Mathematical Foundations of Transformers

This note surveys the mathematical foundations of transformer models, synthesizing existing literature from discrete token dynamics to mean-field limits on probability spaces. It compiles core established results concerning regularity, well-posedness, variational and optimal-transport formulations of attention, gradient-flow structures, and long-time behavior including consensus, clustering, stability, and phase transitions.

Contents

1 Notation and Conventions

Spaces and indices.

$d \in \mathbb{N}$	embedding dimension
$n \in \mathbb{N}$	sequence length
E	$:= \mathbb{R}^d$ or $\mathbb{S}^{d-1} \subset \mathbb{R}^d$
$[n]$	$:= \{1, \ldots, n\}$
i, j, k	$\in [n]$ (token indices)
h	$\in [H]$ (head index)
ℓ	$\in \{0, \ldots, L\}$ (layer index)
t	≥ 0 (continuous time)

Configurations and measures.

$\mathbf{1}_N$	$:= (1, \ldots, 1) \in \mathbb{R}^N$ (vector of ones)
x_i	$\in E$ (token i)
$X = (x_1, \ldots, x_n)$	$\in E^n = (\mathbb{R}^d)^n = \mathbb{R}^{n \times d}$ (token configuration)
$B(0, R)$ or B_R	$:= \{x \in E : \\|x\\| \leq R\}$
$B^n(0, R)$ or B_R^n	$:= \{X \in E^n : \\|x_i\\| \leq R \ \forall i\}$ (bounded inputs)
μ_X	$:= \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \in \mathcal{P}(E)$ (empirical measure)
$\mathcal{M}_n(E)$	$:= \{\mu_X : X \in E^n\}$ (empirical measures with n atoms)
$\mathcal{P}(E)$	Borel probability measures on E
$\mathcal{P}_p(E)$	$:= \left\{\mu \in \mathcal{P}(E) : \int \\|x\\|^p \, d\mu(x) < \infty\right\}$
$\mathcal{P}_{p,R}(E)$	$:= \{\mu \in \mathcal{P}_p(E) : \mathrm{supp}(\mu) \subset B(0, R)\}$
$f_\# \mu$	$:= \mu \circ f^{-1}$ (pushforward of μ by map f)

Linear maps and parameters.

$\mathcal{L}(E)$	$:= \{A : E \to E \text{ linear}\}$
Id_E	Identity map on E
I_N	Identity matrix on $\mathbb{R}^N$
Q, K, V, O	$\in \mathcal{L}(E)$
Q_h, K_h, V_h, O_h	$\in \mathcal{L}(E) \quad (h = 1, \ldots, H)$
θ	generic collection of parameters
Θ	$:= (\theta_1, \ldots, \theta_L)$ (depth-L model)

Norms, products, projections.

$A \otimes B$	Kronecker product of matrices A and B
$\langle x, y\rangle$	$:= x^\top y$ (Euclidean inner product)
$\|x\|$	$:= \sqrt{\langle x, x\rangle}$
$\|X\|_F$	$:= \Big(\sum_{i=1}^n \|x_i\|^2\Big)^{1/2}$
$\|A\|_{\mathrm{op}}$	$:= \sup_{\|x\|=1} \|Ax\|$
$P_x y$	$:= y - \langle x, y\rangle x$ (proj. onto $T_x\mathbb{S}^{d-1}$)

Optimal transport distances.

$W_p(\mu,\nu)$	$:= \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int_{E\times E} \|x-y\|^p \, d\pi(x,y)\Big)^{1/p}$
$W(\mu,\nu)$	$:= W_1(\mu,\nu)$
$W_2(\mu,\nu)$	standard 2-Wasserstein distance
$W_{m,2}(\mu,\nu)$	generic "twisted" W_2 (used when needed)
$c(x,y)$	generic ground cost (e.g. $\|x-y\|^p$)

Static self-attention (discrete and mean-field).

$Q(X), K(X), V(X)$	matrices $(Qx_i)_{i=1}^n, (Kx_i)_{i=1}^n, (Vx_i)_{i=1}^n \in E^n$
$L_{ij}(X)$	$:= \frac{1}{\sqrt{d}}\langle Qx_i, Kx_j\rangle$
$A_{ij}(X)$	$:= \frac{\exp(L_{ij}(X))}{\sum_{k=1}^n \exp(L_{ik}(X))}$
$S_\theta(X)$	$:= A(X)V(X) \in E^n$ (single-head output)
$S_\theta^{\mathrm{MH}}(X)$	$:= \big[S_{\theta_1}(X) \mid \dots \mid S_{\theta_H}(X)\big]O$
$\kappa_{\theta,\mu}(x,y)$	$:= \exp\big(\beta\langle Qx, Ky\rangle\big)$ (kernel)
$Z_{\theta,\mu}(x)$	$:= \int_E \kappa_{\theta,\mu}(x,y)\,\mu(dy)$ (partition function)
$a_{\theta,\mu}(x,dy)$	$:= \frac{\kappa_{\theta,\mu}(x,y)}{Z_{\theta,\mu}(x)}\,\mu(dy)$ (attention kernel)
$\Gamma_{\theta,\mu}$	$: E \to E$ (attention-induced map)
$F_\theta(\mu)$	$:= (\Gamma_{\theta,\mu})_\#\mu$ (mean-field map)

Transformer block.

$\mathrm{LN}, \mathrm{FFN}$	layer normalization and feed-forward maps ($E^n \to E^n$)
X^{att}	$:= X + S_\theta^{\mathrm{MH}}(\mathrm{LN}(X))$
$T_\theta(X)$	$:= X^{\mathrm{att}} + \mathrm{FFN}(\mathrm{LN}(X^{\mathrm{att}}))$ (pre-LN block)
$T_\Theta^{(L)}$	$:= T_{\theta_L} \circ \cdots \circ T_{\theta_1}$ (depth-L)

Local regularity.

$JF(X)$	Jacobian of $F : E^n \to E^n$ at X
$\mathrm{Lip}_{\mathrm{loc}}(F; X)$	$:= \|JF(X)\|_{\mathrm{op}}$
$\mathrm{Lip}_{\mathrm{loc}}^{W_p}(G;\mu)$	$:= \limsup_{\nu\to\mu} \frac{W_p(G(\mu),G(\nu))}{W_p(\mu,\nu)}$

Dynamics and flows.

$$
\begin{aligned}
x_i(t) &\quad \text{particle trajectory } [0,\infty) \to E \\
\dot{x}_i(t) &\quad = v(t, x_i(t); X(t)) \\
\mu_t &\quad := \tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_{x_i(t)} \text{ (empirical law)} \\
\text{PDE} &\quad \partial_t \mu_t + \nabla \cdot (v[\mu_t]\, \mu_t) = 0 \\
\mathcal{E} &\quad : \mathcal{P}(E) \to \mathbb{R} \text{ (energy functional)} \\
\partial_t \mu_t &\quad = -\operatorname{grad}_{\mathsf{d}} \mathcal{E}(\mu_t) \text{ (gradient flow)}
\end{aligned}
$$

2 Static Self-Attention Maps

2.1 Objects and Assumptions

Ambient space and metrics.

$$
\begin{aligned}
&E \subset \mathbb{R}^d, \quad \text{either } E = \mathbb{R}^d \text{ or } E = \mathbb{S}^{d-1}, \\
&\|\cdot\| \text{ as in 1}, \\
&W_p \text{ as in 1 (by default } p = 2 \text{ in this setting).}
\end{aligned}
$$

Discrete single-head SA.

$$
\begin{aligned}
&S_\theta : E^n \to E^n, \quad X \mapsto S_\theta(X), \\
&S_\theta(X) = A(X)V(X), \quad A_{ij}(X) = \frac{\exp(L_{ij}(X))}{\sum_{k=1}^n \exp(L_{ik}(X))}, \\
&L_{ij}(X) = \frac{1}{\sqrt{d}} \langle Qx_i, Kx_j \rangle \quad \text{(cf. 1).}
\end{aligned}
$$

Multi-head SA.

$$
S_\theta^{\mathrm{MH}} : E^n \to E^n, \quad S_\theta^{\mathrm{MH}}(X) := \big[S_{\theta_1}(X) \mid \ \dots \ \mid S_{\theta_H}(X)\big] O.
$$

Mean-field SA.

$$
F_\theta : \mathcal{P}_p(E) \to \mathcal{P}_p(E), \quad F_\theta(\mu) := (\Gamma_{\theta,\mu})_{\#}\mu,
$$

where $\Gamma_{\theta,\mu} : E \to E$ is the attention-induced map.

Standing assumptions.

(H1) *Bounded configurations.* There exists $R > 0$ such that $X \in B_R^n$, that is

$$
\|x_i\| \le R \quad \forall i \in [n]
$$

(equivalently, $\mathrm{supp}(\mu_X) \subset B(0,R)$) unless explicitly stated otherwise.

(H2) *Linear maps with bounded norms.*

$$
Q, K, V, O \in \mathcal{L}(E), \quad \|Q\|_{\mathrm{op}}, \|K\|_{\mathrm{op}}, \|V\|_{\mathrm{op}}, \|O\|_{\mathrm{op}} < \infty.
$$

(H3) *Softmax / temperature.* Logits are of the form

$$
L_{ij}(X) = \beta \langle Qx_i, Kx_j \rangle,
$$

with fixed $\beta > 0$ (by default $\beta = 1/\sqrt{d}$). Masked variants (causal, padding) are obtained by

$$
L_{ij}^M(X) = L_{ij}(X) + M_{ij},
$$

with a fixed mask matrix $M \in (\mathbb{R} \cup \{-\infty\})^{n \times n}$.

(H4) *Discrete–mean-field consistency.* For any $X = (x_1, \dots, x_n) \in E^n$ with $\mu_X = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$,

$$
F_\theta(\mu_X) = \mu_{S_\theta(X)}.
$$

2.2 Discrete Self-Attention on E^n: Euclidean Regularity

Definition 2.1. Single-head SA with temperature $\beta > 0$:

$$
\begin{aligned}
Q(X) &:= (Qx_i)_{i=1}^n, \quad K(X) := (Kx_i)_{i=1}^n, \quad V(X) := (Vx_i)_{i=1}^n, \\
L_{ij}(X) &:= \beta\langle Qx_i, Kx_j\rangle, \\
A_{ij}(X) &:= \frac{\exp(L_{ij}(X))}{\sum_{k=1}^n \exp(L_{ik}(X))}, \\
S_\theta(X) &:= A(X)V(X) \in E^n.
\end{aligned}
$$

Masked variant (e.g. causal mask M):

$$
\begin{aligned}
L_{ij}^M(X) &:= L_{ij}(X) + M_{ij}, \\
A_{ij}^M(X) &:= \frac{\exp(L_{ij}^M(X))}{\sum_{k=1}^n \exp(L_{ik}^M(X))}, \\
S_\theta^M(X) &:= A^M(X)V(X).
\end{aligned}
$$

Multi-head SA:

$$
\begin{aligned}
S_{\theta_h}(X) &:= A_h(X)V_h(X), \quad h = 1, \dots, H, \\
S_\theta^{\mathrm{MH}}(X) &:= \big[S_{\theta_1}(X) \mid \dots \mid S_{\theta_H}(X)\big]O.
\end{aligned}
$$

Definition 2.2 (Jacobian and local Lipschitz constant)**.** Directional derivative in direction $Y = (y_1, \dots, y_n) \in E^n$:

$$
\begin{aligned}
JS_\theta(X)[Y] &= J\big(A(X)V(X)\big)[Y] \\
&= JA(X)[Y]V(X) + A(X)JV(X)[Y].
\end{aligned}
$$

With $V(X) = (Vx_j)_j$,

$$JV(X)[Y] = (Vy_j)_{j=1}^n.$$

For the softmax weights,

$$JA_{ij}(X)[Y] = A_{ij}(X)\Big(JL_{ij}(X)[Y] - \sum_{k=1}^n A_{ik}(X)JL_{ik}(X)[Y]\Big),$$

and

$$JL_{ij}(X)[Y] = \beta\big(\langle Qy_i, Kx_j\rangle + \langle Qx_i, Ky_j\rangle\big).$$

Define the (local) Jacobian operator

$$JS_\theta(X) : E^n \to E^n, \quad Y \mapsto JS_\theta(X)[Y].$$

Local Lipschitz constant (Frobenius norm on E^n):

$$\mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) := \|JS_\theta(X)\|_{\mathrm{op}} = \sup_{\|Y\|_F = 1} \|JS_\theta(X)[Y]\|_F.$$

Global Lipschitz constant on the ball B_R^n (configurations with $\|x_i\| \le R$):

$$\mathrm{Lip}_{B_R^n}(S_\theta) := \sup_{X \in B_R^n} \mathrm{Lip}_{\mathrm{loc}}(S_\theta; X).$$

Theorem 2.3 (General Euclidean Upper Bound [3, Theorem 3.3])**.** *Consider single-head self-attention $S_\theta : E^n \to E^n$ in the Euclidean case $E = \mathbb{R}^d$, with logits*

$$L_{ij}(X) = \frac{1}{\sqrt{d}}\langle Qx_i, Kx_j\rangle \quad (i, j \in [n]).$$

Define the linear operator

$$B := \frac{1}{\sqrt{d}} Q^\top K \in \mathcal{L}(E),$$

so that $L_{ij}(X) = x_i^\top B x_j$. Let $R > 0$ and $n \in \mathbb{N}$. Then, for every configuration $X = (x_1, \ldots, x_n) \in B_R^n$, the local Lipschitz constant of S_θ satisfies

$$\mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) \;\le\; \sqrt{3}\,\|V\|_{\mathrm{op}}\Big(\|B\|_{\mathrm{op}}^2 R^4(4n+1) + n\Big)^{1/2}. \tag{1}$$

In particular, the global Lipschitz constant on B_R^n obeys the same bound:

$$\mathrm{Lip}_{B_R^n}(S_\theta) \;\le\; \sqrt{3}\,\|V\|_{\mathrm{op}}\Big(\|B\|_{\mathrm{op}}^2 R^4(4n+1) + n\Big)^{1/2}. \tag{2}$$

Proof. Appendix A.2.1. □

Proposition 2.4 (Tightness of $\sqrt{n}$ Bound [3, Proposition 3.4])**.** *Work in the Euclidean case $E = \mathbb{R}^d$ and assume single-head self-attention $S_\theta : E^n \to E^n$ with $V = \mathrm{Id}_E$ and logits*

$$L_{ij}(X) = \frac{1}{\sqrt{d}}\langle Qx_i, Kx_j\rangle \quad (i, j \in [n]).$$

Define

$$B := \frac{1}{\sqrt{d}} Q^\top K \in \mathcal{L}(E),$$

so that $L_{ij}(X) = x_i^\top B x_j$, and let $\gamma_1 \ge \cdots \ge \gamma_r$ be the real eigenvalues of B ($r \le d$). Set

$$\gamma := \max\big\{-\gamma_r,\ \gamma_1/8\big\}.$$

Then, for any $R > 0$ and any $n \ge 2$, the global Lipschitz constant $\mathrm{Lip}_{B_R^n}(S_\theta)$ in (2) satisfies

$$\mathrm{Lip}_{B_R^n}(S_\theta) \ge \frac{\sqrt{n-1}}{1 + (n-1)e^{-2R^2\gamma}}. \tag{3}$$

Proof. Appendix A.2.2 □

Theorem 2.5 (Large Radius Regime (Fixed n) [3, Theorem 3.7])**.** *Work in the Euclidean case $E = \mathbb{R}^d$ and consider single-head self-attention $S_\theta : E^n \to E^n$ with logits*

$$L_{ij}(X) = x_i^\top B x_j, \quad B := \frac{1}{\sqrt{d}} Q^\top K \in \mathcal{L}(E),$$

and a non-zero output projection $V \in \mathcal{L}(E)$.

Let $B(0,1) \subset E$ denote the open unit ball, and define

$$E_B := \Big\{X = (x_1, \ldots, x_n) \in B(0,1)^n : \forall i \in [n]\ \exists!\ j_i^\star \in [n]\ \text{s.t. } x_i^\top B x_{j_i^\star} = \max_{1 \le j \le n} x_i^\top B x_j\Big\}.$$

The complement of E_B in $B(0,1)^n$ has Lebesgue measure zero. Moreover, for any $X \in E_B$ there exists a function $\vartheta_X : [0, +\infty) \to [0, +\infty)$ such that $\vartheta_X(R) \to 1$ exponentially fast as $R \to +\infty$ and, for all $R > 0$,

$$\big\|JS_\theta(RX)\big\|_{\mathrm{op}} \le \vartheta_X(R)\|V\|_{\mathrm{op}}\sqrt{n}, \tag{4}$$

where $RX := (Rx_1, \ldots, Rx_n)$.

Proof. Appendix A.2.3 ☐

Theorem 2.6 (Finite mass splitting via replication [4, Theorem 4.2])**.** *Work in the Euclidean case $E = \mathbb{R}^d$ and consider single-head self-attention $S_\theta : E^n \to E^n$ with logits*

$$L_{ij}(X) = x_i^\top B x_j, \quad B \in \mathcal{L}(E),$$

and output projection $V = \mathrm{Id}_E$ (so $S_\theta(X) = A(X)X$ in matrix form), where

$$A_{ij}(X) = \frac{\exp(L_{ij}(X))}{\sum_{k=1}^n \exp(L_{ik}(X))}, \quad A(X) = (A_{ij}(X))_{i,j=1}^n.$$

For every integer $N \geq 1$, define the N-replication of $X = (x_1, \dots, x_n) \in E^n$ by

$$Y_N := \mathbf{1}_N \otimes X \in E^{Nn},$$

i.e. Y_N is the configuration obtained by repeating each token of X exactly N times.

For each $i \in [n]$, define an E-valued random variable $Z^{(i)}$ supported on $\{x_1, \dots, x_n\}$ by

$$\mathbb{P}\big(Z^{(i)} = x_j\big) = A_{ij}(X), \quad j \in [n],$$

and let its covariance operator be

$$\Sigma_i(X) := \mathbb{E}\Big[(Z^{(i)} - \mathbb{E}Z^{(i)})(Z^{(i)} - \mathbb{E}Z^{(i)})^\top\Big] \in \mathcal{L}(E).$$

Then:

1. *For every $N \geq 2$,*
$$\big\|JS_\theta(Y_N)\big\|_{\mathrm{op}} = \big\|JS_\theta(Y_2)\big\|_{\mathrm{op}}.$$

2. *If*
$$\sup_{N\geq 1}\big\|JS_\theta(Y_N)\big\|_{\mathrm{op}} > \big\|JS_\theta(X)\big\|_{\mathrm{op}},$$
then
$$\big\|JS_\theta(Y_2)\big\|_{\mathrm{op}}^2 = \max_{1\leq i\leq n}\big\|\Sigma_i(X)B^T\big\|_{\mathrm{op}}^2 > \big\|JS_\theta(X)\big\|_{\mathrm{op}}^2.$$
Writing
$$i^\star \in \arg\max_{1\leq i\leq n}\big\|\Sigma_i(X)B^T\big\|_{\mathrm{op}},$$
and letting $u^\star$ be a leading right singular vector of $\Sigma_{i^\star}(X)B^T$, the maximal increase under duplication can be achieved by splitting the $i^\star$-th token equally along $u^\star$, i.e. by replacing $x_{i^\star}$ with two atoms $x_{i^\star} \pm \varepsilon u^\star$ of equal weight.

Proof. Appendix A.2.4. ☐

Theorem 2.7 (Masked Attention Upper Bound [3, Theorem 4.3])**.** *Work in the Euclidean case $E = \mathbb{R}^d$. Let $\beta = 1/\sqrt{d}$ and define*

$$B := \beta Q^\top K \in \mathcal{L}(E), \quad L_{ij}(X) := \beta\langle Qx_i, Kx_j\rangle = \langle x_i, Bx_j\rangle, \quad i,j \in [n].$$

Let $M \in (\mathbb{R} \cup \{-\infty\})^{n\times n}$ be the standard left-causal mask,

$$M_{ij} := \begin{cases} 0 & \text{if } j \leq i, \\ -\infty & \text{if } j > i, \end{cases}$$

and let $S_\theta^M : E^n \to E^n$ be the masked single-head self-attention map. For $R > 0$ and $n \in \mathbb{N}$, define

$$B_R^n := \{X = (x_1, \dots, x_n) \in E^n : \|x_i\| \leq R \quad \forall i \in [n]\}.$$

Then S_θ^M is globally Lipschitz on B_R^n, and

$$\mathrm{Lip}_{B_R^n}(S_\theta^M) \leq \sqrt{3}\|V\|_{\mathrm{op}}\Big(\|B\|_{\mathrm{op}}^2 R^4(n+1) + n\Big)^{1/2}. \tag{5}$$

Proof. Appendix A.2.5. □

Theorem 2.8 (Masked Large Radius Regime (Fixed n) [3, Theorem 4.5]). *Work in the Euclidean case $E = \mathbb{R}^d$ and consider single-head masked self-attention $S^M_\theta : E^n \to E^n$ with the standard left-causal mask, logits*

$$L_{ij}(X) = x_i^\top B x_j, \qquad B := \frac{1}{\sqrt{d}} Q^\top K \in \mathcal{L}(E),$$

and a non-zero output projection $V \in \mathcal{L}(E)$.

Let $B(0,1) \subset E$ denote the open unit ball, and define

$$E^M_B := \Big\{ X = (x_1, \dots, x_n) \in B(0,1)^n : \forall i \in [n] \ \exists! \ j_i^\star \in \{1, \dots, i\} \ \text{s.t.} \ x_i^\top B x_{j_i^\star} = \max_{1 \le j \le i} x_i^\top B x_j \Big\}.$$

The complement of E^M_B in $B(0,1)^n$ has Lebesgue measure zero. Moreover, for any $X \in E^M_B$ there exists a function $\vartheta_X : [0, +\infty) \to [0, +\infty)$ such that $\vartheta_X(R) \to 1$ exponentially fast as $R \to +\infty$ and, for all $R > 0$,

$$\big\| J S^M_\theta(RX) \big\|_{\mathrm{op}} \le \vartheta_X(R) \|V\|_{\mathrm{op}} \sqrt{n}, \tag{6}$$

where $RX := (Rx_1, \dots, Rx_n)$.

Proof. Appendix A.2.6. □

2.3 Mean-Field Self-Attention on $\mathcal{P}_p(E)$: Wasserstein Regularity

Definition 2.9 (Mean-field self-attention). We work on $\mathcal{P}_p(E)$ (default $p = 2$). Queries/keys/values:

$$q(x) := Qx, \quad k(y) := Ky, \quad v(y) := Vy, \quad x, y \in E.$$

Unnormalized kernel (SDPA-type):

$$\begin{aligned} \kappa_\theta(x,y) &:= \exp\big(\beta \langle q(x), k(y) \rangle\big), \quad \beta > 0, \\ \kappa^M_\theta(x,y) &:= \exp\big(\beta \langle q(x), k(y) \rangle + M(x,y)\big) \quad \text{(masked variant)}. \end{aligned}$$

Partition function (for a fixed $\mu \in \mathcal{P}_p(E)$):

$$\begin{aligned} Z_{\theta,\mu}(x) &:= \int_E \kappa_\theta(x,y)\mu(dy), \\ Z^M_{\theta,\mu}(x) &:= \int_E \kappa^M_\theta(x,y)\mu(dy), \end{aligned}$$

(assume $0 < Z_{\theta,\mu}(x), Z^M_{\theta,\mu}(x) < \infty$ for the μ under consideration).

Attention kernel (probability kernel on E):

$$\begin{aligned} a_{\theta,\mu}(x,dy) &:= \frac{\kappa_\theta(x,y)}{Z_{\theta,\mu}(x)} \mu(dy), \\ a^M_{\theta,\mu}(x,dy) &:= \frac{\kappa^M_\theta(x,y)}{Z^M_{\theta,\mu}(x)} \mu(dy). \end{aligned}$$

Barycentric map (single-head mean-field SA update):

$$\begin{aligned} \Gamma_{\theta,\mu}(x) &:= \int_E v(y) a_{\theta,\mu}(x,dy), \\ \Gamma^M_{\theta,\mu}(x) &:= \int_E v(y) a^M_{\theta,\mu}(x,dy). \end{aligned}$$

Mean-field self-attention map:

$$F_\theta : \mathcal{P}_p(E) \to \mathcal{P}_p(E),$$
$$F_\theta(\mu) := (\Gamma_{\theta,\mu})_\# \mu \quad \text{(unmasked)},$$
$$F_\theta^M(\mu) := (\Gamma^M_{\theta,\mu})_\# \mu \quad \text{(masked)}.$$

Multi-head mean-field SA:

$$F_{\theta_h}(\mu) := (\Gamma_{\theta_h,\mu})_\# \mu, \quad h = 1, \dots, H,$$
$$F_\theta^{\mathrm{MH}}(\mu) := \left(\Gamma^{\mathrm{MH}}_{\theta,\mu}\right)_\# \mu,$$

where $\Gamma^{\mathrm{MH}}_{\theta,\mu}$ stacks the H heads and applies O (as in the discrete case). Precise formula omitted here; we treat it via single-head bounds and $\|O\|_{\mathrm{op}}$ when needed.

Remark (Discrete–mean-field consistency on empirical measures)**.** *For any $X = (x_1, \dots, x_n) \in E^n$ with $\mu_X = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$,*

$$F_\theta(\mu_X) = \mu_{S_\theta(X)},$$
$$F_\theta^M(\mu_X) = \mu_{S_\theta^M(X)},$$
$$F_\theta^{\mathrm{MH}}(\mu_X) = \mu_{S_\theta^{\mathrm{MH}}(X)}.$$

Definition 2.10 (Wasserstein regularity quantities)**.** Fix $p \geq 1$ (default $p = 2$). The local W_p-Lipschitz constant at $\mu \in \mathcal{P}_p(E)$ is

$$\mathrm{Lip}^{W_p}_{\mathrm{loc}}(F_\theta; \mu) := \limsup_{\nu \to \mu} \frac{W_p\big(F_\theta(\mu), F_\theta(\nu)\big)}{W_p(\mu, \nu)}.$$

For $R > 0$, define the radius-R class

$$\mathcal{P}_{p,R}(E) := \{\mu \in \mathcal{P}_p(E) : \mathrm{supp}(\mu) \subset B(0, R)\}.$$

Uniform local Lipschitz bound on $\mathcal{P}_{p,R}(E)$:

$$\mathrm{Lip}_{\mathcal{P}_{p,R}(E)}(F_\theta) := \sup_{\mu \in \mathcal{P}_{p,R}(E)} \mathrm{Lip}^{W_p}_{\mathrm{loc}}(F_\theta; \mu).$$

Analogous definitions apply to F_θ^M and F_θ^{MH}.

Remark (Local versus global W_p-Lipschitz constants)**.** *By definition,*

$$\sup_{\mu \in A} \mathrm{Lip}^{W_p}_{\mathrm{loc}}(F_\theta; \mu) \leq \sup_{\substack{\mu \neq \nu \\ \mu, \nu \in A}} \frac{W_p\big(F_\theta(\mu), F_\theta(\nu)\big)}{W_p(\mu, \nu)} \quad \textit{for any } A \subset \mathcal{P}_p(E).$$

If A is a geodesic (length) subset of $(\mathcal{P}_p(E), W_p)$ and F_θ is locally W_p-Lipschitz on A, then the reverse inequality also holds, hence the two notions coincide. In particular, this applies to $A = \mathcal{P}_{p,R}(E)$ in the Euclidean setting (and more generally on W_p-geodesic spaces where A is geodesically convex). Analogous remarks apply to F_θ^M and F_θ^{MH}.

Lemma 2.11 (Matrix–Measure Bridge ($p = 2$) [4, Lemma 3.3])**.** *Let $E = \mathbb{R}^d$ with its Euclidean norm. Let*

$$S_\theta : E^n \to E^n, \quad F_\theta : \mathcal{P}_2(E) \to \mathcal{P}_2(E)$$

denote respectively discrete self-attention and its mean-field counterpart. Assume consistency on empirical measures: for all $X \in E^n$,

$$F_\theta(\mu_X) = \mu_{S_\theta(X)}, \quad \mu_X := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}.$$

Let $\mathcal{M}_n(E) := \{\mu_X : X \in E^n\} \subset \mathcal{P}_2(E)$ be the set of all n-point empirical measures. Define the empirical-measure local Lipschitz constant as

$$\mathrm{Lip}^{W_2}_{\mathrm{loc},\mathcal{M}_n(E)}(F_\theta;\mu) := \limsup_{\substack{\nu\to\mu \\ \nu\in\mathcal{M}_n(E)}} \frac{W_2(F_\theta(\mu),F_\theta(\nu))}{W_2(\mu,\nu)}.$$

Then for every $X \in E^n$,

$$\mathrm{Lip}_{\mathrm{loc}}(S_\theta;X) = \mathrm{Lip}^{W_2}_{\mathrm{loc},\mathcal{M}_n(E)}(F_\theta;\mu_X) \le \mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_X),$$

where $\mathrm{Lip}_{\mathrm{loc}}(S_\theta;X) = \|JS_\theta(X)\|_{\mathrm{op}}$ is computed with respect to the Frobenius norm on E^n.

Proof. Appendix A.2.7. The same argument extends to any $p \ge 1$ by replacing W_2 with W_p and the Frobenius norm by the corresponding ℓ_p-product norm on E^n. ☐

Theorem 2.12 (Mean-field upper bound [3, Theorem 3.5])**.** *Assume $E = \mathbb{R}^d$ and $p = 2$. Consider the (unmasked) mean-field self-attention map $F_\theta : \mathcal{P}_2(E) \to \mathcal{P}_2(E)$ with parameters $\theta = (B, V)$, where $B, V \in \mathcal{L}(E)$ and the kernel is*

$$\kappa_\theta(x,y) = \exp(\langle x, By\rangle), \quad B = \beta Q^\top K.$$

Then F_θ is W_2-Lipschitz on $\mathcal{P}_{2,R}(E)$, and its Lipschitz constant satisfies

$$\mathrm{Lip}_{\mathcal{P}_{2,R}(E)}(F_\theta) \le \|V\|_{\mathrm{op}}\big(1 + 3\|B\|_{\mathrm{op}}R^2\big)\exp\big(2\|B\|_{\mathrm{op}}R^2\big). \tag{7}$$

Proof. Appendix A.2.8. ☐

Theorem 2.13 (Exponential lower bound [4, Theorem 3.4])**.** *Let $p \ge 1$ and $R > 0$. Consider the (unmasked) mean-field self-attention map $F_\theta : \mathcal{P}_p(E) \to \mathcal{P}_p(E)$ where $F_\theta(\mu) = (\Gamma_{\theta,\mu})_\#\mu$. Then F_θ is W_p-Lipschitz on $\mathcal{P}_{p,R}(E)$.*

Assume in addition that $E = \mathbb{R}^d$, $p = 2$, $V = \mathrm{Id}_E$ and $B = \beta Q^\top K \in \mathcal{L}(E)$ is symmetric. Let $\gamma_1 \ge \cdots \ge \gamma_d$ denote the ordered eigenvalues of B then the following claim holds:

1. *If $\gamma_1 \ge -8\gamma_d$, there exists a function $\vartheta : [0,+\infty) \to [0,+\infty)$ with $\vartheta(R) \to 1$ as $R \to +\infty$ such that*

$$\mathrm{Lip}_{\mathcal{P}_{2,R}(E)}(F_\theta) \ge \vartheta(R)\frac{\gamma_1}{16}R^2 e^{\gamma_1 R^2/8}.$$

2. *If $\gamma_1 < -8\gamma_d$, there exists a function $\vartheta : [0,+\infty) \to [0,+\infty)$ with $\vartheta(R) \to 1$ as $R \to +\infty$ such that*

$$\mathrm{Lip}_{\mathcal{P}_{2,R}(E)}(F_\theta) \ge \vartheta(R)\frac{|\gamma_d|}{2}R^2 e^{|\gamma_d|R^2}.$$

Moreover, in both cases the right-hand side is asymptotic, as $R \to +\infty$, to the local W_2-Lipschitz constant of F_θ at a suitably chosen two-Dirac measure $\mu_R = p_R\delta_{x_R} + (1-p_R)\delta_{y_R}$ supported in $B(0,R)$.

Proof. Appendix A.2.9. ☐

Proposition 2.14 (Worst-case two-Dirac measures [4, Proposition 3.5])**.** *Assume $E = \mathbb{R}^d$, $p = 2$, $V = \mathrm{Id}_E$, and consider the (unmasked) mean-field self-attention map $F_\theta : \mathcal{P}_2(E) \to \mathcal{P}_2(E)$ defined by*

$$F_\theta(\mu) = (\Gamma_{\theta,\mu})_\#\mu,$$

with linear parameters $\theta = (Q, K, V, \beta)$ and $B = \beta Q^\top K \in \mathcal{L}(E)$. Assume that B is symmetric. Denote by $\gamma_1 \ge \cdots \ge \gamma_d$ the eigenvalues of B and by $u_1, \ldots, u_d$ an orthonormal eigenbasis.

Let $\mathcal{P}_{2\delta}(E)$ be the set of two-Dirac probability measures on E.

1. *If $\gamma_1 \ge -8\gamma_d$, set $C := \gamma_1/8$ and for each $R > 0$ define*

$$p_R := e^{-2CR^2}, \quad \mu_R := p_R\delta_{Ru_1} + (1-p_R)\delta_{(R/2)u_1}.$$

Then $\mu_R \in \mathcal{P}_{2\delta}(E)$, and there exists a function $\vartheta_1 : [0,+\infty) \to (0,+\infty)$ with $\vartheta_1(R) \to 1$ as $R \to +\infty$ such that

$$\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R) \;\ge\; \vartheta_1(R)\frac{C}{2}R^2 e^{CR^2},$$

and moreover the right-hand side is asymptotic (as $R \to +\infty$) to $\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R)$.

2. *If $\gamma_1 < -8\gamma_d$, set $C' := |\gamma_d|$ and for each $R > 0$ define*

$$p_R := e^{-2C'R^2}, \quad \mu_R := p_R\delta_{Ru_d} + (1-p_R)\delta_{-Ru_d}.$$

Then $\mu_R \in \mathcal{P}_{2\delta}(E)$, and there exists a function $\vartheta_2 : [0,+\infty) \to (0,+\infty)$ with $\vartheta_2(R) \to 1$ as $R \to +\infty$ such that

$$\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R) \;\geq\; \vartheta_2(R)\frac{C'}{2}R^2 e^{C'R^2},$$

and moreover the right-hand side is asymptotic (as $R \to +\infty$) to $\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R)$.

