WASEDA University

NEURAL INFORMATION PROCESSING SYSTEMS

Chain of Dual Structures in Transformer Attention

Masato Uchida

Waseda University, Japan

$\psi \longleftrightarrow \psi^*$

Query-key similarity

$\varphi \longleftrightarrow \varphi^*$

Score normalization

$\pi \longleftrightarrow \pi^*$

Value aggregation

Why These Three Operations?

Queries, keys, and values belong to distinct representation spaces.

Query space

$$\boldsymbol{q}^\top \in \mathcal{Q}$$
$$\boldsymbol{q}^\top = W_q \boldsymbol{x}^\top$$

Key space

$$\boldsymbol{k}^\top \in \mathcal{K}$$
$$\boldsymbol{k}^\top = W_k \boldsymbol{x}^\top$$

Value space

$$\boldsymbol{v}^\top \in \mathcal{V}$$
$$\boldsymbol{v}^\top = W_v \boldsymbol{x}^\top$$

Compare

Queries and keys → scores

$$s_j^i = \frac{\langle \boldsymbol{q}^i, \boldsymbol{k}^j \rangle}{\sqrt{d_{qk}}}$$

Inner product

Normalize

Scores → weights

$$a_j^i = \frac{\exp(s_j^i)}{\sum_{\ell=1}^{n} \exp(s_\ell^i)}$$

Softmax

Aggregate

Weights and values → output

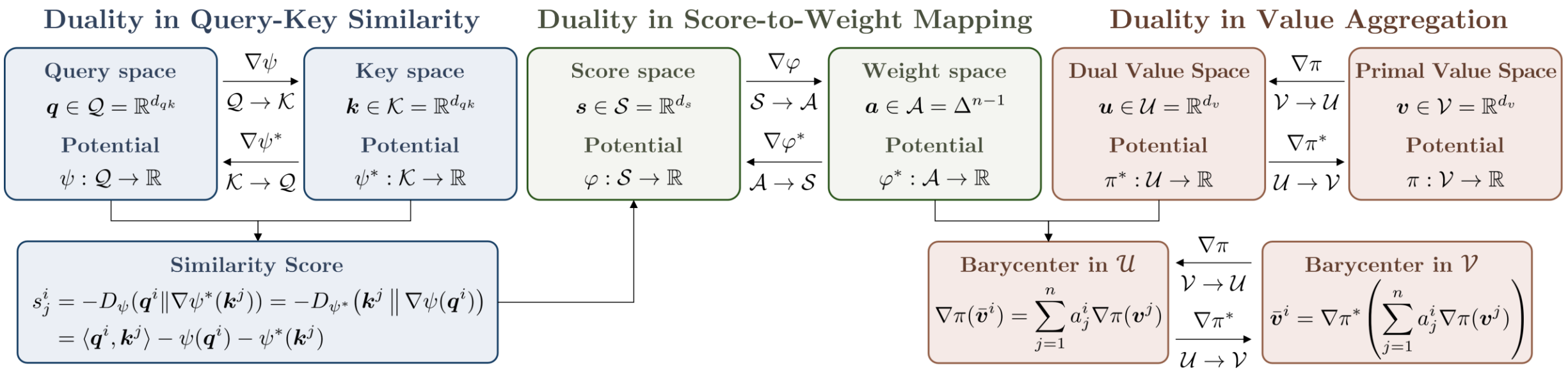
$$\boldsymbol{z}^i = \sum_{j=1}^{n} a_j^i \boldsymbol{v}^j$$

Weighted sum

Why do these three operations take these mathematical forms?

A Chain of Dual Structures

Convex potentials and their Legendre conjugates connect each stage.



Comparison

Bregman scores recover inner-product attention in a limit.

Normalization

Structural requirements characterize the gradient form.

Aggregation

Consistency requirements characterize the barycentric form.

Does this structure follow from fundamental requirements?

Query–Key Similarity and Normalization

Recovering the standard score; characterizing the score-to-weight map.

Query-Key Similarity

THEOREM 5.1

A score defined by a Bregman divergence

$$s_j^i = -D_\psi(\boldsymbol{q}^i \| \nabla\psi^*(\boldsymbol{k}^j)) = \langle \boldsymbol{q}^i, \boldsymbol{k}^j \rangle - \psi(\boldsymbol{q}^i) - \psi^*(\boldsymbol{k}^j)$$

Scale the potential; apply softmax.

