Conservative Entropic Transport Attention: A Mass-Balanced and Stability-Certified Attention Primitive

Yu Murakami
New York General Group

Abstract

Scaled dot-product attention is the central operation of the Transformer [1]. Its attention matrix is row-stochastic: each query distributes one unit of probability mass across keys. This guarantees that each output token is a convex combination of value tokens, but it does not constrain the total mass received by each key. Consequently, many queries may concentrate on the same key, producing highly imbalanced token participation. This paper studies *Conservative Entropic Transport Attention* (CETA), an attention primitive in which the attention matrix is the unique entropy-regularized optimal transport plan over the Birkhoff polytope of doubly stochastic matrices. CETA is closely related to Sinkhorn-normalized attention and Sinkformers [10], but the focus here is different: we isolate the conservation and perturbation-stability consequences of replacing row-stochastic attention by entropy-regularized doubly stochastic transport. We prove existence, uniqueness, Sinkhorn representability, permutation equivariance, mean conservation, attention-receiver balance, convex-hull safety, fixed-plan non-expansiveness, residual non-expansiveness, and an explicit score-perturbation certificate. We also distinguish which guarantees follow from row-stochasticity, which follow from double stochasticity, and which require the entropic optimal-transport formulation. The resulting theory clarifies what reliability guarantees are obtained by mass-balanced attention and where those guarantees remain limited.

1 Introduction

The Transformer replaced recurrent and convolutional sequence transduction with stacked self-attention and feed-forward layers [1]. Its basic attention operation is

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\Big(QK^\top/\sqrt{d_k}\Big)\,V, \qquad (1)$$

where softmax is applied row-wise. Thus the attention matrix is row-stochastic: every row sums to one. This makes each output token a convex combination of value vectors.

Row-stochasticity, however, is one-sided. It says how much mass each query spends, but not how much mass each key receives. A single key may receive nearly all attention mass from many queries. This phenomenon is related to attention concentration and, in autoregressive streaming settings, to empirically observed attention sinks [13]. The present paper asks a theoretical question: what changes if self-attention is made *mass-balanced*, so that every key receives exactly the same total attention mass?

We answer this by defining attention as an entropy-regularized optimal transport problem over the Birkhoff polytope. The resulting attention matrix is doubly stochastic. Thus every query sends one unit of mass and every key receives one unit of mass. The method can be computed by Sinkhorn scaling, following the same mathematical foundation as entropic optimal transport [11].

The idea of Sinkhorn-normalized attention is not new. Sinkformers introduced doubly stochastic attention via Sinkhorn normalization and connected it to optimal transport, Wasserstein gradient flows, and heat diffusion limits [10]. The contribution of this paper is narrower and explicitly theoretical. We give a self-contained conservation-stability account of entropy-regularized doubly stochastic attention, prove a score-perturbation bound for the attention plan, and classify the resulting guarantees according to the assumptions they actually require.

1.1 Contributions

This paper makes four contributions. First, it formulates self-attention as a strictly concave entropy-regularized transport problem over the Birkhoff polytope and proves the precise Sinkhorn representation of the unique optimizer. Second, it separates three classes of guarantees:

Row-stochastic	:	convex-hull safety, fixed-plan non-expansiveness;
Doubly stochastic	:	mean conservation, equal receiver mass;
Entropic OT	:	uniqueness, Sinkhorn form, score stability.

Third, it proves an explicit score perturbation certificate

$$\|P_\tau(S) - P_\tau(T)\|_1 \le \frac{n}{\tau}\|S-T\|_\infty, \qquad (2)$$

where $P_\tau(S)$ is the CETA attention plan. This bound is dimension-free in the feature dimension but grows linearly with sequence length and inversely with temperature. Fourth, it discusses practical scope: multi-head attention, causal masking, long-sequence limitations, and diagnostic tests for receiver-mass imbalance.