Proof. Appendix A.2.10. □

Proposition 2.15 (Mean-field edge regime [3, Proposition 3.6])**.** *Let $E = \mathbb{R}^d$ and consider the single-head self-attention map $S_\theta : E^n \to E^n$ with temperature $\beta > 0$. Assume $V = \mathrm{Id}_E$. Denote by $\gamma_1 \geq \cdots \geq \gamma_r$ the real eigenvalues of B (assume $r \geq 1$). Set*

$$\gamma := \max\{-\gamma_r,\ \gamma_1/8\}.$$

Let $n = n(R)$ depend on R and satisfy

$$n(R) \sim_{R\to+\infty} \exp\big(2\gamma R^2\big).$$

Then there exists a function $\vartheta : [0,+\infty) \to [0,+\infty)$ with $\vartheta(R) \to 1$ as $R \to +\infty$ such that

$$\mathrm{Lip}_{B^n_R}(S_\theta) \geq \vartheta(R)\frac{\gamma}{2}R^2 e^{\gamma R^2}.$$

Proof. Appendix A.2.11. □

Theorem 2.16 (Masked Attention Mean-Field Bound [3, Theorem 4.4])**.** *Let $E = \mathbb{R}^d$, fix $p \geq 1$ and $R > 0$, and fix parameters $\theta = (Q,K,V,\beta)$ with $Q,K,V \in \mathbb{R}^{d\times d}$ and $\beta > 0$. Recall*

$$B = \beta Q^\top K \in \mathbb{R}^{d\times d} \quad \text{so that} \quad \beta\langle Qx, Ky\rangle = \langle x, By\rangle.$$

Let $\mathrm{pr}_1 : [0,1]\times E \to [0,1]$ denote the projection $\mathrm{pr}_1(s,x) = s$. For $\bar\mu,\bar\nu \in \mathcal{P}_p([0,1]\times E)$, set $d_p(\bar\mu,\bar\nu) = +\infty$ if $(\mathrm{pr}_1)_\#\bar\mu \neq (\mathrm{pr}_1)_\#\bar\nu$. Otherwise, letting $\vartheta := (\mathrm{pr}_1)_\#\bar\mu = (\mathrm{pr}_1)_\#\bar\nu$ and disintegrating

$$\bar\mu(ds,dx) = \vartheta(ds)\mu^s(dx), \quad \bar\nu(ds,dx) = \vartheta(ds)\nu^s(dx),$$

define

$$d_p(\bar\mu,\bar\nu) := \Big(\int_0^1 W_p(\mu^s,\nu^s)^p \vartheta(ds)\Big)^{1/p},$$

where W_p is the Wasserstein distance on E induced by the cost $\|x-y\|^p$.

Let $M : ([0,1]\times E)^2 \to \mathbb{R}\cup\{-\infty\}$ be the causal mask

$$M((s,x),(t,y)) := \begin{cases} 0, & t \leq s, \\ -\infty, & t > s. \end{cases}$$

Define, for $\bar\mu \in \mathcal{P}_p([0,1]\times E)$, the masked kernel and partition function

$$\kappa^M_\theta\big((s,x),(t,y)\big) := \exp\big(\langle x, By\rangle + M((s,x),(t,y))\big) = \exp(\langle x, By\rangle)\mathbf{1}_{\{t\leq s\}},$$

$$Z^M_{\theta,\bar\mu}(s,x) := \int_{[0,1]\times E} \kappa^M_\theta\big((s,x),(t,y)\big)\bar\mu(dt,dy).$$

For measures supported in $[0,1]\times B(0,R)$, one has $0 < Z^M_{\theta,\bar\mu}(s,x) < \infty$ for $\bar\mu$-a.e. (s,x), hence the kernel below is well-defined $\bar\mu$-a.e. Define the attention kernel and the update map (time unchanged):

$$a^M_{\theta,\bar\mu}\big((s,x),(dt,dy)\big) := \frac{\kappa^M_\theta\big((s,x),(t,y)\big)}{Z^M_{\theta,\bar\mu}(s,x)}\bar\mu(dt,dy),$$

$$\bar{\Gamma}^M_{\theta,\bar{\mu}}(s,x) := \Big(s,\ \int_{[0,1]\times E} Vy a^M_{\theta,\bar{\mu}}\big((s,x),(dt,dy)\big)\Big), \quad F^M_\theta(\bar{\mu}) := (\bar{\Gamma}^M_{\theta,\bar{\mu}})_\# \bar{\mu}.$$

Let

$$\mathcal{P}_{p,R}([0,1]\times E) := \{\bar{\mu} \in \mathcal{P}_p([0,1]\times E) : \operatorname{supp}(\bar{\mu}) \subset [0,1]\times B(0,R)\}.$$

Then F^M_θ is Lipschitz on $\mathcal{P}_{p,R}([0,1]\times E)$ with respect to d_p, with

$$\operatorname{Lip}^{d_p}_{\mathcal{P}_{p,R}([0,1]\times E)}(F^M_\theta) := \sup_{\substack{\bar{\mu}\neq\bar{\nu} \\ \bar{\mu},\bar{\nu}\in\mathcal{P}_{p,R}([0,1]\times E)}} \frac{d_p(F^M_\theta(\bar{\mu}), F^M_\theta(\bar{\nu}))}{d_p(\bar{\mu},\bar{\nu})} \ \le\ \|V\|_{\mathrm{op}}\Big(1+3\|B\|_{\mathrm{op}}R^2\Big)e^{2\|B\|_{\mathrm{op}}R^2}.$$

Assume in addition that $p=2$ for the remainder of this statement. For $n\in\mathbb{N}$, define the discrete causal mask on indices by $M_{ij}=0$ if $j\le i$ and $M_{ij}=-\infty$ otherwise, and let $S^M_\theta : E^n \to E^n$ be the masked self-attention map obtained by replacing $L_{ij}(X)$ with $L_{ij}(X)+M_{ij}$ in the softmax. For $X=(x_1,\ldots,x_n)\in E^n$, define the time-stamped empirical measure

$$\bar{\mu}_X := \frac{1}{n}\sum_{i=1}^n \delta_{(i/n,\ x_i)} \ \in\ \mathcal{P}_{2,R}([0,1]\times E).$$

Then $F^M_\theta(\bar{\mu}_X) = \bar{\mu}_{S^M_\theta(X)}$, and in particular

$$\operatorname{Lip}_{B^n_R}(S^M_\theta) \ \le\ \operatorname{Lip}^{d_2}_{\mathcal{P}_{2,R}([0,1]\times E)}(F^M_\theta).$$

2.4 Variational and OT Formulations of SDPA

In this subsection, parameters θ are fixed and we use the shorthand

$$\kappa_\mu := \kappa_{\theta,\mu}, \quad Z_\mu := Z_{\theta,\mu}, \quad a_\mu := a_{\theta,\mu}, \quad \Gamma_\mu := \Gamma_{\theta,\mu}.$$

Definition 2.17 (One-query variational problem (row-wise SDPA))**.** Fix $\mu\in\mathcal{P}(E)$ (key distribution) and $x\in E$ (query). We consider probability measures $\alpha(\cdot|x)\in\mathcal{P}(E)$ ("attention weights for query x"). Use the same q,k,v and a cost

$$c(x,y) := -\langle q(x), k(y)\rangle \quad \text{or a variant (e.g. with scaling).}$$

For $\varepsilon>0$ (entropic parameter, related to β), define the functional

$$\mathcal{J}_\varepsilon(\alpha(\cdot|x);x,\mu) := \int_E c(x,y)\alpha(dy|x) + \varepsilon\int_E \log\left(\frac{d\alpha(\cdot|x)}{d\mu}(y)\right)\alpha(dy|x),$$

whenever $\alpha(\cdot|x)\ll\mu$ (and $+\infty$ otherwise). Variational problem (one-sided EOT for a single query):

$$\alpha_\varepsilon(\cdot|x) \in \arg\min_{\alpha(\cdot|x)\in\mathcal{P}(E)} \mathcal{J}_\varepsilon(\alpha(\cdot|x);x,\mu).$$

By formal first-order conditions (Gibbs form),

$$\alpha_\varepsilon(dy|x) \propto \exp\Big(-\tfrac{1}{\varepsilon}c(x,y)\Big)d\mu(y).$$

For $c(x,y)=-\langle q(x),k(y)\rangle$ and $\varepsilon=1/\beta$, this coincides with the SDPA kernel from I.3:

$$\alpha_\varepsilon(dy|x) = a_\mu(x,dy) = \frac{\kappa_\mu(x,y)}{Z_\mu(x)}d\mu(y).$$

Definition 2.18 (Mean-field SDPA as barycentric projection)**.** The mean-field SA map can be rewritten as a barycentric projection of the one-sided EOT plans:

$$\begin{aligned}\Gamma_\mu(x) &= \int_E v(y)a_\mu(x,dy)\\ &= \int_E v(y)\alpha_\varepsilon(dy|x),\end{aligned}$$

and

$$F_\theta(\mu) = (\Gamma_\mu)_\# \mu = (x \mapsto \Gamma_\mu(x))_\# \mu.$$

Masked variant:

$$\begin{aligned} c_M(x,y) &:= c(x,y) - M(x,y), \\ \kappa_\mu^M(x,y) &:= \exp\big(-\tfrac{1}{\varepsilon} c_M(x,y)\big), \\ a_\mu^M(x,dy) &\propto \kappa_\mu^M(x,y) d\mu(y), \\ \Gamma_\mu^M(x) &:= \int_E v(y) a_\mu^M(x,dy), \\ F_\theta^M(\mu) &:= (\Gamma_\mu^M)_\# \mu. \end{aligned}$$

Definition 2.19 (Global SDPA as a one-sided EOT problem on $\mathcal{P}(E)$)**.** Heuristically, for a query distribution $\nu \in \mathcal{P}(E)$ (e.g. empirical queries $\nu = \mu_X$) and key distribution μ:

$$\Pi(\nu,\mu) := \{\pi \in \mathcal{P}(E \times E) : (\mathrm{pr}_1)_\# \pi = \nu, (\mathrm{pr}_2)_\# \pi = \mu\},$$

and a reference coupling $\nu \otimes \mu$. Entropic one-sided OT functional:

$$\mathcal{J}_\varepsilon(\pi;\nu,\mu) := \int_{E\times E} c(x,y) d\pi(x,y) + \varepsilon \int_{E\times E} \log\left(\frac{d\pi}{d(\nu\otimes\mu)}(x,y)\right) d\pi(x,y),$$

with constraint $(\mathrm{pr}_2)_\# \pi = \mu$ fixed (one-sided). Minimizer π_ε has disintegration

$$\pi_\varepsilon(dx,dy) = \nu(dx)\alpha_\varepsilon(dy|x).$$

The SDPA mean-field update $F_\theta(\mu)$ corresponds to the barycentric projection of the second marginal through v:

$$\begin{aligned} \Gamma_\mu(x) &= \int_E v(y)\alpha_\varepsilon(dy|x), \\ F_\theta(\nu) &= (\Gamma_\mu)_\# \nu. \end{aligned}$$

(For the discrete case, take $\nu = \mu_X = \frac{1}{n}\sum_i \delta_{x_i}$ and $\mu = \mu_X$ or a separate key distribution, and recover the usual SDPA matrix.)

Forward Pass: SDPA as Entropic Optimal Transport (EOT)

Definition 2.20 (The One-Sided EOT Problem [7, Definition 3.2])**.** Given scores $s_j = \langle q, k_j\rangle$ and costs $C_j := -s_j$, define $\Delta_{m-1} := \{p \in \mathbb{R}^m_{\geq 0} : \sum_j p_j = 1\}$. For $\varepsilon > 0$, the one-sided entropic OT problem is

$$p^* \in \arg\min_{p\in\Delta_{m-1}} \Big(\sum_{j=1}^m p_j C_j - \varepsilon H(p)\Big), \qquad H(p) := -\sum_{j=1}^m p_j \log p_j.$$

Equivalently,

$$p^* \in \arg\min_{p\in\Delta_{m-1}} \Big(-\langle p, s\rangle + \varepsilon \sum_{j=1}^m p_j \log p_j\Big).$$

Proof. Appendix A.5.1. ☐

Theorem 2.21 (Attention as the Unique EOT Solution [7, Theorem 3.3])**.** *For the problem in Definition 2.20, if $\varepsilon = \tau > 0$, there is a unique minimizer and it is exactly scaled-dot-product attention:*

$$p_j^* = \frac{\exp(s_j/\tau)}{\sum_{\ell=1}^m \exp(s_\ell/\tau)} =: \alpha_j.$$

Hence SDPA is the primal optimizer of one-sided entropy-regularized OT.

Proof. Appendix A.5.2. ☐

Remark (Interpretation of τ [7, Remark 3.4])**.** *In the EOT view, τ is the entropy-regularization strength: small τ favors concentrated plans, large τ favors diffuse plans. The Transformer choice $\tau = \sqrt{d_k}$ can be read as variance normalization of query-key dot products and thus as a dimension-aware regularization schedule.*

Proof. Appendix A.5.3. ☐

Proposition 2.22 (Unifying Attention Mechanisms [7, Proposition 4.1])**.** *In the variational family*

$$p^* \in \arg\min_{p \in \Delta_{m-1}} \{-\langle p, s\rangle + \Omega(p)\},$$

choosing $\Omega(p) = -\tau H(p)$ yields uniquely

$$p_j^* \propto e^{s_j/\tau},$$

i.e. softmax attention.

Proof. Appendix A.5.4. ☐

Proposition 2.23 (Unifying Attention Mechanisms [7, Proposition 4.2])**.** *Choosing the quadratic regularizer*

$$\Omega(p) = \frac{1}{2}\|p\|_2^2$$

in the same variational program yields Sparsemax, i.e. projection of s onto the simplex:

$$p_j^* = (s_j - \theta)_+, \qquad \sum_j (s_j - \theta)_+ = 1.$$

Proof. Appendix A.5.5. ☐

Proposition 2.24 (Unifying Attention Mechanisms [7, Proposition 4.3])**.** *For $\alpha > 1$, choose Tsallis regularization*

$$\Omega(p) = \frac{1}{\alpha(\alpha-1)} \sum_j (p_j^\alpha - p_j).$$

Then the optimizer is the α-entmax family:

$$p_j^* = \Big[(\alpha-1)s_j - \theta\Big]_+^{\frac{1}{\alpha-1}} \Big/ \sum_\ell \Big[(\alpha-1)s_\ell - \theta\Big]_+^{\frac{1}{\alpha-1}},$$

with threshold θ set by normalization.

Proof. Appendix A.5.6. ☐

Proposition 2.25 (ALiBi as Locality-Biased EOT [7, Proposition 4.4])**.** *For query index i, augment the regularizer by a linear locality cost:*

$$\Omega_L(p_i) = \tau \sum_j p_{ij} \log p_{ij} + \gamma \sum_j p_{ij}|i-j|, \qquad \gamma \geq 0.$$

Then

$$p_{ij}^* = \frac{\exp((s_{ij} - \gamma|i-j|)/\tau)}{\sum_\ell \exp((s_{i\ell} - \gamma|i-\ell|)/\tau)},$$

which is exactly ALiBi-style linear logit biasing.

Proof. Appendix A.5.7. ☐

Proposition 2.26 (PriorSoftmax (Bayesian Generalization) [7, Proposition 4.5])**.** *Let $\pi \in \Delta_{m-1}$ with $\pi_j > 0$. Consider*

$$p^* \in \arg\min_{p \in \Delta_{m-1}} \{-\langle p, s\rangle + \tau \operatorname{KL}(p\|\pi)\}.$$

The unique solution is

$$p_j^* = \frac{\pi_j e^{s_j/\tau}}{\sum_\ell \pi_\ell e^{s_\ell/\tau}}.$$

Thus standard softmax is recovered at uniform prior π.

Proof. Appendix A.5.8. □

Backward Pass: Optimal Control Signal

Theorem 2.27 (Gradient as Advantage-Based Policy Update [7, Theorem 5.2])**.** *Let $p^* = \alpha(s)$ be softmax attention with temperature τ. For a downstream loss L and marginal utilities $u_j := -\partial L/\partial p_j$,*

$$\frac{\partial L}{\partial s_j} = -\frac{p_j^*}{\tau}\Big(u_j - \mathbb{E}_{p^*}[u]\Big), \qquad \mathbb{E}_{p^*}[u] = \sum_k p_k^* u_k.$$

So score updates are weighted advantage estimates.

Proof. Appendix A.5.9. □

The Dual Problem and Information Geometry

Theorem 2.28 (The Log-Sum-Exp (LSE) Potential [7, Theorem 6.3])**.** *The dual optimal value of the one-sided EOT problem defines*

$$\phi^*(s) = \tau \log\Big(\sum_{\ell=1}^m e^{s_\ell/\tau}\Big).$$

Its gradient equals attention:

$$\nabla_s \phi^*(s) = p^*(s).$$

Proof. Appendix A.5.10. □

Theorem 2.29 (Hessian is the Fisher Information Matrix (FIM) [7, Theorem 7.3])**.** *Let $F(s)$ be the Fisher information matrix of the attention family $p(s) = \operatorname{softmax}(s/\tau)$ and let $u_j = -\partial L/\partial p_j$. Then the Euclidean score-gradient satisfies*

$$\nabla_s L = -\tau F(s) u.$$

Hence the standard gradient is the Fisher-metric image of the utility direction, exhibiting duality with natural-gradient geometry.

Proof. Appendix A.5.11. □

Corollary 2.30 (Hessian is the FIM [7, Theorem 7.5])**.** *Let ϕ^* be the dual potential from Theorem 2.28. Then*

$$\nabla_s^2 \phi^*(s) = \tau F(s).$$

Equivalently, the dual curvature of one-sided EOT is proportional to the attention Fisher information.

Proof. Appendix A.5.12. □

3 Self-Attention Dynamics

3.1 Objects and Assumptions

Ambient space, time, and static ingredients.

$$\begin{aligned}
&E \subset \mathbb{R}^d, \quad \text{in this setting typically } E = \mathbb{S}^{d-1},\\
&X(t) = (x_1(t), \ldots, x_n(t)) \in E^n,\\
&t \in [0, T] \text{ or } t \in [0, \infty),\\
&q(x) := Qx, \quad k(x) := Kx, \quad v(x) := Vx,\\
&S_\theta : E^n \to E^n \quad \text{(static discrete SA)},\\
&F_\theta : \mathcal{P}(E) \to \mathcal{P}(E) \quad \text{(mean-field SA)}.
\end{aligned}$$

Empirical law.

$$\mu_t := \mu_{X(t)} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i(t)} \in \mathcal{P}(E).$$

Self-attention vector field on E^n. We denote by

$$G_\theta^{\mathrm{SA}} : E^n \to T(E^n)$$

a self-attention–induced vector field on configurations, of the form

$$G_\theta^{\mathrm{SA}}(X) = \big(g_\theta(x_i; X)\big)_{i=1}^{n},$$

where for each $X = (x_1, \ldots, x_n)$,

$$g_\theta(\cdot; X) : E \to TE$$

is a vector field built from the same q, k, v and attention weights (e.g. via $S_\theta(X)$, attention kernels, gradients of an energy, etc.).

In the spherical case $E = \mathbb{S}^{d-1}$ we implicitly require

$$g_\theta(x_i; X) \in T_{x_i}\mathbb{S}^{d-1}, \quad \text{equivalently } \langle x_i, g_\theta(x_i; X)\rangle = 0,$$

often implemented via the projection P_{x_i} from 1.

(Concrete formulas for g_θ in the various SA dynamic models will be given; here we only fix the notation G_θ^{SA}.)

Particle system (autonomous SA dynamics). Self-attention dynamics on E^n:

$$\begin{aligned}
&\dot{X}(t) = G_\theta^{\mathrm{SA}}(X(t)), \quad t \geq 0,\\
&X(0) = X^0 \in E^n.
\end{aligned}$$

Component-wise:

$$\dot{x}_i(t) = g_\theta\big(x_i(t); X(t)\big), \quad i \in [n].$$

We view this as an autonomous ODE on E^n:

$$\dot{X}(t) = G_\theta^{\mathrm{SA}}(X(t)).$$

Mean-field velocity and continuity equation (SA-only). Associated mean-field velocity functional

$$v_\theta^{\mathrm{SA}}[\mu] : E \to TE$$

is understood as the large-n limit of $g_\theta(\cdot; X)$ when $\mu_X \to \mu$:

$$v_\theta^{\mathrm{SA}}[\mu](x) \approx g_\theta(x; X) \quad \text{whenever } \mu_X \approx \mu.$$

Mean-field (formal) SA continuity equation:

$$\partial_t \mu_t + \nabla \cdot \big(v_\theta^{\mathrm{SA}}[\mu_t]\mu_t\big) = 0.$$

Standing assumptions (SA dynamics).

(D1) *Bounded configurations / support.* There exists $R > 0$ such that along the flow,

$$\|x_i(t)\| \le R \quad \forall i \in [n],\ \forall t \in [0, T],$$

equivalently $\mathrm{supp}(\mu_t) \subset B(0, R)$ (or $\subset \mathbb{S}^{d-1}$).

(D2) *Regularity of G_θ^{SA}.* For any $R > 0$, there exists $L_G(R) < \infty$ such that

$$\|G_\theta^{\mathrm{SA}}(X) - G_\theta^{\mathrm{SA}}(X')\|_F \le L_G(R)\|X - X'\|_F$$

for all $X, X' \in E^n$ with $\|x_i\|, \|x_i'\| \le R$. (Local Lipschitz on bounded sets.)

(D3) *Tangent structure (spherical case).* If $E = \mathbb{S}^{d-1}$ and $X^0 \in (\mathbb{S}^{d-1})^n$, then

$$x_i(t) \in \mathbb{S}^{d-1} \quad \forall i, t,$$

i.e. G_θ^{SA} is tangent: $g_\theta(x_i; X) \in T_{x_i}\mathbb{S}^{d-1}$.

(D4) *Mean-field well-posedness (formal).* Under appropriate conditions on $v_\theta^{\mathrm{SA}}[\cdot]$, the continuity equation has (at least) a weak solution $(\mu_t)_{t\ge 0}$ in $\mathcal{P}(E)$ for any initial $\mu_0 \in \mathcal{P}(E)$.

3.2 Discrete Self-Attention Dynamics on E^n

Definition 3.1 (SA-Induced Drift via S_θ)**.** We start from the static single-head SA map

$$S_\theta : E^n \to E^n, \quad X \mapsto S_\theta(X)$$

or its masked / multi-head variants. Define the (Euclidean) SA drift

$$\begin{aligned} B_\theta(X) &:= S_\theta(X) - X \in E^n, \\ B_\theta(X) &= \big(b_\theta(x_i; X)\big)_{i=1}^n, \end{aligned}$$

with

$$b_\theta(x_i; X) := S_\theta(X)_i - x_i \in E.$$

In the spherical case, $E = \mathbb{S}^{d-1}$ we consider the tangent-projected drift

$$B_\theta^{\mathrm{sph}}(X) := \big(P_{x_i} b_\theta(x_i; X)\big)_{i=1}^n.$$

(In concrete models, B_θ or B_θ^{sph} is replaced by the exact vector field given in the paper; this reference form keeps the link to S_θ explicit.)

Definition 3.2 (SA-Only ODE on E^n)**.** Euclidean version:

$$\begin{aligned} \dot{X}(t) &= B_\theta\big(X(t)\big), \quad t \ge 0, \\ X(0) &= X^0 \in E^n. \end{aligned}$$

Spherical version:

$$\begin{aligned} \dot{X}(t) &= B_\theta^{\mathrm{sph}}\big(X(t)\big), \quad t \ge 0, \\ X(0) &= X^0 \in (\mathbb{S}^{d-1})^n. \end{aligned}$$

Component-wise (spherical case):

$$\dot{x}_i(t) = P_{x_i(t)}\big(S_\theta(X(t))_i - x_i(t)\big), \quad i \in [n].$$

(This ODE can be viewed as a continuous-depth limit of repeatedly applying $X \mapsto X + \Delta t B_\theta(X)$, i.e. an Euler discretization of the SA drift.)

Observation (Link to the generic vector field G_θ^{SA})**.** *With*

$$G_\theta^{\mathrm{SA}}(X) := B_\theta(X) \quad \textit{(Euclidean case)},$$
$$G_\theta^{\mathrm{SA}}(X) := B_\theta^{\mathrm{sph}}(X) \quad \textit{(spherical case)}.$$

Then the SA dynamics

$$\dot{X}(t) = G_\theta^{\mathrm{SA}}(X(t))$$

is a special case of the general framework

Remark (Discrete-time vs continuous-time (Euler link))**.** *For a time step $\Delta t > 0$, the explicit Euler scheme for the ODE $\dot{X}(t) = B_\theta(X(t))$ is*

$$\begin{aligned} X^{k+1} &= X^k + \Delta t B_\theta(X^k) \\ &= (1 - \Delta t)X^k + \Delta t S_\theta(X^k). \end{aligned}$$

For $\Delta t = 1$ this reduces to

$$X^{k+1} = S_\theta(X^k),$$

i.e. iterations of the static SA map. For $0 < \Delta t < 1$, we obtain a convex combination of the identity and S_θ. (In the spherical case, the Euler step is followed by a projection onto $(\mathbb{S}^{d-1})^n$ if needed.)

Theorem 3.3 (Asymptotic Low-Rankness ($d = 1$) [6, Theorem 2.1])**.** *Assume $d = 1$, $V > 0$, and $QK > 0$. For any initial sequence of pairwise distinct tokens $(x_1(0), \dots, x_n(0)) \in \mathbb{R}^n$, there exists a matrix P^* in the Boolean low-rank class described in [6] such that the self-attention matrix $P(t)$ converges:*

$$P(t) \longrightarrow P^* \qquad (t \to +\infty).$$

In particular, the limiting attention is typically rank 1 or 2 for generic data.

Proof. Appendix A.1.1. □

Theorem 3.4 (Polytope Clustering ($V = Id$) [6, Theorem 3.1])**.** *Assume $V = \mathrm{Id}$ and $Q^\top K \succ 0$. For any initial tokens $(z_i(0))_{i=1}^n \subset \mathbb{R}^d$ in the rescaled dynamics, there exists a convex polytope $\mathcal{K} \subset \mathbb{R}^d$ such that each trajectory $z_i(t)$ converges either to 0 or to a point on $\partial\mathcal{K}$ as $t \to +\infty$. Thus asymptotic cluster representatives lie on the boundary of a data-dependent convex polytope.*

Proof. Appendix A.1.2. □

Theorem 3.5 (Hyperplane Clustering (V simple positive leading eigenvalue) [6, Theorem 4.2])**.** *Let (Q, K, V) be a good triple in the sense of Definition 4.1 of [6]. Then for any initial sequence $(z_i(0))_{i=1}^n \subset \mathbb{R}^d$, there exist at most three parallel hyperplanes in $\mathbb{R}^d$ such that each $z_i(t)$ converges in distance to one of these hyperplanes as $t \to +\infty$.*

Proof. Appendix A.1.3. □

Theorem 3.6 (Mixed Clustering (V dominant eigenvalue with multiplicity) [6, Theorem 5.2])**.** *Assume (Q, K, V) is a good triple with multiplicity (Definition 5.1 in [6]), with splitting $\mathbb{R}^d = F \oplus G$ and dominant expanding action on F. Then there exists a bounded convex polytope $\mathcal{K} \subset F$ such that, with $\mathcal{H} := (\partial\mathcal{K} \cup \{0\}) \times G$,*

$$\mathrm{dist}(z_i(t), \mathcal{H}) \to 0 \qquad (t \to +\infty), \ \forall i.$$

So dynamics exhibit polytope clustering in F and subspace transport in G.

Proof. Appendix A.1.4. □

Proposition 3.7 ([6, Proposition 6.2])**.** *For every initial configuration $X_0 = (x_1^0, \dots, x_n^0) \in (\mathbb{R}^d)^n$, the self-attention ODE system admits a unique global Lipschitz trajectory $t \mapsto X(t) = (x_1(t), \dots, x_n(t))$ with $X(0) = X_0$.*

Proof. Appendix A.1.5. □

Proposition 3.8 ([6, Proposition 6.3])**.** *For every initial configuration $Z_0 = (z_1^0, \ldots, z_n^0) \in (\mathbb{R}^d)^n$, the rescaled system admits a unique global Lipschitz solution $t \mapsto Z(t) = (z_1(t), \ldots, z_n(t))$ with $Z(0) = Z_0$.*

Proof. Appendix A.1.6. $\square$

Theorem 3.9 (Collapse to Origin ($V = -\mathrm{Id}$) [6, Theorem 8.5])**.** *Assume $V = -\mathrm{Id}$ and $Q^\top K = \mathrm{Id}$. For any initial token family $(x_i(0))_{i=1}^n \subset \mathbb{R}^d$, the solution of the self-attention dynamics satisfies*

$$\|x_i(t)\| \to 0 \qquad (t \to +\infty),\ \forall i \in [n].$$

Hence all particles collapse to the origin and attention converges to the uniform row-stochastic matrix.

Proof. Appendix A.1.8. $\square$

Theorem 3.10 (Generic Consensus [5, Theorem 4.1 ($\beta = 0$)])**.** *Fix $d, n \geq 2$ and consider either normalized or unnormalized dynamics at $\beta = 0$ on S^{d-1}. For Lebesgue-a.e. initial configuration $(x_i(0))_{i=1}^n \in (S^{d-1})^n$, there exists $x_* \in S^{d-1}$ such that*

$$x_i(t) \to x_* \qquad (t \to +\infty),\ \forall i \in [n].$$

Proof. Appendix A.1.12. $\square$

Theorem 3.11 (Generic Consensus [5, Theorem 4.3 (small β)])**.** *There exists a universal constant $C > 0$ such that for*

$$\beta \leq C n^{-1},$$

for Lebesgue-a.e. initial condition in $(S^{d-1})^n$, both (SA) and (USA) converge to consensus:

$$\exists x_* \in S^{d-1} \text{ with } x_i(t) \to x_* \ \forall i.$$

In dimension $d = 2$, one can take the full range $\beta \leq 1$.

Proof. Appendix A.1.13. $\square$

Theorem 3.12 (Generic Consensus [5, Theorem 5.1 (large β)])**.** *For each $d \geq 2$ there exists $C(d) > 0$ such that whenever*

$$\beta \geq C(d) n^2,$$

for Lebesgue-a.e. initial condition in $(S^{d-1})^n$, both normalized and unnormalized dynamics converge to a single consensus cluster.

Proof. Appendix A.1.14. $\square$

Theorem 3.13 (General Consensus in High Dimension [5, Theorem 6.1 ($d \geq 3$)])**.** *Fix $n \geq 2$, $d \geq 3$, and $\beta \geq 0$. Then the consensus conclusion of Theorem 3.11 holds for both (SA) and (USA): for almost every initial condition, all particles converge to one common limit point on S^{d-1}.*

Proof. Appendix A.1.15. $\square$

Theorem 3.14 (Exponential Rate in High Dimension [5, Theorem 6.3])**.** *Let $n \geq 1$, $\beta > 0$, and $d \geq n$. If initial tokens are i.i.d. uniform on S^{d-1}, then almost surely there exist $x_* \in S^{d-1}$ and constants $C, \lambda > 0$ such that for all $i \in [n]$,*

$$\|x_i(t) - x_*\| \leq C e^{-\lambda t}, \qquad t \geq 0.$$

The same statement holds for general Q, K in the corresponding identity-V model.

Proof. Appendix A.1.16. $\square$

Theorem 3.15 (Quantitative Convergence [5, Theorem 6.9]). *Fix $\beta \geq 0$ and $n \geq 2$. There exists $d_*(n, \beta) \geq n$ such that for all $d \geq d_*(n, \beta)$: if $(x_i(0))_{i=1}^n$ are i.i.d. uniform on S^{d-1} and $x_i(t)$ solves (SA), then there exist constants $C(n, \beta), \lambda(n, \beta) > 0$ with probability at least $1 - 2n^2 d^{-1/64}$,*

$$\left|\langle x_i(t), x_j(t)\rangle - \gamma_\beta(t)\right| \leq \min\left\{2c(\beta)\frac{nt\sqrt{\log d}}{d},\ Ce^{-\lambda t}\right\},$$

for all $i \neq j$ and all $t \geq 0$, where γ_β solves the reduced scalar model.

Proof. Appendix A.1.17. □

Theorem 3.16 (Global Synchronization / Clustering [8, Theorem 1]). *For both (SA) and (USA) dynamics with $n \geq 2$, $d \geq 3$, and any $\beta \geq 0$, for almost every initial condition $(x_1(0), \ldots, x_n(0)) \in (S^{d-1})^n$, trajectories exist globally and satisfy*

$$\lim_{t\to\infty} \|x_i(t) - x_j(t)\| = 0, \qquad \forall i, j \in [n].$$

Equivalently, the empirical measure converges weakly to a Dirac mass on S^{d-1}.

Proof. Appendix A.6.1. □

Theorem 3.17 (Local Exponential Rates (Hemisphere Condition) [8, Theorem 3]). *Let $n \geq 1$, $d \geq 2$, $\beta > 0$, and assume initial tokens satisfy*

$$\exists w \in S^{d-1} \text{ such that } \langle x_i(0), w\rangle > 0 \quad \forall i.$$

Then for SA or USA, there exist $x_* \in S^{d-1}$ and constants $C, \lambda > 0$ (depending on initialization, n, β) such that

$$\|x_i(t) - x_*\| \leq Ce^{-\lambda t}, \qquad \forall i \in [n],\ t \geq 0.$$

Proof. Appendix A.6.2. □

Corollary 3.18 (Random Initialization Clustering [8, Corollary 4]). *If $x_1(0), \ldots, x_n(0)$ are i.i.d. uniform on S^{d-1} and $d \geq n$, then almost surely they lie in some open hemisphere. Consequently, Theorem 3.17 applies and yields exponential convergence to one cluster.*

Proof. Appendix A.6.3. □

Theorem 3.19 (Long-Context Phase Transition [8, Theorem 7]). *Assume equiangular inputs: $\|x_i\| = 1$ and $\langle x_i, x_j\rangle = \rho \in (0, 1)$ for $i \neq j$. Set attention scale $\beta_n = \gamma \log n$ and denote output directions by $\theta_i \in S^{d-1}$ after one attention layer. Then for $i \neq j$,*

$$\lim_{n\to\infty} \langle \theta_i, \theta_j\rangle = \begin{cases} 1, & \gamma < \frac{1}{1-\rho}, \\ \frac{4\rho}{1+3\rho}, & \gamma = \frac{1}{1-\rho}, \\ \rho, & \gamma > \frac{1}{1-\rho}. \end{cases}$$

Hence attention exhibits a sharp subcritical/critical/supercritical phase transition.

Proof. Appendix A.6.6. □

Definition 3.20 (Noisy Transformer SDE [8, Equation (8)]). Let $\kappa > 0$. The mean-field noisy attention limit is the McKean–Vlasov SDE on S^{d-1}:

$$dX(t) = P^\perp_{X(t)}\left(\int_{S^{d-1}} e^{\beta\langle X(t), y\rangle} y\, \mu_t(dy)\right) dt + \sqrt{2\kappa^{-1}}\, dW(t),$$

where $\mu_t = \mathcal{L}(X(t))$.

Proof. Appendix A.6.7. □

3.3 Mean-Field Self-Attention Dynamics: Continuity Equation and Energy

Definition 3.21 (Mean-field SA velocity via the barycentric map)**.** We reuse the mean-field SA objects:

$$
\begin{aligned}
&q(x) := Qx, \quad k(y) := Ky, \quad v(y) := Vy,\\
&\kappa_\mu(x,y) := \exp\big(\beta\langle q(x), k(y)\rangle\big),\\
&Z_\mu(x) := \int_E \kappa_\mu(x,y)d\mu(y),\\
&a_\mu(x,dy) := \frac{\kappa_\mu(x,y)}{Z_\mu(x)}d\mu(y),\\
&\Gamma_\mu(x) := \int_E v(y)a_\mu(x,dy).
\end{aligned}
$$

Prototype mean-field SA velocity (Euclidean):

$$v_\theta^{\mathrm{SA}}[\mu](x) := \Gamma_\mu(x) - x.$$

Spherical version ($E = \mathbb{S}^{d-1}$):

$$v_{\theta,\mathrm{sph}}^{\mathrm{SA}}[\mu](x) := P_x\big(\Gamma_\mu(x) - x\big) \in T_x\mathbb{S}^{d-1}.$$

(Concrete models may modify $v_\theta^{\mathrm{SA}}[\mu]$ by adding damping factors, normalization constants, or extra terms; those will be specified below on a case-by-case basis.)

Definition 3.22 (Mean-field SA continuity equation)**.** Mean-field SA dynamics (Euclidean form):

$$
\begin{aligned}
\partial_t\mu_t + \nabla\cdot\big(v_\theta^{\mathrm{SA}}[\mu_t]\mu_t\big) = 0, \quad t \geq 0,\\
\mu_{t=0} = \mu_0.
\end{aligned}
$$

Spherical form:

$$
\begin{aligned}
\partial_t\mu_t + \nabla_{\mathbb{S}^{d-1}}\cdot\big(v_{\theta,\mathrm{sph}}^{\mathrm{SA}}[\mu_t]\mu_t\big) = 0, \quad t \geq 0,\\
\mu_{t=0} = \mu_0 \in \mathcal{P}(\mathbb{S}^{d-1}),
\end{aligned}
$$

where $\nabla_{\mathbb{S}^{d-1}}\cdot$ denotes the divergence on the sphere.

