pi(\boldsymbol{v})-\pi(\boldsymbol{v}^j)-\left\langle\nabla\pi(\boldsymbol{v}^j),\boldsymbol{v}-\boldsymbol{v}^j\right\rangle. \tag{197}$$

Hence

$$\begin{aligned}F(\boldsymbol{v})&=\sum_{j=1}^n a_j\left[\pi(\boldsymbol{v})-\pi(\boldsymbol{v}^j)-\left\langle\nabla\pi(\boldsymbol{v}^j),\boldsymbol{v}-\boldsymbol{v}^j\right\rangle\right]\\&=\pi(\boldsymbol{v})-\left\langle\sum_{j=1}^n a_j\nabla\pi(\boldsymbol{v}^j),\boldsymbol{v}\right\rangle+C,\end{aligned} \tag{198}$$

where C is independent of $\boldsymbol{v}$. Therefore,

$$\nabla F(\boldsymbol{v})=\nabla\pi(\boldsymbol{v})-\sum_{j=1}^n a_j\nabla\pi(\boldsymbol{v}^j). \tag{199}$$

Thus, the first-order condition $\nabla F(\boldsymbol{v})=0$ is equivalent to

$$\nabla\pi(\boldsymbol{v})=\sum_{j=1}^n a_j\nabla\pi(\boldsymbol{v}^j), \tag{200}$$

which is exactly the dual barycentric relation (30). If the right-hand side belongs to $\nabla\pi(\mathcal{V})$, there exists $\boldsymbol{v}^\star\in\mathcal{V}$ satisfying this equation. Since π is strictly convex, F is strictly convex in $\boldsymbol{v}$, and this stationary point is the unique global minimizer. Consequently,

$$A_\pi(\boldsymbol{a};\boldsymbol{v}^1,\dots,\boldsymbol{v}^n)=\arg\min_{\boldsymbol{v}\in\mathcal{V}}\sum_{j=1}^n a_j D_\pi(\boldsymbol{v}\|\boldsymbol{v}^j). \tag{201}$$

The variational characterization in (201) is a standard property of Bregman barycenters. Its role here is to provide a variational interpretation of the aggregation rule characterized in Theorem 7.1. The theorem itself does not assume this minimization problem; it derives the dual-coordinate barycentric form from the stated consistency requirements. Thus, (201) provides an interpretation of the resulting aggregation rule, while the characterization itself follows from the consistency requirements.

K Examples of Value Aggregation with Non-Quadratic Potentials

On the other hand, when a non-quadratic potential is employed, the gradient mapping $\nabla\pi$ is no longer the identity, and the attention output is given as a generalized mean determined via the gradient mapping. As a concrete example other than the quadratic potential, following the same construction as in (14), consider, for $p > 1$, $p' > 1$ with $1/p + 1/p' = 1$,

$$\pi(\boldsymbol{v}) = \frac{1}{p}\sum_{l=1}^{d_v} |v_l|^p, \quad \pi^*(\boldsymbol{u}) = \frac{1}{p'}\sum_{l=1}^{d_v} |u_l|^{p'}. \tag{202}$$

The corresponding gradient mapping and its inverse are given componentwise by

$$(\nabla\pi(\boldsymbol{v}))_l = |v_l|^{p-2}v_l, \quad (\nabla\pi^*(\boldsymbol{u}))_l = |u_l|^{p'-2}u_l. \tag{203}$$

Accordingly, the attention output $\bar{\boldsymbol{v}}^i$ can be written componentwise as

$$\bar{v}_l^i = \operatorname{sgn}\left(\sum_{j=1}^n a_j^i |v_l^j|^{p-2}v_l^j\right)\left|\sum_{j=1}^n a_j^i |v_l^j|^{p-2}v_l^j\right|^{\frac{1}{p-1}}. \tag{204}$$

In particular, when $p = 2$, we have $p' = 2$, and the standard linear weighted sum is recovered as a special case.

As another concrete example, following the same construction as in (16), consider for $\beta > 0$

$$\pi(\boldsymbol{v}) = \sum_{l=1}^{d_v} \frac{1}{\beta^2}\big(\cosh(\beta v_l) - 1\big), \quad \pi^*(\boldsymbol{u}) = \sum_{l=1}^{d_v} \frac{1}{\beta^2}\left[(\beta u_l)\,\operatorname{arcsinh}(\beta u_l) - \sqrt{1 + (\beta u_l)^2} + 1\right]. \tag{205}$$

The corresponding gradient mapping and its inverse are given componentwise by

$$(\nabla\pi(\boldsymbol{v}))_l = \frac{1}{\beta}\sinh(\beta v_l), \quad (\nabla\pi^*(\boldsymbol{u}))_l = \frac{1}{\beta}\operatorname{arcsinh}(\beta u_l). \tag{206}$$

Accordingly, the attention output $\bar{\boldsymbol{v}}^i$ can be written componentwise as

$$\bar{v}_l^i = \frac{1}{\beta}\operatorname{arcsinh}\left(\sum_{j=1}^n a_j^i \sinh(\beta v_l^j)\right). \tag{207}$$

In particular, in the limit $\beta \to 0$, the standard linear weighted sum is recovered as a special case.