1.2 Relation to Sinkformers and Entropic OT Stability

Sinkformers [10] already propose replacing row-wise softmax attention with Sinkhorn-normalized doubly stochastic attention. They also identify the connection to entropic optimal transport and prove properties related to permutation equivariance, continuity, Wasserstein gradient-flow structure, and heat-diffusion limits. The present paper should therefore not be read as claiming that Sinkhorn attention itself is new. Instead, the technical emphasis is different: an explicit

conservation-stability decomposition and an elementary perturbation certificate tailored to finite Transformer attention matrices.

The score-stability theorem is related to a broader entropic optimal transport stability literature, including work on Schrödinger potentials and Sinkhorn convergence [14, 15]. The bound proved here is intentionally elementary and is not claimed to be tight; sharper bounds exist in more specialized OT settings.

2 Notation

Let n be the sequence length and d_k the key/query dimension. Let $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^n$. For a matrix A, define

$$\|A\|_1 = \sum_{i,j} |A_{ij}|, \qquad \|A\|_\infty = \max_{i,j} |A_{ij}|. \tag{3}$$

For token matrices $X \in \mathbb{R}^{n\times d}$, define

$$\|X\|_{\infty,2} = \max_{1\le i\le n} \|X_i\|_2. \tag{4}$$

The Birkhoff polytope is

$$\mathcal{B}_n = \{P \in \mathbb{R}_+^{n\times n} : P\mathbf{1} = \mathbf{1},\ P^\top \mathbf{1} = \mathbf{1}\}. \tag{5}$$

Matrices in $\mathcal{B}_n$ are doubly stochastic. The Birkhoff-von Neumann theorem states that $\mathcal{B}_n$ is the convex hull of the $n \times n$ permutation matrices [12]. For $S \in \mathbb{R}^{n\times n}$ and $\tau > 0$, define

$$\Phi_\tau(P; S) = \langle S, P\rangle - \tau \sum_{i,j} P_{ij} \log P_{ij}, \tag{6}$$

with the continuous extension $0 \log 0 = 0$.

3 Conservative Entropic Transport Attention

Given queries, keys, and values

$$Q \in \mathbb{R}^{n\times d_k}, \quad K \in \mathbb{R}^{n\times d_k}, \quad V \in \mathbb{R}^{n\times d_v}, \tag{7}$$

define the score matrix $S = QK^\top/\sqrt{d_k}$. CETA replaces rowwise softmax by the entropy-regularized transport plan

$$P_\tau(S) = \arg\max_{P\in\mathcal{B}_n} \left\{ \langle S, P\rangle - \tau \sum_{i,j} P_{ij} \log P_{ij} \right\}. \tag{8}$$

The CETA head is

$$\mathrm{CETA}_\tau(Q, K, V) = P_\tau(S)V. \tag{9}$$

For a Transformer block, one may use

$$Y = X + \alpha P_\tau(S) V W_O, \tag{10}$$

where $V = XW_V$, $Q = XW_Q$, $K = XW_K$, W_O is an output projection, and α is a residual scaling parameter. For the cleanest conservation theory, it is useful to study the identity-value residual core

$$T_\alpha(X) = (1-\alpha)X + \alpha P_\tau(S)X, \quad \alpha \in [0, 1]. \tag{11}$$

The projected form is the practical neural module; the identity-value form isolates the exact conservation and nonexpansiveness properties.

4 Computation by Sinkhorn Scaling

Let $G = \exp(S/\tau)$ entrywise. Since $G_{ij} > 0$, the CETA plan is the unique matrix of the form

$$P_\tau(S) = \mathrm{diag}(a) G\, \mathrm{diag}(b) \tag{12}$$

with positive scaling vectors $a, b \in \mathbb{R}^n_{++}$ such that

$$P_\tau(S)\mathbf{1} = \mathbf{1}, \qquad P_\tau(S)^\top \mathbf{1} = \mathbf{1}. \tag{13}$$

The standard Sinkhorn iteration alternates row and column scaling:

$$b^{(t)} = 1/(G^\top a^{(t)}), \qquad a^{(t+1)} = 1/(Gb^{(t)}), \tag{14}$$

where division is coordinatewise. The finite-iteration approximation is

$$P^{(t)} = \mathrm{diag}(a^{(t)}) G\, \mathrm{diag}(b^{(t)}). \tag{15}$$

For strictly positive kernels, Sinkhorn scaling converges geometrically in Hilbert's projective metric under standard assumptions; this is the classical convergence theory of matrix scaling [11, 16].