Definition 3.23 (Energy functional and gradient-flow form (SA-only))**.** We consider an energy functional

$$\mathcal{E}_\theta^{\mathrm{SA}} : \mathcal{P}(E) \to \mathbb{R}\cup\{+\infty\},$$

and a metric d on $\mathcal{P}(E)$ (e.g. $\mathsf{d} = W_2$ or a twisted variant $W_{m,2}$). Formal gradient-flow form:

$$\partial_t\mu_t = -\operatorname{grad}_{\mathsf{d}}\mathcal{E}_\theta^{\mathrm{SA}}(\mu_t),$$

which, in PDE form, corresponds to

$$\partial_t\mu_t + \nabla\cdot\left(-\nabla\frac{\delta\mathcal{E}_\theta^{\mathrm{SA}}}{\delta\mu}(\mu_t)(x)\mu_t\right) = 0,$$

with

$$v_\theta^{\mathrm{SA}}[\mu_t](x) = -\nabla\frac{\delta\mathcal{E}_\theta^{\mathrm{SA}}}{\delta\mu}(\mu_t)(x)$$

(or the corresponding twisted-gradient expression when $\mathsf{d} \neq W_2$).

Typical structure:

$$\mathcal{E}_\theta^{\mathrm{SA}}(\mu) = \int_E \Phi(x)d\mu(x) + \frac{1}{2}\iint_{E\times E} W(x,y)d\mu(x)d\mu(y),$$

for some "potential" Φ and interaction kernel W constructed from q, k, v, β.

Definition 3.24 (Equilibria, clusters, and Lyapunov structure)**.** Equilibrium measures (stationary solutions):

$$\mu_\star \in \mathcal{P}(E) \quad \text{s.t.} \quad v_\theta^{\mathrm{SA}}[\mu_\star](x) = 0 \text{ for } \mu_\star\text{-a.e. } x \quad \Leftrightarrow \quad \partial_t \mu_t\big|_{\mu_t=\mu_\star} = 0.$$

Cluster-type equilibria:

$$\mu_\star = \sum_{m=1}^{M} \alpha_m \delta_{z_m}, \quad \alpha_m \geq 0, \ \sum_m \alpha_m = 1,$$

with $(z_m)_m$ local consensus points.

Energy dissipation (Lyapunov) identity/inequality (formal):

$$\frac{d}{dt}\mathcal{E}_\theta^{\mathrm{SA}}(\mu_t) \leq 0,$$

with equality iff μ_t is an equilibrium (under suitable conditions).

Proposition 3.25 (Well-Posedness and W_2-Stability [6, Proposition 6.6])**.** *For any $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$, the continuity equation*

$$\partial_t \mu + \nabla \cdot (\mathcal{X}[\mu]\, \mu) = 0, \qquad \mu_{|t=0} = \mu_0,$$

admits a unique global solution in $C^0_{co}(\mathbb{R}_+; \mathcal{P}_c(\mathbb{R}^d))$. Moreover, for every $R, T > 0$ there exists $C(T,R) > 0$ such that if $\mathrm{supp}(\mu_0), \mathrm{supp}(\nu_0) \subset B(0,R)$, then

$$W_2(\mu(t), \nu(t)) \leq e^{C(T,R)t} W_2(\mu_0, \nu_0), \qquad t \in [0,T].$$

Proof. Appendix A.1.7. □

Remark (Relation to Unnormalized Dynamics [5, Remark 3.1])**.** *For the mean-field continuity equation on S^{d-1},*

$$\partial_t \mu + \mathrm{div}(\mathcal{X}[\mu]\mu) = 0,$$

global weak measure-valued well-posedness for arbitrary initial data $\mu(0) \in \mathcal{P}(S^{d-1})$ follows from the same Lipschitz/Dobrushin argument as in [6].

Proof. Appendix A.1.9. □

Proposition 3.26 (Continuity Equation Derivation [5, Proposition 3.4])**.** *Let $\beta > 0$ and $d \geq 2$, and let*

$$E_\beta(\mu) = \frac{1}{2} \iint_{S^{d-1}\times S^{d-1}} e^{\beta\langle x,y\rangle}\, d\mu(x)\, d\mu(y).$$

Then:

the unique global minimizer of E_β over $\mathcal{P}(S^{d-1})$ is the uniform law σ_d,
every global maximizer is a Dirac mass δ_{x_*}.

Proof. Appendix A.1.10. □

Lemma 3.27 (Gradient Flow Structure [5, Lemma 3.6])**.** *For the unnormalized spherical dynamics,*

$$\mathcal{X}[\mu](x) = P_x\Big(\int_{S^{d-1}} e^{\beta\langle x,y\rangle} y\, d\mu(y)\Big),$$

one has

$$\mathcal{X}[\mu](x) = \nabla \frac{\delta E_\beta}{\delta \mu}(x).$$

Hence the corresponding continuity equation is the Wasserstein gradient flow of E_β on $\mathcal{P}(S^{d-1})$.

Proof. Appendix A.1.11. □

Theorem 3.28 (Metric Structure of $W_{m,2}$ [1, Theorem 2.2]). *Let M be a compact Riemannian manifold without boundary and $K \in C(M \times M)$ satisfy $0 < c_K \le K \le C_K$. For $\mu \in \mathcal{P}(M)$ define*

$$m_\mu(x) := \int_M K(x,y)\,\mu(dy),$$

and for $\mu_0, \mu_1 \in \mathcal{P}(M)$ define

$$W_{m,2}^2(\mu_0,\mu_1) := \inf_{(\mu_t,v_t)\in CE(0,1;\mu_0\to\mu_1)} \int_0^1 \int_M m_{\mu_t}(x)\,|v_t(x)|^2\,\mu_t(dx)\,dt.$$

Then:

(i) For every pair with finite $W_{m,2}$, the infimum is attained;

(ii) equivalently, $W_{m,2}$ admits the length representation

$$W_{m,2}(\mu_0,\mu_1) = \inf_{(\mu_t,v_t)\in CE(0,T;\mu_0\to\mu_1)} \int_0^T \left(\int_M m_{\mu_t}(x)\,|v_t(x)|^2\,\mu_t(dx) \right)^{1/2} dt.$$

Proof. Appendix A.3.1. □

Theorem 3.29 (Geodesic Structure of $W_{m,2}$ [1, Theorem 2.3]). *Let M be a compact Riemannian manifold without boundary, and let $K \in C(M \times M)$ satisfy*

$$0 < c_K \le K(x,y) \le C_K < \infty \quad \forall (x,y) \in M \times M.$$

For $\mu \in \mathcal{P}(M)$ define the mobility

$$m_\mu(x) := \int_M K(x,y)\,\mu(dy),$$

and define the modified transport distance by

$$W_{m,2}^2(\mu_0,\mu_1) := \inf_{(\mu_t,v_t)} \int_0^1 \int_M m_{\mu_t}(x)\,|v_t(x)|^2\,\mu_t(dx)\,dt,$$

where the infimum is taken over all distributional solutions of the continuity equation connecting μ_0 to μ_1. Then $(\mathcal{P}(M), W_{m,2})$ is a geodesic metric space: for every $\mu_0, \mu_1 \in \mathcal{P}(M)$ there exists a constant-speed geodesic $(\mu_t)_{t\in[0,1]}$ such that

$$W_{m,2}(\mu_s,\mu_t) = |t-s|\,W_{m,2}(\mu_0,\mu_1) \quad \forall\, 0 \le s \le t \le 1.$$

Proof. Appendix A.3.2. □

Theorem 3.30 (Gradient Flow Identification [1, Theorem 2.7]). *Let M be a compact Riemannian manifold without boundary, let $K \in C(M \times M)$ satisfy $0 < c_K \le K \le C_K$, and let $W \in C^1(M \times M)$ be symmetric. Define*

$$m_\mu(x) := \int_M K(x,y)\,\mu(dy), \qquad W[\mu](x) := \int_M W(x,y)\,\mu(dy),$$

and the interaction energy

$$\mathcal{E}(\mu) := \frac{1}{2}\int_{M\times M} W(x,y)\,\mu(dx)\,\mu(dy).$$

Then weak solutions of

$$\partial_t \mu_t + \operatorname{div}_M\left(\frac{1}{m_{\mu_t}} \operatorname{grad}_M W[\mu_t]\,\mu_t \right) = 0 \tag{8}$$

coincide with curves of maximal slope of $\mathcal{E}$ in $(\mathcal{P}(M), W_{m,2})$. Equivalently, they are exactly the absolutely continuous curves satisfying the energy-dissipation equality

$$\begin{aligned} \mathcal{E}(\mu_T) - \mathcal{E}(\mu_0) &+ \frac{1}{2}\int_0^T \int_M m_{\mu_t}|v_t|^2\,d\mu_t\,dt \\ &+ \frac{1}{2}\int_0^T \int_M \frac{1}{m_{\mu_t}}\left|\operatorname{grad}_M W[\mu_t]\right|^2 d\mu_t\,dt = 0. \end{aligned}$$

Proof. Appendix A.3.4. □

Corollary 3.31 (Topology of $W_{m,2}$ [1, Corollary 2.5])**.** *Under the assumptions of Theorem 3.29, there exist constants $c_1, c_2 > 0$ such that for all $\mu, \nu \in \mathcal{P}(M)$,*

$$c_1 \, W_2(\mu,\nu) \le W_{m,2}(\mu,\nu) \le c_2 \, W_2(\mu,\nu).$$

Consequently, $W_{m,2}$ and W_2 induce the same topology on $\mathcal{P}(M)$; on compact M, this is equivalent to narrow convergence. In particular, $(\mathcal{P}(M), W_{m,2})$ is complete.

Proof. Appendix A.3.3. □

Lemma 3.32 (Energy Dissipation Inequality [1, Lemma 2.9])**.** *Let $W \in C^1(M \times M)$ be symmetric and define*

$$\mathcal{E}(\mu) := \frac{1}{2} \int_{M \times M} W(x,y) \, \mu(dx) \, \mu(dy), \qquad W[\mu](x) := \int_M W(x,y) \, \mu(dy).$$

For every absolutely continuous curve $t \mapsto \mu_t$ in $(\mathcal{P}(M), W_{m,2})$ with continuity-equation velocity v_t,

$$\begin{aligned} \mathcal{E}(\mu_T) - \mathcal{E}(\mu_0) &+ \frac{1}{2} \int_0^T \int_M m_{\mu_t}(x) \, |v_t(x)|^2 \, \mu_t(dx) \, dt \\ &+ \frac{1}{2} \int_0^T \int_M \frac{1}{m_{\mu_t}(x)} \big| \operatorname{grad}_M W[\mu_t](x) \big|^2 \, \mu_t(dx) \, dt \ge 0. \end{aligned}$$

Moreover, equality holds if and only if μ_t is a weak solution of the gradient-flow PDE (8).

Proof. Appendix A.3.5. □

Corollary 3.33 (Convergence to Stationary Measures [1, Corollary 2.11])**.** *Let M be a compact manifold without boundary, and let $W \in C^{1,\alpha}(M \times M)$ $(\alpha > 0)$ be symmetric. If $(\mu_t)_{t \ge 0}$ is a weak solution of the gradient-flow continuity equation (8), then there exists a sequence $t_n \to \infty$ and a measure $\mu_\star \in \mathcal{P}(M)$ such that*

$$\mu_{t_n} \rightharpoonup \mu_\star \quad \text{(narrowly)},$$

and $\mu_\star$ is stationary, i.e.

$$\operatorname{div}_M \left(\frac{1}{m_{\mu_\star}} \operatorname{grad}_M W[\mu_\star] \, \mu_\star \right) = 0.$$

Proof. Appendix A.3.6. □

Theorem 3.34 (Characterization of Maximizers [1, Theorem 3.1])**.** *Let $D = D^\top$ and let $\lambda_{\max}$ be an eigenvalue of maximal absolute value with $\lambda_{\max} > 0$. Denote by*

$$Z_{\lambda_{\max}} := \{ z \in \mathbb{S}^{d-1} : Dz = \lambda_{\max} z \}.$$

For the interaction energy

$$\mathcal{E}_D(\mu) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} e^{\langle x, Dy \rangle} \, \mu(dx) \, \mu(dy),$$

the maximizers are exactly the Dirac measures δ_z with $z \in Z_{\lambda_{\max}}$.

Proof. Appendix A.3.7. □

Theorem 3.35 (Characterization of Minimizers [1, Theorem 3.4])**.** *Let $D = D^\top$ be not positive definite, and let $\lambda_{\min} \le 0$ be its smallest eigenvalue.*

$$\begin{aligned} &\text{If } \lambda_{\min} < 0, \text{ then } \mu \text{ minimizes } \mathcal{E}_D \text{ iff } \mu = \delta_z \text{ with } Dz = \lambda_{\min} z, \\ &\text{If } \lambda_{\min} = 0, \text{ then } \mu \text{ minimizes } \mathcal{E}_D \text{ iff } \operatorname{supp}(\mu) \subset \ker D. \end{aligned}$$

Proof. Appendix A.3.8. ☐

Corollary 3.36 (Symmetry of Minimizers [1, Corollary 3.10]). *Assume $D = D^\top \succ 0$, and let $\mu_\star$ be a minimizer of $\mathcal{E}_D$. Then $\mu_\star$ is symmetric with respect to reflections across every eigendirection of D. Equivalently, for each unit eigenvector z of D,*

$$(H_z)_\# \mu_\star = \mu_\star, \qquad H_z(x) := x - 2\langle x, z\rangle z.$$

Proof. Appendix A.3.9. ☐

Theorem 3.37 (Maximizers for Negative Semidefinite D [1, Theorem 3.17]). *Let $D = D^\top \preceq 0$ and let $\lambda_{\min} < 0$ be the smallest eigenvalue. Then μ maximizes $\mathcal{E}_D$ if and only if*

$$\mu = \frac{1}{2}\big(\delta_z + \delta_{-z}\big), \qquad Dz = \lambda_{\min} z, \quad z \in \mathbb{S}^{d-1}.$$

Proof. Appendix A.3.10. ☐

Proposition 3.38 (Variational Optima are Stationary Solutions [1, Proposition 4.2]). *Let E_D be the interaction energy on $\mathcal{P}(S^{d-1})$ and let $dE_D(\mu; W)$ denote its first variation along a Lipschitz tangent field W. If μ_* is a global minimizer or global maximizer of E_D, then*

$$dE_D(\mu_*; W) = 0 \quad \textit{for every Lipschitz } W : TS^{d-1} \to \mathbb{R}^d.$$

Hence every global extremizer is a stationary point of the mean-field dynamics.

Proof. Appendix A.3.11. ☐

Lemma 3.39 (Stationarity of Dirac Measures [1, Lemma 4.3]). *Let $\mu \in \mathcal{P}(S^{d-1})$. The following are equivalent:*

(i) μ is stationary for $\partial_t \mu + \nabla \cdot (\mu \Gamma_\mu) = 0$,
(ii) $dE_D(\mu; W) = 0$ for every Lipschitz W.

In particular, stationarity of the PDE is exactly the Euler–Lagrange condition for E_D under Wasserstein perturbations.

Proof. Appendix A.3.12. ☐

Lemma 3.40 (Stationarity of Symmetric 2-Point Measures [1, Lemma 4.5]). *For $z \in S^{d-1}$, the Dirac measure δ_z is a stationary point of E_D if and only if z is an eigenvector of D. Equivalently,*

$$\delta_z \textit{ is stationary} \iff P_z^\perp D z = 0 \iff Dz = (z^\top D z) z.$$

Proof. Appendix A.3.13. ☐

Lemma 3.41 (Stationarity Condition for D [1, Lemma 4.6]). *Let $z \in S^{d-1}$ and $t \in [0, 1]$. The two-point measure*

$$\mu_t = t\delta_z + (1 - t)\delta_{-z}$$

is a stationary point of E_D if and only if z is an eigenvector of D. In particular, all antipodal mixtures generated by an eigenvector are stationary.

Proof. Appendix A.3.14. ☐

Lemma 3.42 (Instability of Unaligned States [1, Lemma 4.7]). *Let $Z_D \subset S^{d-1}$ be a finite set of eigenvectors of D such that*

$$w \cdot z = 0 \quad \text{for all distinct } w, z \in Z_D.$$

For any weights $t : Z_D \to \mathbb{R}_+$ with $\sum_{z \in Z_D} t(z) = 1$, define

$$\mu = \frac{1}{2} \sum_{z \in Z_D} t(z)(\delta_z + \delta_{-z}).$$

Then μ is a stationary point of E_D.

Proof. Appendix A.3.15. □

Lemma 3.43 (Linear Stability Analysis [1, Lemma 4.10]). *Let σ be the uniform probability measure on S^{d-1}. Then σ is a stationary point of E_D if and only if all eigenvalues $\{\lambda_i\}_{i=1}^d$ of D have equal absolute value:*

$$|\lambda_1| = \cdots = |\lambda_d|.$$

Proof. Appendix A.3.16. □

Corollary 3.44 (Local Stability Criteria [1, Corollary 4.11]). *The uniform measure σ minimizes E_D if and only if*

$$D = \lambda I_d \quad \text{for some } \lambda \geq 0.$$

Equivalently, isotropic nonnegative interaction is exactly the regime where the uniform state is globally energy-minimizing.

Proof. Appendix A.3.17. □

Theorem 3.45 (Global Well-Posedness and Stability [2, Theorem 3.1]). *Let $d, k \in \mathbb{N}^*$ with $k \leq d$, $p \geq 1$, and let $Q, K : [0, +\infty) \to \mathbb{R}^{k \times d}$, $V : [0, +\infty) \to \mathbb{R}^{d \times d}$ be continuous. For any compactly supported initial datum $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ and any attention field $\Gamma \in \{\Gamma^{(SM)}, \Gamma^{(L2)}, \Gamma^{(sink)}, \Gamma^{(\sigma)}, \Gamma^{(MH)}\}$, the Transformer PDE*

$$\partial_t \mu + \nabla \cdot (\mu \Gamma_\mu) = 0, \qquad \mu_{|t=0} = \mu_0,$$

has a unique global weak solution in $C([0, +\infty), \mathcal{P}_c(\mathbb{R}^d))$. If $\operatorname{supp}(\mu_0) \subset B_{R_0}$ and

$$R(t) := \exp\Big(\int_0^t \|V(s)\|_2 \, ds \Big) R_0,$$

then $\operatorname{supp}(\mu(t)) \subset B_{R(t)}$ for all $t \geq 0$. Moreover, for each $t > 0$, there exists $C(t, R_0) > 0$ such that any two solutions μ, ν with initial data in B_{R_0} satisfy

$$W_p(\mu(t), \nu(t)) \leq C(t, R_0) W_p(\mu_0, \nu_0).$$

Proof. Appendix A.4.1. □

Gaussian Dynamics (Softmax).

Lemma 3.46 (Gaussian Velocity Field (Softmax) [2, Lemma 4.1]). *Let $\mu = \mathcal{N}(\alpha, \Sigma)$ on $\mathbb{R}^d$ and set $A := K^\top Q$. Then for all $x \in \mathbb{R}^d$,*

$$\Gamma_\mu^{(SM)}(x) = V\big(\alpha + \Sigma A x\big).$$

Hence the Softmax mean-field velocity is affine on Gaussian states.

Proof. Appendix A.4.3. □

Proposition 3.47 (Gaussian ODEs (Softmax) [2, Proposition 4.2])**.** *Assume continuous parameter paths $Q, K, V : [0, +\infty) \to \mathbb{R}^{d\times d}$ and Gaussian initial data $\mu_0 = \mathcal{N}(\alpha_0, \Sigma_0)$ with $\Sigma_0 \succ 0$. For the Softmax PDE*

$$\partial_t \mu + \nabla \cdot (\mu \Gamma_\mu^{(SM)}) = 0,$$

there exists a unique maximal solution $\mu(t)$ that remains Gaussian: $\mu(t) = \mathcal{N}(\alpha(t), \Sigma(t))$. With $A := K^\top Q$, (α, Σ) solve

$$\dot{\Sigma} = V\Sigma A\Sigma + \Sigma A^\top \Sigma V^\top, \qquad \dot{\alpha} = V(\mathrm{Id} + \Sigma A)\alpha.$$

Proof. Appendix A.4.4. □

Proposition 3.48 (Finite-Time Blow-Up vs Convergence (Softmax) [2, Proposition 4.4])**.** *Consider*

$$\dot{\Sigma} = V\Sigma A\Sigma + \Sigma A^\top \Sigma V^\top, \qquad \Sigma(0) = \Sigma_0 \succ 0,$$

and assume V commutes with Σ_0 and with $VA + A^\top V^\top$. Then

$$\Sigma(t) = \left(\Sigma_0^{-1} - t(VA + A^\top V^\top)\right)^{-1}$$

on its maximal interval of existence. If $VA + A^\top V^\top \preceq 0$, the solution is global and converges to a limit $\Sigma_* \succeq 0$ with rank loss along strictly negative eigendirections. If $VA + A^\top V^\top$ has a positive eigenvalue, the largest eigenvalue of $\Sigma(t)$ blows up in finite time.

Proof. Appendix A.4.5. □

Proposition 3.49 (Low-Rank Stationary Points (Softmax) [2, Proposition 4.8])**.** *Let $A \in \mathbb{R}^{d\times d}$ be symmetric and set $V = \mathrm{Id}$. If a symmetric matrix Σ satisfies the stationarity equation*

$$\Sigma A \Sigma + \Sigma A^\top \Sigma = 0,$$

then

$$\mathrm{rk}(\Sigma) \le \dim\ker A + \min\left(\#\{\lambda_i(A) > 0\}, \#\{\lambda_i(A) < 0\}\right).$$

Therefore stationary Gaussian covariances are necessarily low rank.

Proof. Appendix A.4.6. □

Proposition 3.50 (Gaussian Dynamics for Multi-Head Softmax [2, Proposition 4.9])**.** *Let $Q, K, V : [0, +\infty) \to \mathbb{R}^{d\times d}$ define H heads with matrices $A^{(h)}$ and weights $W^{(h)}V^{(h)}$ (notation of Eq. (2.9) in [2]). For Gaussian initial data $\mu_0 = \mathcal{N}(\alpha_0, \Sigma_0)$, the multi-head PDE keeps Gaussian form $\mu(t) = \mathcal{N}(\alpha(t), \Sigma(t))$ and*

$$\dot{\Sigma} = \sum_{h=1}^{H} \Big(W^{(h)}V^{(h)}\Sigma A^{(h)}\Sigma + \Sigma A^{(h)\top}\Sigma (W^{(h)}V^{(h)})^\top \Big),$$
$$\dot{\alpha} = \sum_{h=1}^{H} W^{(h)}V^{(h)}\big(\mathrm{Id} + \Sigma A^{(h)}\big)\alpha.$$

Proof. Appendix A.4.7. □

Gaussian Dynamics (L2).

Lemma 3.51 (Gaussian Velocity Field (L2) [2, Lemma 4.10])**.** *For $\mu = \mathcal{N}(\alpha, \Sigma)$ and $x \in \mathbb{R}^d$,*

$$\Gamma_\mu^{(L2)}(x) = V(\Sigma^{-1} + 2K^\top K)^{-1}(\Sigma^{-1}\alpha + 2K^\top Q x).$$

Hence the L2 attention velocity is affine on Gaussian measures.

Proof. Appendix A.4.8. □

Proposition 3.52 (Gaussian ODEs (L2) [2, Proposition 4.11])**.** *Let $Q, K, V : [0, +\infty) \to \mathbb{R}^{d\times d}$ be continuous and $\mu_0 = \mathcal{N}(\alpha_0, \Sigma_0)$ with $\Sigma_0 \succ 0$. For $\partial_t \mu + \nabla \cdot (\mu \Gamma_\mu^{(L2)}) = 0$, $\mu(t)$ stays Gaussian: $\mu(t) = \mathcal{N}(\alpha(t), \Sigma(t))$, with*

$$\begin{aligned}\dot{\Sigma} &= 2V\Sigma(\mathrm{Id} + 2K^\top K\Sigma)^{-1}A\Sigma + 2\Sigma A^\top(\mathrm{Id} + 2K^\top K\Sigma)^{-1}\Sigma V^\top,\\ \dot{\alpha} &= V(\Sigma^{-1} + 2K^\top K)^{-1}(\Sigma^{-1} + 2A)\alpha,\end{aligned}$$

where $A = K^\top Q$.

Proof. Appendix A.4.9. □

Lemma 3.53 (Global Existence for L2 ODEs [2, Lemma 4.12])**.** *For any $\Sigma_0 \succ 0$, the matrix ODE*

$$\dot{\Sigma} = 2V\Sigma(\mathrm{Id} + 2K^\top K\Sigma)^{-1}A\Sigma + 2\Sigma A^\top(\mathrm{Id} + 2K^\top K\Sigma)^{-1}\Sigma V^\top$$

admits a unique global solution. Hence L2 Gaussian covariance dynamics cannot blow up in finite time.

Proof. Appendix A.4.10. □

Gaussian Dynamics (Sinkhorn).

Lemma 3.54 (Gaussian Velocity Field (Sinkhorn) [2, Lemma 4.14])**.** *Let $\mu = \mathcal{N}(\alpha, \Sigma)$ on $\mathbb{R}^d$, with $A := K^\top Q$ and assume A, Σ invertible. Then*

$$\Gamma_{\mu,\varepsilon}^{(sink)}(x) = \frac{1}{\varepsilon}V(\mathrm{Id} - A^{-\top}\Sigma^{-1}C)\alpha + \frac{1}{\varepsilon}VA^{-\top}\Sigma^{-1}Cx,$$

where

$$C := \Sigma^{1/2}\Big(\Sigma^{1/2}A^\top\Sigma A\Sigma^{1/2} + \frac{\varepsilon^2}{4}\,\mathrm{Id}\Big)^{1/2}\Sigma^{-1/2} - \frac{\varepsilon}{2}\,\mathrm{Id}\,.$$

Proof. Appendix A.4.11. □

Proposition 3.55 (Gaussian ODEs (Sinkhorn) [2, Proposition 4.15])**.** *Let $Q, K, V : [0, +\infty) \to \mathbb{R}^{d\times d}$ be continuous and $\mu_0 = \mathcal{N}(\alpha_0, \Sigma_0)$ with $\Sigma_0 \succ 0$. For the Sinkhorn PDE*

$$\partial_t \mu + \nabla \cdot (\mu \Gamma_{\mu,\varepsilon}^{(sink)}) = 0,$$

$\mu(t)$ remains Gaussian: $\mu(t) = \mathcal{N}(\alpha(t), \Sigma(t))$, and

$$\begin{aligned}\dot{\Sigma} &= \frac{1}{\varepsilon}\Big(VA^{-\top}\Sigma^{-1}C\Sigma + \Sigma C^\top\Sigma^{-1}A^{-1}V^\top\Big),\\ \dot{\alpha} &= \frac{1}{\varepsilon}V\alpha,\end{aligned}$$

with

$$C = \Sigma^{1/2}\Big(\Sigma^{1/2}A\Sigma A^\top\Sigma^{1/2} + \frac{\varepsilon^2}{4}\,\mathrm{Id}\Big)^{1/2}\Sigma^{-1/2} - \frac{\varepsilon}{2}\,\mathrm{Id}\,.$$

In particular, mean and covariance evolve independently.

Proof. Appendix A.4.12. □

Gradient Flow Structures.

Theorem 3.56 (Sinkformer as Bures-Wasserstein Flow [2, Section 5.3])**.** *Assume Sinkhorn attention with $A := K^\top Q$ satisfying $A = A^\top = -V$. Then the Sinkformer PDE is a Wasserstein gradient flow:*

$$\partial_t \mu + \nabla \cdot \big(\mu \Gamma^{(sink)}_{\mu,\varepsilon}\big) = 0 \quad \iff \quad \partial_t \mu + \nabla \cdot \Big(\mu \nabla \frac{\delta F_\varepsilon}{\delta \mu}(\mu)\Big) = 0,$$

for the energy (Eq. (5.11) in [2])

$$F_\varepsilon(\mu) = -\frac{1}{2}\,\kappa^\infty_{\mu,\varepsilon} \log \frac{\kappa^\infty_{\mu,\varepsilon}}{\kappa^0_{\mu,\varepsilon}} + \frac{1}{4\varepsilon} \int \big(|Qx|^2 + |Kx|^2\big)\, d\mu(x).$$

Restricted to Gaussian states, this induces the Bures-Wasserstein Riemannian flow for (α, Σ) described in Section 5.3.

Proof. Appendix A.4.13. ☐

Theorem 3.57 (Softmax as Twisted Wasserstein Gradient Flow [2, Section 5.4])**.** *Let $A, V \in \mathbb{R}^{d\times d}$ and define $B := -V A^{-\top}$. If B is symmetric positive definite, define the twisted mobility operator*

$$G_\mu(v)(x) := \frac{Bv(x)}{\int e^{Ax\cdot y}\, d\mu(y)}$$

and the associated distance $d_{A,V}$. Then the Softmax Transformer PDE can be written as a gradient flow in this metric for the quadratic interaction energy

$$F(\mu) = \frac{1}{2} \iint e^{Ax\cdot y}\, d\mu(x)\, d\mu(y).$$

Moreover, Section 5.4 proves that, even in this formulation, F is generally not geodesically convex for $d_{A,V}$.

Proof. Appendix A.4.14. ☐

Theorem 3.58 (Global Rates in Mean-Field Limit [8, Theorem 5])**.** *Let $d \geq 2$ and let μ_t solve the mean-field equation from initial law μ_0 with density $f_0 \in L^2(S^{d-1})$ and nonzero first moment*

$$R_0 := \Big\| \int_{S^{d-1}} x\, d\mu_0(x) \Big\| > 0.$$

Then there exist constants $\beta_0, C_0, T_0 > 0$ (depending on μ_0) such that if $|\beta| < \beta_0$, there is $x_\infty \in S^{d-1}$ with

$$W_2(\mu_t, \delta_{x_\infty}) \leq C_0 e^{-t/100}, \qquad t \geq T_0.$$

Proof. Appendix A.6.4. ☐

Theorem 3.59 (Metastability and Saddle-to-Saddle Dynamics [8, Theorem 6])**.** *Consider a discrete multi-cluster initial datum with a unique closest pair $(\bar{\imath}, \bar{\jmath})$ maximizing $\langle x_i(0), x_j(0)\rangle$. Under the rescaled time*

$$dt = e^{\beta(1-\langle x_{\bar{\imath}}(0), x_{\bar{\jmath}}(0)\rangle)} ds,$$

as $\beta \to \infty$, trajectories converge uniformly on finite intervals (before contact) to a reduced system where only the closest pair moves along their connecting geodesic, all other clusters remaining frozen. The pair merges in finite rescaled time.

Proof. Appendix A.6.5. ☐

Definition 3.60 (Noisy Transformer Fokker-Planck Equation [8, Equation (9)])**.** The law μ_t of Definition 3.20 satisfies

$$\partial_t \mu_t + \kappa^{-1} \Delta \mu_t = \mathrm{div}\bigg(\mu_t\, P_x^\perp \bigg(\int_{S^{d-1}} e^{\beta\langle x,y\rangle} y\, \mu_t(dy)\bigg)\bigg),$$

which reduces to the deterministic mean-field equation as $\kappa \to \infty$.

Proof. Appendix A.6.8. ☐

4 Transformer Dynamics

4.1 Objects and Assumptions

Ambient space, time, and static building blocks.

$$
\begin{aligned}
&E \subset \mathbb{R}^d, \quad \text{typically } E = \mathbb{S}^{d-1},\\
&X(t) = (x_1(t), \dots, x_n(t)) \in E^n,\\
&t \in [0, T] \text{ or } t \in [0, \infty),\\
&\mu_t := \mu_{X(t)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)} \in \mathcal{P}(E).
\end{aligned}
$$

Static ingredients (from 1):

$$
\begin{aligned}
&S_\theta^{\mathrm{MH}} : E^n \to E^n \quad \text{(multi-head SA)},\\
&\mathrm{LN} : E^n \to E^n \quad \text{(layer normalisation)},\\
&\mathrm{FFN} : E^n \to E^n \quad \text{(feed-forward network)},\\
&T_\theta : E^n \to E^n \quad \text{(pre-LN Transformer block)},\\
&T_\Theta^{(L)} := T_{\theta_L} \circ \cdots \circ T_{\theta_1} \quad \text{(depth-}L\text{ Transformer)}.
\end{aligned}
$$

Block map and residual structure. Recall the pre-LN block (1):

$$
\begin{aligned}
X^{\mathrm{att}} &:= X + S_\theta^{\mathrm{MH}}(\mathrm{LN}(X)),\\
T_\theta(X) &:= X^{\mathrm{att}} + \mathrm{FFN}(\mathrm{LN}(X^{\mathrm{att}})).
\end{aligned}
$$

We write T_θ as a residual map

$$
T_\theta(X) = X + G_\theta^{\mathrm{blk}}(X),
$$

where the block drift is

$$
\begin{aligned}
G_\theta^{\mathrm{blk}}(X) &:= S_\theta^{\mathrm{MH}}(\mathrm{LN}(X)) + \mathrm{FFN}(\mathrm{LN}(X^{\mathrm{att}}))\\
&= S_\theta^{\mathrm{MH}}(\mathrm{LN}(X)) + \mathrm{FFN}\big(\mathrm{LN}\big(X + S_\theta^{\mathrm{MH}}(\mathrm{LN}(X))\big)\big).
\end{aligned}
$$

Component-wise:

$$
G_\theta^{\mathrm{blk}}(X) = \big(g_\theta^{\mathrm{blk}}(x_i; X)\big)_{i=1}^{n}.
$$

Discrete-depth Transformer dynamics. Given parameters $\Theta = (\theta_1, \dots, \theta_L)$, the discrete-depth evolution is

$$
\begin{aligned}
X^{\ell+1} &= T_{\theta_{\ell+1}}(X^\ell), \quad \ell = 0, \dots, L-1,\\
X^0 &= X^{\mathrm{in}} \in E^n.
\end{aligned}
$$

In residual form:

$$
X^{\ell+1} = X^\ell + G_{\theta_{\ell+1}}^{\mathrm{blk}}(X^\ell).
$$

Empirical laws:

$$
\mu^\ell := \mu_{X^\ell} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^\ell}.
$$

Continuous-depth Transformer vector field on E^n. We introduce a (possibly time-dependent) Transformer vector field

$$G_\Theta^{\mathrm{Tr}} : [0,T] \times E^n \to T(E^n),$$

of the form

$$G_\Theta^{\mathrm{Tr}}(t,X) = \big(g_\Theta^{\mathrm{Tr}}(t,x_i;X)\big)_{i=1}^n.$$

Prototype: piecewise-constant in t using block drifts:

$$G_\Theta^{\mathrm{Tr}}(t,X) = G_{\theta_{\ell(t)}}^{\mathrm{blk}}(X),$$

where $t \mapsto \ell(t)$ maps each time interval $[t_\ell, t_{\ell+1})$ to layer index ℓ (layer-wise time slicing).

Spherical case $E = \mathbb{S}^{d-1}$:

$$G_\Theta^{\mathrm{Tr}}(t,X) := \big(P_{x_i} g_\Theta^{\mathrm{Tr}}(t,x_i;X)\big)_{i=1}^n,$$

so that $G_\Theta^{\mathrm{Tr}}(t,X) \in T(E^n)$ preserves $(\mathbb{S}^{d-1})^n$.