L Numerical Example of Value Aggregation Potentials

L.1 Purpose of the Experiment

This section presents a numerical experiment for examining how the choice of the potential in value aggregation affects prediction behavior. The experiment in Appendix I varies the score-to-weight mapping while using the same token representations and scores, and examines how the resulting attention weights affect prediction performance. In contrast, the present experiment uses the same synthetic data generation process, score computation, and parameter values, while fixing the attention weights to the standard softmax weights. We then vary only the potential used for value aggregation. Thus, all value aggregation rules compared in this section use the same token representations, the same scores, and the same softmax attention weights. The purpose of the experiment is to examine whether different geometries on the value representation space can lead to different aggregated representations and prediction behavior, even when the attention weights are held fixed. Taken together, Appendices I and L provide complementary analyses of two distinct stages of attention under the same synthetic setting, separately examining the effects of the score-to-weight mapping and the value aggregation rule on prediction behavior.

This experiment is not intended to show that non-quadratic potentials are uniformly superior to the standard linear aggregation. Rather, it serves as a numerical illustration of the viewpoint developed in this paper: the usual weighted sum is not the only possible form of value aggregation, but can be viewed as the special case induced by the quadratic potential. By keeping the standard softmax attention weights fixed and changing only the value-side potential, we examine whether the choice of this potential can produce observable differences in prediction performance.

L.2 Value Aggregation Potentials

We compare value aggregation rules based on the viewpoint described in Theorem 7.1. The three aggregation rules considered in this experiment are the standard linear aggregation used in ordinary attention, the power aggregation based on the power-potential example in Appendix K, and the hyperbolic-cosine aggregation based on the cosh-potential example in Appendix K.

Standard linear aggregation. Standard linear aggregation corresponds to the value aggregation used in ordinary attention. It is the special case induced by the quadratic potential. We use this aggregation as the baseline for evaluating the effect of non-quadratic value-side potentials.

Power aggregation. Power aggregation is based on the power-potential example in Appendix K. Specifically, it uses the aggregation rule induced by the potential in Eq. (202). In the experiment, we set the power parameter to $p = 3.5$.

Hyperbolic-cosine aggregation. Hyperbolic-cosine aggregation is based on the cosh-potential example in Appendix K. Specifically, it uses the aggregation rule induced by the potential in Eq. (16). In the experiment, we set the scale parameter to $\beta = 3.5$. For numerical stability, the implementation clips βx_l^j to the range $[-20, 20]$ before applying $\sinh$.

L.3 Prediction Rule and Evaluation Metrics

Let $\boldsymbol{z}_\pi$ denote the aggregated representation obtained by the aggregation rule induced by the potential π, where π is one of the quadratic, power, or hyperbolic-cosine potentials described above. The quadratic case corresponds to the standard linear aggregation. For each aggregation rule, we compute the prediction logit by projecting $\boldsymbol{z}_\pi$ onto the fixed label direction $\boldsymbol{u}$:

$$\ell_\pi = \langle \boldsymbol{z}_\pi, \boldsymbol{u} \rangle. \tag{208}$$

The predicted label is

$$\hat{y}_\pi = \begin{cases} +1, & \ell_\pi \geq 0, \\ -1, & \ell_\pi < 0. \end{cases} \tag{209}$$

As in Appendix I, we do not train a classifier. Instead, the label direction $\boldsymbol{u}$ used in data generation is used as a fixed classifier. This removes the effect of classifier training. For each pair (m, σ_s), we generate 5000 samples and repeat the experiment over 10 random seeds.

We compute the accuracy of the standard linear aggregation, power aggregation, and hyperbolic-cosine aggregation. We also report the differences from the standard linear aggregation:

$$\Delta_{\mathrm{pow}} = \mathrm{Acc}_{\mathrm{pow}} - \mathrm{Acc}_{\mathrm{std}}, \tag{210}$$

$$\Delta_{\cosh} = \mathrm{Acc}_{\cosh} - \mathrm{Acc}_{\mathrm{std}}. \tag{211}$$

Positive and negative values indicate improvement and degradation relative to the standard linear aggregation, respectively.

L.4 Experimental Results

Figure 4 shows the accuracy differences obtained by changing the value aggregation potential. Both panels report differences from the standard setting shown in the top panel of Figure 3, namely the setting with softmax attention weights and standard linear value aggregation. The top panel shows the result for power aggregation, and the bottom panel shows the result for hyperbolic-cosine aggregation.

For both power aggregation and hyperbolic-cosine aggregation, the accuracy difference is not uniformly positive or uniformly negative relative to the standard setting. When the number of relevant tokens is small, there are regions where non-quadratic aggregation achieves higher accuracy than the standard setting. In contrast, when the number of relevant tokens is large, the standard linear aggregation tends to perform better. This suggests that non-quadratic potentials affect prediction differently depending on whether the useful signal is concentrated in a small number of tokens or distributed across many tokens.