5 Main Results

Theorem 1 (Existence and uniqueness)**.** *For every $S \in \mathbb{R}^{n\times n}$ and every $\tau > 0$, the CETA problem has a unique maximizer $P_\tau(S) \in \mathcal{B}_n$.*

Proof. The set $\mathcal{B}_n$ is nonempty because $n^{-1}\mathbf{1}\mathbf{1}^\top \in \mathcal{B}_n$. It is closed and bounded in finite-dimensional Euclidean space, hence compact. The function $x \mapsto -x \log x$ has a continuous extension to $[0, \infty)$ by setting $0 \log 0 = 0$. Therefore $P \mapsto \Phi_\tau(P; S)$ is continuous on $\mathcal{B}_n$. By Weierstrass' theorem, a maximizer exists.

To prove uniqueness, take distinct $P, Q \in \mathcal{B}_n$ and $\lambda \in (0, 1)$. Since $P \ne Q$, at least one coordinate differs. The function $x \mapsto -x \log x$ is strictly concave on $[0, \infty)$ in the sense that for $x \ne y$,

$$-\big(\lambda x + (1-\lambda)y\big) \log\big(\lambda x + (1-\lambda)y\big) > -\lambda x \log x - (1-\lambda) y \log y. \tag{16}$$

Summing over coordinates gives $H(\lambda P + (1-\lambda)Q) > \lambda H(P) + (1-\lambda)H(Q)$, where $H(P) = -\sum_{i,j} P_{ij} \log P_{ij}$. Since $P \mapsto \langle S, P\rangle$ is linear, $\Phi_\tau(\cdot; S)$ is strictly concave on the convex set $\mathcal{B}_n$. A strictly concave function has at most one maximizer. Hence the maximizer exists and is unique. □

Theorem 2 (Full support and Sinkhorn representation)**.** *Let $G = \exp(S/\tau)$. The unique optimizer satisfies*

$$P_\tau(S) = \mathrm{diag}(a) G\, \mathrm{diag}(b) \tag{17}$$

for positive vectors $a, b \in \mathbb{R}^n_{++}$. The vectors are unique up to the reciprocal rescaling $a \mapsto ca$, $b \mapsto b/c$.

Proof. Let $P = P_\tau(S)$. We first prove that $P_{ij} > 0$ for all i, j. Assume for contradiction that $P_{ij} = 0$. Since row i sums to one, there exists $l \ne j$ such that $P_{il} > 0$. Since column j

sums to one, there exists $k \neq i$ such that $P_{kj} > 0$. Consider the 2×2 cycle perturbation

$$D = e_i e_j^\top + e_k e_l^\top - e_i e_l^\top - e_k e_j^\top. \tag{18}$$

For sufficiently small $\varepsilon > 0$, the matrix $P_\varepsilon = P + \varepsilon D$ has the same row and column sums as P. Also, all entries remain nonnegative because the only decreased entries are P_{il} and P_{kj}, both positive. Therefore $P_\varepsilon \in \mathcal{B}_n$.

The linear term changes by $O(\varepsilon)$. The entropy contribution from increasing P_{ij} from zero to ε is $-\tau\varepsilon \log \varepsilon$, whose ratio to ε tends to $+\infty$ as $\varepsilon \downarrow 0$. The entropy changes in the other three involved coordinates are $O(\varepsilon)$, because those coordinates are positive at P. Hence, for sufficiently small $\varepsilon > 0$, $\Phi_\tau(P_\varepsilon; S) > \Phi_\tau(P; S)$, contradicting optimality. Thus P has full support.