Continuous-depth Transformer dynamics:

$$\begin{aligned} \dot X(t) &= G_\Theta^{\mathrm{Tr}}(t,X(t)), \quad t \in [0,T], \\ X(0) &= X^0 \in E^n. \end{aligned}$$

Mean-field Transformer velocity and continuity equation. We associate a (possibly time-dependent) mean-field velocity functional

$$v_\Theta^{\mathrm{Tr}} : [0,T] \times \mathcal{P}(E) \to \{\text{vector fields on } E\},$$

so that, in the mean-field regime,

$$v_\Theta^{\mathrm{Tr}}[t,\mu](x) \approx g_\Theta^{\mathrm{Tr}}(t,x;X) \quad \text{whenever } \mu_X \approx \mu.$$

Mean-field Transformer continuity equation (Euclidean version):

$$\begin{aligned} \partial_t \mu_t + \nabla \cdot \big(v_\Theta^{\mathrm{Tr}}[t,\mu_t]\mu_t\big) &= 0, \quad t \in [0,T], \\ \mu_{t=0} &= \mu_0. \end{aligned}$$

Spherical version:

$$\begin{aligned} \partial_t \mu_t + \nabla_{\mathbb{S}^{d-1}} \cdot \big(v_{\Theta,\mathrm{sph}}^{\mathrm{Tr}}[t,\mu_t]\mu_t\big) &= 0, \quad t \in [0,T], \\ \mu_{t=0} &= \mu_0 \in \mathcal{P}(\mathbb{S}^{d-1}), \end{aligned}$$

where

$$v_{\Theta,\mathrm{sph}}^{\mathrm{Tr}}[t,\mu](x) := P_x v_\Theta^{\mathrm{Tr}}[t,\mu](x).$$

Standing assumptions.

(T1) *Bounded configurations / support.* There exists $R > 0$ such that along the flow,

$$\|x_i(t)\| \le R \quad \forall i \in [n],\ \forall t \in [0,T],$$

equivalently $\mathrm{supp}(\mu_t) \subset B(0,R)$ (or $\subset \mathbb{S}^{d-1}$).

(T2) *Local Lipschitzness of G_Θ^{Tr}.* For any $R > 0$, there exists $L_{\mathrm{Tr}}(R) < \infty$ such that

$$\|G_\Theta^{\mathrm{Tr}}(t,X) - G_\Theta^{\mathrm{Tr}}(t,X')\|_F \le L_{\mathrm{Tr}}(R)\|X - X'\|_F$$

for all $t \in [0,T]$ and all $X, X' \in E^n$ with $\|x_i\|, \|x_i'\| \le R$.

(T3) *Tangent structure (spherical case).* If $E = \mathbb{S}^{d-1}$ and $X^0 \in (\mathbb{S}^{d-1})^n$, then $x_i(t) \in \mathbb{S}^{d-1}$ for all i, t, i.e. $G_\Theta^{\mathrm{Tr}}(t,X) \in T(E^n)$ for all (t,X).

(T4) *Mean-field well-posedness (formal).* Under suitable regularity of $v_\Theta^{\mathrm{Tr}}[\cdot]$, the continuity equation admits (at least) a weak solution $(\mu_t)_{t\in[0,T]}$ in $\mathcal{P}(E)$ for any initial $\mu_0 \in \mathcal{P}(E)$.

4.2 Discrete Transformer Dynamics on E^n

Definition 4.1 (Block drift induced by T_θ)**.** Recall the pre-LN block from 1:

$$
\begin{aligned}
X^{\mathrm{att}} &:= X + S_\theta^{\mathrm{MH}}(\mathrm{LN}(X)),\\
T_\theta(X) &:= X^{\mathrm{att}} + \mathrm{FFN}(\mathrm{LN}(X^{\mathrm{att}})).
\end{aligned}
$$

Residual form:

$$T_\theta(X) = X + G_\theta^{\mathrm{blk}}(X),$$

with block drift

$$
\begin{aligned}
G_\theta^{\mathrm{blk}}(X) &:= S_\theta^{\mathrm{MH}}(\mathrm{LN}(X)) + \mathrm{FFN}\big(\mathrm{LN}(X^{\mathrm{att}})\big)\\
&= S_\theta^{\mathrm{MH}}(\mathrm{LN}(X)) + \mathrm{FFN}\Big(\mathrm{LN}\big(X + S_\theta^{\mathrm{MH}}(\mathrm{LN}(X))\big)\Big).
\end{aligned}
$$

Component-wise:

$$G_\theta^{\mathrm{blk}}(X) = \big(g_\theta^{\mathrm{blk}}(x_i; X)\big)_{i=1}^n.$$

Definition 4.2 (Continuous-Depth Block ODE on E^n)**.** Euclidean version:

$$
\begin{aligned}
&\dot X(t) = G_\theta^{\mathrm{blk}}(X(t)), \quad t \ge 0,\\
&X(0) = X^0 \in E^n.
\end{aligned}
$$

Spherical version ($E = \mathbb{S}^{d-1}$):

$$
\begin{aligned}
&\dot X(t) = G_{\theta,\mathrm{sph}}^{\mathrm{blk}}(X(t)), \quad t \ge 0,\\
&X(0) = X^0 \in (\mathbb{S}^{d-1})^n,
\end{aligned}
$$

where

$$G_{\theta,\mathrm{sph}}^{\mathrm{blk}}(X) := \big(P_{x_i} g_\theta^{\mathrm{blk}}(x_i; X)\big)_{i=1}^n.$$

This gives a continuous-depth Transformer block dynamics, obtained formally as the limit of repeatedly applying

$$X \mapsto X + \Delta t G_\theta^{\mathrm{blk}}(X)$$

with small Δt.

Definition 4.3 (Layer-wise continuous-depth model with time-dependent parameters)**.** Let $\Theta(t)$ denote a (piecewise constant or smooth) parameter schedule over $t \in [0, T]$, with

$$\Theta(t) \approx \theta_\ell \quad \text{for } t \in [t_\ell, t_{\ell+1}),$$

where $(t_\ell)_{\ell=0}^L$ is a time partition. Define the time-dependent block drift

$$G_\Theta^{\mathrm{Tr}}(t, X) := G_{\Theta(t)}^{\mathrm{blk}}(X) = G_{\theta_{\ell(t)}}^{\mathrm{blk}}(X) \quad \text{(piecewise-constant schedule)}.$$

Continuous-depth Transformer dynamics:

$$
\begin{aligned}
&\dot X(t) = G_\Theta^{\mathrm{Tr}}(t, X(t)), \quad t \in [0, T],\\
&X(0) = X^0 \in E^n.
\end{aligned}
$$

Spherical case:

$$
\begin{aligned}
G_{\Theta,\mathrm{sph}}^{\mathrm{Tr}}(t, X) &:= \big(P_{x_i} g_\Theta^{\mathrm{Tr}}(t, x_i; X)\big)_{i=1}^n,\\
\dot X(t) &= G_{\Theta,\mathrm{sph}}^{\mathrm{Tr}}(t, X(t)).
\end{aligned}
$$

Definition 4.4 (Discrete-depth Transformer as an Euler scheme)**.** Consider the ODE

$$\dot{X}(t) = G^{\mathrm{blk}}_{\theta_\ell}(X(t)), \quad t \in [t_\ell, t_{\ell+1}),$$

with step size $\Delta t_\ell := t_{\ell+1} - t_\ell$. Explicit Euler scheme on $[t_\ell, t_{\ell+1}]$:

$$X^{\ell+1} = X^\ell + \Delta t_\ell G^{\mathrm{blk}}_{\theta_\ell}(X^\ell).$$

For $\Delta t_\ell = 1$,

$$X^{\ell+1} = X^\ell + G^{\mathrm{blk}}_{\theta_\ell}(X^\ell) = T_{\theta_\ell}(X^\ell),$$

i.e. the standard discrete-depth Transformer block. For $0 < \Delta t_\ell < 1$, we obtain a convex-combination-like update between the identity and T_{θ_ℓ}:

$$X^{\ell+1} = (1 - \Delta t_\ell)X^\ell + \Delta t_\ell T_{\theta_\ell}(X^\ell).$$

Note. Concrete full-Transformer dynamics are paper-dependent and will be introduced in the corresponding result files.

4.3 Mean-Field Transformer Dynamics: Continuity Equation and Energy

Definition 4.5 (Mean-field Transformer velocity)**.** We consider a (possibly time-dependent) mean-field Transformer velocity functional

$$v^{\mathrm{Tr}}_\Theta[t, \mu] : E \to TE,$$

built from the same ingredients as in Setting I and II: multi-head self-attention, layer norm, and FFN. Euclidean form:

$$v^{\mathrm{Tr}}_\Theta[t, \mu](x) := \Gamma^{\mathrm{Tr}}_\mu(x) - x,$$

where

$$\Gamma^{\mathrm{Tr}}_\mu(x) := (\text{one-step Transformer block update at } x \text{ under input law } \mu).$$

Spherical version ($E = \mathbb{S}^{d-1}$):

$$v^{\mathrm{Tr}}_{\Theta,\mathrm{sph}}[t, \mu](x) := P_x\big(\Gamma^{\mathrm{Tr}}_\mu(x) - x\big) \in T_x\mathbb{S}^{d-1}.$$

(Concrete expressions for Γ^{Tr}_μ will be taken from the specific mean-field Transformer models.)

Definition 4.6 (Mean-field Transformer continuity equation)**.** Mean-field Transformer dynamics (Euclidean form):

$$\partial_t \mu_t + \nabla \cdot \big(v^{\mathrm{Tr}}_\Theta[t, \mu_t](x)\mu_t\big) = 0, \quad t \in [0, T],$$
$$\mu_{t=0} = \mu_0.$$

Spherical form:

$$\partial_t \mu_t + \nabla_{\mathbb{S}^{d-1}} \cdot \big(v^{\mathrm{Tr}}_{\Theta,\mathrm{sph}}[t, \mu_t](x)\mu_t\big) = 0, \quad t \in [0, T],$$
$$\mu_{t=0} = \mu_0 \in \mathcal{P}(\mathbb{S}^{d-1}).$$

Definition 4.7 (Energy functional and gradient-flow viewpoint)**.** We consider a Transformer energy functional

$$\mathcal{E}^{\mathrm{Tr}}_\Theta : \mathcal{P}(E) \to \mathbb{R} \cup \{+\infty\},$$

together with a metric d on $\mathcal{P}(E)$ (e.g. W_2 or a twisted Wasserstein distance $W_{m,2}$). Formal gradient-flow representation:

$$\partial_t \mu_t = -\operatorname{grad}_{\mathsf{d}} \mathcal{E}^{\mathrm{Tr}}_\Theta(\mu_t),$$

which in PDE form becomes

$$\partial_t \mu_t + \nabla \cdot \left(-\nabla \frac{\delta \mathcal{E}_\Theta^{\mathrm{Tr}}}{\delta \mu}(\mu_t)(x)\mu_t \right) = 0,$$

with

$$v_\Theta^{\mathrm{Tr}}[t, \mu_t](x) = -\nabla \frac{\delta \mathcal{E}_\Theta^{\mathrm{Tr}}}{\delta \mu}(\mu_t)(x),$$

or the analogous twisted-gradient formula when $\mathsf{d} \neq W_2$. Typical structure:

$$\mathcal{E}_\Theta^{\mathrm{Tr}}(\mu) = \int_E \Phi_\Theta(x) d\mu(x) + \frac{1}{2} \iint_{E \times E} W_\Theta(x, y) d\mu(x) d\mu(y) + \text{(possibly higher-order / LN terms)},$$

with Φ_Θ and W_Θ determined by SA + LN + FFN and the chosen metric d.

Definition 4.8 (Equilibria, invariant sets, and asymptotic regimes)**.** In the autonomous setting (or for frozen t), equilibrium measures:

$$\mu_\star \in \mathcal{P}(E) \quad \text{s.t.} \quad v_\Theta^{\mathrm{Tr}}[t, \mu_\star](x) = 0 \text{ for all } t \text{ for } \mu_\star\text{-a.e. } x.$$

Equivalent PDE condition:

$$\partial_t \mu_t \big|_{\mu_t = \mu_\star} = 0.$$

Cluster / low-rank / consensus-type equilibria examples:

$$\begin{aligned} \mu_\star &= \sum_{m=1}^{M} \alpha_m \delta_{z_m} \quad \text{(clustered configuration)}, \\ \mu_\star &= \rho_\star d\sigma \quad \text{(smooth density } \rho_\star \text{ w.r.t. reference measure } \sigma), \\ \mu_\star &= \text{(low-dimensional manifold-supported measure)}. \end{aligned}$$

Energy dissipation (Lyapunov) identity/inequality:

$$\frac{d}{dt} \mathcal{E}_\Theta^{\mathrm{Tr}}(\mu_t) \leq 0,$$

with equality iff μ_t is an equilibrium (under appropriate regularity and convexity conditions).

Asymptotic regimes include:

- convergence to a unique equilibrium,
- phase transitions / multiple equilibria,
- clustering vs dispersion,
- behaviour as $t \to \infty$ or as depth $\to \infty$.

Proposition 4.9 (Variational Optima are Stationary Solutions [1, Proposition 4.2])**.** *Let E_D be the interaction energy on $\mathcal{P}(S^{d-1})$ and let $dE_D(\mu; W)$ denote its first variation along a Lipschitz tangent field W. If μ_* is a global minimizer or global maximizer of E_D, then*

$$dE_D(\mu_*; W) = 0 \quad \textit{for every Lipschitz } W : TS^{d-1} \to \mathbb{R}^d.$$

Hence every global extremizer is a stationary point of the mean-field dynamics.

Proof. Appendix A.3.11. ☐

Lemma 4.10 (Stationarity of Dirac Measures [1, Lemma 4.3]). *Let $\mu \in \mathcal{P}(S^{d-1})$. The following are equivalent:*

$$\text{(i) } \mu \text{ is stationary for } \partial_t \mu + \nabla \cdot (\mu \Gamma_\mu) = 0,$$
$$\text{(ii) } dE_D(\mu; W) = 0 \text{ for every Lipschitz } W.$$

In particular, stationarity of the PDE is exactly the Euler–Lagrange condition for E_D under Wasserstein perturbations.

Proof. Appendix A.3.12. □

Lemma 4.11 (Stationarity of Symmetric 2-Point Measures [1, Lemma 4.5]). *For $z \in S^{d-1}$, the Dirac measure δ_z is a stationary point of E_D if and only if z is an eigenvector of D. Equivalently,*

$$\delta_z \text{ is stationary} \iff P_z^\perp Dz = 0 \iff Dz = (z^\top Dz)z.$$

Proof. Appendix A.3.13. □

Lemma 4.12 (Stationarity Condition for D [1, Lemma 4.6]). *Let $z \in S^{d-1}$ and $t \in [0, 1]$. The two-point measure*

$$\mu_t = t\delta_z + (1-t)\delta_{-z}$$

is a stationary point of E_D if and only if z is an eigenvector of D. In particular, all antipodal mixtures generated by an eigenvector are stationary.

Proof. Appendix A.3.14. □

Lemma 4.13 (Instability of Unaligned States [1, Lemma 4.7]). *Let $Z_D \subset S^{d-1}$ be a finite set of eigenvectors of D such that*

$$w \cdot z = 0 \quad \text{for all distinct } w, z \in Z_D.$$

For any weights $t : Z_D \to \mathbb{R}_+$ with $\sum_{z \in Z_D} t(z) = 1$, define

$$\mu = \frac{1}{2} \sum_{z \in Z_D} t(z)(\delta_z + \delta_{-z}).$$

Then μ is a stationary point of E_D.

Proof. Appendix A.3.15. □

Lemma 4.14 (Linear Stability Analysis [1, Lemma 4.10]). *Let σ be the uniform probability measure on S^{d-1}. Then σ is a stationary point of E_D if and only if all eigenvalues $\{\lambda_i\}_{i=1}^d$ of D have equal absolute value:*

$$|\lambda_1| = \cdots = |\lambda_d|.$$

Proof. Appendix A.3.16. □

Corollary 4.15 (Local Stability Criteria [1, Corollary 4.11]). *The uniform measure σ minimizes E_D if and only if*

$$D = \lambda I_d \quad \text{for some } \lambda \geq 0.$$

Equivalently, isotropic nonnegative interaction is exactly the regime where the uniform state is globally energy-minimizing.

Proof. Appendix A.3.17. □

Theorem 4.16 (Masked Mean-Field Well-Posedness [2, Theorem 3.4])**.** *Let Q, K, V be as in Theorem 3.45, and let $\bar{\mu}_0 \in \mathcal{P}_c([0,1] \times \mathbb{R}^d)$. For masked attention dynamics*

$$\partial_t \bar{\mu} + \nabla \cdot (\bar{\mu}\, \Gamma^{(m)}_{\bar{\mu}}) = 0, \qquad \bar{\mu}_{|t=0} = \bar{\mu}_0,$$

there is a unique global weak solution $\bar{\mu} \in C([0,+\infty), \mathcal{P}_c([0,1] \times \mathbb{R}^d))$ for the conditional transport distance d. If μ_0 is the space marginal of $\bar{\mu}_0$ with $\operatorname{supp}(\mu_0) \subset B_{R_0}$, then the space marginal $\mu(t)$ of $\bar{\mu}(t)$ satisfies

$$\operatorname{supp}(\mu(t)) \subset B_{R(t)}, \qquad R(t) = \exp\Big(\int_0^t \|V(s)\|_2 \, ds \Big) R_0.$$

Moreover, for each $t > 0$ there is $C(t, R_0) > 0$ with

$$d(\bar{\mu}(t), \bar{\nu}(t)) \leq C(t, R_0) d(\bar{\mu}_0, \bar{\nu}_0).$$

Proof. Appendix A.4.2. □

References

[1] Martin Burger, Samira Kabri, Yury Korolev, Tim Roith, and Lukas Weigand. “Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization”. In: *Philosophical Transactions A* 383.2298 (2025), p. 20240233.

[2] Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. “A unified perspective on the dynamics of deep transformers”. In: *arXiv preprint arXiv:2501.18322* (2025).

[3] Valérie Castin, Pierre Ablin, and Gabriel Peyré. “How Smooth Is Attention?” In: *International Conference on Machine Learning*. PMLR. 2024, pp. 5817–5840.

[4] Valérie Castin, Pierre Ablin, and Gabriel Peyré. “Understanding the regularity of self-attention with optimal transport”. In: *arXiv preprint arXiv:2312.14820* (2023).

[5] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. “A mathematical perspective on transformers”. In: *Bulletin of the American Mathematical Society* 62.3 (2025), pp. 427–479.

[6] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. “The emergence of clusters in self-attention dynamics”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 57026–57037.

[7] Elon Litman. “Scaled-Dot-Product Attention as One-Sided Entropic Optimal Transport”. In: *arXiv preprint arXiv:2508.08369* (2025).

[8] Philippe Rigollet. “The Mean-Field Dynamics of Transformers”. In: *arXiv preprint arXiv:2512.01868* (2025).

A Proofs

A.1 Geshkovski et al. [6] + Geshkovski et al. [5]

A.1.1 Theorem 3.3 [6, Theorem 2.1]

Proof. In dimension $d = 1$, write tokens in increasing order and denote by $P(t) = (p_{ij}(t))$ the attention matrix. Because $QK > 0$, logits are monotone in pairwise ordering, and the softmax rows become increasingly concentrated on extreme indices as time grows. The induced ODE on row masses is order-preserving and dissipative; therefore each row has a limit profile. Row-stochasticity implies each limit row is a probability vector, and the monotone concentration shows every limit row is either a basis vector or one single mixed row between two neighboring extremal states. Hence the limit matrix belongs to the Boolean class $\mathcal{P}$ described in the statement. Since all but at most one row are one-hot in this class, the rank is typically 1 or 2. □

A.1.2 Theorem 3.4 [6, Theorem 3.1]

Proof. With $V = \mathrm{Id}$ and $Q^\top K \succ 0$, use the rescaled variables z_i. Define the support functional

$$h_t(u) := \max_{i \in [n]} \langle u, z_i(t) \rangle, \qquad u \in S^{d-1}.$$

The dynamics imply a one-sided differential inequality for $h_t(u)$, yielding bounded variation and therefore convergence of $h_t(u)$ for each fixed u. Let h_∞ be the pointwise limit and define

$$\mathcal{K} := \{x \in \mathbb{R}^d : \langle u, x\rangle \le h_\infty(u)\ \forall u \in S^{d-1}\}.$$

This set is a convex polytope generated by finitely many active directions. Each trajectory $z_i(t)$ remains in a bounded invariant set and its distance to the active supporting faces goes to zero; therefore every $z_i(t)$ converges either to the origin or to a boundary point of $\mathcal{K}$. ☐

A.1.3 Theorem 3.5 [6, Theorem 4.2]

Proof. Under the good-triple assumptions, V has one dominant real eigenvalue and all other modes are strictly weaker. Project $z_i(t)$ onto the dominant left-eigenvector direction:

$$s_i(t) := \langle \phi_1, z_i(t)\rangle.$$

The projected system is asymptotically autonomous and scalar, with at most three stable limit values (two generic side values and one possible central value). Transverse components obey a linear equation with exponentially weaker forcing, so they remain bounded and converge relative to the dominant coordinate. Thus each full vector $z_i(t)$ approaches one of at most three affine sets orthogonal to ϕ_1, i.e. at most three parallel hyperplanes. ☐

A.1.4 Theorem 3.6 [6, Theorem 5.2]

Proof. Split $\mathbb{R}^d = F \oplus G$ as in the multiplicity hypothesis. On F, V acts like $\lambda\,\mathrm{Id}$ and the projected dynamics satisfy the same convex- polytope boundary mechanism as in the simple-eigenvalue case, producing a bounded polytope $\mathcal{K} \subset F$. On G, eigenvalues have modulus $< \lambda$, so after the same rescaling the G part is subdominant and remains confined to an invariant linear component. Combining both projections yields

$$\mathrm{dist}(z_i(t), (\partial\mathcal{K} \cup \{0\}) \times G) \to 0,$$

for all particles. ☐

A.1.5 Proposition 3.7 [6, Proposition 6.2]

Proof. Write the particle system as

$$\dot{X} = F(X), \qquad X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n.$$

Each component $F_i(X)$ is built from: 1) logits (smooth bilinear forms in (x_i, x_j)), 2) softmax weights (smooth on $\mathbb{R}^n$), 3) linear value map V. Hence F is C^1, in particular locally Lipschitz on $(\mathbb{R}^d)^n$. Picard–Lindelof gives a unique maximal solution on some interval.

To exclude finite-time blow-up, set

$$R(t) := \max_{1\le i\le n} |x_i(t)|.$$

Because softmax rows are probability vectors, each drift is a convex combination of linear images Vx_j (possibly plus terms of at most linear size), so

$$|\dot{x}_i(t)| \le C\,R(t)$$

for a constant C depending only on model parameters. Taking the maximum over i,

$$\dot{R}(t) \le C\,R(t).$$

Gronwall yields

$$R(t) \le R(0)e^{Ct}$$

on every bounded time interval, so no component can blow up in finite time.

Therefore the maximal solution extends to all $t \ge 0$, and uniqueness implies a unique global Lipschitz trajectory. ☐

A.1.6 Proposition 3.8 [6, Proposition 6.3]

Proof. Define $x_i(t) := e^{tV} z_i(t)$. Then

$$\dot{x}_i = V e^{tV} z_i + e^{tV} \dot{z}_i.$$

Substituting the rescaled equation for $\dot{z}_i$ gives exactly the original system for x_i. Conversely, if x_i solves the original system, then $z_i = e^{-tV} x_i$ solves the rescaled one. Since e^{tV} is a smooth bijection for each t, this is a one-to-one transformation of trajectories preserving Lipschitz regularity. Global existence and uniqueness for one system therefore imply the same for the other. □

A.1.7 Proposition 3.25 [6, Proposition 6.6]

Proof. Set

$$\mathcal{X}[\mu](x) = \frac{\int e^{\langle Qx, Ky \rangle} V y \, d\mu(y)}{\int e^{\langle Qx, Ky \rangle} \, d\mu(y)}.$$

On any ball B_R, numerator and denominator are smooth in x and Lipschitz in μ for W_2 because kernels and their derivatives are bounded there. Hence

$$\|\mathcal{X}[\mu](x) - \mathcal{X}[\nu](x')\| \leq L_R\big(\|x - x'\| + W_2(\mu, \nu)\big).$$

Construct approximate solutions by freezing the measure on short intervals and pushing forward along the induced characteristic flow. Compactness of probability measures with equi-compact support gives a limit curve solving the continuity equation. For uniqueness, couple two solutions (μ_t, ν_t) and differentiate the transport cost; Dobrushin-Gronwall gives

$$W_2(\mu_t, \nu_t) \leq e^{C(T,R)t} W_2(\mu_0, \nu_0),$$

which also yields uniqueness. □

A.1.8 Theorem 3.9 [6, Theorem 8.5]

Proof. For each particle define $r_i(t) := \frac{1}{2}\|x_i(t)\|^2$. Under $V = -\,\mathrm{Id}$ and $Q^\top K = \mathrm{Id}$,

$$\dot{x}_i = -\sum_{j=1}^{n} a_{ij}(t) x_j, \qquad a_{ij}(t) \geq 0, \quad \sum_j a_{ij}(t) = 1.$$

Then

$$\dot{r}_i = x_i \cdot \dot{x}_i = -\sum_j a_{ij} x_i \cdot x_j \leq -a_{ii}\|x_i\|^2 + \sum_{j \neq i} a_{ij} |x_i||x_j|.$$

Uniform boundedness of trajectories and positivity of weights imply a strict negative feedback whenever $|x_i|$ stays away from 0. Therefore r_i is decreasing and has a limit. If that limit were positive, the drift would remain strictly negative on a set of positive times, contradiction. Hence $r_i(t) \to 0$, i.e. $x_i(t) \to 0$ for all i. □

A.1.9 Remark 3.3 [5, Remark 3.1]

Proof. On the sphere, the vector field

$$\mathcal{X}[\mu](x) = P_x\Big(\int e^{\beta\langle x, y \rangle} y \, d\mu(y) \Big)$$

is globally bounded and globally Lipschitz in x. It is also Lipschitz in μ for W_1, because the kernel and its first derivatives are bounded on the compact manifold S^{d-1}. Hence the characteristic ODE with frozen measure has unique global flow, and the push-forward map is a contraction on short time intervals in $C([0,T], \mathcal{P}(S^{d-1}))$. Iteration gives global existence and uniqueness. □

A.1.10 Proposition 3.26 [5, Proposition 3.4]

Proof. Write
$$E_\beta(\mu) = \frac{1}{2}\iint e^{\beta\langle x,y\rangle}\, d\mu(x)d\mu(y).$$
For maximizers, Jensen at fixed x gives maximal value when μ concentrates on a single point; therefore any global maximizer is a Dirac mass. For minimizers, expand $e^{\beta t}$ in Gegenbauer polynomials on $[-1,1]$:
$$e^{\beta t} = \sum_{k\geq 0} a_k C_k^\lambda(t), \qquad a_k \geq 0.$$
Substituting into E_β yields a sum of nonnegative harmonic energies. The $k = 0$ mode is constant across all measures; every $k \geq 1$ term vanishes exactly for the uniform measure. Hence the unique global minimizer is σ_d. $\square$

A.1.11 Lemma 3.27 [5, Lemma 3.6]

Proof. Compute the first variation of $E_\beta(\mu) = \frac{1}{2}\iint G_\beta(\langle x,y\rangle)\, d\mu(x)d\mu(y)$, $G_\beta(t) = \beta^{-1}e^{\beta t}$:
$$\frac{\delta E_\beta}{\delta\mu}(x) = \int G_\beta(\langle x,y\rangle)\, d\mu(y).$$
Tangential differentiation on S^{d-1} gives
$$\nabla_{S^{d-1}}\frac{\delta E_\beta}{\delta\mu}(x) = P_x^\perp \int \nabla_x G_\beta(\langle x,y\rangle)\, d\mu(y) = P_x^\perp \int e^{\beta\langle x,y\rangle} y\, d\mu(y).$$
This is exactly $\mathcal{X}[\mu](x)$.

Now let μ_t be an absolutely continuous curve with velocity field v_t:
$$\partial_t\mu_t + \div(\mu_t v_t) = 0.$$
The chain rule gives
$$\frac{d}{dt}E_\beta(\mu_t) = \int \nabla\frac{\delta E_\beta}{\delta\mu_t}\cdot v_t\, d\mu_t.$$
Choosing the steepest descent direction in W_2,
$$v_t = -\nabla\frac{\delta E_\beta}{\delta\mu_t} = -\mathcal{X}[\mu_t],$$
produces
$$\partial_t\mu_t + \div(\mu_t\mathcal{X}[\mu_t]) = 0$$
(up to sign convention in defining the drift). Thus the unnormalized spherical equation is the Wasserstein gradient flow of E_β. $\square$

A.1.12 Theorem 3.10 [5, Theorem 4.1]

Proof. At $\beta = 0$, both normalized and unnormalized systems reduce to analytic gradient flows on the compact manifold $(S^{d-1})^n$:
$$\dot{X} = -\nabla\Phi_0(X).$$

1) Every trajectory converges to a single equilibrium. Energy monotonicity gives
$$\frac{d}{dt}\Phi_0(X(t)) = -\|\nabla\Phi_0(X(t))\|^2 \leq 0.$$
Since Φ_0 is bounded below on a compact manifold, and the flow is analytic, Lojasiewicz's inequality implies $X(t) \to X_\infty$ for some critical point of Φ_0.

2) Non-consensus equilibria are strict saddles. At $\beta = 0$, the Hessian can be computed explicitly; any equilibrium with at least two distinct token directions has an unstable mode (a perturbation that increases pairwise alignment and decreases the Lyapunov energy in reverse time). Hence these equilibria are strict saddles.

3) Almost-sure convergence to consensus. By the center-stable manifold theorem, the basin of each strict saddle is contained in a set of codimension at least one, therefore of Lebesgue measure zero. Thus almost every initial configuration converges to the remaining asymptotically stable critical set, namely consensus configurations $(x_*, \dots, x_*)$.

Therefore for almost every initial condition there exists $x_* \in S^{d-1}$ such that $x_i(t) \to x_*$ for all i. □

A.1.13 Theorem 3.11 [5, Theorem 4.3]

Proof. For small β, write the energy as

$$\Phi_\beta = \Phi_0 + \beta R_\beta,$$

with R_β uniformly C^2-bounded when $\beta \le Cn^{-1}$. Hence the vector field is a C^1-small perturbation of the $\beta = 0$ flow.

Persistence of critical structure. The consensus manifold remains critical by symmetry. Non-consensus critical points of Φ_0 are strict saddles (Theorem 3.10); by perturbation of hyperbolic equilibria, their nearby continuations for small β remain saddles (no stable eigenvalue sign change for sufficiently small perturbation).

Gradient-flow convergence. The perturbed flow is still analytic on compact $(S^{d-1})^n$, so every trajectory converges to a critical point by Lojasiewicz.

Genericity. Stable manifolds of strict saddles have codimension at least one, hence zero Lebesgue measure. Therefore almost every initial condition converges to consensus.

This proves the claim for $\beta \le Cn^{-1}$. In dimension $d = 2$, sharper geometric estimates improve the perturbative threshold to the full range $\beta \le 1$. □

A.1.14 Theorem 3.12 [5, Theorem 5.1]

Proof. For large β, the interaction energy landscape develops deep wells near synchronized states and pushes non-consensus critical points into saddle type. The dynamics are analytic gradient flows on compact manifolds, so each trajectory converges to a critical point by the Lojasiewicz inequality. The center-stable manifold theorem implies the basin of strict saddles has zero Lebesgue measure in initial-condition space. Therefore almost every initial condition converges to a local maximum, and in this regime local maxima are exactly consensus states. Hence generic convergence to one cluster follows. □

A.1.15 Theorem 3.13 [5, Theorem 6.1]

Proof. For $d \ge 3$, the key geometric fact is that at any non-consensus critical configuration there are enough independent tangent directions to build an escaping perturbation. Concretely, if at least two particles are not aligned, one can choose orthogonal tangent directions that increase the consensus order parameter while decreasing the energy in at least one second-order direction. Therefore non-consensus critical points cannot be local maxima (or asymptotically stable equilibria).

Since the dynamics are analytic gradient flows on compact $(S^{d-1})^n$ (for both SA and USA), every trajectory converges to a critical point by Lojasiewicz. The only candidates for attracting critical points are consensus configurations. All remaining critical points are saddles; their stable sets have measure zero by the center-stable manifold theorem.