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

$$\psi^*(\tau \boldsymbol{k}^j) \to c, \ (\tau \to 0, \forall j)$$

⬇

Standard inner-product softmax attention

Score-to-Weight Mapping

THEOREM 6.1

Structural requirements on normalization

Translation invariance	$g(\boldsymbol{s} + c\mathbf{1}) = g(\boldsymbol{s})$
Positive, normalized	$\sum_{i=1}^n g_i = 1, \ g_i > 0$
Monotonicity	$\partial_i g_i > 0 > \partial_j g_i, \ (i \neq j)$
Symmetric Jacobian	$\partial_j g_i = \partial_i g_j$

⬇

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

for a convex potential φ

For Softmax: $\varphi(\boldsymbol{s}) = \log \sum_j \exp(s_j)$

Value Aggregation from Consistency

Deriving linear aggregation in dual coordinates from consistency requirements.

Consistency Requirements in Dual Coordinates

THEOREM 7.1

Unanimity	$B(\boldsymbol{a}; \boldsymbol{u}, \boldsymbol{u}, \ldots, \boldsymbol{u}) = \boldsymbol{u}$
Permutation invariance	$B(\boldsymbol{a}_\sigma; \boldsymbol{u}^{\sigma(1)}, \ldots, \boldsymbol{u}^{\sigma(n)}) = B(\boldsymbol{a}; \boldsymbol{u}^1, \ldots, \boldsymbol{u}^n)$
Affine equivariance	$B(\boldsymbol{a}; T\boldsymbol{u}^1, \ldots, T\boldsymbol{u}^n) = TB(\boldsymbol{a}; \boldsymbol{u}^1, \ldots, \boldsymbol{u}^n)$
Continuity in weights	$\boldsymbol{a}^{(m)} \to \boldsymbol{a} \;\Rightarrow\; B(\boldsymbol{a}^{(m)}; \boldsymbol{u}^1, \ldots, \boldsymbol{u}^n) \to B(\boldsymbol{a}; \boldsymbol{u}^1, \ldots, \boldsymbol{u}^n)$
Zero-weight invariance	$B((1,0); \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}, \quad B((0,1); \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}$
Grouping and reaggregation	$B(\boldsymbol{a}; \boldsymbol{u}^1, \ldots, \boldsymbol{u}^n) = B\left(\alpha; B(\boldsymbol{a}_{\cdot\|1}; U_1), \ldots, B(\boldsymbol{a}_{\cdot\|m}; U_m)\right)$ $\alpha_r = \sum_{j \in G_r} a_j, \;\; a_{j\|r} = \frac{a_j}{\alpha_r} \;\; U_r = (\boldsymbol{u}^j)_{j \in G_r}$



$$B(\boldsymbol{a}; \boldsymbol{u}^1, \ldots, \boldsymbol{u}^n) = \sum_{j=1}^n a_j \boldsymbol{u}^j, \quad \boldsymbol{z} = \nabla\pi^*\left(\sum_{j=1}^n a_j \nabla\pi(\boldsymbol{v}^j)\right)$$

Euclidean case: $\pi(\boldsymbol{v}) = \|\boldsymbol{v}\|^2/2 \;\Rightarrow\; \boldsymbol{z} = \sum_{j=1}^n a_j \boldsymbol{v}^j$

From Dual Structure to Local Output Behavior

The keys, values, and normalization determine the directions and dimension of local output changes.

Local Response in Dual Coordinates THEOREM 7.2

First-order change of the dual output

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

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

Attainable Directions $\operatorname{Im}\mathcal{C}_\pi^i$

Local Value-side Dimension $\operatorname{rank}\mathcal{C}_\pi^i$

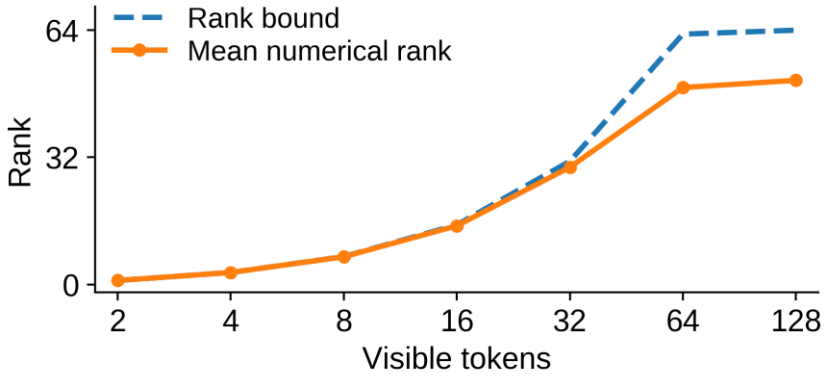
Rank bound: $\operatorname{rank}\mathcal{C}_\pi^i \le \min\{d_{qk}, d_v, n-1\}$

For Standard Attention

Softmax + linear value aggregation

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

GPT-2 small



Local output dimensions in pretrained GPT-2