Since P is an interior point relative to the affine row/column constraints, the Karush-Kuhn-Tucker stationarity conditions apply. Introduce Lagrange multipliers λ_i, μ_j for the equality constraints. The Lagrangian for the maximization problem is

$$\mathcal{L}(P, \lambda, \mu) = \sum_{i,j} S_{ij} P_{ij} - \tau \sum_{i,j} P_{ij} \log P_{ij} + \sum_i \lambda_i \left(1 - \sum_j P_{ij}\right) + \sum_j \mu_j \left(1 - \sum_i P_{ij}\right). \tag{19}$$

At the optimum,

$$0 = \frac{\partial \mathcal{L}}{\partial P_{ij}} = S_{ij} - \tau(\log P_{ij} + 1) - \lambda_i - \mu_j. \tag{20}$$

Thus

$$P_{ij} = a_i \exp(S_{ij}/\tau) b_j = a_i G_{ij} b_j, \tag{21}$$

where $a_i = \exp(-1 - \lambda_i/\tau)$ and $b_j = \exp(-\mu_j/\tau)$. Since $P \in \mathcal{B}_n$, a, b scale G to have all row and column sums equal to one. If (a, b) and (a', b') produce the same positive scaled matrix, then $a'_i/a_i = b_j/b'_j$ for all i, j, so the ratio is a positive constant. Hence the scaling vectors are unique up to reciprocal rescaling. □

Theorem 3 (Permutation equivariance)**.** *For every permutation matrix* Π*,*

$$P_\tau(\Pi S \Pi^\top) = \Pi P_\tau(S) \Pi^\top. \tag{22}$$

Consequently,

$$\mathrm{CETA}_\tau(\Pi Q, \Pi K, \Pi V) = \Pi\, \mathrm{CETA}_\tau(Q, K, V). \tag{23}$$

Proof. A matrix R belongs to $\mathcal{B}_n$ if and only if $\Pi^\top R \Pi \in \mathcal{B}_n$. Also, $\langle \Pi S \Pi^\top, R \rangle = \langle S, \Pi^\top R \Pi \rangle$, and entropy is invariant under coordinate permutation. Thus maximizing the objective with score $\Pi S \Pi^\top$ over $R \in \mathcal{B}_n$ is equivalent to maximizing the objective with score S over $\Pi^\top R \Pi \in \mathcal{B}_n$. By uniqueness, $P_\tau(\Pi S \Pi^\top) = \Pi P_\tau(S) \Pi^\top$.

If Q, K, V are permuted by Π, then $(\Pi Q)(\Pi K)^\top/\sqrt{d_k} = \Pi S \Pi^\top$. Therefore,

$$\mathrm{CETA}_\tau(\Pi Q, \Pi K, \Pi V) = P_\tau(\Pi S \Pi^\top) \Pi V = \Pi P_\tau(S) V. \tag{24}$$

□

Theorem 4 (Equal receiver mass)**.** *For every CETA attention matrix* $P_\tau(S)$*,*

$$\sum_{i=1}^n P_\tau(S)_{ij} = 1 \tag{25}$$

for every key index j*. Thus every token receives exactly equal total attention mass.*

Proof. Because $P_\tau(S) \in \mathcal{B}_n$, $P_\tau(S)^\top \mathbf{1} = \mathbf{1}$. The j-th coordinate of this equality is $\sum_i P_\tau(S)_{ij} = 1$. □

This theorem is not specific to the entropic objective. It holds for every doubly stochastic attention mechanism. It is nevertheless the central conservation property that distinguishes doubly stochastic attention from ordinary row-softmax attention.

Theorem 5 (Mean conservation)**.** *For every* $V \in \mathbb{R}^{n \times d_v}$*,*

$$\frac{1}{n} \mathbf{1}^\top P_\tau(S) V = \frac{1}{n} \mathbf{1}^\top V. \tag{26}$$

Proof. Since $P_\tau(S)^\top \mathbf{1} = \mathbf{1}$, we have $\mathbf{1}^\top P_\tau(S) = \mathbf{1}^\top$. Multiplying by V and dividing by n proves the claim. □

This is again a doubly stochastic consequence, not an entropic-OT-specific consequence.