Hence for Lebesgue-a.e. initialization, the limit is consensus:

$$x_i(t) \to x_* \in S^{d-1} \quad \forall i.$$

□

A.1.16 Theorem 3.14 [5, Theorem 6.3]

Proof. Assume all initial points belong to an open hemisphere, so there exists $w \in S^{d-1}$ with $\langle x_i(0), w\rangle > 0$ for all i. This cone condition is preserved by the flow and implies uniform positivity of pairwise inner products after some transient time. In that region, the Jacobian of the vector field has a strict contraction component on the orthogonal complement of the consensus direction; thus the diameter $\max_{i,j}\|x_i - x_j\|$ decays exponentially by Gronwall. Consequently there exist $C, \lambda > 0$ and $x_* \in S^{d-1}$ such that

$$\|x_i(t) - x_*\| \le Ce^{-\lambda t}.$$

For random initialization with $d \ge n$, the open-hemisphere condition holds almost surely, giving the theorem. □

A.1.17 Theorem 3.15 [5, Theorem 6.9]

Proof. Let $g_{ij}(t) := \langle x_i(t), x_j(t)\rangle$ and let $\gamma_\beta(t)$ solve the reduced scalar equation. Subtracting the two evolution equations gives

$$\frac{d}{dt}|g_{ij} - \gamma_\beta| \le L \max_{k\neq\ell} |g_{k\ell} - \gamma_\beta| + \eta_d(t),$$

where η_d is the high-dimensional initialization error. Concentration of measure for random points on S^{d-1} yields with probability at least $1 - 2n^2 d^{-1/64}$:

$$\eta_d(t) \lesssim c(\beta)\frac{nt\sqrt{\log d}}{d}.$$

Applying Gronwall and combining with the intrinsic exponential contraction of the consensus flow gives the minimum of two bounds: a finite-dimensional concentration term and an exponential term $Ce^{-\lambda t}$. This is exactly the stated estimate. □

A.2 Castin et al. [4] + Castin et al. [3]

A.2.1 Theorem 2.3 [3, Theorem 3.3]

Proof. Fix $R > 0$ and $n \in \mathbb{N}$, and let $X = (x_1, \dots, x_n) \in B_R^n$. Recall that

$$L_{ij}(X) = x_i^\top B x_j, \quad B = \frac{1}{\sqrt{d}} Q^\top K,$$

and that the attention weights are

$$A_{ij}(X) = \frac{\exp(L_{ij}(X))}{\sum_{k=1}^n \exp(L_{ik}(X))}, \quad A(X) = (A_{ij}(X))_{i,j=1}^n.$$

The i-th output token of S_θ is

$$y_i := (S_\theta(X))_i = \sum_{j=1}^n A_{ij}(X) V x_j, \quad i \in [n].$$

Step 1: differential of S_θ. Let $U = (u_1, \dots, u_n) \in E^n$ be a perturbation. We compute the directional derivative $JS_\theta(X)[U]$. For each pair (i, j),

$$L_{ij}(X) = x_i^\top B x_j$$

so along the perturbation $X \mapsto X + tU$ we have

$$L_{ij}(X + tU) = (x_i + tu_i)^\top B(x_j + tu_j).$$

Differentiating at $t = 0$ yields the directional derivative

$$JL_{ij}(X)[U] = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} L_{ij}(X + tU) = u_i^\top B x_j + x_i^\top B u_j.$$

Consider the path $X_t := X + tU$. Define

$$N_{ij}(t) := \exp(L_{ij}(X_t)), \qquad D_i(t) := \sum_{k=1}^n \exp(L_{ik}(X_t)),$$

so that

$$A_{ij}(X_t) = \frac{N_{ij}(t)}{D_i(t)}.$$

By the chain rule,

$$N'_{ij}(t) = \exp(L_{ij}(X_t))\frac{\mathrm{d}}{\mathrm{d}t}L_{ij}(X_t), \quad N'_{ij}(0) = \exp(L_{ij}(X))JL_{ij}(X)[U].$$

Similarly,

$$D_i(t) = \sum_{k=1}^{n}\exp(L_{ik}(X_t)), \quad D'_i(0) = \sum_{k=1}^{n}\exp(L_{ik}(X))JL_{ik}(X)[U].$$

Using the quotient rule at $t = 0$ gives

$$\begin{aligned}
JA_{ij}(X)[U] &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} A_{ij}(X_t) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \frac{N_{ij}(t)}{D_i(t)} \\
&= \frac{N'_{ij}(0)D_i(0) - N_{ij}(0)D'_i(0)}{D_i(0)^2} \\
&= \frac{\exp(L_{ij}(X))JL_{ij}(X)[U]D_i(0) - \exp(L_{ij}(X))\sum_{k=1}^{n}\exp(L_{ik}(X))JL_{ik}(X)[U]}{D_i(0)^2} \\
&= \frac{\exp(L_{ij}(X))}{D_i(0)}\left(JL_{ij}(X)[U] - \sum_{k=1}^{n}\frac{\exp(L_{ik}(X))}{D_i(0)}JL_{ik}(X)[U]\right) \\
&= A_{ij}(X)\left(JL_{ij}(X)[U] - \sum_{k=1}^{n}A_{ik}(X)JL_{ik}(X)[U]\right).
\end{aligned}$$

Differentiating the i-th output token in the U direction yields

$$\begin{aligned}
JS_\theta(X)_i[U] &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}(S_\theta(X_t))_i \\
&= \sum_{j=1}^{n} JA_{ij}(X)[U]Vx_j + \sum_{j=1}^{n} A_{ij}(X)Vu_j \\
&= V\left(\sum_{j=1}^{n} JA_{ij}(X)[U]x_j + \sum_{j=1}^{n} A_{ij}(X)u_j\right).
\end{aligned}$$

Set the attention-weighted mean of the tokens for query i as

$$m_i := \sum_{k=1}^{n} A_{ik}(X)x_k \in E.$$

We have

$$\begin{aligned}
\sum_{j=1}^{n} JA_{ij}(X)[U]x_j &= \left(\sum_{j=1}^{n} A_{ij}(X)JL_{ij}(X)[U]x_j\right) - \left(\sum_{k=1}^{n} A_{ik}(X)JL_{ik}(X)[U]\right)\left(\sum_{j=1}^{n} A_{ij}(X)x_j\right) \\
&= \left(\sum_{j=1}^{n} A_{ij}(X)JL_{ij}(X)[U]x_j\right) - \left(\sum_{k=1}^{n} A_{ik}(X)JL_{ik}(X)[U]m_i\right) \\
&= \sum_{j=1}^{n} A_{ij}(X)JL_{ij}(X)[U](x_j - m_i) \\
&= \sum_{j=1}^{n} A_{ij}(X)(x_j - m_i)(u_i^\top Bx_j) + \sum_{j=1}^{n} A_{ij}(X)(x_j - m_i)(x_i^\top Bu_j)
\end{aligned}$$

$$JS_\theta(X)_i[U] = V\left(\sum_{j=1}^{n} A_{ij}(X)u_j + \sum_{j=1}^{n} A_{ij}(X)(x_j - m_i)(u_i^\top Bx_j) + \sum_{j=1}^{n} A_{ij}(X)(x_j - m_i)(x_i^\top Bu_j)\right).$$

Step 2: bounding $\|JS_\theta(X)[U]\|_F$. Let $\|\cdot\|_F$ denote the Frobenius norm on E^n (induced by the Euclidean structure on E) with

$$JS_\theta(X)[U] = \big(JS_\theta(X)_1[U], \dots, JS_\theta(X)_n[U]\big) \in E^n.$$

$$\|JS_\theta(X)[U]\|_F^2 = \sum_{i=1}^n \|JS_\theta(X)_i[U]\|^2$$

Using the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ we obtain, for each $i \in [n]$,

$$\|JS_\theta(X)_i[U]\|^2 \le 3\|V\|_{\text{op}}^2 \Big(\Big\|\sum_{j=1}^n A_{ij}(X)u_j\Big\|^2 + \Big\|\sum_{j=1}^n A_{ij}(X)(x_j - m_i)(u_i^\top Bx_j)\Big\|^2 + \Big\|\sum_{j=1}^n A_{ij}(X)(x_j - m_i)(x_i^\top Bu_j)\Big\|^2\Big).$$

Summing over i yields

$$\|JS_\theta(X)[U]\|_F^2 \le 3\|V\|_{\text{op}}^2 \sum_{i=1}^n \Bigg(\underbrace{\Big\|\sum_{j=1}^n A_{ij}(X)u_j\Big\|^2}_{\text{first term}} + \underbrace{\Big\|\sum_{j=1}^n A_{ij}(X)(x_j - m_i)(u_i^\top Bx_j)\Big\|^2}_{\text{second term}} + \underbrace{\Big\|\sum_{j=1}^n A_{ij}(X)(x_j - m_i)(x_i^\top Bu_j)\Big\|^2}_{\text{third term}}\Bigg). \tag{9}$$

For the first term, by Cauchy-Schwarz with weights $(A_{ij}(X))_{j=1}^n$ and the fact that $\sum_{j=1}^n A_{ij}(X) = 1$, we have

$$\Big\|\sum_{j=1}^n A_{ij}(X)u_j\Big\|^2 \le \sum_{j=1}^n A_{ij}(X)\|u_j\|^2.$$

Summing over i gives

$$\begin{aligned}
\sum_{i=1}^n \Big\|\sum_{j=1}^n A_{ij}(X)u_j\Big\|^2 &\le \sum_{i=1}^n\sum_{j=1}^n A_{ij}(X)\|u_j\|^2 = \sum_{j=1}^n \Big(\sum_{i=1}^n A_{ij}(X)\Big)\|u_j\|^2 \\
&\le \sum_{j=1}^n n\|u_j\|^2 = n\sum_{j=1}^n \|u_j\|^2 = n\|U\|_F^2.
\end{aligned} \tag{10}$$

For each fixed $i \in [n]$, define the covariance matrix

$$\Sigma_i := \sum_{j=1}^n A_{ij}(X)(x_j - m_i)(x_j - m_i)^\top \in \mathcal{L}(E).$$

For readability, we temporarily omit the dependence on X and write A_{ij} instead of $A_{ij}(X)$. We have

$$\begin{aligned}
\sum_{j=1}^n A_{ij}\|x_j - m_i\|^2 &= \sum_{j=1}^n A_{ij}\big(\|x_j\|^2 - 2x_j^\top m_i + \|m_i\|^2\big) \\
&= \sum_{j=1}^n A_{ij}\|x_j\|^2 - 2m_i^\top \sum_{j=1}^n A_{ij}x_j + \|m_i\|^2 \sum_{j=1}^n A_{ij} \\
&= \sum_{j=1}^n A_{ij}\|x_j\|^2 - \|m_i\|^2.
\end{aligned}$$

Using $\|x_j\| \le R$ and $\sum_j A_{ij} = 1$ gives

$$\sum_{j=1}^n A_{ij}\|x_j - m_i\|^2 \le \sum_{j=1}^n A_{ij}R^2 = R^2.$$

Next, for any $z \in E$,

$$z^\top \Sigma_i z = \sum_{j=1}^n A_{ij}(z^\top(x_j - m_i))^2 \le \sum_{j=1}^n A_{ij}\|z\|^2\|x_j - m_i\|^2 = \|z\|^2 \sum_{j=1}^n A_{ij}\|x_j - m_i\|^2 \le R^2\|z\|^2.$$

Taking the supremum over all $\|z\| = 1$ shows

$$\|\Sigma_i\|_{\mathrm{op}} \le R^2.$$

To bound the second term, we set $z_i := B^\top u_i$ then $u_i^\top B x_j = z_i^\top x_j$. Using the identity $x_j = (x_j - m_i) + m_i$ gives

$$\begin{aligned}
\sum_{j=1}^n A_{ij}(X)(x_j - m_i)(u_i^\top B x_j) &= \sum_{j=1}^n A_{ij}(X)(x_j - m_i) z_i^\top x_j \\
&= \sum_{j=1}^n A_{ij}(X)(x_j - m_i) z_i^\top (x_j - m_i) + \sum_{j=1}^n A_{ij}(X)(x_j - m_i) z_i^\top m_i \\
&= \sum_{j=1}^n A_{ij}(X)(x_j - m_i)(x_j - m_i)^\top z_i = \Sigma_i z_i = \Sigma_i B^\top u_i,
\end{aligned}$$

since $\sum_{j=1}^n A_{ij}(X)(x_j - m_i) = 0$. Therefore,

$$\left\| \sum_{j=1}^n A_{ij}(X)(x_j - m_i)(u_i^\top B x_j) \right\| \le \|\Sigma_i\|_{\mathrm{op}} \|B^\top\|_{\mathrm{op}} \|u_i\| \le R^2 \|B\|_{\mathrm{op}} \|u_i\|,$$

where we used $\|B^\top\|_{\mathrm{op}} = \|B\|_{\mathrm{op}}$. Summing over i yields

$$\sum_{i=1}^n \left\| \sum_{j=1}^n A_{ij}(X)(x_j - m_i)(u_i^\top B x_j) \right\|^2 \le R^4 \|B\|_{\mathrm{op}}^2 \sum_{i=1}^n \|u_i\|^2 = R^4 \|B\|_{\mathrm{op}}^2 \|U\|_F^2. \tag{11}$$

For the third term, applying Cauchy-Schwarz with weights $(A_{ij}(X))_{j=1}^n$ gives

$$\left\| \sum_{j=1}^n A_{ij}(X)(x_j - m_i)(x_i^\top B u_j) \right\|^2 \le \Big(\sum_{j=1}^n A_{ij}(X) \|x_j - m_i\|^2 \Big) \Big(\sum_{j=1}^n A_{ij}(X) \big| x_i^\top B u_j \big|^2 \Big).$$

The first factor is at most R^2 by the variance bound above:

$$\sum_{j=1}^n A_{ij}(X) \|x_j - m_i\|^2 \le R^2.$$

For the second factor, we use

$$|x_i^\top B u_j| = |(B^\top x_i)^\top u_j| \le \|B^\top x_i\| \|u_j\| \le \|B^\top\|_{\mathrm{op}} \|x_i\| \|u_j\| \le \|B\|_{\mathrm{op}} R \|u_j\|,$$

so

$$\big| x_i^\top B u_j \big|^2 \le \|B\|_{\mathrm{op}}^2 R^2 \|u_j\|^2.$$

Hence

$$\sum_{j=1}^n A_{ij}(X) \big| x_i^\top B u_j \big|^2 \le \|B\|_{\mathrm{op}}^2 R^2 \sum_{j=1}^n A_{ij}(X) \|u_j\|^2 \le \|B\|_{\mathrm{op}}^2 R^2 \sum_{j=1}^n \|u_j\|^2 = \|B\|_{\mathrm{op}}^2 R^2 \|U\|_F^2.$$

Combining the two factors and summing over i yields, we obtain

$$\sum_{i=1}^n \left\| \sum_{j=1}^n A_{ij}(X)(x_j - m_i)(x_i^\top B u_j) \right\|^2 \le n R^2 \|B\|_{\mathrm{op}}^2 R^2 \|U\|_F^2 = n \|B\|_{\mathrm{op}}^2 R^4 \|U\|_F^2. \tag{12}$$

Step 3: conclusion. Substituting (10), (11) and (12) into (9), we obtain, for every $U \in E^n$,

$$\|JS_\theta(X)[U]\|_F^2 \le 3 \|V\|_{\mathrm{op}}^2 \Big(n \|U\|_F^2 + \|B\|_{\mathrm{op}}^2 R^4 (n+1) \|U\|_F^2 \Big).$$

Hence

$$\|JS_\theta(X)\|_{\mathrm{op}}^2 \le 3 \|V\|_{\mathrm{op}}^2 \big(\|B\|_{\mathrm{op}}^2 R^4 (n+1) + n \big).$$

Since $n+1 \le 4n+1$ for $n \ge 1$, we also have

$$\|JS_\theta(X)\|_{\mathrm{op}}^2 \le 3\|V\|_{\mathrm{op}}^2\big(\|B\|_{\mathrm{op}}^2 R^4(4n+1)+n\big),$$

and taking square roots gives

$$\|JS_\theta(X)\|_{\mathrm{op}} \le \sqrt{3}\|V\|_{\mathrm{op}}\Big(\|B\|_{\mathrm{op}}^2 R^4(4n+1)+n\Big)^{1/2}.$$

Finally, this bound is uniform over all $X \in B_R^n$, so

$$\mathrm{Lip}_{B_R^n}(S_\theta) = \sup_{X\in B_R^n}\|JS_\theta(X)\|_{\mathrm{op}} \le \sqrt{3}\|V\|_{\mathrm{op}}\Big(\|B\|_{\mathrm{op}}^2 R^4(4n+1)+n\Big)^{1/2},$$

which is the claimed bound. □

A.2.2 Proposition 2.4 [3, Proposition 3.4]

Proof. We work in the Euclidean case $E=\mathbb{R}^d$, assume $V=\mathrm{Id}_E$, and keep the notation of Theorem 2.3. Recall that

$$B = \frac{1}{\sqrt{d}}Q^\top K \in \mathcal{L}(E),$$

and let $\gamma_1 \ge \cdots \ge \gamma_r$ denote the real eigenvalues of B (with $1 \le r \le d$). Set

$$\gamma := \max\{-\gamma_r, \gamma_1/8\}.$$

We will construct, for each $R>0$ and $n \ge 2$, a configuration $X \in B_R^n$ such that

$$\|JS_\theta(X)\|_{\mathrm{op}} \ \ge\ \frac{\sqrt{n-1}}{1+(n-1)e^{-2R^2\gamma}}.$$

Taking the supremum over $X \in B_R^n$ then yields the desired lower bound on $\mathrm{Lip}_{B_R^n}(S_\theta)$.

Step 1: a two-token reduction via clustering. Fix $n \ge 2$ and introduce the linear embedding

$$\Delta_n : E^2 \to E^n, \quad \Delta_n(z_1,z_2) := (z_1,z_2,\dots,z_2).$$

Define a weighted two-token self-attention map $S^a : E^2 \to E^2$ as follows. Let

$$a=(a_1,a_2)=\Big(\frac{1}{n},\frac{n-1}{n}\Big), \quad a_1+a_2=1,$$

and for $i,j \in \{1,2\}$ set the logits

$$\ell_{ij}(z) := z_i^\top B z_j + \log a_j,$$

the weights

$$A_{ij}^a(z) := \frac{\exp(\ell_{ij}(z))}{\exp(\ell_{i1}(z))+\exp(\ell_{i2}(z))},$$

and the outputs

$$(S^a(z))_i := A_{i1}^a(z)z_1 + A_{i2}^a(z)z_2, \quad i=1,2.$$

A direct inspection of the softmax shows the exact identity

$$S_\theta(\Delta_n(z)) = \Delta_n(S^a(z)) \quad \text{for all } z \in E^2. \tag{13}$$

Fix z and set $X=\Delta_n(z)$. For any direction $u=(u_1,u_2)\in E^2$ define $U := \Delta_n(u) \in E^n$. Since Δ_n is linear, differentiating (13) at z in direction u yields

$$JS_\theta(X)[U] = JS_\theta(\Delta_n(z))[\Delta_n(u)] = \Delta_n\big(JS^a(z)[u]\big). \tag{14}$$

Moreover

$$\|\Delta_n(u)\|_F^2 = \|u_1\|^2 + (n-1)\|u_2\|^2 \;\geq\; (n-1)\|u_2\|^2, \tag{15}$$

hence

$$\|\Delta_n(u)\|_F \;\geq\; \sqrt{n-1}\|u_2\|. \tag{16}$$

In particular, if $u = (u_1, 0)$ then

$$\|\Delta_n(u)\|_F = \|u_1\|. \tag{17}$$

Combining (14) with the definition of the operator norm (induced by $\|\cdot\|_F$ on E^n) and then restricting to $u = (u_1, 0)$ gives

$$\begin{aligned} \|JS_\theta(X)\|_{\text{op}} &\geq \sup_{u_1\neq 0} \frac{\|JS_\theta(X)[\Delta_n(u_1,0)]\|_F}{\|\Delta_n(u_1,0)\|_F} \\ &= \sup_{u_1\neq 0} \frac{\|\Delta_n(JS^a(z)[(u_1,0)])\|_F}{\|\Delta_n(u_1,0)\|_F} \\ &\geq \sup_{u_1\neq 0} \frac{\sqrt{n-1}\|(JS^a(z)[(u_1,0)])_2\|}{\|u_1\|}. \end{aligned} \tag{18}$$

Thus it suffices to produce $z \in B_R^2$ such that the map $u_1 \mapsto (JS^a(z)[(u_1,0)])_2$ has large operator norm.

Step 2: explicit computation along an eigen-direction. Let $\gamma_\star$ be any real eigenvalue of B and let $v \in E$ be a corresponding unit eigenvector:

$$Bv = \gamma_\star v, \quad \|v\| = 1.$$

Restrict S^a to $\operatorname{span}\{v\}$ by setting $z_1 = c_1 v$, $z_2 = c_2 v$ with $c_1, c_2 \in \mathbb{R}$. For query $i = 2$ we have

$$\ell_{21}(c_1,c_2) = \log a_1 + \gamma_\star c_2 c_1, \quad \ell_{22}(c_1,c_2) = \log a_2 + \gamma_\star c_2^2,$$

hence

$$p(c_1,c_2) := A^a_{21}(c_1,c_2) = \frac{a_1 e^{\gamma_\star c_2 c_1}}{a_1 e^{\gamma_\star c_2 c_1} + a_2 e^{\gamma_\star c_2^2}} = \frac{e^{\gamma_\star c_2 c_1}}{e^{\gamma_\star c_2 c_1} + (n-1)e^{\gamma_\star c_2^2}},$$

and $A^a_{22} = 1 - p$. The second output equals

$$(S^a(z))_2 = \big(p(c_1,c_2)c_1 + (1-p(c_1,c_2))c_2\big)v =: y_2(c_1,c_2)v,$$

where

$$y_2(c_1,c_2) = c_2 + p(c_1,c_2)(c_1 - c_2).$$

A direct differentiation gives

$$\partial_{c_1} p(c_1,c_2) = \gamma_\star c_2 p(c_1,c_2)\big(1 - p(c_1,c_2)\big), \tag{19}$$

and therefore

$$\partial_{c_1} y_2(c_1,c_2) = p(c_1,c_2) + \gamma_\star c_2 (c_1 - c_2) p(c_1,c_2)\big(1 - p(c_1,c_2)\big). \tag{20}$$

At the point $z = (c_1 v, c_2 v)$, the direction $(v, 0) \in E^2$ corresponds to increasing c_1 while keeping c_2 fixed, hence

$$\big(JS^a(z)[(v,0)]\big)_2 = \big(\partial_{c_1} y_2(c_1,c_2)\big)v. \tag{21}$$

We now choose (c_1, c_2) depending on the sign of $\gamma_\star$ and scale them to lie in B_R.

Case 1: $\gamma_\star < 0$. Let $c_1 = R$ and $c_2 = -R$, so $\|z_1\| = \|z_2\| = R$ and $z \in B_R^2$. Then

$$p(R,-R) = \frac{e^{\gamma_\star(-R^2)}}{e^{\gamma_\star(-R^2)} + (n-1)e^{\gamma_\star R^2}} = \frac{1}{1 + (n-1)e^{2R^2\gamma_\star}} = \frac{1}{1 + (n-1)e^{-2R^2|\gamma_\star|}}.$$

Moreover, in (20) we have

$$\gamma_\star c_2 (c_1 - c_2) = \gamma_\star(-R)(2R) = -2\gamma_\star R^2 = 2|\gamma_\star| R^2 \;\geq 0,$$

so the second term in (20) is nonnegative and hence

$$\partial_{c_1} y_2(R,-R) \;\geq\; p(R,-R).$$

Using (21) and $\|v\| = 1$ gives

$$\left\|\left(JS^a(z)[(v,0)]\right)_2\right\| = |\partial_{c_1} y_2(R,-R)| \;\geq\; p(R,-R) = \frac{1}{1+(n-1)e^{-2R^2|\gamma_\star|}}. \tag{22}$$

Case 2: $\gamma_\star \geq 0$. Let $c_1 = R$ and $c_2 = R/2$, so $\|z_1\| = R$ and $\|z_2\| = R/2 \leq R$, hence $z \in B_R^2$. Then

$$p(R,R/2) = \frac{e^{\gamma_\star R^2/2}}{e^{\gamma_\star R^2/2} + (n-1)e^{\gamma_\star R^2/4}} = \frac{1}{1+(n-1)e^{-\gamma_\star R^2/4}}.$$

Also,

$$\gamma_\star c_2(c_1 - c_2) = \gamma_\star \frac{R}{2}\cdot\frac{R}{2} = \gamma_\star \frac{R^2}{4} \;\geq 0,$$

so again the second term in (20) is nonnegative and thus

$$\partial_{c_1} y_2(R,R/2) \;\geq\; p(R,R/2).$$

Using (21) yields

$$\left\|\left(JS^a(z)[(v,0)]\right)_2\right\| \;\geq\; p(R,R/2) = \frac{1}{1+(n-1)e^{-\gamma_\star R^2/4}}. \tag{23}$$

Step 3: conclusion and choice of γ. Combining (18) with (22)–(23), we obtain: for any real eigenvalue $\gamma_\star$ of B with unit eigenvector v, there exists $z \in B_R^2$ (hence $X = \Delta_n(z) \in B_R^n$) such that

$$\|JS_\theta(X)\|_{\mathrm{op}} \;\geq\; \sqrt{n-1}\,\omega(R,\gamma_\star),$$

where

$$\omega(R,\gamma_\star) := \begin{cases} \dfrac{1}{1+(n-1)e^{-2R^2|\gamma_\star|}}, & \gamma_\star < 0, \\[2ex] \dfrac{1}{1+(n-1)e^{-\gamma_\star R^2/4}}, & \gamma_\star \geq 0. \end{cases}$$

Now recall $\gamma = \max\{-\gamma_r, \gamma_1/8\}$. If $\gamma = -\gamma_r$, choose $\gamma_\star = \gamma_r \leq 0$; then $|\gamma_\star| = \gamma$ and

$$\omega(R,\gamma_\star) = \frac{1}{1+(n-1)e^{-2R^2\gamma}}.$$

If instead $\gamma = \gamma_1/8$, choose $\gamma_\star = \gamma_1 \geq 0$; then

$$\omega(R,\gamma_\star) = \frac{1}{1+(n-1)e^{-\gamma_1 R^2/4}} = \frac{1}{1+(n-1)e^{-2R^2\gamma}}.$$

Therefore, in all cases there exists $X \in B_R^n$ such that

$$\|JS_\theta(X)\|_{\mathrm{op}} \;\geq\; \sqrt{n-1}\frac{1}{1+(n-1)e^{-2R^2\gamma}}.$$

Since $\mathrm{Lip}_{B_R^n}(S_\theta) = \sup_{X\in B_R^n} \|JS_\theta(X)\|_{\mathrm{op}}$, we conclude

$$\mathrm{Lip}_{B_R^n}(S_\theta) \;\geq\; \frac{\sqrt{n-1}}{1+(n-1)e^{-2R^2\gamma}},$$

which is exactly (3). This completes the proof. $\square$

A.2.3 Theorem 2.5 [3, Theorem 3.7]

Proof. We work in the Euclidean case $E = \mathbb{R}^d$ and consider the single-head self-attention map $S_\theta : E^n \to E^n$ with logits

$$l_{ij} := L_{ij}(X) = x_i^\top B x_j, \quad B = \frac{1}{\sqrt{d}} Q^\top K \in \mathcal{L}(E).$$

Fix $R > 0$ and a configuration $X = (x_1, \ldots, x_n) \in E^n$, and set

$$X_R := RX = (Rx_1, \ldots, Rx_n).$$

For $i, j \in [n]$, the attention weights at X_R is

$$A_{ij}(X_R) = \frac{\exp(L_{ij}(X_R))}{\sum_{k=1}^n \exp(L_{ik}(X_R))}.$$

Note that

$$L_{ij}(X_R) = (Rx_i)^\top B(Rx_j) = R^2 l_{ij}.$$

If $B = 0$, then $A_{ij}(X_R) \equiv 1/n$ for all R and $JS_\theta(X_R)$ is uniformly bounded by $\|V\|_{\mathrm{op}}$ (hence (4) holds with $\vartheta_X \equiv 1$). We therefore assume $B \neq 0$ for the remainder of the proof.

We first show that the complement of E_B has Lebesgue measure zero, then compute $JS_\theta(X_R)$ explicitly and analyse its behaviour as $R \to +\infty$.

Step 1: generic configurations and the set E_B. Recall

$$E_B = \Big\{ X = (x_1, \ldots, x_n) \in B(0,1)^n : \forall i \in [n] \ \exists! \ j_i^\star \in [n] \text{ s.t. } l_{ij_i^\star} = \max_{1 \le j \le n} l_{ij} \Big\}.$$

If $X \notin E_B$, then there exist $i \in [n]$ and distinct $j, k \in [n]$ such that $l_{ij} = l_{ik}$, i.e.

$$x_i^\top B(x_j - x_k) = 0.$$

For each (i, j, k) with $i \in [n]$ and $1 \le j < k \le n$, define

$$Z_{i,j,k} := \Big\{ X \in (\mathbb{R}^d)^n : \ x_i^\top B(x_j - x_k) = 0 \Big\}.$$

Let $N := \ker(B^\top) \subset E$. Since $B \neq 0$, N is a proper linear subspace, hence has Lebesgue measure zero in E. If $x_i \notin N$, then $w_i := B^\top x_i \neq 0$ and the constraint $x_i^\top B(x_j - x_k) = 0$ becomes

$$w_i^\top (x_j - x_k) = 0,$$

which is a nontrivial linear constraint on $(x_j, x_k) \in E^2$ and therefore cuts out a codimension-1 affine subspace in E^2 (in particular, a measure-zero subset). By Fubini's theorem, $Z_{i,j,k}$ has Lebesgue measure zero in $(\mathbb{R}^d)^n$. Since

$$B(0,1)^n \setminus E_B \subset \bigcup_{i=1}^n \ \bigcup_{1 \le j < k \le n} Z_{i,j,k},$$

a finite union of measure-zero sets, it follows that $B(0,1)^n \setminus E_B$ has Lebesgue measure zero.

Step 2: explicit expression for the Jacobian at X_R. Fix $X \in E_B$ and $R > 0$, and set $X_R = RX = (Rx_1, \ldots, Rx_n)$. Let $U = (u_1, \ldots, u_n) \in E^n$. First, the directional derivative of the logit at X_R in direction U is

$$JL_{ij}(X_R)[U] = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} (Rx_i + tu_i)^\top B(Rx_j + tu_j) = R\big(u_i^\top B x_j + x_i^\top B u_j\big).$$

Since $A_{i\cdot}$ is a row-wise softmax, its Jacobian satisfies

$$JA_{ij}(X_R)[U] = A_{ij}(X_R)\Big(JL_{ij}(X_R)[U] - \sum_{k=1}^n A_{ik}(X_R) JL_{ik}(X_R)[U] \Big).$$

Define the attention-weighted mean (at X_R)

$$m_i^R := \sum_{k=1}^{n} A_{ik}(X_R)x_k \in E.$$

Applying the chain rule to $(S_\theta(X_R))_i = \sum_j A_{ij}(X_R)V(Rx_j)$, we obtain

$$\begin{aligned}(JS_\theta(X_R)[U])_i &= \sum_{j=1}^{n} JA_{ij}(X_R)[U]V(Rx_j) + \sum_{j=1}^{n} A_{ij}(X_R)Vu_j \\ &= V\sum_{j=1}^{n} A_{ij}(X_R)u_j + VR^2\sum_{j=1}^{n} A_{ij}(X_R)\big(x_j - m_i^R\big)\big(u_i^\top Bx_j + x_i^\top Bu_j\big),\end{aligned}$$

where we used $\sum_{j=1}^n A_{ij}(X_R)(x_j - m_i^R) = 0$ to collect terms.

Thus, for each $i \in [n]$,

$$(JS_\theta(X_R)[U])_i = V\Big(t_i^R(U) + R^2 s_i^R(U)\Big), \tag{24}$$

with

$$t_i^R(U) := \sum_{j=1}^{n} A_{ij}(X_R)u_j, \tag{25}$$

$$s_i^R(U) := \sum_{j=1}^{n} A_{ij}(X_R)\big(x_j - m_i^R\big)\big(u_i^\top Bx_j + x_i^\top Bu_j\big). \tag{26}$$

Step 3: softmax asymptotics for $A_{ij}(X_R)$. Since $X \in E_B$, for each $i \in [n]$ there is a unique index $j_i^\star \in [n]$ such that

$$l_{ij_i^\star} = \max_{1\le j\le n} l_{ij}.$$

Define

$$\Delta_i := \min_{j\ne j_i^\star}\big(l_{ij_i^\star} - l_{ij}\big) > 0.$$

Using $L_{ij}(X_R) = R^2 l_{ij}$, we can write

$$A_{ij}(X_R) = \frac{e^{R^2(l_{ij}-l_{ij_i^\star})}}{\sum_{k=1}^n e^{R^2(l_{ik}-l_{ij_i^\star})}}.$$

Hence, for $j \ne j_i^\star$,

$$l_{ij} - l_{ij_i^\star} \le -\Delta_i \quad\Longrightarrow\quad A_{ij}(X_R) \le e^{-R^2\Delta_i},$$

and therefore

$$1 - A_{ij_i^\star}(X_R) = \sum_{j\ne j_i^\star} A_{ij}(X_R) \le (n-1)e^{-R^2\Delta_i}.$$

In particular,

$$A_{ij}(X_R) \xrightarrow[R\to+\infty]{} \mathbf{1}_{\{j=j_i^\star\}}, \quad 0 \le A_{ij}(X_R)\big(1 - A_{ij}(X_R)\big) \le (n-1)e^{-R^2\Delta_i}. \tag{27}$$

Equivalently, for each i there exist constants $C_i > 0$ and $c_i > 0$ such that

$$A_{ij}(X_R)\big(1 - A_{ij}(X_R)\big) \le C_i e^{-c_iR^2} \quad \text{for all } j \in [n],\ R \ge 0. \tag{28}$$

Step 4: the main term t^R. Let $\|\cdot\|_F$ denote the Frobenius norm on E^n and $\|\cdot\|_{\mathrm{op}}$ the induced operator norm. Define the linear operator $T_R : E^n \to E^n$ by

$$(T_RU)_i := t_i^R(U) = \sum_{j=1}^{n} A_{ij}(X_R)u_j.$$

Using Cauchy–Schwarz with weights $(A_{ij}(X_R))_j$ and $\sum_j A_{ij}(X_R) = 1$, we have

$$\|t_i^R(U)\|^2 \le \sum_{j=1}^n A_{ij}(X_R)\|u_j\|^2.$$

Summing over i gives

$$\|T_R U\|_F^2 = \sum_{i=1}^n \|t_i^R(U)\|^2 \le \sum_{i=1}^n \sum_{j=1}^n A_{ij}(X_R)\|u_j\|^2 = \sum_{j=1}^n \Big(\sum_{i=1}^n A_{ij}(X_R)\Big)\|u_j\|^2 \le n\sum_{j=1}^n \|u_j\|^2 = n\|U\|_F^2.$$

Hence

$$\|T_R\|_{\mathrm{op}} \le \sqrt{n} \quad \text{for all } R > 0. \tag{29}$$

Moreover, by (27),

$$(T_R U)_i = \sum_{j=1}^n A_{ij}(X_R)u_j \xrightarrow[R\to+\infty]{} u_{j_i^\star},$$

so T_R converges pointwise to the linear map $T_\infty : E^n \to E^n$ given by

$$(T_\infty U)_i = u_{j_i^\star}, \quad i \in [n].$$

(We only use the uniform bound (29) below.)