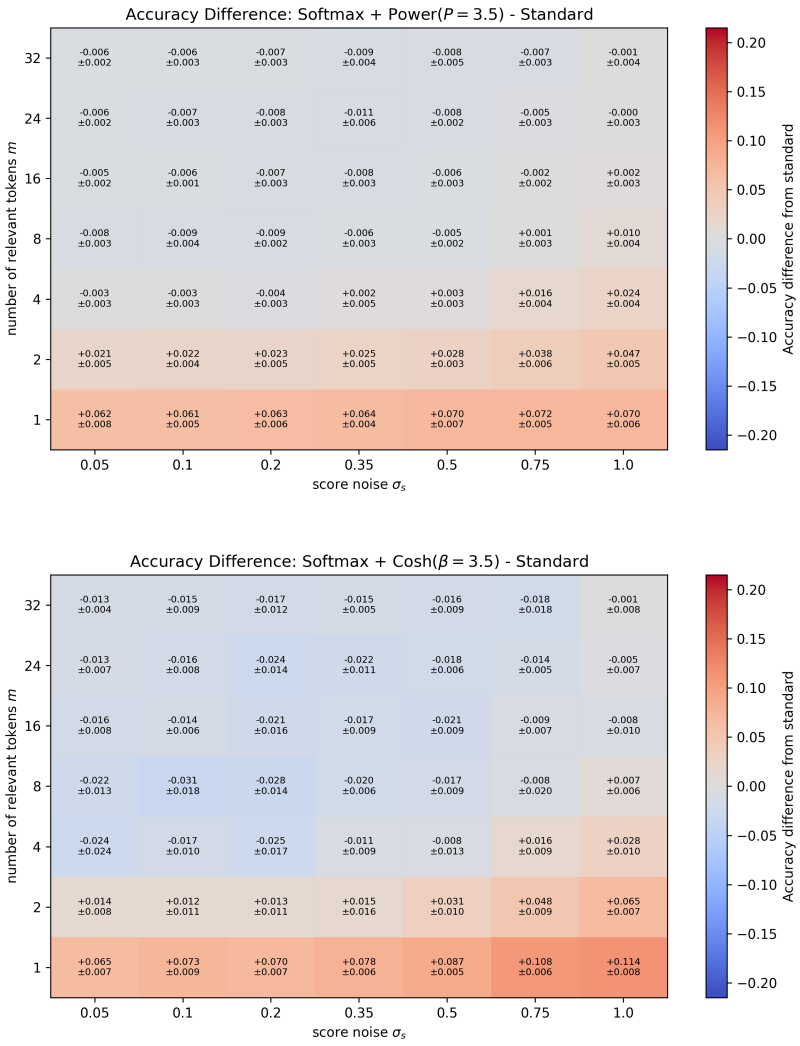

Figure 4: Accuracy differences obtained by changing the value aggregation potential while keeping the attention weights fixed to the standard softmax weights. The top panel shows $\Delta_{\mathrm{pow}} = \mathrm{Acc}_{\mathrm{pow}} - \mathrm{Acc}_{\mathrm{std}}$, and the bottom panel shows $\Delta_{\mathrm{cosh}} = \mathrm{Acc}_{\mathrm{cosh}} - \mathrm{Acc}_{\mathrm{std}}$. In both panels, $\mathrm{Acc}_{\mathrm{std}} = \mathrm{Acc}_{\mathrm{softmax}}$ denotes the accuracy of the standard setting shown in the top panel of Figure 3.

The two non-quadratic aggregations show different patterns under the parameter values used in this experiment. With $p = 3.5$, the power aggregation produces relatively small differences from the standard setting, although positive differences appear when the number of relevant tokens is small. With $\beta = 3.5$, the hyperbolic-cosine aggregation shows larger positive differences in the small-m regime, while negative differences appear for moderate or large values of m. These observations should be interpreted as the behavior under the chosen parameter values, rather than as a general ordering between the two potentials.

Taken together with the score-to-weight experiment in Appendix I, this result highlights a practical implication of the duality-based view developed in this paper. The score-to-weight potential controls how attention mass is assigned to tokens, while the value-side potential controls how the weighted value representations are combined through the dual space. Without this potential-based decomposition, these choices would appear as separate technical modifications; the present framework shows that they are two structurally parallel degrees of freedom governed by the same duality principle.

This point is not merely a matter of mathematical elegance. By identifying the potentials behind both the weight side and the value side, the framework turns attention design into a problem of choosing

geometries: one geometry for producing attention weights and another for aggregating values. The synthetic experiments are limited in scope, but they illustrate that changing these potentials can produce observable differences in prediction behavior. Thus, the duality-based characterization is practically suggestive: it reveals a systematic design space in which attention mechanisms can be adjusted by selecting appropriate potentials, rather than by introducing ad hoc modifications.

M Proof of Theorem 7.2 and Related Consequences

We first give a detailed statement and proof of Theorem 7.2. The theorem characterizes, for fixed keys and values, the locally attainable directions of the $\nabla\pi$-representation of the attention output when the query is varied. After establishing this characterization, we discuss its direct extension to multi-head attention, and then derive several consequences. These consequences show that the same local dimension is obtained for the ordinary output representation, give a basic rank constraint on the attainable directions, and show that, under softmax attention, the resulting operator takes the form of an attention-weighted covariance between keys and value-side dual coordinates.

Theorem M.1 (Detailed version of Theorem 7.2)**.** *Fix the keys $\boldsymbol{k}^1, \dots, \boldsymbol{k}^n$ and values $\boldsymbol{v}^1, \dots, \boldsymbol{v}^n$, and consider the single-head attention output $\boldsymbol{z}^i$ as a function of the query $\boldsymbol{q}^i$. Assume that φ is C^2 and satisfies $\nabla\varphi(\boldsymbol{s} + c\mathbf{1}) = \nabla\varphi(\boldsymbol{s})$ for all $\boldsymbol{s}$ and $c \in \mathbb{R}$. Define*

$$\mathcal{C}_\pi^i = \sum_{\ell,m=1}^{n} \nabla\pi(\boldsymbol{v}^\ell) \left[\nabla^2\varphi(\boldsymbol{s}^i)\right]_{\ell m} (\boldsymbol{k}^m)^\top. \tag{212}$$