Theorem 6 (Convex-hull safety)**.** *Let* $A \in \mathbb{R}_+^{n \times n}$ *be any row-stochastic matrix. Then every row of* AV *lies in the convex hull of the rows of* V*. In particular, CETA and ordinary softmax attention both satisfy this property.*

Proof. Since A is row-stochastic, $A_{ij} \geq 0$ and $\sum_j A_{ij} = 1$. Therefore $(AV)_i = \sum_j A_{ij} V_j$ is a convex combination of $V_1, \ldots, V_n$. □

Theorem 7 (Fixed-plan non-expansiveness)**.** *Let* $A \in \mathbb{R}_+^{n \times n}$ *be row-stochastic. Then for all* $V, W \in \mathbb{R}^{n \times d_v}$*,*

$$\|AV - AW\|_{\infty,2} \leq \|V - W\|_{\infty,2}. \tag{27}$$

Thus this property holds for CETA, Sinkhorn attention, and ordinary row-softmax attention whenever the attention matrix is held fixed.

Proof. Let $D = V - W$. For row i, $(AD)_i = \sum_j A_{ij} D_j$. By the triangle inequality,

$$\|(AD)_i\|_2 \leq \sum_j A_{ij} \|D_j\|_2 \leq \max_j \|D_j\|_2 = \|D\|_{\infty,2}. \tag{28}$$

Taking the maximum over i gives the claim. □

Theorem 8 (Fixed-plan residual non-expansiveness)**.** *Let $A \in \mathbb{R}_+^{n\times n}$ be row-stochastic and let $\alpha \in [0,1]$. Define $T_\alpha(X) = (1-\alpha)X + \alpha AX$. Then*

$$\|T_\alpha(X) - T_\alpha(Y)\|_{\infty,2} \le \|X - Y\|_{\infty,2}. \tag{29}$$

Proof. Let $D = X - Y$. Then $T_\alpha(X) - T_\alpha(Y) = (1-\alpha)D + \alpha AD$. For each row i,

$$\|(1-\alpha)D_i + \alpha(AD)_i\|_2 \le (1-\alpha)\|D_i\|_2 + \alpha\|(AD)_i\|_2. \tag{30}$$

Using the previous theorem and $\|D_i\|_2 \le \|D\|_{\infty,2}$ gives the result. □

Theorem 9 (Score perturbation stability)**.** *Let $S, T \in \mathbb{R}^{n\times n}$, $\tau > 0$, and $P = P_\tau(S)$, $Q = P_\tau(T)$. Then*

$$\|P - Q\|_1 \le \frac{n}{\tau}\|S - T\|_\infty. \tag{31}$$

Proof. Define $\psi(R) = \sum_{i,j} R_{ij} \log R_{ij}$. The matrix P minimizes $F_S(R) = \tau\psi(R) - \langle S, R\rangle$ over $\mathcal{B}_n$, and Q minimizes $F_T(R) = \tau\psi(R) - \langle T, R\rangle$. Since P and Q have full support, the gradients are well-defined: $(\nabla\psi(P))_{ij} = \log P_{ij} + 1$ and $(\nabla\psi(Q))_{ij} = \log Q_{ij} + 1$.

The variational inequality for P gives

$$\langle \tau\nabla\psi(P) - S, Q - P\rangle \ge 0. \tag{32}$$

The variational inequality for Q gives

$$\langle \tau\nabla\psi(Q) - T, P - Q\rangle \ge 0. \tag{33}$$

Adding these inequalities and rearranging yields

$$\tau\langle \nabla\psi(P) - \nabla\psi(Q), P - Q\rangle \le \langle S - T, P - Q\rangle. \tag{34}$$

By Holder's inequality,

$$\langle S - T, P - Q\rangle \le \|S - T\|_\infty \|P - Q\|_1. \tag{35}$$

We now lower-bound the entropy term. Let $p = \mathrm{vec}(P)/n$ and $q = \mathrm{vec}(Q)/n$. Because $P, Q \in \mathcal{B}_n$, their entries sum to n, so p and q are probability vectors. For the negative entropy $\varphi(p) = \sum_a p_a \log p_a$ on the probability simplex,

$$\langle \nabla\varphi(p) - \nabla\varphi(q), p - q\rangle = \mathrm{KL}(p\|q) + \mathrm{KL}(q\|p). \tag{36}$$