Step 5: bounding the remainder $R^2 s^R$. We now control the contribution of $R^2 s_i^R(U)$ in (24). Since $X \in B(0,1)^n$, we have $\|x_j\| \le 1$ and hence $\|m_i^R\| \le 1$, so

$$\|x_j - m_i^R\| \le 2 \quad \text{for all } i, j.$$

A sharper bound uses

$$x_j - m_i^R = x_j - \sum_{k=1}^n A_{ik}(X_R)x_k = \sum_{k\ne j} A_{ik}(X_R)(x_j - x_k),$$

hence

$$\|x_j - m_i^R\| \le \sum_{k\ne j} A_{ik}(X_R)\|x_j - x_k\| \le 2\sum_{k\ne j} A_{ik}(X_R) = 2\big(1 - A_{ij}(X_R)\big),$$

and therefore

$$A_{ij}(X_R)\|x_j - m_i^R\| \le 2A_{ij}(X_R)\big(1 - A_{ij}(X_R)\big). \tag{30}$$

Using (26), Cauchy–Schwarz, and $\|x_i\|, \|x_j\| \le 1$,

$$\begin{aligned}\|s_i^R(U)\| &\le \sum_{j=1}^n A_{ij}(X_R)\|x_j - m_i^R\|\big(|u_i^\top B x_j| + |x_i^\top B u_j|\big)\\ &\le \|B\|_{\mathrm{op}} \sum_{j=1}^n A_{ij}(X_R)\|x_j - m_i^R\|\big(\|u_i\| + \|u_j\|\big).\end{aligned}$$

Applying (30) and then (28), we get

$$\|s_i^R(U)\| \le 2\|B\|_{\mathrm{op}} \sum_{j=1}^n A_{ij}(X_R)(1 - A_{ij}(X_R))\big(\|u_i\| + \|u_j\|\big) \le 2\|B\|_{\mathrm{op}} C_i e^{-c_i R^2}\Big(n\|u_i\| + \sum_{j=1}^n \|u_j\|\Big).$$

Using $\sum_{j=1}^n \|u_j\| \le \sqrt{n}\|U\|_F$, we obtain

$$\|s_i^R(U)\| \le 2\|B\|_{\mathrm{op}} C_i e^{-c_i R^2}\big(n\|u_i\| + \sqrt{n}\|U\|_F\big).$$

We now absorb the factor R^2 into the exponential: for any $c > 0$ there exists $R_0 > 0$ such that $R^2 \le e^{cR^2}$ for all $R \ge R_0$, hence $R^2 e^{-c_i R^2} \le e^{-(c_i - c)R^2}$. Choosing $c = c_i/2$ gives, for $R \ge R_0$,

$$R^2 e^{-c_i R^2} \le e^{-(c_i/2)R^2}.$$

Thus, for $R \ge R_0$,

$$\|R^2 s_i^R(U)\| \le 2\|B\|_{\mathrm{op}} C_i e^{-c_i' R^2}\big(n\|u_i\| + \sqrt{n}\|U\|_F\big), \quad c_i' := \tfrac{1}{2}c_i.$$

Summing over i and using $\sum_i (a_i + b)^2 \le 2\sum_i a_i^2 + 2nb^2$, we obtain

$$\sum_{i=1}^n \|R^2 s_i^R(U)\|^2 \le C_X e^{-c_X R^2}\|U\|_F^2,$$

for some constants $C_X, c_X > 0$ depending only on X, B and n. Equivalently, if we define $S_R : E^n \to E^n$ by $(S_R U)_i := s_i^R(U)$, then

$$\|R^2 S_R\|_{\mathrm{op}} \le \kappa_X(R), \quad \kappa_X(R) := C_X e^{-c_X R^2} \xrightarrow[R\to+\infty]{} 0. \tag{31}$$

Step 6: combining everything. From (24), (29) and (31), for all $U \in E^n$ and all $R \ge R_0$ we have

$$\begin{aligned}
\|JS_\theta(X_R)[U]\|_F &= \big\|V\big(T_R U + R^2 S_R U\big)\big\|_F \\
&\le \|V\|_{\mathrm{op}}\big(\|T_R U\|_F + \|R^2 S_R U\|_F\big) \\
&\le \|V\|_{\mathrm{op}}\big(\|T_R\|_{\mathrm{op}} + \|R^2 S_R\|_{\mathrm{op}}\big)\|U\|_F \\
&\le \|V\|_{\mathrm{op}}\big(\sqrt{n} + \kappa_X(R)\big)\|U\|_F.
\end{aligned}$$

Hence

$$\|JS_\theta(X_R)\|_{\mathrm{op}} \le \big(\sqrt{n} + \kappa_X(R)\big)\|V\|_{\mathrm{op}}.$$

Define

$$\vartheta_X(R) := 1 + \frac{\kappa_X(R)}{\sqrt{n}} = 1 + \frac{C_X}{\sqrt{n}} e^{-c_X R^2}.$$

Then $\vartheta_X(R) \to 1$ exponentially fast as $R \to +\infty$, and for all $R \ge R_0$,

$$\|JS_\theta(X_R)\|_{\mathrm{op}} \le \vartheta_X(R)\|V\|_{\mathrm{op}}\sqrt{n}.$$

Adjusting constants on $[0, R_0]$ if needed, the same inequality holds for all $R > 0$, which is precisely (4). This completes the proof. ☐

A.2.4 Theorem 2.6 [4, Theorem 4.2]

Proof. Work in the Euclidean case $E = \mathbb{R}^d$ and consider single-head self-attention $S_\theta : E^n \to E^n$ with $V = \mathrm{Id}_E$ and logits

$$L_{ij}(X) = x_i^\top B x_j, \quad X = (x_1, \dots, x_n) \in E^n.$$

Thus

$$(S_\theta(X))_i = \sum_{j=1}^n A_{ij}(X)x_j, \quad A_{ij}(X) = \frac{e^{L_{ij}(X)}}{\sum_{k=1}^n e^{L_{ik}(X)}}.$$

For each $i \in [n]$ define the attention barycentre

$$m_i(X) := \sum_{j=1}^n A_{ij}(X)x_j \in E,$$

and the covariance matrix

$$\Sigma_i(X) := \sum_{j=1}^n A_{ij}(X)(x_j - m_i(X))(x_j - m_i(X))^\top \in \mathbb{R}^{d\times d}.$$

Step 1: Jacobian formula. Let $U = (u_1, \dots, u_n) \in E^n$ be a perturbation. Computation gives, for each $i \in [n]$,

$$(JS_\theta(X)[U])_i = \sum_{j=1}^{n} A_{ij}(X)u_j + \Sigma_i(X)B^\top u_i + \sum_{j=1}^{n} A_{ij}(X)(x_j - m_i(X))(x_i^\top B u_j). \tag{32}$$

Step 2: replication and mean/fluctuation decomposition. Fix $X = (x_1, \dots, x_n) \in E^n$. For $N \geq 1$, define the N-replication

$$Y_N := \mathbf{1}_N \otimes X \in E^{Nn},$$

i.e. token x_i is repeated N times. We index tokens of Y_N by pairs (α, i) with $\alpha \in [N]$ and $i \in [n]$, and write

$$(Y_N)_{(\alpha,i)} = x_i.$$

Let $U \in E^{Nn}$ be a perturbation of Y_N, with components $u_{(\alpha,i)} \in E$. Define the replica-mean and replica-fluctuation parts:

$$\bar{u}_i := \frac{1}{N}\sum_{\alpha=1}^{N} u_{(\alpha,i)} \in E, \quad \tilde{u}_{(\alpha,i)} := u_{(\alpha,i)} - \bar{u}_i,$$

so that $\sum_{\alpha=1}^{N} \tilde{u}_{(\alpha,i)} = 0$ for each i. Set $\bar{U} := (\bar{u}_1, \dots, \bar{u}_n) \in E^n$.

Because all replicas share identical queries/keys/values, the attention weights at Y_N satisfy

$$A_{(\alpha,i),(\beta,j)}(Y_N) = \frac{1}{N}A_{ij}(X) \quad (\alpha, \beta \in [N],\ i, j \in [n]).$$

Consequently,

$$m_{(\alpha,i)}(Y_N) = m_i(X), \quad \Sigma_{(\alpha,i)}(Y_N) = \Sigma_i(X) \quad \text{for all } \alpha \in [N], i \in [n].$$

Insert these identities into (32) (with X replaced by Y_N). Using $\sum_{\beta=1}^{N} \tilde{u}_{(\beta,j)} = 0$, one checks that for each (α, i),

$$(JS_\theta(Y_N)[U])_{(\alpha,i)} = (JS_\theta(X)[\bar{U}])_i + \Sigma_i(X)B^\top \tilde{u}_{(\alpha,i)}. \tag{33}$$

Step 3: operator norm and independence of $N \geq 2$. Equip E^{Nn} with the Frobenius norm $\|U\|_F^2 = \sum_{\alpha=1}^{N}\sum_{i=1}^{n} \|u_{(\alpha,i)}\|^2$. The decomposition $u_{(\alpha,i)} = \bar{u}_i + \tilde{u}_{(\alpha,i)}$ is orthogonal, and the two terms on the right-hand side of (33) act on orthogonal subspaces (replica-means vs zero-sum fluctuations). Therefore, for $N \geq 2$,

$$\|JS_\theta(Y_N)\|_{\text{op}} = \max\Big\{\|JS_\theta(X)\|_{\text{op}}, \max_{1\leq i\leq n} \|\Sigma_i(X)B^\top\|_{\text{op}}\Big\}. \tag{34}$$

The right-hand side does not depend on N (once $N \geq 2$), hence

$$\|JS_\theta(Y_N)\|_{\text{op}} = \|JS_\theta(Y_2)\|_{\text{op}}, \quad \forall N \geq 2,$$

which proves item (1). Squaring (34) yields

$$\|JS_\theta(Y_N)\|_{\text{op}}^2 = \max\left\{\|JS_\theta(X)\|_{\text{op}}^2, \max_{1\leq i\leq n} \|\Sigma_i(X)B^\top\|_{\text{op}}^2\right\}.$$

Step 4: splitting regime and optimal direction. Assume the finite-splitting condition in the statement, i.e. $\sup_{N\geq 1} \|JS_\theta(Y_N)\|_{\text{op}} > \|JS_\theta(X)\|_{\text{op}}$. By item (1) this is equivalent to $\|JS_\theta(Y_2)\|_{\text{op}} > \|JS_\theta(X)\|_{\text{op}}$, and by (34) this holds iff

$$\max_{1\leq i\leq n} \|\Sigma_i(X)B^\top\|_{\text{op}} > \|JS_\theta(X)\|_{\text{op}}.$$

In that case,

$$\|JS_\theta(Y_2)\|_{\text{op}}^2 = \max_{1\leq i\leq n} \|\Sigma_i(X)B^\top\|_{\text{op}}^2 > \|JS_\theta(X)\|_{\text{op}}^2,$$

which is the desired strict inequality.

Let

$$i^\star \in \arg\max_{1\leq i\leq n} \|\Sigma_i(X)B^\top\|_{\text{op}},$$

and let $u^\star \in E$ be a unit right singular vector attaining $\|\Sigma_{i^\star}(X)B^\top\|_{\mathrm{op}} = \|\Sigma_{i^\star}(X)B^\top u^\star\|$. Take $N = 2$ and choose a perturbation $U \in E^{2n}$ supported only on the two replicas of token $i^\star$, with opposite signs (hence zero replica-mean):

$$\tilde{u}_{(1,i^\star)} = u^\star, \quad \tilde{u}_{(2,i^\star)} = -u^\star, \quad \tilde{u}_{(\alpha,i)} = 0 \text{ otherwise.}$$

Then $(JS_\theta(X)[\bar{U}]) = 0$ and (33) reduces to

$$(JS_\theta(Y_2)[U])_{(\alpha,i^\star)} = \Sigma_{i^\star}(X)B^\top \tilde{u}_{(\alpha,i^\star)},$$

which achieves the top singular value. Geometrically, this corresponds to splitting the mass of $x_{i^\star}$ into two equal atoms $x_{i^\star} \pm \varepsilon u^\star$, proving the optimal direction claim. This completes the proof. □

A.2.5 Theorem 2.7 [3, Theorem 4.3]

Proof. Fix $R > 0$ and $n \in \mathbb{N}$, and let $X = (x_1, \ldots, x_n) \in B_R^n$. Recall that

$$L_{ij}(X) = x_i^\top B x_j, \quad B = \frac{1}{\sqrt{d}} Q^\top K,$$

and that under the standard left-causal mask,

$$A_{ij}^M(X) = \begin{cases} \dfrac{\exp(L_{ij}(X))}{\sum_{k=1}^i \exp(L_{ik}(X))}, & j \le i, \\ 0, & j > i. \end{cases}$$

The i-th output token of S_θ^M is

$$y_i := (S_\theta^M(X))_i = \sum_{j=1}^i A_{ij}^M(X) V x_j, \quad i \in [n].$$

Step 1: differential of S_θ^M. Let $U = (u_1, \ldots, u_n) \in E^n$ be a perturbation and set $X_t := X + tU$. For each (i, j),

$$L_{ij}(X_t) = (x_i + tu_i)^\top B(x_j + tu_j),$$

hence

$$JL_{ij}(X)[U] = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} L_{ij}(X_t) = u_i^\top B x_j + x_i^\top B u_j.$$

For each fixed i, define

$$N_{ij}(t) := \exp(L_{ij}(X_t)), \quad D_i(t) := \sum_{k=1}^i \exp(L_{ik}(X_t)),$$

so that for $j \le i$,

$$A_{ij}^M(X_t) = \frac{N_{ij}(t)}{D_i(t)}.$$

The chain rule gives

$$N'_{ij}(0) = \exp(L_{ij}(X)) JL_{ij}(X)[U], \quad D'_i(0) = \sum_{k=1}^i \exp(L_{ik}(X)) JL_{ik}(X)[U].$$

Applying the quotient rule at $t = 0$, for $j \le i$,

$$JA_{ij}^M(X)[U] = A_{ij}^M(X)\Big(JL_{ij}(X)[U] - \sum_{k=1}^i A_{ik}^M(X) JL_{ik}(X)[U]\Big),$$

and $JA_{ij}^M(X)[U] = 0$ for $j > i$.

Differentiate the output:

$$\begin{aligned}
JS_\theta^M(X)_i[U] &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0}(S_\theta^M(X_t))_i \\
&= \sum_{j=1}^{i} JA_{ij}^M(X)[U]Vx_j + \sum_{j=1}^{i} A_{ij}^M(X)Vu_j \\
&= V\left(\sum_{j=1}^{i} JA_{ij}^M(X)[U]x_j + \sum_{j=1}^{i} A_{ij}^M(X)u_j\right).
\end{aligned}$$

Define the masked attention-weighted mean

$$m_i := \sum_{k=1}^{i} A_{ik}^M(X)x_k \in E.$$

Then we have

$$\sum_{j=1}^{i} JA_{ij}^M(X)[U]x_j = \sum_{j=1}^{i} A_{ij}^M(X)JL_{ij}(X)[U](x_j - m_i).$$

Using $JL_{ij}(X)[U] = u_i^\top Bx_j + x_i^\top Bu_j$, we obtain

$$JS_\theta^M(X)_i[U] = V\Big(u_i^{(0)} + v_i^{(0)} + w_i^{(0)}\Big),$$

where

$$u_i^{(0)} := \sum_{j=1}^{i} A_{ij}^M(X)u_j, \tag{35}$$

$$v_i^{(0)} := \sum_{j=1}^{i} A_{ij}^M(X)(x_j - m_i)(u_i^\top Bx_j), \tag{36}$$

$$w_i^{(0)} := \sum_{j=1}^{i} A_{ij}^M(X)(x_j - m_i)(x_i^\top Bu_j). \tag{37}$$

Step 2: bounding $\|JS_\theta^M(X)[U]\|_F$. Let $\|\cdot\|_F$ be the Frobenius norm on E^n. Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ and $\|Vz\| \le \|V\|_{\mathrm{op}}\|z\|$, we get

$$\|JS_\theta^M(X)[U]\|_F^2 \le 3\|V\|_{\mathrm{op}}^2 \sum_{i=1}^{n}\Big(\|u_i^{(0)}\|^2 + \|v_i^{(0)}\|^2 + \|w_i^{(0)}\|^2\Big). \tag{38}$$

First term. By Cauchy–Schwarz with weights $(A_{ij}^M)_{j\le i}$ and $\sum_{j\le i} A_{ij}^M = 1$,

$$\|u_i^{(0)}\|^2 = \Big\|\sum_{j=1}^{i} A_{ij}^M u_j\Big\|^2 \le \sum_{j=1}^{i} A_{ij}^M\|u_j\|^2.$$

Summing over i yields

$$\sum_{i=1}^{n}\|u_i^{(0)}\|^2 \le \sum_{i=1}^{n}\sum_{j=1}^{i} A_{ij}^M\|u_j\|^2 = \sum_{j=1}^{n}\Big(\sum_{i=j}^{n} A_{ij}^M\Big)\|u_j\|^2 \le n\sum_{j=1}^{n}\|u_j\|^2 = n\|U\|_F^2. \tag{$*$}$$

Variance bound (masked). For each i, define

$$\Sigma_i := \sum_{j=1}^{i} A_{ij}^M(X)(x_j - m_i)(x_j - m_i)^\top.$$

As in Lecture 1,

$$\sum_{j=1}^{i} A^M_{ij} \|x_j - m_i\|^2 = \sum_{j=1}^{i} A^M_{ij} \|x_j\|^2 - \|m_i\|^2 \le R^2,$$

hence $\|\Sigma_i\|_{\mathrm{op}} \le R^2$.

Second term. Set $z_i := B^\top u_i$, so $u_i^\top B x_j = z_i^\top x_j$. Then

$$\begin{aligned} v_i^{(0)} &= \sum_{j=1}^{i} A^M_{ij} (x_j - m_i) z_i^\top x_j \\ &= \sum_{j=1}^{i} A^M_{ij} (x_j - m_i) z_i^\top (x_j - m_i) \quad \Big(\text{since } \sum_{j\le i} A^M_{ij}(x_j - m_i) = 0\Big) \\ &= \Sigma_i z_i = \Sigma_i B^\top u_i, \end{aligned}$$

so

$$\|v_i^{(0)}\| \le \|\Sigma_i\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \|u_i\| \le R^2 \|B\|_{\mathrm{op}} \|u_i\|.$$

Therefore

$$\sum_{i=1}^{n} \|v_i^{(0)}\|^2 \le R^4 \|B\|_{\mathrm{op}}^2 \sum_{i=1}^{n} \|u_i\|^2 = R^4 \|B\|_{\mathrm{op}}^2 \|U\|_F^2. \tag{$**$}$$

Third term. By Cauchy–Schwarz with weights $(A^M_{ij})_{j\le i}$,

$$\|w_i^{(0)}\|^2 \le \Big(\sum_{j=1}^{i} A^M_{ij} \|x_j - m_i\|^2\Big)\Big(\sum_{j=1}^{i} A^M_{ij} |x_i^\top B u_j|^2\Big).$$

The first factor is $\le R^2$. For the second,

$$|x_i^\top B u_j| = |(B^\top x_i)^\top u_j| \le \|B\|_{\mathrm{op}} \|x_i\| \|u_j\| \le \|B\|_{\mathrm{op}} R \|u_j\|,$$

hence

$$\sum_{j=1}^{i} A^M_{ij} |x_i^\top B u_j|^2 \le \|B\|_{\mathrm{op}}^2 R^2 \sum_{j=1}^{i} A^M_{ij} \|u_j\|^2 \le \|B\|_{\mathrm{op}}^2 R^2 \sum_{j=1}^{n} \|u_j\|^2 = \|B\|_{\mathrm{op}}^2 R^2 \|U\|_F^2.$$

Thus $\|w_i^{(0)}\|^2 \le \|B\|_{\mathrm{op}}^2 R^4 \|U\|_F^2$, and summing over i gives

$$\sum_{i=1}^{n} \|w_i^{(0)}\|^2 \le n \|B\|_{\mathrm{op}}^2 R^4 \|U\|_F^2. \tag{$***$}$$

Step 3: conclusion. Insert $(*)$, $(**)$, $(***)$ into (38):

$$\|JS^M_\theta(X)[U]\|_F^2 \le 3\|V\|_{\mathrm{op}}^2 \Big(n + \|B\|_{\mathrm{op}}^2 R^4 (n+1)\Big) \|U\|_F^2.$$

Hence

$$\|JS^M_\theta(X)\|_{\mathrm{op}} \le \sqrt{3} \|V\|_{\mathrm{op}} \Big(\|B\|_{\mathrm{op}}^2 R^4 (n+1) + n\Big)^{1/2}.$$

This bound is uniform over $X \in B_R^n$, so it yields exactly (5). $\square$

A.2.6 Theorem 2.8 [3, Theorem 4.5]

Proof. We work in the Euclidean case $E = \mathbb{R}^d$ and consider the single-head masked self-attention map $S_\theta^M : E^n \to E^n$ with the standard left-causal mask and logits

$$l_{ij} := L_{ij}(X) = x_i^\top B x_j, \qquad B = \frac{1}{\sqrt{d}} Q^\top K \in \mathcal{L}(E).$$

Fix $R > 0$ and a configuration $X = (x_1, \ldots, x_n) \in E^n$, and set

$$X_R := RX = (Rx_1, \ldots, Rx_n).$$

For $i, j \in [n]$, the masked attention weights at X_R are

$$A_{ij}^M(X_R) = \begin{cases} \dfrac{\exp(L_{ij}(X_R))}{\sum_{k=1}^i \exp(L_{ik}(X_R))}, & j \le i, \\ 0, & j > i, \end{cases}$$

so that each row is stochastic: $\sum_{j=1}^n A_{ij}^M(X_R) = 1$. Note that

$$L_{ij}(X_R) = (Rx_i)^\top B (Rx_j) = R^2 l_{ij}.$$

If $B = 0$, then $A_{ij}^M(X_R) = \mathbf{1}_{\{j \le i\}}/i$ for all R, and $JS_\theta^M(X_R)$ is uniformly bounded by $\|V\|_{\mathrm{op}}$. We therefore assume $B \neq 0$ for the remainder of the proof.

We first show that the complement of E_B^M has Lebesgue measure zero, then compute $JS_\theta^M(X_R)$ explicitly and analyze its behaviour as $R \to +\infty$.

Step 1: generic configurations and the set E_B^M. Recall

$$E_B^M := \Big\{ X = (x_1, \ldots, x_n) \in B(0,1)^n : \forall i \in [n] \ \exists! \ j_i^\star \in \{1, \ldots, i\} \text{ s.t. } l_{ij_i^\star} = \max_{1 \le j \le i} l_{ij} \Big\}.$$

If $X \notin E_B^M$, then there exist $i \in [n]$ and distinct $j, k \in \{1, \ldots, i\}$ such that $l_{ij} = l_{ik}$, i.e.

$$x_i^\top B (x_j - x_k) = 0.$$

For each triple (i, j, k) with $i \in [n]$ and $1 \le j < k \le i$, define

$$Z_{i,j,k} := \Big\{ X \in (\mathbb{R}^d)^n : \ x_i^\top B(x_j - x_k) = 0 \Big\}.$$

As in the unmasked proof (replace $k \le n$ by $k \le i$), each $Z_{i,j,k}$ has Lebesgue measure zero in $(\mathbb{R}^d)^n$ (use that $B \neq 0$ and Fubini). Since

$$B(0,1)^n \setminus E_B^M \subset \bigcup_{i=1}^n \ \bigcup_{1 \le j < k \le i} Z_{i,j,k},$$

a finite union of measure-zero sets, it follows that $B(0,1)^n \setminus E_B^M$ has Lebesgue measure zero.

Step 2: explicit expression for the Jacobian at X_R. Fix $X \in E_B^M$ and $R > 0$, and set $X_R = RX$. Let $U = (u_1, \ldots, u_n) \in E^n$. The directional derivative of the logit at X_R in direction U is

$$JL_{ij}(X_R)[U] = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} (Rx_i + tu_i)^\top B (Rx_j + tu_j) = R\big(u_i^\top B x_j + x_i^\top B u_j\big).$$

Row-wise, $A_{i\cdot}^M$ is a softmax over the allowed indices $\{1, \ldots, i\}$, hence for $j \le i$ its Jacobian satisfies

$$JA_{ij}^M(X_R)[U] = A_{ij}^M(X_R)\Big(JL_{ij}(X_R)[U] - \sum_{k=1}^i A_{ik}^M(X_R) JL_{ik}(X_R)[U] \Big),$$

and $JA^M_{ij}(X_R)[U] = 0$ for $j > i$.

Define the masked attention-weighted mean (at X_R)

$$m_i(X_R) := \sum_{k=1}^{i} A^M_{ik}(X_R) x_k \in E \qquad \Big(= \sum_{k=1}^{n} A^M_{ik}(X_R) x_k \text{ since } A^M_{ik} = 0 \text{ for } k > i\Big).$$

Applying the chain rule to $(S^M_\theta(X_R))_i = \sum_{j \le i} A^M_{ij}(X_R) V(Rx_j)$, we obtain

$$\begin{aligned}
(JS^M_\theta(X_R)[U])_i &= \sum_{j=1}^{i} JA^M_{ij}(X_R)[U] V(Rx_j) + \sum_{j=1}^{i} A^M_{ij}(X_R) V u_j \\
&= V \sum_{j=1}^{i} A^M_{ij}(X_R) u_j + V R^2 \sum_{j=1}^{i} A^M_{ij}(X_R) \big(x_j - m_i(X_R)\big)\big(u_i^\top B x_j + x_i^\top B u_j\big),
\end{aligned}$$

where we used $\sum_{j \le i} A^M_{ij}(X_R)(x_j - m_i(X_R)) = 0$.

Thus, for each $i \in [n]$,

$$(JS^M_\theta(X_R)[U])_i = V\Big(t^R_i(U) + R^2 s^R_i(U)\Big), \tag{39}$$

with

$$t^R_i(U) := \sum_{j=1}^{i} A^M_{ij}(X_R) u_j, \tag{40}$$

$$s^R_i(U) := \sum_{j=1}^{i} A^M_{ij}(X_R)\big(x_j - m_i(X_R)\big)\big(u_i^\top B x_j + x_i^\top B u_j\big). \tag{41}$$

Step 3: softmax asymptotics for $A^M_{ij}(X_R)$. Since $X \in E^M_B$, for each $i \in [n]$ there is a unique index $j^\star_i \in \{1, \dots, i\}$ such that

$$l_{ij^\star_i} = \max_{1 \le j \le i} l_{ij}.$$

Define

$$\Delta_i := \min_{\substack{1 \le j \le i \\ j \ne j^\star_i}} \big(l_{ij^\star_i} - l_{ij}\big) > 0.$$

Using $L_{ij}(X_R) = R^2 l_{ij}$ and the restricted normalization, we can write for $j \le i$

$$A^M_{ij}(X_R) = \frac{e^{R^2(l_{ij} - l_{ij^\star_i})}}{\sum_{k=1}^{i} e^{R^2(l_{ik} - l_{ij^\star_i})}}.$$

Hence, for $j \ne j^\star_i$ (with $j \le i$),

$$l_{ij} - l_{ij^\star_i} \le -\Delta_i \quad \Longrightarrow \quad A^M_{ij}(X_R) \le e^{-R^2 \Delta_i},$$

and therefore

$$1 - A^M_{ij^\star_i}(X_R) = \sum_{\substack{1 \le j \le i \\ j \ne j^\star_i}} A^M_{ij}(X_R) \le (i-1) e^{-R^2 \Delta_i} \le (n-1) e^{-R^2 \Delta_i}.$$

In particular,

$$A^M_{ij}(X_R) \xrightarrow[R \to +\infty]{} \mathbf{1}_{\{j = j^\star_i\}}, \quad 0 \le A^M_{ij}(X_R)\big(1 - A^M_{ij}(X_R)\big) \le (n-1) e^{-R^2 \Delta_i}, \tag{42}$$

for all $j \le i$ (and trivially also for $j > i$ since $A^M_{ij} \equiv 0$). Equivalently, for each i there exist constants $C_i, c_i > 0$ such that

$$A^M_{ij}(X_R)\big(1 - A^M_{ij}(X_R)\big) \le C_i e^{-c_i R^2} \quad \text{for all } j \in [n],\ R \ge 0. \tag{43}$$

Step 4: the main term t^R. Let $\|\cdot\|_F$ denote the Frobenius norm on E^n and $\|\cdot\|_{\mathrm{op}}$ the induced operator norm. Define $T_R : E^n \to E^n$ by

$$(T_R U)_i := t_i^R(U) = \sum_{j=1}^{i} A_{ij}^M(X_R) u_j.$$

Using Cauchy–Schwarz with weights $(A_{ij}^M(X_R))_{j\le i}$ and $\sum_{j\le i} A_{ij}^M(X_R) = 1$, we have

$$\|t_i^R(U)\|^2 \le \sum_{j=1}^{i} A_{ij}^M(X_R)\|u_j\|^2.$$

Summing over i gives

$$\begin{aligned}\|T_R U\|_F^2 = \sum_{i=1}^{n} \|t_i^R(U)\|^2 &\le \sum_{i=1}^{n}\sum_{j=1}^{i} A_{ij}^M(X_R)\|u_j\|^2 \\ &= \sum_{j=1}^{n}\Big(\sum_{i=j}^{n} A_{ij}^M(X_R)\Big)\|u_j\|^2 \le n\sum_{j=1}^{n}\|u_j\|^2 = n\|U\|_F^2.\end{aligned}$$

Hence

$$\|T_R\|_{\mathrm{op}} \le \sqrt{n} \quad \text{for all } R > 0. \tag{44}$$

(We only use the uniform bound (44) below.)

Step 5: bounding the remainder $R^2 s^R$. We now control the contribution of $R^2 s_i^R(U)$ in (39). Since $X \in B(0,1)^n$, we have $\|x_j\| \le 1$ and hence $\|m_i(X_R)\| \le 1$, so

$$\|x_j - m_i(X_R)\| \le 2 \quad \text{for all } i \in [n],\ 1 \le j \le i.$$

A sharper bound uses

$$x_j - m_i(X_R) = x_j - \sum_{k=1}^{i} A_{ik}^M(X_R) x_k = \sum_{\substack{1\le k\le i\\ k\ne j}} A_{ik}^M(X_R)(x_j - x_k),$$

hence

$$\|x_j - m_i(X_R)\| \le 2\sum_{k\ne j} A_{ik}^M(X_R) = 2\big(1 - A_{ij}^M(X_R)\big),$$

and therefore

$$A_{ij}^M(X_R)\|x_j - m_i(X_R)\| \le 2A_{ij}^M(X_R)\big(1 - A_{ij}^M(X_R)\big). \tag{45}$$

Using (41), Cauchy–Schwarz, and $\|x_i\|, \|x_j\| \le 1$,

$$\begin{aligned}\|s_i^R(U)\| &\le \sum_{j=1}^{i} A_{ij}^M(X_R)\|x_j - m_i(X_R)\|\big(|u_i^\top B x_j| + |x_i^\top B u_j|\big) \\ &\le \|B\|_{\mathrm{op}} \sum_{j=1}^{i} A_{ij}^M(X_R)\|x_j - m_i(X_R)\|\big(\|u_i\| + \|u_j\|\big).\end{aligned}$$

Applying (45) and then (43), we get

$$\|s_i^R(U)\| \le 2\|B\|_{\mathrm{op}} \sum_{j=1}^{i} A_{ij}^M(X_R)(1 - A_{ij}^M(X_R))\big(\|u_i\| + \|u_j\|\big) \le 2\|B\|_{\mathrm{op}} C_i e^{-c_i R^2}\Big(n\|u_i\| + \sum_{j=1}^{n}\|u_j\|\Big),$$

where we used $i \le n$ and $\sum_{j=1}^{i}\|u_j\| \le \sum_{j=1}^{n}\|u_j\|$. Using $\sum_{j=1}^{n}\|u_j\| \le \sqrt{n}\|U\|_F$, we obtain

$$\|s_i^R(U)\| \le 2\|B\|_{\mathrm{op}} C_i e^{-c_i R^2}\big(n\|u_i\| + \sqrt{n}\|U\|_F\big).$$

We now absorb the factor R^2 into the exponential: for any $c > 0$ there exists $R_0 > 0$ such that $R^2 \le e^{cR^2}$ for all $R \ge R_0$, hence $R^2 e^{-c_i R^2} \le e^{-(c_i - c)R^2}$. Choosing $c = c_i/2$ gives, for $R \ge R_0$,

$$R^2 e^{-c_i R^2} \le e^{-(c_i/2)R^2}.$$

Thus, for $R \ge R_0$,

$$\|R^2 s_i^R(U)\| \le 2\|B\|_{\mathrm{op}} C_i e^{-c_i' R^2}\big(n\|u_i\| + \sqrt{n}\|U\|_F\big), \quad c_i' := \tfrac{1}{2}c_i.$$

Summing over i and using $\sum_i (a_i + b)^2 \le 2\sum_i a_i^2 + 2nb^2$, we obtain

$$\sum_{i=1}^n \|R^2 s_i^R(U)\|^2 \le C_X e^{-c_X R^2}\|U\|_F^2,$$

for some constants $C_X, c_X > 0$ depending only on X, B and n. Equivalently, if we define $S_R : E^n \to E^n$ by $(S_R U)_i := s_i^R(U)$, then

$$\|R^2 S_R\|_{\mathrm{op}} \le \kappa_X(R), \quad \kappa_X(R) := C_X e^{-c_X R^2} \xrightarrow[R \to +\infty]{} 0. \tag{46}$$

Step 6: combining everything. From (39), (44) and (46), for all $U \in E^n$ and all $R \ge R_0$ we have

$$\begin{aligned}
\|JS_\theta^M(X_R)[U]\|_F &= \big\|V\big(T_R U + R^2 S_R U\big)\big\|_F \\
&\le \|V\|_{\mathrm{op}}\big(\|T_R U\|_F + \|R^2 S_R U\|_F\big) \\
&\le \|V\|_{\mathrm{op}}\big(\|T_R\|_{\mathrm{op}} + \|R^2 S_R\|_{\mathrm{op}}\big)\|U\|_F \\
&\le \|V\|_{\mathrm{op}}\big(\sqrt{n} + \kappa_X(R)\big)\|U\|_F.
\end{aligned}$$

Hence

$$\|JS_\theta^M(X_R)\|_{\mathrm{op}} \le \big(\sqrt{n} + \kappa_X(R)\big)\|V\|_{\mathrm{op}}.$$

Define

$$\vartheta_X(R) := 1 + \frac{\kappa_X(R)}{\sqrt{n}} = 1 + \frac{C_X}{\sqrt{n}} e^{-c_X R^2}.$$

Then $\vartheta_X(R) \to 1$ exponentially fast as $R \to +\infty$, and for all $R \ge R_0$,

$$\|JS_\theta^M(X_R)\|_{\mathrm{op}} \le \vartheta_X(R)\|V\|_{\mathrm{op}}\sqrt{n}.$$

Adjusting constants on $[0, R_0]$ if needed, the same inequality holds for all $R > 0$, which is precisely (6). This completes the proof. □

A.2.7 Lemma 2.11 [4, Lemma 3.3]

Proof. We prove the statement for $p = 2$ (Frobenius norm on E^n and W_2 on $\mathcal{P}_2(E)$). The extension to any $p \ge 1$ is identical after replacing W_2 by W_p and the Frobenius norm by the ℓ_p-product norm on E^n (see the last paragraph).

Fix $X = (x_1, \dots, x_n) \in E^n$ and set

$$\mu := \mu_X = \frac{1}{n}\sum_{i=1}^n \delta_{x_i} \in \mathcal{M}_n(E).$$

By the assumed consistency on empirical measures,

$$F_\theta(\mu_Y) = \mu_{S_\theta(Y)} \quad \forall\, Y \in E^n.$$

Also note that permuting the coordinates of Y does not change μ_Y, and hence does not change $F_\theta(\mu_Y) = \mu_{S_\theta(Y)}$.

Step 1: W_2 versus Frobenius norm on a small neighbourhood. Define the separation of X by

$$\Delta_X := \min_{x_i \neq x_j} \|x_i - x_j\| \quad (\Delta_X := +\infty \text{ if all } x_i \text{ coincide}).$$

Choose $\varepsilon_1 > 0$ such that $\varepsilon_1 < \Delta_X/4$. If $\|X - Y\|_F \leq \varepsilon_1$, then for every i we have $\|x_i - y_i\| \leq \|X - Y\|_F \leq \varepsilon_1$, hence for $x_j \neq x_i$,

$$\|y_i - x_j\| \geq \|x_i - x_j\| - \|x_i - y_i\| > 4\varepsilon_1 - \varepsilon_1 = 3\varepsilon_1 > \varepsilon_1 \geq \|y_i - x_i\|.$$

Thus, among the support points $\{x_1, \ldots, x_n\}$ of μ_X, the nearest neighbour of y_i is x_i. It follows that in the optimal transport problem between μ_X and μ_Y with cost $c(x, y) = \|x - y\|^2$, the optimal coupling matches y_i with x_i for each i, and therefore

$$W_2(\mu_X, \mu_Y)^2 = \frac{1}{n} \sum_{i=1}^{n} \|x_i - y_i\|^2 = \frac{1}{n} \|X - Y\|_F^2.$$

Equivalently, for all Y with $\|X - Y\|_F \leq \varepsilon_1$,

$$W_2(\mu_X, \mu_Y) = \frac{1}{\sqrt{n}} \|X - Y\|_F. \tag{47}$$

We also need the analogous relation after applying S_θ. Set

$$Z := S_\theta(X) \in E^n, \quad \Delta_Z := \min_{z_i \neq z_j} \|z_i - z_j\| \quad (\Delta_Z := +\infty \text{ if all } z_i \text{ coincide}),$$

and pick $\varepsilon_2 > 0$ such that $\varepsilon_2 < \Delta_Z/4$. By continuity of S_θ, there exists $\varepsilon \in (0, \varepsilon_1]$ such that

$$\|X - Y\|_F \leq \varepsilon \quad \Longrightarrow \quad \|S_\theta(X) - S_\theta(Y)\|_F \leq \varepsilon_2.$$

Then, for all Y with $\|X - Y\|_F \leq \varepsilon$, we may apply (47) both to (X, Y) and to $(S_\theta(X), S_\theta(Y))$:

$$W_2(\mu_X, \mu_Y) = \frac{1}{\sqrt{n}} \|X - Y\|_F, \quad W_2(\mu_{S_\theta(X)}, \mu_{S_\theta(Y)}) = \frac{1}{\sqrt{n}} \|S_\theta(X) - S_\theta(Y)\|_F. \tag{48}$$

Step 2: rewriting the Lipschitz ratio in measure form. Fix $\eta \in (0, \varepsilon]$. Using the standard (metric) definition of the local Lipschitz constant,

$$\mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) = \limsup_{Y \to X} \frac{\|S_\theta(X) - S_\theta(Y)\|_F}{\|X - Y\|_F} = \limsup_{\substack{Y \to X \\ \|X - Y\|_F \leq \eta}} \frac{\|S_\theta(X) - S_\theta(Y)\|_F}{\|X - Y\|_F}.$$

For $\|X - Y\|_F \leq \eta \leq \varepsilon$, we can replace the numerator and denominator using (48) and the consistency $F_\theta(\mu_Y) = \mu_{S_\theta(Y)}$:

$$\begin{aligned} \frac{\|S_\theta(X) - S_\theta(Y)\|_F}{\|X - Y\|_F} &= \frac{\sqrt{n} W_2(\mu_{S_\theta(X)}, \mu_{S_\theta(Y)})}{\sqrt{n} W_2(\mu_X, \mu_Y)} \\ &= \frac{W_2\big(F_\theta(\mu_X), F_\theta(\mu_Y)\big)}{W_2(\mu_X, \mu_Y)}. \end{aligned}$$

Hence,

$$\mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) = \limsup_{\substack{Y \to X \\ Y \in E^n}} \frac{W_2\big(F_\theta(\mu_X), F_\theta(\mu_Y)\big)}{W_2(\mu_X, \mu_Y)}. \tag{49}$$

Finally, the map $Y \mapsto \mu_Y$ takes values in $\mathcal{M}_n(E)$, and $\mu_Y \to \mu_X$ in W_2 as $Y \to X$ in $\|\cdot\|_F$ by (47). Therefore (49) is exactly

$$\mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) = \mathrm{Lip}^{W_2}_{\mathrm{loc}, \mathcal{M}_n(E)}(F_\theta; \mu_X).$$

Step 3: comparison with the unrestricted local constant. Since $\mathcal{M}_n(E) \subset \mathcal{P}_2(E)$, restricting the set of approaching measures can only decrease the limsup, hence

$$\mathrm{Lip}^{W_2}_{\mathrm{loc},\mathcal{M}_n(E)}(F_\theta;\mu_X) \le \mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_X).$$

Combining with Step 2 yields the desired chain of inequalities.