Let

$$\mathcal{T}_\pi^i = \left\{ \frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\Big|_{t=0} : \boldsymbol{q}^i(t) \text{ is smooth and } \boldsymbol{q}^i(0) = \boldsymbol{q}^i \right\}, \tag{213}$$

where $\boldsymbol{z}^i(t)$ denotes the attention output obtained by replacing $\boldsymbol{q}^i$ with $\boldsymbol{q}^i(t)$, while keeping the keys and values fixed. Then

$$\mathcal{T}_\pi^i = \operatorname{Im}\mathcal{C}_\pi^i = \{\mathcal{C}_\pi^i \boldsymbol{u} : \boldsymbol{u} \in \mathbb{R}^{d_{qk}}\}. \tag{214}$$

Hence, the local value-side dimension of this head is

$$\dim\mathcal{T}_\pi^i = \operatorname{rank}\mathcal{C}_\pi^i. \tag{215}$$

Proof. Fix one curve of queries $\boldsymbol{q}^i(t)$. Assume that

$$\boldsymbol{q}^i(0) = \boldsymbol{q}^i \tag{216}$$

and denote its derivative at $t = 0$ by

$$\dot{\boldsymbol{q}}^i = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}^i(t)\Big|_{t=0}. \tag{217}$$

The keys and values are fixed.

First, the score is

$$s_j^i(t) = \langle \boldsymbol{q}^i(t), \boldsymbol{k}^j\rangle - \psi(\boldsymbol{q}^i(t)) - \psi^*(\boldsymbol{k}^j). \tag{218}$$

Differentiating this with respect to t, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}s_j^i(t)\Big|_{t=0} = \langle \dot{\boldsymbol{q}}^i, \boldsymbol{k}^j\rangle - \langle \nabla\psi(\boldsymbol{q}^i), \dot{\boldsymbol{q}}^i\rangle. \tag{219}$$

Therefore, in vector form,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{s}^i(t)\Big|_{t=0} = \left(\langle \dot{\boldsymbol{q}}^i, \boldsymbol{k}^1\rangle, \dots, \langle \dot{\boldsymbol{q}}^i, \boldsymbol{k}^n\rangle\right)^\top - \langle \nabla\psi(\boldsymbol{q}^i), \dot{\boldsymbol{q}}^i\rangle\mathbf{1}. \tag{220}$$

Here, the second term

$$-\langle \nabla\psi(\boldsymbol{q}^i), \dot{\boldsymbol{q}}^i\rangle \mathbf{1} \tag{221}$$

is the direction that adds the same value to all score components.

Next,

$$\boldsymbol{a}^i(t) = \nabla\varphi(\boldsymbol{s}^i(t)). \tag{222}$$

By the ordinary chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a}^i(t)\bigg|_{t=0} = \nabla^2\varphi(\boldsymbol{s}^i)\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{s}^i(t)\bigg|_{t=0}. \tag{223}$$

On the other hand, differentiating the translation invariance

$$\nabla\varphi(\boldsymbol{s} + c\mathbf{1}) = \nabla\varphi(\boldsymbol{s}) \tag{224}$$

with respect to c, we obtain

$$\nabla^2\varphi(\boldsymbol{s})\mathbf{1} = \mathbf{0}. \tag{225}$$

Therefore, the common component appearing in the score derivative,

$$-\langle \nabla\psi(\boldsymbol{q}^i), \dot{\boldsymbol{q}}^i\rangle \mathbf{1}, \tag{226}$$

is eliminated by $\nabla^2\varphi(\boldsymbol{s}^i)$. Thus, for each ℓ,

$$\frac{\mathrm{d}}{\mathrm{d}t}a^i_\ell(t)\bigg|_{t=0} = \sum_{m=1}^{n} \left[\nabla^2\varphi(\boldsymbol{s}^i)\right]_{\ell m} \langle \dot{\boldsymbol{q}}^i, \boldsymbol{k}^m\rangle. \tag{227}$$

Next, the value-side dual coordinate is

$$\nabla\pi(\boldsymbol{z}^i(t)) = \sum_{\ell=1}^{n} a^i_\ell(t)\nabla\pi(\boldsymbol{v}^\ell). \tag{228}$$

Since the values are fixed, only $a^i_\ell(t)$ changes with t. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\bigg|_{t=0} = \sum_{\ell=1}^{n} \frac{\mathrm{d}}{\mathrm{d}t}a^i_\ell(t)\bigg|_{t=0} \nabla\pi(\boldsymbol{v}^\ell). \tag{229}$$

Substituting the above expression gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\bigg|_{t=0} = \sum_{\ell=1}^{n}\sum_{m=1}^{n} \nabla\pi(\boldsymbol{v}^\ell) \left[\nabla^2\varphi(\boldsymbol{s}^i)\right]_{\ell m} (\boldsymbol{k}^m)^\top \dot{\boldsymbol{q}}^i. \tag{230}$$

By the definition of $\mathcal{C}^i_\pi$, this is

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\bigg|_{t=0} = \mathcal{C}^i_\pi \dot{\boldsymbol{q}}^i. \tag{231}$$

Therefore, no matter how the query is moved along a smooth curve, the first-order movement of the value-side dual coordinate is of the form

$$\mathcal{C}^i_\pi \dot{\boldsymbol{q}}^i. \tag{232}$$

Hence,

$$\mathcal{T}^i_\pi \subseteq \mathrm{Im}\,\mathcal{C}^i_\pi. \tag{233}$$

Conversely, take any $\boldsymbol{u} \in \mathbb{R}^{d_{qk}}$. Consider the straight line

$$\boldsymbol{q}^i(t) = \boldsymbol{q}^i + t\boldsymbol{u}. \tag{234}$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}^i(t)\Big|_{t=0} = \boldsymbol{u}. \tag{235}$$