Indeed,

$$\langle \nabla\varphi(p) - \nabla\varphi(q), p - q\rangle = \sum_a (\log p_a - \log q_a)(p_a - q_a) \tag{37}$$

$$= \mathrm{KL}(p\|q) + \mathrm{KL}(q\|p), \tag{38}$$

where the $+1$ terms cancel. By Pinsker's inequality,

$$\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p) \ge \|p - q\|_1^2. \tag{39}$$

Because $p_{ij} = P_{ij}/n$ and $q_{ij} = Q_{ij}/n$,

$$\log p_{ij} - \log q_{ij} = \log P_{ij} - \log Q_{ij}, \qquad p_{ij} - q_{ij} = \frac{P_{ij} - Q_{ij}}{n}. \tag{40}$$

Therefore,

$$\langle \nabla\varphi(p) - \nabla\varphi(q), p - q\rangle = \frac{1}{n}\langle \nabla\psi(P) - \nabla\psi(Q), P - Q\rangle. \tag{41}$$

Also, $\|p - q\|_1 = \|P - Q\|_1/n$. Hence

$$\langle \nabla\psi(P) - \nabla\psi(Q), P - Q\rangle \ge \frac{1}{n}\|P - Q\|_1^2. \tag{42}$$

Combining the inequalities gives

$$\frac{\tau}{n}\|P - Q\|_1^2 \le \|S - T\|_\infty \|P - Q\|_1. \tag{43}$$

If $P = Q$, the theorem is immediate. Otherwise divide by $\|P - Q\|_1$ to obtain the claim. □

The bound is intentionally simple. It is independent of the feature dimension but deteriorates linearly with n and as $1/\tau$. For long sequences or low temperatures, this certificate may be loose or vacuous.

Theorem 10 (Output perturbation certificate)**.** *Let $P = P_\tau(S)$ and $Q = P_\tau(T)$. For values $V, W \in \mathbb{R}^{n\times d_v}$,*

$$\|PV - QW\|_{\infty,2} \le \|V - W\|_{\infty,2} + \frac{n}{\tau}\|S - T\|_\infty \|W\|_{\infty,2}. \tag{44}$$

Proof. Decompose $PV - QW = P(V - W) + (P - Q)W$. Since P is row-stochastic, fixed-plan non-expansiveness gives $\|P(V - W)\|_{\infty,2} \le \|V - W\|_{\infty,2}$. For the second term,

$$\left\|\sum_j (P_{ij} - Q_{ij})W_j\right\|_2 \le \|W\|_{\infty,2}\sum_j |P_{ij} - Q_{ij}|. \tag{45}$$

Taking the maximum over rows gives $\|(P - Q)W\|_{\infty,2} \le \|W\|_{\infty,2}\|P - Q\|_1$. Applying the score perturbation theorem completes the proof. □

6 Attention-Sink Diagnostic

For a row-stochastic attention matrix A, define the receiver-mass vector

$$m(A) = A^\top \mathbf{1}. \tag{46}$$

For ordinary attention, $\sum_j m_j(A) = n$, but the coordinates of $m(A)$ can be highly unequal. A natural imbalance diagnostic is

$$\Delta(A) = \|A^\top \mathbf{1} - \mathbf{1}\|_\infty. \tag{47}$$

For CETA, $\Delta(P_\tau(S)) = 0$ exactly. For ordinary row-softmax attention, $\Delta(A)$ can be as large as $n - 1$. Indeed, if every row concentrates on the first key, then $A_{i1} = 1$ and $A_{ij} = 0$ for $j \ne 1$. Hence $A^\top \mathbf{1} = (n, 0, \ldots, 0)^\top$ and $\Delta(A) = n - 1$. This is not an empirical result about trained models. It is a worst-case diagnostic showing that row-stochastic attention permits receiver-mass collapse, whereas CETA forbids it by construction.