Extension to $p \ge 1$. The argument above only uses (i) the identification of W_p between empirical measures with the ℓ_p-product norm on E^n on a small neighborhood (up to the fixed factor $n^{-1/p}$, which cancels in the ratio), and (ii) continuity of S_θ. Replacing squared distances by $\|\cdot\|^p$ and $\|\cdot\|_F$ by the ℓ_p-product norm yields the general case. □

A.2.8 Theorem 2.12 [3, Theorem 3.5]

Proof. We work in $E = \mathbb{R}^d$ with $p = 2$. Let $\theta = (Q, K, V, \beta)$ and recall $q(x) = Qx$, $k(y) = Ky$, $v(y) = Vy$, and

$$\kappa_\theta(x,y) = \exp\big(\beta\langle q(x), k(y)\rangle\big).$$

We may also write $\beta\langle Qx, Ky\rangle = \langle x, By\rangle$ with $B = \beta Q^\top K$, hence $\kappa_\theta(x,y) = \exp(\langle x, By\rangle)$. For $\mu \in \mathcal{P}_{2,R}(E)$ and $x \in E$, recall

$$Z_{\theta,\mu}(x) := \int_E \kappa_\theta(x,y)\mu(dy), \quad a_{\theta,\mu}(x,dy) := \frac{\kappa_\theta(x,y)}{Z_{\theta,\mu}(x)}\mu(dy),$$
$$\Gamma_{\theta,\mu}(x) := \int_E v(y) a_{\theta,\mu}(x,dy), \quad F_\theta(\mu) := (\Gamma_{\theta,\mu})_\#\mu.$$

Fix $R > 0$ and set $c := \|B\|_{\mathrm{op}} R^2$ (used only as shorthand).

Step 1: bounds on $Z_{\theta,\mu}$ and $\Gamma_{\theta,\mu}$. If $x, y \in B(0,R)$ then $|\langle x, By\rangle| \le \|x\|\|B\|_{\mathrm{op}}\|y\| \le c$, so

$$e^{-c} \le \kappa_\theta(x,y) \le e^c. \tag{50}$$

Therefore, for all $\mu \in \mathcal{P}_{2,R}(E)$ and all $x \in B(0,R)$,

$$e^{-c} \le Z_{\theta,\mu}(x) \le e^c. \tag{51}$$

Moreover $a_{\theta,\mu}(x,\cdot)$ is a probability measure supported on $B(0,R)$, hence $\|v(y)\| \le \|V\|_{\mathrm{op}}\|y\| \le \|V\|_{\mathrm{op}} R$ on its support, and

$$\|\Gamma_{\theta,\mu}(x)\| = \left\|\int_E v(y) a_{\theta,\mu}(x,dy)\right\| \le \int_E \|v(y)\| a_{\theta,\mu}(x,dy) \le \|V\|_{\mathrm{op}} R. \tag{52}$$

Step 2: Lipschitz dependence of $\Gamma_{\theta,\mu}(x)$ on μ (fixed x). Fix $x \in B(0,R)$ and let $\mu, \nu \in \mathcal{P}_{2,R}(E)$. Write

$$N_{\theta,\mu}(x) := \int_E \kappa_\theta(x,y) v(y)\mu(dy) \quad \text{so that} \quad \Gamma_{\theta,\mu}(x) = \frac{N_{\theta,\mu}(x)}{Z_{\theta,\mu}(x)}.$$

Define, for this fixed x,

$$g_x(y) := \kappa_\theta(x,y), \quad h_x(y) := \kappa_\theta(x,y) v(y).$$

On $B(0,R)$, using (50) and that $t \mapsto e^t$ is e^c-Lipschitz on $[-c,c]$, we have for $y_1, y_2 \in B(0,R)$:

$$|g_x(y_1) - g_x(y_2)| \le e^c |\langle x, B(y_1 - y_2)\rangle| \le e^c \|B\|_{\mathrm{op}} R \|y_1 - y_2\|.$$

Also,

$$\begin{aligned}\|h_x(y_1) - h_x(y_2)\| &\le \kappa_\theta(x,y_1)\|v(y_1) - v(y_2)\| + |g_x(y_1) - g_x(y_2)|\|v(y_2)\| \\ &\le e^c\|V\|_{\mathrm{op}}\|y_1 - y_2\| + e^c\|B\|_{\mathrm{op}} R\|y_1 - y_2\|(\|V\|_{\mathrm{op}} R) \\ &\le e^c\|V\|_{\mathrm{op}}\big(1 + \|B\|_{\mathrm{op}} R^2\big)\|y_1 - y_2\|.\end{aligned}$$

Hence, by Kantorovich–Rubinstein and $W_1 \le W_2$,

$$|Z_{\theta,\mu}(x) - Z_{\theta,\nu}(x)| \le e^c \|B\|_{\mathrm{op}} R W_2(\mu,\nu), \tag{53}$$

$$\|N_{\theta,\mu}(x) - N_{\theta,\nu}(x)\| \le e^{c}\|V\|_{\mathrm{op}}\big(1 + \|B\|_{\mathrm{op}}R^2\big)W_2(\mu,\nu). \tag{54}$$

Now decompose the difference of quotients using $\Gamma_{\theta,\nu}(x) = N_{\theta,\nu}(x)/Z_{\theta,\nu}(x)$:

$$\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(x) = \frac{N_{\theta,\mu}(x) - N_{\theta,\nu}(x)}{Z_{\theta,\mu}(x)} + \Gamma_{\theta,\nu}(x)\frac{Z_{\theta,\nu}(x) - Z_{\theta,\mu}(x)}{Z_{\theta,\mu}(x)}.$$

Using (51) (so $1/Z_{\theta,\mu}(x) \le e^{c}$), (52), (53), and (54), we obtain for all $x \in B(0,R)$:

$$\|\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(x)\| \le \|V\|_{\mathrm{op}}e^{2c}\big(1 + 2\|B\|_{\mathrm{op}}R^2\big)W_2(\mu,\nu). \tag{55}$$

Step 3: Lipschitz dependence of $\Gamma_{\theta,\mu}(x)$ on x (fixed μ). Fix $\mu \in \mathcal{P}_{2,R}(E)$ and consider $x_1, x_2 \in B(0,R)$. We show that $x \mapsto \Gamma_{\theta,\mu}(x)$ is Lipschitz on $B(0,R)$ with a constant controlled by $\|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}}R^2$.

We denote by $J\Gamma_{\theta,\mu}(x)$ the Jacobian operator

$$J\Gamma_{\theta,\mu}(x) : E \to E, \quad u \mapsto J\Gamma_{\theta,\mu}(x)[u].$$

Let $u \in E$. Differentiating $a_{\theta,\mu}(x,dy) = \kappa_\theta(x,y)Z_{\theta,\mu}(x)^{-1}\mu(dy)$ in the direction u gives

$$Ja_{\theta,\mu}(x,dy)[u] = a_{\theta,\mu}(x,dy)\Big(\langle u, By\rangle - \int_E \langle u, Bz\rangle a_{\theta,\mu}(x,dz)\Big).$$

Therefore,

$$J\Gamma_{\theta,\mu}(x)[u] = \int_E v(y)Ja_{\theta,\mu}(x,dy)[u].$$

By Cauchy–Schwarz under the probability measure $a_{\theta,\mu}(x,\cdot)$,

$$\|J\Gamma_{\theta,\mu}(x)[u]\| \le \Big(\int_E \|v(y)\|^2 a_{\theta,\mu}(x,dy)\Big)^{1/2}\Big(\mathrm{Var}_{a_{\theta,\mu}(x,\cdot)}(\langle u, By\rangle)\Big)^{1/2}.$$

Since $a_{\theta,\mu}(x,\cdot)$ is supported on $B(0,R)$, we have $\|v(y)\| \le \|V\|_{\mathrm{op}}R$ and $|\langle u, By\rangle| \le \|u\|\|B\|_{\mathrm{op}}R$, hence

$$\|J\Gamma_{\theta,\mu}(x)[u]\| \le \|V\|_{\mathrm{op}}R \cdot \|u\|\|B\|_{\mathrm{op}}R = \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}}R^2\|u\|.$$

Equivalently,

$$\|J\Gamma_{\theta,\mu}(x)\|_{\mathrm{op}} \le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}}R^2 \quad \text{for all } x \in B(0,R).$$

By the mean value theorem, for all $x_1, x_2 \in B(0,R)$,

$$\|\Gamma_{\theta,\mu}(x_1) - \Gamma_{\theta,\mu}(x_2)\| \le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}}R^2\|x_1 - x_2\| \le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}}R^2e^{2c}\|x_1 - x_2\|. \tag{56}$$

(The last inequality uses $e^{2c} \ge 1$ to match the final exponential form.)

Step 4: pushforward coupling and conclusion. Let $\mu, \nu \in \mathcal{P}_{2,R}(E)$ and let π be an optimal coupling, so that $W_2^2(\mu,\nu) = \int \|x-y\|^2\pi(dx,dy)$. Pushing π forward by $(x,y) \mapsto (\Gamma_{\theta,\mu}(x), \Gamma_{\theta,\nu}(y))$ yields a coupling of $(F_\theta(\mu), F_\theta(\nu))$, hence

$$W_2(F_\theta(\mu), F_\theta(\nu)) \le \Big(\int \|\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(y)\|^2\pi(dx,dy)\Big)^{1/2}.$$

Decompose

$$\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(y) = \big(\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\mu}(y)\big) + \big(\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\big),$$

and apply Minkowski's inequality in $L^2(\pi)$ to get

$$\begin{aligned} W_2(F_\theta(\mu), F_\theta(\nu)) &\le \Big(\int \|\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\mu}(y)\|^2\pi(dx,dy)\Big)^{1/2} + \Big(\int \|\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\|^2\pi(dx,dy)\Big)^{1/2} \\ &\le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}}R^2e^{2c}W_2(\mu,\nu) + \sup_{y \in B(0,R)} \|\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\|. \end{aligned}$$

Using (55) for the supremum term, we obtain

$$W_2(F_\theta(\mu), F_\theta(\nu)) \le \|V\|_{\mathrm{op}}e^{2c}\big(\|B\|_{\mathrm{op}}R^2 + 1 + 2\|B\|_{\mathrm{op}}R^2\big)W_2(\mu,\nu) = \|V\|_{\mathrm{op}}e^{2c}\big(1 + 3\|B\|_{\mathrm{op}}R^2\big)W_2(\mu,\nu).$$

Since $c = \|B\|_{\mathrm{op}}R^2$, this is exactly

$$W_2(F_\theta(\mu), F_\theta(\nu)) \le \|V\|_{\mathrm{op}}\big(1 + 3\|B\|_{\mathrm{op}}R^2\big)\exp\big(2\|B\|_{\mathrm{op}}R^2\big)W_2(\mu,\nu),$$

which proves the claimed W_2-Lipschitz bound on $\mathcal{P}_{2,R}(E)$. $\square$

A.2.9 Theorem 2.13 [3, Theorem 3.4]

Proof. Fix $p \geq 1$ and $R > 0$. We first prove that F_θ is W_p-Lipschitz on $\mathcal{P}_{p,R}(E)$.

Part I: W_p-Lipschitzness of F_θ on $\mathcal{P}_{p,R}(E)$. Work in $E = \mathbb{R}^d$ and let $\theta = (Q, K, V, \beta)$ with $q(x) = Qx$, $k(y) = Ky$, $v(y) = Vy$ and

$$\kappa_\theta(x, y) = \exp\bigl(\beta\langle q(x), k(y)\rangle\bigr).$$

Equivalently, writing $B = \beta Q^\top K \in \mathcal{L}(E)$, we have $\kappa_\theta(x, y) = \exp(\langle x, By\rangle)$. For $\mu \in \mathcal{P}_{p,R}(E)$ and $x \in E$, recall

$$Z_{\theta,\mu}(x) := \int_E \kappa_\theta(x, y)\mu(dy), \quad a_{\theta,\mu}(x, dy) := \frac{\kappa_\theta(x, y)}{Z_{\theta,\mu}(x)}\mu(dy),$$

$$\Gamma_{\theta,\mu}(x) := \int_E v(y)a_{\theta,\mu}(x, dy), \quad F_\theta(\mu) := (\Gamma_{\theta,\mu})_\#\mu.$$

Set $c_R := \|B\|_{\text{op}}R^2$.

Step 1: basic bounds on $Z_{\theta,\mu}$ and $\Gamma_{\theta,\mu}$. If $x, y \in B(0, R)$ then $|\langle x, By\rangle| \leq \|x\|\|B\|_{\text{op}}\|y\| \leq c_R$, hence

$$e^{-c_R} \leq \kappa_\theta(x, y) \leq e^{c_R}. \tag{57}$$

Therefore for all $\mu \in \mathcal{P}_{p,R}(E)$ and $x \in B(0, R)$,

$$e^{-c_R} \leq Z_{\theta,\mu}(x) \leq e^{c_R}, \quad \text{hence} \quad \frac{1}{Z_{\theta,\mu}(x)} \leq e^{c_R}. \tag{58}$$

Moreover $a_{\theta,\mu}(x, \cdot)$ is a probability measure supported on $B(0, R)$, so $\|v(y)\| \leq \|V\|_{\text{op}}\|y\| \leq \|V\|_{\text{op}}R$ on its support and thus

$$\|\Gamma_{\theta,\mu}(x)\| \leq \int_E \|v(y)\|a_{\theta,\mu}(x, dy) \leq \|V\|_{\text{op}}R. \tag{59}$$

Step 2: dependence on μ at fixed x. Fix $x \in B(0, R)$ and let $\mu, \nu \in \mathcal{P}_{p,R}(E)$. Introduce

$$N_{\theta,\mu}(x) := \int_E \kappa_\theta(x, y)v(y)\mu(dy), \quad \text{so that} \quad \Gamma_{\theta,\mu}(x) = \frac{N_{\theta,\mu}(x)}{Z_{\theta,\mu}(x)}.$$

Define $g_x(y) := \kappa_\theta(x, y)$ and $h_x(y) := \kappa_\theta(x, y)v(y)$. For $y_1, y_2 \in B(0, R)$, using that $t \mapsto e^t$ is e^{c_R}-Lipschitz on $[-c_R, c_R]$,

$$|g_x(y_1) - g_x(y_2)| \leq e^{c_R}|\langle x, B(y_1 - y_2)\rangle| \leq e^{c_R}\|B\|_{\text{op}}R\|y_1 - y_2\|.$$

Also,

$$\begin{aligned}\|h_x(y_1) - h_x(y_2)\| &\leq \kappa_\theta(x, y_1)\|v(y_1) - v(y_2)\| + |g_x(y_1) - g_x(y_2)|\|v(y_2)\| \\ &\leq e^{c_R}\|V\|_{\text{op}}\|y_1 - y_2\| + e^{c_R}\|B\|_{\text{op}}R\|y_1 - y_2\|(\|V\|_{\text{op}}R) \\ &\leq e^{c_R}\|V\|_{\text{op}}\bigl(1 + \|B\|_{\text{op}}R^2\bigr)\|y_1 - y_2\|.\end{aligned}$$

Hence, for any coupling π of (μ, ν),

$$\Bigl|\int g_x d\mu - \int g_x d\nu\Bigr| \leq e^{c_R}\|B\|_{\text{op}}R\int \|y_1 - y_2\|d\pi \leq e^{c_R}\|B\|_{\text{op}}RW_1(\mu, \nu) \leq e^{c_R}\|B\|_{\text{op}}RW_p(\mu, \nu),$$

and similarly componentwise for h_x. We obtain

$$|Z_{\theta,\mu}(x) - Z_{\theta,\nu}(x)| \leq e^{c_R}\|B\|_{\text{op}}RW_p(\mu, \nu), \tag{60}$$

$$\|N_{\theta,\mu}(x) - N_{\theta,\nu}(x)\| \leq e^{c_R}\|V\|_{\text{op}}\bigl(1 + \|B\|_{\text{op}}R^2\bigr)W_p(\mu, \nu). \tag{61}$$

Now decompose the quotient using $\Gamma_{\theta,\nu}(x) = N_{\theta,\nu}(x)/Z_{\theta,\nu}(x)$:

$$\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(x) = \frac{N_{\theta,\mu}(x) - N_{\theta,\nu}(x)}{Z_{\theta,\mu}(x)} + \Gamma_{\theta,\nu}(x)\frac{Z_{\theta,\nu}(x) - Z_{\theta,\mu}(x)}{Z_{\theta,\mu}(x)}.$$

Using (58)–(59) together with (60)–(61), we obtain for all $x \in B(0,R)$

$$\|\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(x)\| \le \|V\|_{\mathrm{op}} e^{2c_R}\big(1 + 2\|B\|_{\mathrm{op}} R^2\big) W_p(\mu,\nu). \tag{62}$$

Step 3: dependence on x at fixed μ. Fix $\mu \in \mathcal{P}_{p,R}(E)$. As in Lecture 1, denote by $J\Gamma_{\theta,\mu}(x)$ the Jacobian operator $u \mapsto J\Gamma_{\theta,\mu}(x)[u]$. Differentiating $a_{\theta,\mu}(x,dy) = \kappa_\theta(x,y) Z_{\theta,\mu}(x)^{-1}\mu(dy)$ in the direction $u \in E$ gives

$$Ja_{\theta,\mu}(x,dy)[u] = a_{\theta,\mu}(x,dy)\Big(\langle u, By\rangle - \int_E \langle u, Bz\rangle a_{\theta,\mu}(x,dz)\Big),$$

hence

$$J\Gamma_{\theta,\mu}(x)[u] = \int_E v(y) Ja_{\theta,\mu}(x,dy)[u].$$

By Cauchy–Schwarz under the probability measure $a_{\theta,\mu}(x,\cdot)$,

$$\|J\Gamma_{\theta,\mu}(x)[u]\| \le \Big(\int_E \|v(y)\|^2 a_{\theta,\mu}(x,dy)\Big)^{1/2}\Big(\mathrm{Var}_{a_{\theta,\mu}(x,\cdot)}(\langle u, By\rangle)\Big)^{1/2}.$$

Since $a_{\theta,\mu}(x,\cdot)$ is supported on $B(0,R)$, $\|v(y)\| \le \|V\|_{\mathrm{op}} R$ and $|\langle u, By\rangle| \le \|u\|\|B\|_{\mathrm{op}} R$, hence

$$\|J\Gamma_{\theta,\mu}(x)[u]\| \le \|V\|_{\mathrm{op}} R \cdot \|u\|\|B\|_{\mathrm{op}} R = \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}} R^2\|u\|.$$

Therefore $\|J\Gamma_{\theta,\mu}(x)\|_{\mathrm{op}} \le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}} R^2$ for all $x \in B(0,R)$, and by the mean value theorem,

$$\|\Gamma_{\theta,\mu}(x_1) - \Gamma_{\theta,\mu}(x_2)\| \le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}} R^2\|x_1 - x_2\| \le \|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}} R^2 e^{2c_R}\|x_1 - x_2\| \tag{63}$$

for all $x_1, x_2 \in B(0,R)$ (the last inequality uses $e^{2c_R} \ge 1$).

Step 4: pushforward coupling in W_p. Let $\mu,\nu \in \mathcal{P}_{p,R}(E)$ and let π be an optimal coupling: $W_p^p(\mu,\nu) = \int \|x - y\|^p \pi(dx,dy)$. Push π forward by $(x,y) \mapsto (\Gamma_{\theta,\mu}(x), \Gamma_{\theta,\nu}(y))$ to obtain a coupling of $(F_\theta(\mu), F_\theta(\nu))$. Then

$$W_p(F_\theta(\mu), F_\theta(\nu)) \le \Big(\int \|\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(y)\|^p \pi(dx,dy)\Big)^{1/p}.$$

Decompose

$$\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\nu}(y) = \big(\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\mu}(y)\big) + \big(\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\big),$$

and apply Minkowski's inequality in $L^p(\pi)$:

$$W_p(F_\theta(\mu), F_\theta(\nu)) \le \Big(\int \|\Gamma_{\theta,\mu}(x) - \Gamma_{\theta,\mu}(y)\|^p \pi(dx,dy)\Big)^{1/p} + \Big(\int \|\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\|^p \pi(dx,dy)\Big)^{1/p}.$$

The first term is bounded using (63) by $\|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}} R^2 e^{2c_R} W_p(\mu,\nu)$. The second term depends only on the second marginal of π, i.e. ν, so it is

$$\Big(\int \|\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\|^p \nu(dy)\Big)^{1/p} \le \sup_{y \in B(0,R)} \|\Gamma_{\theta,\mu}(y) - \Gamma_{\theta,\nu}(y)\| \overset{(62)}{\le} \|V\|_{\mathrm{op}} e^{2c_R}\big(1 + 2\|B\|_{\mathrm{op}} R^2\big) W_p(\mu,\nu).$$

Combining yields

$$W_p(F_\theta(\mu), F_\theta(\nu)) \le \|V\|_{\mathrm{op}} e^{2c_R}\big(1 + 3\|B\|_{\mathrm{op}} R^2\big) W_p(\mu,\nu),$$

so F_θ is W_p-Lipschitz on $\mathcal{P}_{p,R}(E)$.

Part II: exponential lower bound when $p = 2$, $V = \mathrm{Id}_E$, and B is symmetric. Assume now $p = 2$, $V = \mathrm{Id}_E$, and B is symmetric with eigenvalues $\gamma_1 \ge \cdots \ge \gamma_d$ and orthonormal eigenvectors $u_1, \ldots, u_d$. Recall

$$\mathrm{Lip}_{\mathcal{P}_{2,R}(E)}(F_\theta) = \sup_{\mu \in \mathcal{P}_{2,R}(E)} \mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu),$$

hence for any $\mu_R \in \mathcal{P}_{2,R}(E)$,

$$\mathrm{Lip}_{\mathcal{P}_{2,R}(E)}(F_\theta) \;\ge\; \mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R). \tag{64}$$

Case 1: $\gamma_1 \geq -8\gamma_d$**.** Define

$$p_R := e^{-\gamma_1 R^2/4}, \quad \mu_R := p_R \delta_{Ru_1} + (1-p_R)\delta_{(R/2)u_1}.$$

Then $\mu_R \in \mathcal{P}_{2,R}(E)$. Proposition 2.14 yields a function $\vartheta : [0,+\infty) \to [0,+\infty)$ with $\vartheta(R) \to 1$ as $R \to +\infty$ such that

$$\mathrm{Lip}_{\mathrm{loc}}^{W_2}(F_\theta; \mu_R) \geq \vartheta(R) \frac{\gamma_1}{16} R^2 e^{\gamma_1 R^2/8}.$$

Combining with (64) gives item (1).

Case 2: $\gamma_1 < -8\gamma_d$**.** Define

$$p_R := e^{-2|\gamma_d| R^2}, \quad \mu_R := p_R \delta_{Ru_d} + (1-p_R)\delta_{-Ru_d}.$$

Then $\mu_R \in \mathcal{P}_{2,R}(E)$. Proposition 2.14 yields a function $\vartheta : [0,+\infty) \to [0,+\infty)$ with $\vartheta(R) \to 1$ as $R \to +\infty$ such that

$$\mathrm{Lip}_{\mathrm{loc}}^{W_2}(F_\theta; \mu_R) \geq \vartheta(R) \frac{|\gamma_d|}{2} R^2 e^{|\gamma_d| R^2}.$$

Combining with (64) gives item (2).

Asymptotics. Proposition 2.14 also states that the lower bounds above are asymptotic toward $\mathrm{Lip}_{\mathrm{loc}}^{W_2}(F_\theta; \mu_R)$ as $R \to +\infty$ for the chosen two-Dirac measures. This yields the final sentence of the theorem. □

A.2.10 Proposition 2.14 [3, Proposition 3.5]

Proof. We work in $E = \mathbb{R}^d$ with $p = 2$ and assume $V = \mathrm{Id}_E$. Let $\theta = (Q, K, V, \beta)$ and set $B := \beta Q^\top K \in \mathcal{L}(E)$. Assume that B is symmetric, with eigenvalues $\gamma_1 \geq \cdots \geq \gamma_d$ and orthonormal eigenvectors $u_1, \ldots, u_d$.

Step 1: two-Dirac coordinates and a local W_2 **chart.** Fix a weight $p \in (0,1)$ and points $x \neq y$ in $\mathbb{R}^d$, and consider the two-Dirac measure

$$\mu = p\delta_x + (1-p)\delta_y.$$

Write $X := (x, y) \in E^2 \simeq \mathbb{R}^{2d}$ and define the weighted inner product and norm on E^2:

$$\langle (h_1, h_2), (k_1, k_2) \rangle_a := p\langle h_1, k_1 \rangle + (1-p)\langle h_2, k_2 \rangle, \quad \|(h_1, h_2)\|_a^2 := p\|h_1\|^2 + (1-p)\|h_2\|^2,$$

where $a := (p, 1-p)$.

For $X' = (x', y')$ in a neighbourhood of $X = (x, y)$, define

$$\mu' := p\delta_{x'} + (1-p)\delta_{y'}.$$

Any coupling $\pi \in \Pi(\mu, \mu')$ is supported on $\{x, y\} \times \{x', y'\}$ and hence is determined by a 2×2 matrix of nonnegative masses with prescribed row and column sums $(p, 1-p)$. In particular, $\Pi(\mu, \mu')$ is a compact convex polytope with exactly two extreme points, and since the quadratic transport cost is linear in π, an optimal coupling is attained at an extreme point. Consequently, $W_2^2(\mu, \mu')$ is the minimum of the transport costs corresponding to these two extreme couplings; one of them is the identity plan sending $x \mapsto x'$ and $y \mapsto y'$.

Define the (continuous) cost gap function

$$\Delta(X') := \Big[\text{cost of the non-identity extreme plan}\Big] - \Big[p\|x - x'\|^2 + (1-p)\|y - y'\|^2\Big].$$

At $X' = X$ we have

$$\Delta(X) = \|x - y\|^2 > 0 \quad (\text{since } x \neq y).$$

By continuity of Δ, there exists $\varepsilon > 0$ such that

$$\|x - x'\| + \|y - y'\| < \varepsilon \quad \Longrightarrow \quad \Delta(X') > 0,$$

i.e. the identity plan has strictly smaller cost than the other extreme plan. Therefore, for all X' in a sufficiently small neighbourhood of X, the identity plan is optimal and

$$W_2^2(\mu, \mu') = p\|x - x'\|^2 + (1-p)\|y - y'\|^2 = \|X - X'\|_a^2. \tag{65}$$

Step 2: the induced two-point map f_a and its Jacobian. For $\mu = p\delta_x + (1-p)\delta_y$, recall the mean-field kernel $\kappa_\theta(\xi,\eta) = \exp(\langle \xi, B\eta\rangle)$ and define, for the two atoms,

$$Z_{\theta,\mu}(x) = pe^{\langle x,Bx\rangle} + (1-p)e^{\langle x,By\rangle}, \quad Z_{\theta,\mu}(y) = pe^{\langle y,Bx\rangle} + (1-p)e^{\langle y,By\rangle}.$$

Set the two attention weights (for query x and query y)

$$\pi_1 := \frac{pe^{\langle x,Bx\rangle}}{Z_{\theta,\mu}(x)}, \quad \pi_2 := \frac{(1-p)e^{\langle y,By\rangle}}{Z_{\theta,\mu}(y)}.$$

Since $V = \mathrm{Id}_E$, the pushforward map $\Gamma_{\theta,\mu}$ evaluated at the atoms is

$$\Gamma_{\theta,\mu}(x) = \pi_1 x + (1-\pi_1)y, \quad \Gamma_{\theta,\mu}(y) = (1-\pi_2)x + \pi_2 y.$$

Define the induced map on E^2 (with fixed weights $a = (p, 1-p)$)

$$f_a : E^2 \to E^2, \quad f_a(x,y) := \big(\Gamma_{\theta,\mu}(x), \Gamma_{\theta,\mu}(y)\big), \quad \mu = p\delta_x + (1-p)\delta_y.$$

By (65), for $X' = (x', y')$ close to $X = (x,y)$ and μ' as above, we also have

$$W_2\big(F_\theta(\mu), F_\theta(\mu')\big) = \|f_a(X) - f_a(X')\|_a, \quad W_2(\mu,\mu') = \|X - X'\|_a.$$

Therefore, restricting the $\limsup_{\nu\to\mu}$ in $\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu)$ to such two-Dirac perturbations yields the lower bound

$$\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu) \;\geq\; \mathrm{Lip}_{\mathrm{loc}}(f_a;X) \;\geq\; \|Jf_a(X)\|_{\mathrm{op},a}, \tag{66}$$

where $\|Jf_a(X)\|_{\mathrm{op},a} := \sup_{U\neq 0} \|Jf_a(X)[U]\|_a / \|U\|_a$.

A direct differentiation gives the block form

$$Jf_a(X) = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix} \quad \text{as an operator on } E^2,$$

where (as linear maps $E \to E$)

$$B_1 = \pi_1 I + \pi_1(1-\pi_1)(x-y)(2x-y)^\top B, \tag{67}$$
$$B_2 = (1-\pi_1)I - \pi_1(1-\pi_1)(x-y)x^\top B, \tag{68}$$
$$B_3 = (1-\pi_2)I - \pi_2(1-\pi_2)(y-x)y^\top B, \tag{69}$$
$$B_4 = \pi_2 I + \pi_2(1-\pi_2)(y-x)(2y-x)^\top B. \tag{70}$$

(These are obtained by differentiating π_1, π_2 using $d\pi_1 = \pi_1(1-\pi_1)d(\langle x,Bx\rangle - \langle x,By\rangle)$ and $d\pi_2 = \pi_2(1-\pi_2)d(\langle y,By\rangle - \langle y,Bx\rangle)$, and then differentiating the affine combinations defining f_a.)

Step 3: Case 1 ($\gamma_1 \geq -8\gamma_d$). Assume $\gamma_1 \geq -8\gamma_d$ and set $C := \gamma_1/8 > 0$. For each $R > 0$, define

$$p_R := e^{-2CR^2}, \quad x_R := Ru_1, \quad y_R := \frac{R}{2}u_1, \quad \mu_R := p_R\delta_{x_R} + (1-p_R)\delta_{y_R}.$$

Let $X_R := (x_R, y_R)$ and $a_R := (p_R, 1-p_R)$. Since $Bu_1 = \gamma_1 u_1$, the relevant inner products are

$$\langle x_R, Bx_R\rangle = \gamma_1 R^2, \quad \langle x_R, By_R\rangle = \frac{\gamma_1}{2}R^2, \quad \langle y_R, Bx_R\rangle = \frac{\gamma_1}{2}R^2, \quad \langle y_R, By_R\rangle = \frac{\gamma_1}{4}R^2.$$

Compute π_1, π_2 at (p_R, x_R, y_R):

$$\pi_1 = \frac{p_R e^{\gamma_1 R^2}}{p_R e^{\gamma_1 R^2} + (1-p_R)e^{\gamma_1 R^2/2}} = \frac{1}{1 + \frac{1-p_R}{p_R}e^{-\gamma_1 R^2/2}},$$

so using $p_R = e^{-\gamma_1 R^2/4}$ we have

$$1 - \pi_1 \sim e^{-\gamma_1 R^2/4}, \quad \pi_1(1-\pi_1) \sim e^{-\gamma_1 R^2/4}. \tag{71}$$

Similarly,

$$\pi_2 = \frac{(1-p_R)e^{\gamma_1 R^2/4}}{p_R e^{\gamma_1 R^2/2} + (1-p_R)e^{\gamma_1 R^2/4}} = \frac{1}{1+\frac{p_R}{1-p_R}e^{\gamma_1 R^2/4}} \longrightarrow \frac{1}{2},$$

hence

$$\pi_2 \to \frac{1}{2}, \quad \pi_2(1-\pi_2) \to \frac{1}{4}. \tag{72}$$

Now evaluate the blocks (67)–(70) at X_R. Using $x_R - y_R = \frac{R}{2}u_1$, $2x_R - y_R = \frac{3R}{2}u_1$, $y_R - x_R = -\frac{R}{2}u_1$, and $2y_R - x_R = 0$, we obtain:

$$\begin{aligned}
B_1 &= \pi_1 I + \pi_1(1-\pi_1)\frac{3R^2}{4}u_1u_1^\top B = I + O\left(R^2 e^{-\gamma_1 R^2/4}\right),\\
B_2 &= (1-\pi_1)I - \pi_1(1-\pi_1)\frac{R^2}{2}u_1u_1^\top B = O\left(R^2 e^{-\gamma_1 R^2/4}\right),\\
B_4 &= \pi_2 I \to \frac{1}{2}I,\\
B_3 &= (1-\pi_2)I - \pi_2(1-\pi_2)(y_R - x_R)y_R^\top B = (1-\pi_2)I + \pi_2(1-\pi_2)\frac{\gamma_1 R^2}{4}u_1u_1^\top.
\end{aligned}$$

In particular, on the line $\operatorname{span}\{u_1\}$ the map B_3 has eigenvalue

$$\lambda_R := (1-\pi_2) + \pi_2(1-\pi_2)\frac{\gamma_1 R^2}{4} = \frac{\gamma_1 R^2}{16} + O(1), \tag{73}$$

while on $u_1^\perp$ we have $u_1u_1^\top = 0$ and hence $B_3 = (1-\pi_2)I$ is bounded.