In this case, from the formula above,

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\Big|_{t=0} = \mathcal{C}_\pi^i \boldsymbol{u}. \tag{236}$$

Therefore, any direction in $\operatorname{Im}\mathcal{C}_\pi^i$ is actually attainable. Hence,

$$\operatorname{Im}\mathcal{C}_\pi^i \subseteq \mathcal{T}_\pi^i. \tag{237}$$

Combining the two inclusions, we obtain

$$\mathcal{T}_\pi^i = \operatorname{Im}\mathcal{C}_\pi^i. \tag{238}$$

The right-hand side is the column space of the matrix $\mathcal{C}_\pi^i$, and hence its dimension is equal to the rank of the matrix. Therefore,

$$\dim\mathcal{T}_\pi^i = \operatorname{rank}\mathcal{C}_\pi^i. \tag{239}$$

This completes the proof. □

The same argument applies head by head in a multi-head block. If $\mathcal{C}_{\pi_h}^{i,h}$ denotes the operator for head h, then the locally attainable directions of the concatenated dual-coordinate head representation are contained in

$$\operatorname{Im}\mathcal{C}_{\pi_1}^{i,1} \oplus \cdots \oplus \operatorname{Im}\mathcal{C}_{\pi_H}^{i,H}. \tag{240}$$

Thus, the local dimension is at most

$$\sum_{h=1}^{H} \operatorname{rank}\mathcal{C}_{\pi_h}^{i,h}. \tag{241}$$

The bound is attained when the head-wise image subspaces are independent, and a subsequent linear output projection cannot increase this dimension.

The following corollary translates the local characterization in Theorem M.1 from the $\nabla\pi$-representation back to the ordinary output vector $\boldsymbol{z}^i$. If the inverse map is locally nondegenerate, this change of representation preserves the number of available directions. Hence, $\operatorname{rank}\mathcal{C}_\pi^i$ also gives the local dimension of the attainable variations of the attention output itself.

Corollary M.2. *Assume the conditions of Theorem M.1. In addition, assume that $\nabla^2\pi^*(\nabla\pi(\boldsymbol{z}^i))$ is nonsingular. Let*

$$\mathcal{S}_\pi^i = \left\{ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{z}^i(t)\Big|_{t=0} : \boldsymbol{q}^i(t) \text{ is smooth and } \boldsymbol{q}^i(0) = \boldsymbol{q}^i \right\}. \tag{242}$$

Then

$$\dim\mathcal{S}_\pi^i = \operatorname{rank}\mathcal{C}_\pi^i. \tag{243}$$

Proof. Since

$$\boldsymbol{z}^i(t) = \nabla\pi^*(\nabla\pi(\boldsymbol{z}^i(t))), \tag{244}$$

the ordinary chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{z}^i(t)\Big|_{t=0} = \nabla^2\pi^*(\nabla\pi(\boldsymbol{z}^i))\ \frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\Big|_{t=0}. \tag{245}$$

By Theorem M.1, the attainable direction space of

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\pi(\boldsymbol{z}^i(t))\Big|_{t=0} \tag{246}$$

is $\operatorname{Im}\mathcal{C}_\pi^i$. Since $\nabla^2\pi^*(\nabla\pi(\boldsymbol{z}^i))$ is nonsingular, left multiplication by this matrix does not change the dimension of a vector space. Therefore

$$\dim\mathcal{S}_\pi^i = \dim\operatorname{Im}\mathcal{C}_\pi^i = \operatorname{rank}\mathcal{C}_\pi^i. \tag{247}$$

□

The following corollary gives a simple upper bound on the local dimension of the attainable directions of $\nabla\pi(\boldsymbol{z}^i)$. The number of independent directions available to a single head is limited by the query/key dimension, the value dimension, and the local rank of the score-to-weight map. Moreover, since adding a common constant to all scores does not change the attention weights, one score direction is always ineffective. This gives a basic rank constraint on the local directions in which the head can change its output.

Corollary M.3. *Under the assumptions of Theorem 7.2,*

$$\operatorname{rank} \mathcal{C}_\pi^i \le \min\left\{d_{qk}, d_v, \operatorname{rank} \nabla^2\varphi(\boldsymbol{s}^i)\right\}. \tag{248}$$

Moreover, translation invariance implies

$$\nabla^2\varphi(\boldsymbol{s}^i)\mathbf{1} = \mathbf{0}, \tag{249}$$

and therefore

$$\operatorname{rank} \mathcal{C}_\pi^i \le n-1. \tag{250}$$

Thus, a single attention head can locally control at most

$$\operatorname{rank} \mathcal{C}_\pi^i \tag{251}$$

independent value-side dual directions.