7 Multi-Head CETA

For h heads, define $P^{(r)} = P_{\tau_r}(S^{(r)})$, $r = 1, \dots, h$. Each head is individually doubly stochastic:

$$P^{(r)}\mathbf{1} = \mathbf{1}, \qquad (P^{(r)})^\top \mathbf{1} = \mathbf{1}. \tag{48}$$

Thus conservation holds head-wise before output projection. The concatenated multi-head output is

$$\mathrm{Concat}(P^{(1)}V^{(1)}, \dots, P^{(h)}V^{(h)})W_O. \tag{49}$$

The output projection W_O mixes feature channels, not token positions. Therefore it does not change token-level receiver-mass balance inside each head, although it can alter feature-level mean conservation unless the value/output maps are constrained. Exact token-feature mean conservation is guaranteed only for the identity-value conservative core or for projections that preserve the relevant feature averages.

8 Causal and Masked Attention

Standard autoregressive attention uses a lower-triangular mask. A fully doubly stochastic matrix with a strict causal support constraint generally cannot exist. For example, if token 1 may attend only to itself, row one forces $P_{11} = 1$. Column one then already has total mass one, so all later tokens must assign zero mass to token one. Continuing inductively forces the identity matrix as the only feasible exactly causal doubly stochastic plan under a strict triangular mask.

Thus exact CETA is naturally suited to bidirectional or non-causal attention. Causal variants require a different feasible set, such as rectangular prefix transport, unbalanced optimal transport, column-cap constraints, or blockwise bidirectional windows. These variants may preserve partial receiver-mass control, but not full Birkhoff conservation.

9 Limitations

CETA is not presented as a universal replacement for softmax attention. First, dense CETA has quadratic score construction and iterative Sinkhorn normalization. It is not a linear-attention method. Second, the score-stability certificate scales as n/τ. For long sequences and small temperatures, the bound can be loose. This limitation is mathematical, not merely analytical: low-temperature optimal transport approaches hard matching, where small score perturbations can change the selected permutation. Third, exact double stochasticity is incompatible with strict causal triangular attention except in degenerate cases. This limits direct use in standard autoregressive decoding. Fourth, several useful properties–convex-hull safety and fixed-plan non-expansiveness–are not unique to CETA. They are properties of row-stochastic attention generally. Fifth, no empirical superiority claim is made here. The theory predicts exact receiver-mass balance and provides a perturbation certificate, but practical gains must be established by experiments.

10 Conclusion

Conservative Entropic Transport Attention defines self-attention as the unique entropy-regularized transport plan over the Birkhoff polytope. It gives a mass-balanced alternative to row-softmax attention: every query sends one unit of attention mass and every key receives one unit. The resulting attention matrix is permutation equivariant, mean-conserving, Sinkhorn-representable, and stable under score perturbations in an explicit finite-dimensional norm.

The main conceptual point is not that Sinkhorn attention is new. It is that the reliability guarantees of doubly stochastic attention can be precisely decomposed: some are generic consequences of row-stochasticity, some are consequences of double stochasticity, and some depend on the entropic optimal-transport formulation. This decomposition clarifies both the promise and the limits of conservative attention.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473, 2014.

[3] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Re. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, 2022.

[4] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer. arXiv:2004.05150, 2020.

[5] K. Choromanski et al. Rethinking Attention with Performers. arXiv:2009.14794, 2020.

[6] S. Wang, B. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-Attention with Linear Complexity. arXiv:2006.04768, 2020.

[7] J. Ainslie et al. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of EMNLP*, 2023.

[8] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, 2017.

[9] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of ICML*, 2019.

[10] M. E. Sander, P. Ablin, M. Blondel, and G. Peyre. Sinkformers: Transformers with Doubly Stochastic Attention. In *Proceedings of AISTATS*, 2022.

[11] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. In *Advances in Neural Information Processing Systems*, 2013.

[12] G. Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucuman Revista, Serie A*, 5:147–151, 1946.

[13] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*, 2024.

[14] M. Nutz and J. Wiesel. Entropic Optimal Transport: Convergence of Potentials. *Probability Theory and Related Fields*, 2021.

[15] M. Nutz and J. Wiesel. Stability of Schrodinger Potentials and Convergence of Sinkhorn's Algorithm. arXiv:2201.10059, 2022.

[16] J. Franklin and J. Lorenz. On the Scaling of Multidimensional Matrices. *Linear Algebra and its Applications*, 114/115:717–735, 1989.