Proof. The operator $\mathcal{C}_\pi^i$ is a $d_v \times d_{qk}$ matrix, and it factors through $\nabla^2\varphi(\boldsymbol{s}^i)$. Hence

$$\operatorname{rank} \mathcal{C}_\pi^i \le \min\left\{d_{qk}, d_v, \operatorname{rank} \nabla^2\varphi(\boldsymbol{s}^i)\right\}. \tag{252}$$

By translation invariance,

$$\nabla^2\varphi(\boldsymbol{s}^i)\mathbf{1} = \mathbf{0}. \tag{253}$$

Thus $\nabla^2\varphi(\boldsymbol{s}^i)$ has a nonzero vector in its null space, and

$$\operatorname{rank} \nabla^2\varphi(\boldsymbol{s}^i) \le n-1. \tag{254}$$

Therefore

$$\operatorname{rank} \mathcal{C}_\pi^i \le n-1. \tag{255}$$

□

For softmax attention, the matrix $\mathcal{C}_\pi^i$ admits a concrete interpretation as a weighted covariance matrix. In the expression below, each term contains two deviations: the deviation of the key $\boldsymbol{k}^j$ from the attention-weighted average key $\bar{\boldsymbol{k}}^i$, and the deviation of the transformed value representation $\nabla\pi(\boldsymbol{v}^j)$ from the attention-weighted average transformed value $\bar{\boldsymbol{v}}_\pi^i$. These two deviations are multiplied and then averaged with the attention weight a_j^i. Thus, $\mathcal{C}_\pi^i$ measures how deviations in the key vectors are associated with deviations in the transformed value representations under the current attention distribution. This covariance form explains why the local change of the output depends not only on how the attention weights change, but also on whether the tokens whose weights change have different $\nabla\pi$-transformed value representations.

Corollary M.4. *Assume*

$$\varphi(\boldsymbol{s}) = \log \sum_{j=1}^n \exp(s_j). \tag{256}$$

Then

$$\nabla^2\varphi(\boldsymbol{s}^i) = \operatorname{Diag}(\boldsymbol{a}^i) - \boldsymbol{a}^i(\boldsymbol{a}^i)^\top. \tag{257}$$

Define

$$\bar{\boldsymbol{k}}^i = \sum_{j=1}^n a_j^i \boldsymbol{k}^j, \qquad \bar{\boldsymbol{v}}_\pi^i = \sum_{j=1}^n a_j^i \nabla\pi(\boldsymbol{v}^j) = \nabla\pi(\boldsymbol{z}^i). \tag{258}$$

Then

$$\mathcal{C}_\pi^i = \sum_{j=1}^n a_j^i \left(\nabla\pi(\boldsymbol{v}^j) - \bar{\boldsymbol{v}}_\pi^i\right)\left(\boldsymbol{k}^j - \bar{\boldsymbol{k}}^i\right)^\top. \tag{259}$$

Thus, for softmax attention, $\mathcal{C}_\pi^i$ is the attention-weighted covariance operator between the keys and the value-side dual coordinates.

Proof. For

$$\varphi(\boldsymbol{s}) = \log \sum_{j=1}^{n} \exp(s_j), \tag{260}$$

we have

$$a_\ell^i = \frac{\exp(s_\ell^i)}{\sum_{r=1}^n \exp(s_r^i)}, \tag{261}$$

$$\frac{\partial a_\ell^i}{\partial s_m^i} = a_\ell^i(\delta_{\ell m} - a_m^i), \tag{262}$$

$$\left[\nabla^2 \varphi(\boldsymbol{s}^i)\right]_{\ell m} = \frac{\partial^2 \varphi(\boldsymbol{s}^i)}{\partial s_\ell^i \partial s_m^i} = \frac{\partial a_\ell^i}{\partial s_m^i} = a_\ell^i(\delta_{\ell m} - a_m^i), \tag{263}$$

$$\nabla^2 \varphi(\boldsymbol{s}^i) = \mathrm{Diag}(\boldsymbol{a}^i) - \boldsymbol{a}^i(\boldsymbol{a}^i)^\top. \tag{264}$$

Substituting this into the definition of $\mathcal{C}_\pi^i$, we obtain

$$\mathcal{C}_\pi^i = \sum_{\ell,m=1}^{n} \nabla\pi(\boldsymbol{v}^\ell) \left[\nabla^2 \varphi(\boldsymbol{s}^i)\right]_{\ell m} (\boldsymbol{k}^m)^\top \tag{265}$$

$$= \sum_{\ell,m=1}^{n} \nabla\pi(\boldsymbol{v}^\ell) \left(a_\ell^i \delta_{\ell m} - a_\ell^i a_m^i\right) (\boldsymbol{k}^m)^\top \tag{266}$$

$$= \sum_{\ell,m=1}^{n} a_\ell^i \delta_{\ell m} \nabla\pi(\boldsymbol{v}^\ell)(\boldsymbol{k}^m)^\top - \sum_{\ell,m=1}^{n} a_\ell^i a_m^i \nabla\pi(\boldsymbol{v}^\ell)(\boldsymbol{k}^m)^\top \tag{267}$$

$$= \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\boldsymbol{k}^j)^\top - \left(\sum_{\ell=1}^{n} a_\ell^i \nabla\pi(\boldsymbol{v}^\ell)\right)\left(\sum_{m=1}^{n} a_m^i \boldsymbol{k}^m\right)^\top \tag{268}$$

$$= \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\boldsymbol{k}^j)^\top - \bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top. \tag{269}$$

On the other hand, the centered expression expands as

$$\sum_{j=1}^{n} a_j^i \left(\nabla\pi(\boldsymbol{v}^j) - \bar{\boldsymbol{v}}_\pi^i\right)(\boldsymbol{k}^j - \bar{\boldsymbol{k}}^i)^\top \tag{270}$$

$$= \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\boldsymbol{k}^j)^\top - \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\bar{\boldsymbol{k}}^i)^\top - \sum_{j=1}^{n} a_j^i \bar{\boldsymbol{v}}_\pi^i(\boldsymbol{k}^j)^\top + \sum_{j=1}^{n} a_j^i \bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top \tag{271}$$

$$= \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\boldsymbol{k}^j)^\top - \left(\sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)\right)(\bar{\boldsymbol{k}}^i)^\top - \bar{\boldsymbol{v}}_\pi^i\left(\sum_{j=1}^{n} a_j^i \boldsymbol{k}^j\right)^\top + \left(\sum_{j=1}^{n} a_j^i\right)\bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top \tag{272}$$

$$= \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\boldsymbol{k}^j)^\top - \bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top - \bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top + \bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top \tag{273}$$

$$= \sum_{j=1}^{n} a_j^i \nabla\pi(\boldsymbol{v}^j)(\boldsymbol{k}^j)^\top - \bar{\boldsymbol{v}}_\pi^i(\bar{\boldsymbol{k}}^i)^\top. \tag{274}$$

Combining this with (269), we obtain

$$\mathcal{C}_\pi^i = \sum_{j=1}^{n} a_j^i \left(\nabla\pi(\boldsymbol{v}^j) - \bar{\boldsymbol{v}}_\pi^i\right)\left(\boldsymbol{k}^j - \bar{\boldsymbol{k}}^i\right)^\top. \tag{275}$$

□

The local rank characterization in Theorem 7.2 also constrains how perturbations of the query can propagate to subsequent computations. To make this connection explicit, fix the keys $\boldsymbol{k}^1, \ldots, \boldsymbol{k}^n$ and values $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^n$, and view the output of a single attention head as a function of the query $\boldsymbol{q}^i$. Specifically, let

$$\boldsymbol{z}^i(\boldsymbol{q}^i) = \nabla\pi^* \left(\sum_{j=1}^{n} a_j^i(\boldsymbol{q}^i) \nabla\pi(\boldsymbol{v}^j) \right), \tag{276}$$

$$\boldsymbol{a}^i(\boldsymbol{q}^i) = \nabla\varphi\big(\boldsymbol{s}^i(\boldsymbol{q}^i)\big), \tag{277}$$

where $\boldsymbol{s}^i(\boldsymbol{q}^i)$ denotes the score vector between the query $\boldsymbol{q}^i$ and the fixed keys. Thus, $\boldsymbol{z}^i(\boldsymbol{q}^i)$ is the attention-head output whose local variation with respect to $\boldsymbol{q}^i$ is characterized in Theorem 7.2.

Now let $F : \mathbb{R}^{d_v} \to \mathbb{R}^m$ be any differentiable downstream map applied to this attention output. We examine the local variation of the composed map $F(\boldsymbol{z}^i(\boldsymbol{q}^i))$ with respect to the query $\boldsymbol{q}^i$. By Theorem 7.2 and the chain rule, the derivative of the downstream output with respect to the query is

$$D_{\boldsymbol{q}^i} F\big(\boldsymbol{z}^i(\boldsymbol{q}^i)\big) = DF\big(\boldsymbol{z}^i(\boldsymbol{q}^i)\big)\, D_{\boldsymbol{q}^i} \boldsymbol{z}^i(\boldsymbol{q}^i) \tag{278}$$

$$= DF\big(\boldsymbol{z}^i(\boldsymbol{q}^i)\big)\, D_{\boldsymbol{q}^i} \left[\nabla\pi^* \left(\sum_{j=1}^{n} a_j^i(\boldsymbol{q}^i) \nabla\pi(\boldsymbol{v}^j) \right) \right] \tag{279}$$

$$= DF\big(\boldsymbol{z}^i(\boldsymbol{q}^i)\big)\, \nabla^2\pi^* \left(\sum_{j=1}^{n} a_j^i(\boldsymbol{q}^i) \nabla\pi(\boldsymbol{v}^j) \right) D_{\boldsymbol{q}^i} \left[\sum_{j=1}^{n} a_j^i(\boldsymbol{q}^i) \nabla\pi(\boldsymbol{v}^j) \right] \tag{280}$$

$$= DF\big(\boldsymbol{z}^i(\boldsymbol{q}^i)\big)\, \nabla^2\pi^*\big(\nabla\pi(\boldsymbol{z}^i(\boldsymbol{q}^i))\big)\, C_\pi^i. \tag{281}$$

Here, $D_{\boldsymbol{q}^i}$ denotes differentiation with respect to the query $\boldsymbol{q}^i$, and $DF(z^i(\boldsymbol{q}^i))$ denotes the derivative of F evaluated at $z^i(\boldsymbol{q}^i)$.

Since the rank of a matrix product cannot exceed the rank of any of its factors, (281) immediately gives

$$\operatorname{rank}\left(D_{\boldsymbol{q}^i} F\big(z^i(\boldsymbol{q}^i)\big)\right) \le \min\left\{m,\, \operatorname{rank}(C_\pi^i)\right\}. \tag{282}$$

Thus, $\operatorname{rank}(C_\pi^i)$ bounds not only the number of locally attainable directions at the attention output, but also the number of independent directions in subsequent computations that can be locally affected through the attention head. In particular, if $\operatorname{rank}(C_\pi^i) = r < m$, then a local perturbation of the query can affect at most r independent directions in the output of the downstream computation.

For standard softmax attention with linear value aggregation, $\nabla^2\pi^*$ is the identity and C_π^i reduces to the attention-weighted covariance operator between the keys and values. Consequently, the bound in (282) can be computed directly from the internal representations of a trained Transformer.

N Evaluation on a Pretrained Transformer

Theorem 7.2 characterizes the locally attainable output directions of an attention head through the operator C_π^i, whose rank gives the effective local output dimension. This provides a quantity that can be evaluated directly from the internal representations of a trained Transformer. In particular, as shown in Corollary M.4, for standard softmax attention with linear value aggregation, C_π^i reduces to the attention-weighted covariance operator between the keys and values. Moreover, Corollary M.3 gives an upper bound on its rank.

We therefore evaluate the characterization in Theorem 7.2 using pretrained GPT-2 small [28]. GPT-2 small has 12 layers and 12 attention heads per layer, with key and value dimensions of 64 for each head. We use the test split of WikiText-2 [30] and construct 10 token sequences of length 128. For each sequence, we perform a forward pass with all model parameters fixed. At each layer, we obtain the hidden representation entering the attention module and reconstruct the keys, values, and causal softmax attention weights using the pretrained attention projections. We consider seven query positions corresponding to $n_i \in \{2, 4, 8, 16, 32, 64, 128\}$ visible tokens. This experiment was conducted on the free version of Google Colab without specialized compute resources and completed in about ten minutes.

Table 1: Numerical ranks of C_π^i in pretrained GPT-2 small. Each row summarizes 1,440 operators.

Number of visible tokens n_i	Rank bound	Mean numerical rank	Median numerical rank
2	1	1.00	1
4	3	3.00	3
8	7	6.96	7
16	15	14.76	15
32	31	29.51	31
64	63	49.56	56
128	64	51.38	61

GPT-2 uses softmax normalization and linear value aggregation, corresponding to the log-sum-exp potential and the quadratic potential, respectively. Since $\nabla\pi(\boldsymbol{v}) = \boldsymbol{v}$, for each query position, layer, and head, we compute

$$C_\pi^i = \sum_{j=1}^{n_i} a_j^i \left(\boldsymbol{v}^j - \bar{\boldsymbol{v}}^{\,i}\right)\left(\boldsymbol{k}^j - \bar{\boldsymbol{k}}^{\,i}\right)^\top, \tag{283}$$

where

$$\bar{\boldsymbol{k}}^{\,i} = \sum_{j=1}^{n_i} a_j^i \boldsymbol{k}^j, \qquad \bar{\boldsymbol{v}}^{\,i} = \sum_{j=1}^{n_i} a_j^i \boldsymbol{v}^j. \tag{284}$$

We evaluate the numerical rank of each C_π^i against the bound in Corollary M.3, using `torch.linalg.matrix_rank` in single precision.

For each value of n_i, we evaluate $10 \times 12 \times 12 = 1{,}440$ operators, giving 10,080 operators in total. Since $d_{qk} = d_v = 64$, the rank bound gives

$$\operatorname{rank}(C_\pi^i) \le \min\{64, n_i - 1\}. \tag{285}$$

Table 1 summarizes the numerical ranks for each number of visible tokens.

For shorter visible contexts, the measured ranks are close to the maximum allowed by the rank bound, whereas for longer contexts the bound becomes less tight on average. This indicates that $\operatorname{rank}(C_\pi^i)$ is not determined solely by the dimensional constraints of the attention head, but also reflects the learned arrangement of the keys and values and the input-dependent attention structure. Therefore, C_π^i provides a concrete quantity for characterizing how many independent output directions an individual attention head can locally produce in a trained Transformer.

O Broader Impacts

This work is foundational and does not introduce a deployed system, dataset, or released model. Its potential positive impact is to support a more principled understanding and modification of attention mechanisms. Since attention is a core component of many Transformer-based systems, the insights developed here could indirectly affect a broad range of applications, including both beneficial and potentially harmful ones, but we do not identify a direct path to specific negative societal impacts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the paper's contributions and scope as a structural analysis of attention mechanisms under stated assumptions. The claims are consistent with the theoretical results presented in the paper and do not overstate empirical or practical implications.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations related to the structural assumptions used in the theoretical analysis and the scope of the claims. In particular, it clarifies that the results characterize attention mechanisms under these assumptions and do not by themselves imply empirical superiority or universal practical applicability.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions underlying the theoretical results and provides formal proofs in the appendix. The results are presented within the stated scope of the theoretical framework.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the data generation procedures and parameter settings used in the numerical experiments. It also specifies the evaluation procedure needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code for reproducing the numerical experiments is publicly available at `https://github.com/masato-uchida/chain-of-dual-structures`. The appendix provides the data generation procedures, parameter settings, and other experimental details needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental settings needed to understand the numerical results, including the data generation procedures and parameter settings. Additional details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports standard deviations for the numerical experiments. The reported variability is computed over repeated runs under the specified experimental settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the computational environment used for the numerical experiments and states that they completed within a short execution time without specialized compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The paper does not involve human subjects, sensitive personal data, or high-risk data/model releases.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses broader impacts in the appendix. It states that the work is foundational and does not identify direct negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse. It is a foundational theoretical study and does not release pre-trained models, image generators, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The experiment in Appendix N uses the pretrained GPT-2 small model and the WikiText-2 dataset obtained from Hugging Face, specifically `openai-community/gpt2` and `Salesforce/wikitext`, respectively. The GPT-2 model is distributed under the MIT License, and the WikiText-2 dataset is distributed under the CC BY-SA 3.0 and GFDL licenses. These assets are used in accordance with their respective license terms and are credited in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code for reproducing the numerical experiments is publicly available at `https://github.com/masato-uchida/chain-of-dual-structures`. The repository includes documentation describing the files and their correspondence to the experiments in the paper.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development does not involve LLMs as an important, original, or non-standard component. Any LLM use was limited to writing, editing, formatting, or implementation support.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.