Step 4: weighted operator norm and asymptotics. Consider the two-dimensional subspace

$$E_1 := \operatorname{span}\{(u_1,0),(0,u_1)\} \subset E^2.$$

Let $e_1 := (u_1, 0)$ and $e_2 := (0, u_1)$. From the block asymptotics above, the restriction of $Jf_{a_R}(X_R)$ to E_1, in the basis (e_1, e_2), has the form

$$\begin{pmatrix} 1+o(1) & o(1) \\ \lambda_R + o(\lambda_R) & \frac{1}{2}+o(1) \end{pmatrix}, \quad \lambda_R = \frac{\gamma_1 R^2}{16} + O(1).$$

Now relate the weighted norm to the Euclidean norm: for vectors $(\alpha,\beta) \in \mathbb{R}^2$,

$$\|\alpha e_1 + \beta e_2\|_{a_R}^2 = p_R|\alpha|^2 + (1-p_R)|\beta|^2 = \left\|\operatorname{Diag}(\sqrt{p_R}, \sqrt{1-p_R})(\alpha,\beta)\right\|_2^2.$$

Hence the induced operator norm on E_1 for $\|\cdot\|_{a_R}$ equals the Euclidean operator norm of the conjugated 2×2 matrix

$$\operatorname{Diag}(\sqrt{p_R}, \sqrt{1-p_R})\begin{pmatrix} 1+o(1) & o(1) \\ \lambda_R + o(\lambda_R) & \frac{1}{2}+o(1) \end{pmatrix}\operatorname{Diag}(\tfrac{1}{\sqrt{p_R}}, \tfrac{1}{\sqrt{1-p_R}}).$$

Its $(2,1)$ entry is

$$s_R := \lambda_R\sqrt{\frac{1-p_R}{p_R}} = \left(\frac{\gamma_1 R^2}{16} + O(1)\right)\frac{1+o(1)}{\sqrt{p_R}} = \left(\frac{\gamma_1 R^2}{16} + O(1)\right)e^{CR^2},$$

because $\sqrt{p_R} = e^{-CR^2}$ and $\sqrt{1-p_R} \to 1$. Thus the conjugated matrix on E_1 is

$$\begin{pmatrix} 1+o(1) & o(1) \\ s_R + o(s_R) & \frac{1}{2}+o(1) \end{pmatrix}.$$

For the model matrix $\left(\begin{smallmatrix} 1 & 0 \\ s & 1/2 \end{smallmatrix}\right)$ we have the elementary bounds

$$\left\| \begin{pmatrix} 1 & 0 \\ s & 1/2 \end{pmatrix} \right\|_{\mathrm{op}} \geq \left\| \begin{pmatrix} 1 & 0 \\ s & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|_2 = \sqrt{1+s^2} \ \geq \ |s|,$$

and also

$$\left\| \begin{pmatrix} 1 & 0 \\ s & 1/2 \end{pmatrix} \right\|_{\mathrm{op}} \leq \left\| \begin{pmatrix} 1 & 0 \\ s & 1/2 \end{pmatrix} \right\|_{\mathrm{F}} = \sqrt{1+s^2+\frac{1}{4}} \ \leq \ |s|+1.$$

Applying these with $s = s_R$ and using that the $o(1)$ and $o(s_R)$ perturbations change the operator norm by at most $o(|s_R|)$, we obtain

$$\|Jf_{a_R}(X_R)|_{E_1}\|_{\mathrm{op},a_R} = |s_R|(1+o(1)) = \frac{\gamma_1}{16}R^2 e^{CR^2}(1+o(1)) = \frac{C}{2}R^2 e^{CR^2}(1+o(1)).$$

Next, we show that directions orthogonal to E_1 do not contribute at scale $|s_R|$. Indeed, the only term growing like R^2 in the blocks is the rank-one term $u_1u_1^\top$ inside B_3, which vanishes on $u_1^\perp$. All remaining pieces of B_1, B_2, B_3, B_4 are uniformly bounded (using $0 \leq \pi_i \leq 1$ and (71)–(72)). After the same diagonal conjugation corresponding to $\|\cdot\|_{a_R}$, these bounded pieces remain $O(1)$. Therefore,

$$\|Jf_{a_R}(X_R)\|_{\mathrm{op},a_R} = |s_R|(1+o(1)).$$

Finally, combine with (66) to obtain

$$\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R) \ \geq \ \|Jf_{a_R}(X_R)\|_{\mathrm{op},a_R} = \frac{C}{2}R^2 e^{CR^2}(1+o(1)).$$

This proves case (1) with $\vartheta_1(R) = 1+o(1)$.

Step 5: Case 2 ($\gamma_1 < -8\gamma_d$). Assume $\gamma_1 < -8\gamma_d$ and set $C' := |\gamma_d| > 0$. Define

$$p_R := e^{-2C'R^2}, \quad x_R := Ru_d, \quad y_R := -Ru_d, \quad \mu_R := p_R\delta_{x_R} + (1-p_R)\delta_{y_R}.$$

Then $Bu_d = \gamma_d u_d = -C'u_d$. The logits are

$$\langle x_R, Bx_R\rangle = -C'R^2, \quad \langle x_R, By_R\rangle = C'R^2, \quad \langle y_R, Bx_R\rangle = C'R^2, \quad \langle y_R, By_R\rangle = -C'R^2.$$

A direct check from the definitions gives $\pi_1 \to 0$ (in fact $\pi_1 \sim e^{-4C'R^2}$) and $\pi_2 \to \frac{1}{2}$, hence $\pi_2(1-\pi_2) \to \frac{1}{4}$. Evaluating the blocks (67)–(70) at (x_R, y_R) shows again that the only growing term lies in B_3 along u_d, now with coefficient

$$(y_R - x_R)y_R^\top B = (-2Ru_d)(-Ru_d)^\top B = 2R^2u_du_d^\top B = 2\gamma_d R^2 u_du_d^\top = -2C'R^2u_du_d^\top,$$

so

$$B_3 = (1-\pi_2)I + \pi_2(1-\pi_2)\cdot 2C'R^2u_du_d^\top,$$

and along $\mathrm{span}\{u_d\}$ the eigenvalue is $\frac{C'}{2}R^2 + O(1)$. Repeating Step 4 with u_d in place of u_1 yields

$$\mathrm{Lip}^{W_2}_{\mathrm{loc}}(F_\theta;\mu_R) \ \geq \ \frac{C'}{2}R^2 e^{C'R^2}(1+o(1)),$$

i.e. case (2) with $\vartheta_2(R) = 1+o(1)$. □

A.2.11 Proposition 2.15 [3, Proposition 3.6]

Proof. Fix $R > 0$ and let $n = n(R)$ satisfy $n(R) \sim_{R\to\infty} e^{2\gamma R^2}$. We construct $X \in B_R^n$ and $Y \in E^n$ (depending on R) with $\|Y\|_F = 1$ such that

$$\|JS_\theta(X)[Y]\|_F \geq \vartheta(R)\frac{\gamma}{2}R^2 e^{\gamma R^2},$$

which implies the claim since $\mathrm{Lip}_{B_R^n}(S_\theta) \geq \mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) \geq \|JS_\theta(X)[Y]\|_F$.

Case 1: $\gamma = \gamma_1/8$**.** Let u be a unit vector such that $Bu = \gamma_1 u$ and $B^\top u = \gamma_1 u$. Set $x^\star := Ru$ and $x^\circ := (R/2)u$, and define

$$x_1 := x^\star, \quad x_2 = \cdots = x_n := x^\circ,$$

and the direction

$$y_1 := u, \quad y_2 = \cdots = y_n := 0,$$

so that $\|Y\|_F = 1$.

Fix any common query index $i \in \{2, \dots, n\}$. Since all common keys are equal, the attention output for query i has the two-point form

$$(S_\theta(X))_i = p_R x^\star + (1 - p_R)x^\circ,$$

where p_R is the aggregate attention mass assigned by query x° to the unique key $x^\star$:

$$p_R := \frac{\exp(\langle Bx^\circ, x^\star\rangle)}{(n-1)\exp(\langle Bx^\circ, x^\circ\rangle) + \exp(\langle Bx^\circ, x^\star\rangle)}.$$

Compute the logit gap

$$\langle Bx^\circ, x^\star\rangle - \langle Bx^\circ, x^\circ\rangle = \gamma_1\Big(\frac{R}{2}\Big)R - \gamma_1\Big(\frac{R}{2}\Big)^2 = \frac{\gamma_1}{4}R^2 = 2\gamma R^2.$$

Hence

$$p_R = \frac{1}{1 + (n-1)e^{-2\gamma R^2}}.$$

Since $n(R) \sim e^{2\gamma R^2}$, we have $(n-1)e^{-2\gamma R^2} \to 1$ and thus

$$p_R \to \frac{1}{2}, \quad p_R(1 - p_R) \to \frac{1}{4}.$$

Differentiate the common output $(S_\theta(X))_i$ along Y. Only $x^\star$ varies, so

$$J(S_\theta)_i(X)[Y] = (dp_R)(x^\star - x^\circ) + p_R y_1.$$

Moreover, p_R depends on $x^\star$ only through the single logit $\langle Bx^\circ, x^\star\rangle$, hence by the softmax derivative,

$$dp_R = p_R(1 - p_R)d\langle Bx^\circ, x^\star\rangle = p_R(1 - p_R)\langle B^\top x^\circ, y_1\rangle.$$

Using $B^\top u = \gamma_1 u$, $x^\circ = (R/2)u$, and $y_1 = u$ gives

$$\langle B^\top x^\circ, y_1\rangle = \gamma_1 \frac{R}{2}.$$

Also $x^\star - x^\circ = (R/2)u$. Therefore

$$J(S_\theta)_i(X)[Y] = \Big(p_R(1 - p_R)\frac{\gamma_1}{4}R^2 + p_R\Big)u,$$

and hence, for R large,

$$\|J(S_\theta)_i(X)[Y]\| \geq \Big(p_R(1 - p_R)\frac{\gamma_1}{4}R^2 - p_R\Big) = \Big(\frac{\gamma}{2}R^2\Big)\Big(4p_R(1 - p_R) - \frac{2p_R}{\gamma R^2}\Big).$$

The factor in parentheses tends to 1 as $R \to \infty$.

Finally, the same lower bound holds for every common index $i = 2, \dots, n$, so by Frobenius aggregation,

$$\|JS_\theta(X)[Y]\|_F \geq \sqrt{n-1}\|J(S_\theta)_2(X)[Y]\|.$$

Since $\sqrt{n-1}/e^{\gamma R^2} \to 1$ (from $n \sim e^{2\gamma R^2}$), we obtain

$$\|JS_\theta(X)[Y]\|_F \geq \vartheta(R)\frac{\gamma}{2}R^2 e^{\gamma R^2}, \quad \vartheta(R) \to 1.$$

Case 2: $\gamma = -\gamma_r$ Let $u = u_r$ be a unit vector such that $Bu = \gamma_r u$ and $B^\top u = \gamma_r u$ (with $\gamma_r < 0$), and set $\gamma := |\gamma_r|$. Define

$$x^\circ := Ru, \quad x^\star := -Ru,$$

and again take one rare token at $x^\star$ and $n-1$ common tokens at x°:

$$x_1 := x^\star, \quad x_2 = \cdots = x_n := x^\circ, \quad y_1 := u,\ y_2 = \cdots = y_n := 0.$$

For a common query $i \geq 2$, the aggregate mass p_R assigned to the rare key is

$$p_R = \frac{1}{1 + (n-1)e^{-2\gamma R^2}},$$

because the logit gap is $\langle Bx^\circ, x^\star\rangle - \langle Bx^\circ, x^\circ\rangle = (-\gamma_r R^2) - (\gamma_r R^2) = 2\gamma R^2$. The same differentiation as in Case 1 yields

$$J(S_\theta)_i(X)[Y] = \Big(2\gamma p_R(1-p_R)R^2 + p_R\Big)u,$$

so again

$$\|JS_\theta(X)[Y]\|_F \geq \vartheta(R)\frac{\gamma}{2}R^2 e^{\gamma R^2}, \quad \vartheta(R) \to 1.$$

In both cases, we have exhibited $X \in B_R^n$ such that $\mathrm{Lip}_{\mathrm{loc}}(S_\theta; X) \geq \vartheta(R)\frac{\gamma}{2}R^2 e^{\gamma R^2}$. Taking the supremum over $X \in B_R^n$ proves the claim. $\square$

A.2.12 Theorem 2.16 [3, Theorem 4.4]

Proof. Let $\bar\mu, \bar\nu \in \mathcal{P}_{p,R}([0,1] \times E)$ be distinct. If $(\mathrm{pr}_1)_\#\bar\mu \neq (\mathrm{pr}_1)_\#\bar\nu$, then $d_p(\bar\mu, \bar\nu) = +\infty$ by definition and there is nothing to prove. Hence assume that $\bar\mu$ and $\bar\nu$ have the same first marginal $\vartheta \in \mathcal{P}([0,1])$. By disintegration, there exist Borel families $(\mu^s)_{s\in[0,1]}$ and $(\nu^s)_{s\in[0,1]}$ in $\mathcal{P}(B(0,R))$ such that

$$\bar\mu(ds, dx) = \vartheta(ds)\mu^s(dx), \quad \bar\nu(ds, dx) = \vartheta(ds)\nu^s(dx),$$

and

$$d_p(\bar\mu, \bar\nu)^p = \int_0^1 W_p(\mu^s, \nu^s)^p \vartheta(ds).$$

For $s \in [0,1]$ and $x \in B(0,R)$, define the causal masked barycentric map

$$\Gamma^s_{\bar\mu}(x) := \frac{\displaystyle\int_{[0,1]\times E} Vy e^{\langle x, By\rangle}\mathbf{1}_{\{t\leq s\}}\bar\mu(dt, dy)}{\displaystyle\int_{[0,1]\times E} e^{\langle x, By\rangle}\mathbf{1}_{\{t\leq s\}}\bar\mu(dt, dy)},$$

and define $\Gamma^s_{\bar\nu}$ analogously. Since $\mathrm{supp}(\bar\mu), \mathrm{supp}(\bar\nu) \subset [0,1] \times B(0,R)$ and $|\langle x, By\rangle| \leq \|B\|_{\mathrm{op}}R^2$, the denominators are finite and positive for ϑ-a.e. s and μ^s-a.e. x (and similarly for $\bar\nu$).

By construction of F^M_θ (time coordinate unchanged),

$$F^M_\theta(\bar\mu)(ds, dx) = \vartheta(ds)\sigma^s(dx), \quad \sigma^s = (\Gamma^s_{\bar\mu})_\#\mu^s,$$

and similarly

$$F^M_\theta(\bar\nu)(ds, dx) = \vartheta(ds)\rho^s(dx), \quad \rho^s = (\Gamma^s_{\bar\nu})_\#\nu^s.$$

Therefore

$$d_p\big(F^M_\theta(\bar\mu), F^M_\theta(\bar\nu)\big)^p = \int_0^1 W_p(\sigma^s, \rho^s)^p\vartheta(ds).$$

Step 1: splitting the variation. For each s, the triangle inequality yields

$$W_p\big((\Gamma^s_{\bar\mu})_\#\mu^s, (\Gamma^s_{\bar\nu})_\#\nu^s\big) \leq W_p\big((\Gamma^s_{\bar\mu})_\#\mu^s, (\Gamma^s_{\bar\mu})_\#\nu^s\big) + W_p\big((\Gamma^s_{\bar\mu})_\#\nu^s, (\Gamma^s_{\bar\nu})_\#\nu^s\big).$$

Taking the $L^p(\vartheta)$-norm in s and using Minkowski gives

$$\begin{aligned} d_p\big(F^M_\theta(\bar\mu), F^M_\theta(\bar\nu)\big) \leq{}& \Big(\int_0^1 W_p\big((\Gamma^s_{\bar\mu})_\#\mu^s, (\Gamma^s_{\bar\mu})_\#\nu^s\big)^p\vartheta(ds)\Big)^{1/p} \\ &+ \Big(\int_0^1 W_p\big((\Gamma^s_{\bar\mu})_\#\nu^s, (\Gamma^s_{\bar\nu})_\#\nu^s\big)^p\vartheta(ds)\Big)^{1/p}. \end{aligned}$$

Step 2: first term (pushforward by a Lipschitz map). If $f : B(0,R) \to E$ is L-Lipschitz, then for any $\alpha, \beta \in \mathcal{P}(B(0,R))$,

$$W_p(f_\# \alpha, f_\# \beta) \le L W_p(\alpha, \beta).$$

We next bound $\mathrm{Lip}(\Gamma^s_{\bar{\mu}})$ on $B(0,R)$, uniformly in s. Let (T,Y) have the normalized kernel law used in $\Gamma^s_{\bar{\mu}}(x)$, so that $\Gamma^s_{\bar{\mu}}(x) = \mathbb{E}[VY|x]$. Differentiating in x yields

$$\nabla_x \Gamma^s_{\bar{\mu}}(x) = \mathrm{Cov}\left(VY,\ BY \middle| x\right).$$

Hence, by Cauchy–Schwarz,

$$\|\nabla_x \Gamma^s_{\bar{\mu}}(x)\|_{\mathrm{op}} \le \sqrt{\mathbb{E}[\|VY - \mathbb{E}[VY|x]\|^2|x]}\sqrt{\mathbb{E}[\|BY - \mathbb{E}[BY|x]\|^2|x]} \le \|V\|_{\mathrm{op}} R \cdot \|B\|_{\mathrm{op}} R.$$

Therefore $\mathrm{Lip}(\Gamma^s_{\bar{\mu}}) \le \|V\|_{\mathrm{op}} \|B\|_{\mathrm{op}} R^2$ uniformly in s. Applying the pushforward estimate with $f = \Gamma^s_{\bar{\mu}}$ gives

$$W_p\big((\Gamma^s_{\bar{\mu}})_\# \mu^s, (\Gamma^s_{\bar{\mu}})_\# \nu^s\big) \le \|V\|_{\mathrm{op}} \|B\|_{\mathrm{op}} R^2 W_p(\mu^s, \nu^s).$$

Integrating in s yields

$$\Big(\int_0^1 W_p\big((\Gamma^s_{\bar{\mu}})_\# \mu^s, (\Gamma^s_{\bar{\mu}})_\# \nu^s\big)^p \vartheta(ds)\Big)^{1/p} \le \|V\|_{\mathrm{op}} \|B\|_{\mathrm{op}} R^2 d_p(\bar{\mu}, \bar{\nu}).$$

Step 3: second term (same input measure, two maps). Fix $s \in [0,1]$. Consider the coupling

$$\pi_s := (\Gamma^s_{\bar{\mu}}, \Gamma^s_{\bar{\nu}})_\# \nu^s \in \Pi\big((\Gamma^s_{\bar{\mu}})_\# \nu^s,\ (\Gamma^s_{\bar{\nu}})_\# \nu^s\big).$$

Then

$$W_p\big((\Gamma^s_{\bar{\mu}})_\# \nu^s, (\Gamma^s_{\bar{\nu}})_\# \nu^s\big)^p \le \int \|\Gamma^s_{\bar{\mu}}(x) - \Gamma^s_{\bar{\nu}}(x)\|^p \nu^s(dx),$$

hence

$$W_p\big((\Gamma^s_{\bar{\mu}})_\# \nu^s, (\Gamma^s_{\bar{\nu}})_\# \nu^s\big) \le \|\Gamma^s_{\bar{\mu}} - \Gamma^s_{\bar{\nu}}\|_{L^\infty(B(0,R))}.$$

Fix $x \in B(0,R)$ and write, with $g_x(y) := e^{\langle x, By\rangle}$ and $h_x(y) := Vy e^{\langle x, By\rangle}$,

$$Z^s_{\bar{\mu}}(x) := \int_{[0,1]\times E} g_x(y) \mathbf{1}_{\{t \le s\}} \bar{\mu}(dt, dy), \quad N^s_{\bar{\mu}}(x) := \int_{[0,1]\times E} h_x(y) \mathbf{1}_{\{t \le s\}} \bar{\mu}(dt, dy),$$

so that $\Gamma^s_{\bar{\mu}}(x) = N^s_{\bar{\mu}}(x)/Z^s_{\bar{\mu}}(x)$, and similarly for $\bar{\nu}$.

Set $c_R := \|B\|_{\mathrm{op}} R^2$. Since $|\langle x, By\rangle| \le c_R$ on $B(0,R) \times B(0,R)$,

$$e^{-c_R} \le g_x(y) \le e^{c_R} \quad \Rightarrow \quad Z^s_{\bar{\mu}}(x) \ge e^{-c_R} \vartheta([0,s]), \quad \|\Gamma^s_{\bar{\nu}}(x)\| \le \|V\|_{\mathrm{op}} R.$$

Moreover, for all $y_1, y_2 \in B(0,R)$,

$$|g_x(y_1) - g_x(y_2)| \le e^{c_R} \|B\|_{\mathrm{op}} R \|y_1 - y_2\|, \quad \|h_x(y_1) - h_x(y_2)\| \le e^{c_R} \|V\|_{\mathrm{op}} \big(1 + \|B\|_{\mathrm{op}} R^2\big) \|y_1 - y_2\|.$$

Hence $\mathrm{Lip}(g_x) \le e^{c_R} \|B\|_{\mathrm{op}} R$ and $\mathrm{Lip}(h_x) \le e^{c_R} \|V\|_{\mathrm{op}} (1 + \|B\|_{\mathrm{op}} R^2)$.

Using the disintegration $\bar{\mu}(dt, dy) = \vartheta(dt) \mu^t(dy)$ and $\bar{\nu}(dt, dy) = \vartheta(dt) \nu^t(dy)$ and the Kantorovich–Rubinstein bound for Lipschitz test functions, we obtain

$$|Z^s_{\bar{\mu}}(x) - Z^s_{\bar{\nu}}(x)| \le \int_0^s \mathrm{Lip}(g_x) W_1(\mu^t, \nu^t) \vartheta(dt) \le \mathrm{Lip}(g_x) \int_0^s W_p(\mu^t, \nu^t) \vartheta(dt),$$
$$\|N^s_{\bar{\mu}}(x) - N^s_{\bar{\nu}}(x)\| \le \int_0^s \mathrm{Lip}(h_x) W_1(\mu^t, \nu^t) \vartheta(dt) \le \mathrm{Lip}(h_x) \int_0^s W_p(\mu^t, \nu^t) \vartheta(dt),$$

where we used $W_1 \le W_p$ on the bounded set $B(0,R)$.

Now decompose the quotient:

$$\Gamma^s_{\bar{\mu}}(x) - \Gamma^s_{\bar{\nu}}(x) = \frac{N^s_{\bar{\mu}}(x) - N^s_{\bar{\nu}}(x)}{Z^s_{\bar{\mu}}(x)} + \Gamma^s_{\bar{\nu}}(x) \frac{Z^s_{\bar{\nu}}(x) - Z^s_{\bar{\mu}}(x)}{Z^s_{\bar{\mu}}(x)}.$$

Using $1/Z^s_{\bar\mu}(x) \le e^{c_R}/\vartheta([0,s])$ and the bounds above yields

$$\|\Gamma^s_{\bar\mu}(x) - \Gamma^s_{\bar\nu}(x)\| \le \|V\|_{\mathrm{op}} e^{2c_R}\big(1 + 2\|B\|_{\mathrm{op}} R^2\big) \frac{1}{\vartheta([0,s])} \int_0^s W_p(\mu^t, \nu^t)\vartheta(dt).$$

Taking the supremum over $x \in B(0,R)$ gives

$$\|\Gamma^s_{\bar\mu} - \Gamma^s_{\bar\nu}\|_{L^\infty(B(0,R))} \le C_R \frac{1}{\vartheta([0,s])} \int_0^s W_p(\mu^t, \nu^t)\vartheta(dt), \quad C_R := \|V\|_{\mathrm{op}} e^{2c_R}\big(1 + 2\|B\|_{\mathrm{op}} R^2\big).$$

Consequently,

$$W_p\big((\Gamma^s_{\bar\mu})_\# \nu^s, (\Gamma^s_{\bar\nu})_\# \nu^s\big) \le C_R \frac{1}{\vartheta([0,s])} \int_0^s W_p(\mu^t, \nu^t)\vartheta(dt).$$

Step 4: closing via Hardy (Lebesgue case, $p > 1$). Assume ϑ is Lebesgue measure on $[0,1]$. Define $f(t) := W_p(\mu^t, \nu^t) \ge 0$ and recall $d_p(\bar\mu, \bar\nu) = \|f\|_{L^p([0,1])}$. From Step 3,

$$W_p\big((\Gamma^s_{\bar\mu})_\# \nu^s, (\Gamma^s_{\bar\nu})_\# \nu^s\big) \le C_R (Hf)(s), \quad (Hf)(s) := \frac{1}{s}\int_0^s f(t)dt.$$

By Hardy's inequality, for $p > 1$,

$$\|Hf\|_{L^p([0,1])} \le \frac{p}{p-1}\|f\|_{L^p([0,1])} = \frac{p}{p-1} d_p(\bar\mu, \bar\nu).$$

Therefore

$$\Big(\int_0^1 W_p\big((\Gamma^s_{\bar\mu})_\# \nu^s, (\Gamma^s_{\bar\nu})_\# \nu^s\big)^p ds\Big)^{1/p} \le C_R \frac{p}{p-1} d_p(\bar\mu, \bar\nu).$$

Discrete case. If ϑ is the empirical measure on $\{1/n, 2/n, \dots, n/n\}$, i.e. $\vartheta = \frac{1}{n}\sum_{i=1}^n \delta_{i/n}$, then with $f_i := W_p(\mu^{i/n}, \nu^{i/n})$ and the discrete Hardy operator

$$(H_n f)_i := \frac{1}{i}\sum_{k=1}^{i} f_k,$$

the estimate from Step 3 reads

$$W_p\big((\Gamma^{i/n}_{\bar\mu})_\# \nu^{i/n}, (\Gamma^{i/n}_{\bar\nu})_\# \nu^{i/n}\big) \le C_R (H_n f)_i.$$

For $p > 1$, the discrete Hardy inequality gives

$$\Big(\frac{1}{n}\sum_{i=1}^n |(H_n f)_i|^p\Big)^{1/p} \le \frac{p}{p-1}\Big(\frac{1}{n}\sum_{i=1}^n |f_i|^p\Big)^{1/p},$$

hence the same closing step as in the Lebesgue case (with the same factor $p/(p-1)$).

Step 5: collecting the bounds. Combining Steps 1–4,

$$d_p\big(F^M_\theta(\bar\mu), F^M_\theta(\bar\nu)\big) \le \Big(\|V\|_{\mathrm{op}}\|B\|_{\mathrm{op}} R^2 + C_R \frac{p}{p-1}\Big) d_p(\bar\mu, \bar\nu),$$

with $C_R = \|V\|_{\mathrm{op}} e^{2\|B\|_{\mathrm{op}} R^2}\big(1 + 2\|B\|_{\mathrm{op}} R^2\big)$. This proves Lipschitz continuity (in the Lebesgue-marginal case, $p > 1$). □

A.3 Burger et al. [1]

A.3.1 Theorem 3.28 [1, Theorem 2.2]

Proof. The direct method applies to the dynamic action

$$\mathcal{A}(\mu, v) := \int_0^1 \int_M m_{\mu_t}(x)\, |v_t(x)|^2\, \mu_t(dx)\, dt$$

under the continuity-equation constraint. Because M is compact and K is continuous with positive lower bound, the mobilities satisfy uniform bounds $0 < c_K \leq m_{\mu_t} \leq C_K$. Hence $\mathcal{A}$ is coercive with respect to the usual $L^2(\mu_t)$-control on velocities and is lower semicontinuous under narrow convergence plus weak velocity convergence. Compactness of probability curves on compact M and stability of the continuity equation give a minimizing limit pair $(\mu_t, v_t) \in CE(0, 1; \mu_0 \to \mu_1)$, so the infimum is attained.

For the length representation, set $a(t) := \left(\int_M m_{\mu_t} |v_t|^2\, d\mu_t\right)^{1/2}$. By Cauchy–Schwarz,

$$\int_0^T a(t)\, dt \leq \sqrt{T} \left(\int_0^T a(t)^2\, dt\right)^{1/2}.$$

Conversely, reparameterizing time by normalized cumulative action makes the speed constant and turns the quadratic action minimization into the path-length minimization. This yields the equivalent Benamou–Brenier-type length formula. □

A.3.2 Theorem 3.29 [1, Theorem 2.3]

Proof. By the dynamic definition of $W_{m,2}$ and the existence theorem for minimizers (the preceding metric-structure result), for any $\mu_0, \mu_1 \in \mathcal{P}(M)$ with finite distance there exists $(\mu_t, v_t) \in CE(0, 1; \mu_0 \to \mu_1)$ such that

$$W_{m,2}^2(\mu_0, \mu_1) = \int_0^1 \int_M m_{\mu_t}(x)\, |v_t(x)|^2\, \mu_t(dx)\, dt.$$

Using the equivalent length characterization of the same minimizers and constant-speed reparameterization, one obtains an optimal curve with time-independent metric speed

$$|\dot{\mu}|(t) = W_{m,2}(\mu_0, \mu_1) \quad \text{for a.e. } t \in (0, 1).$$

Hence, for $0 \leq s \leq t \leq 1$,

$$W_{m,2}(\mu_s, \mu_t) \leq \int_s^t |\dot{\mu}|(r)\, dr = (t - s) W_{m,2}(\mu_0, \mu_1).$$

Optimality of subsegments of a constant-speed minimizer gives equality, so

$$W_{m,2}(\mu_s, \mu_t) = |t - s|\, W_{m,2}(\mu_0, \mu_1).$$

Therefore $(\mathcal{P}(M), W_{m,2})$ is a geodesic space. □

A.3.3 Corollary 3.31 [1, Corollary 2.5]

Proof. Since M is compact and $K \in C(M \times M)$ with $0 < c_K \leq K \leq C_K$, for every $\mu \in \mathcal{P}(M)$ and $x \in M$,

$$c_K \leq m_\mu(x) = \int_M K(x, y)\, \mu(dy) \leq C_K.$$

Take any admissible pair (μ_t, v_t) in the continuity equation connecting μ_0 and μ_1. Then

$$c_K \int_0^1 \int_M |v_t|^2\, d\mu_t\, dt \leq \int_0^1 \int_M m_{\mu_t} |v_t|^2\, d\mu_t\, dt \leq C_K \int_0^1 \int_M |v_t|^2\, d\mu_t\, dt.$$

Infimizing over admissible pairs gives two-sided metric comparability

$$\sqrt{c_K}\, W_2(\mu_0, \mu_1) \leq W_{m,2}(\mu_0, \mu_1) \leq \sqrt{C_K}\, W_2(\mu_0, \mu_1).$$

Therefore $W_{m,2}$ and W_2 induce the same convergent sequences and the same topology on $\mathcal{P}(M)$. On compact M, W_2 convergence is equivalent to narrow convergence, so the same holds for $W_{m,2}$. Completeness follows from metric equivalence and completeness of $(\mathcal{P}(M), W_2)$. □

A.3.4 Theorem 3.30 [1, Theorem 2.7]

Proof. Let (μ_t, v_t) be absolutely continuous in $(\mathcal{P}(M), W_{m,2})$ and solve the continuity equation. For

$$\mathcal{E}(\mu) = \frac{1}{2}\int_{M\times M} W(x,y)\,\mu(dx)\mu(dy), \qquad W[\mu](x) = \int_M W(x,y)\,\mu(dy),$$

the chain rule along (μ_t, v_t) gives (for a.e. t)

$$\frac{d}{dt}\mathcal{E}(\mu_t) = \int_M \langle \operatorname{grad}_M W[\mu_t](x),\, v_t(x)\rangle \, d\mu_t(x).$$

Applying Cauchy–Schwarz and Young with weights m_{μ_t} yields

$$\frac{d}{dt}\mathcal{E}(\mu_t) \geq -\frac{1}{2}\int_M m_{\mu_t}|v_t|^2\,d\mu_t - \frac{1}{2}\int_M \frac{1}{m_{\mu_t}}\big|\operatorname{grad}_M W[\mu_t]\big|^2\,d\mu_t.$$

Integrating in time gives the energy-dissipation inequality. Equality holds iff equality holds in Young's inequality, i.e.

$$v_t = \frac{1}{m_{\mu_t}}\operatorname{grad}_M W[\mu_t] \quad \text{for a.e. } t,\ \mu_t\text{-a.e. } x,$$

which is exactly the weak PDE

$$\partial_t\mu_t + \operatorname{div}_M\left(\frac{1}{m_{\mu_t}}\operatorname{grad}_M W[\mu_t]\;\mu_t\right) = 0.$$

Define

$$g(\mu) := \left(\int_M \frac{1}{m_\mu}\big|\operatorname{grad}_M W[\mu]\big|^2\,d\mu\right)^{1/2}.$$

The same chain-rule estimate together with the metric-velocity characterization in $(\mathcal{P}(M), W_{m,2})$ shows that g is a strong upper gradient of $\mathcal{E}$. Hence equality in the above dissipation inequality is equivalent to the curve-of-maximal-slope identity. Therefore weak solutions of the PDE coincide with curves of maximal slope of $\mathcal{E}$ with respect to $W_{m,2}$. ☐

A.3.5 Lemma 3.32 [1, Lemma 2.9]

Proof. For an absolutely continuous curve (μ_t, v_t) in $(\mathcal{P}(M), W_{m,2})$, the chain rule gives

$$\frac{d}{dt}\mathcal{E}(\mu_t) = \int_M \langle \operatorname{grad}_M W[\mu_t](x), v_t(x)\rangle\;\mu_t(dx).$$

Using weighted Cauchy–Schwarz and Young inequalities,

$$\begin{aligned}\int_M \langle \operatorname{grad}_M W[\mu_t], v_t\rangle\,d\mu_t &\geq -\left(\int_M m_{\mu_t}|v_t|^2\,d\mu_t\right)^{1/2}\left(\int_M \frac{1}{m_{\mu_t}}|\operatorname{grad}_M W[\mu_t]|^2\,d\mu_t\right)^{1/2}\\ &\geq -\frac{1}{2}\int_M m_{\mu_t}|v_t|^2\,d\mu_t - \frac{1}{2}\int_M \frac{1}{m_{\mu_t}}|\operatorname{grad}_M W[\mu_t]|^2\,d\mu_t.\end{aligned}$$

Integrating in time from 0 to T yields the stated dissipation inequality. Equality holds iff equality holds in Young's inequality, i.e.

$$v_t = \frac{1}{m_{\mu_t}}\operatorname{grad}_M W[\mu_t] \quad \text{for a.e. } t,\ \mu_t\text{-a.e.},$$

which is equivalent to the weak PDE (8). The converse is immediate by substitution. ☐

A.3.6 Corollary 3.33 [1, Corollary 2.11]

Proof. For weak solutions of the gradient-flow PDE, the energy-dissipation inequality implies

$$\mathcal{E}(\mu_t) + \frac{1}{2}\int_0^t\int_M m_{\mu_s}(x)\,\big|\operatorname{grad}_M W[\mu_s](x)\big|^2\,\mu_s(dx)\,ds \leq \mathcal{E}(\mu_0).$$

Hence the dissipation functional

$$D(s) := \int_M m_{\mu_s}(x) \left| \operatorname{grad}_M W[\mu_s](x) \right|^2 \mu_s(dx)$$

belongs to $L^1(0,\infty)$ and is nonnegative, so there exists a sequence $t_n \to \infty$ with $D(t_n) \to 0$. Since M is compact, $\mathcal{P}(M)$ is narrowly compact; after extraction, $\mu_{t_n} \rightharpoonup \mu_\star$.

Under $W \in C^{1,\alpha}(M \times M)$, the maps

$$\mu \mapsto m_\mu, \qquad \mu \mapsto \operatorname{grad}_M W[\mu]$$

are compact in $C(M)$ along narrowly convergent sequences (Arzela–Ascoli argument used in the source). Passing to the limit along (t_n) gives

$$0 = \lim_{n\to\infty} D(t_n) = \int_M m_{\mu_\star}(x) \left| \operatorname{grad}_M W[\mu_\star](x) \right|^2 \mu_\star(dx).$$

Because $m_{\mu_\star} \geq c_K > 0$, we obtain $\operatorname{grad}_M W[\mu_\star] = 0$ $\mu_\star$-a.e., which is equivalent to the stationary weak form

$$\operatorname{div}_M \left(\frac{1}{m_{\mu_\star}} \operatorname{grad}_M W[\mu_\star]\, \mu_\star \right) = 0.$$

Thus a subsequence converges to a stationary measure. □

A.3.7 Theorem 3.34 [1, Theorem 3.1]

Proof. For $x, y \in \mathbb{S}^{d-1}$ and symmetric D, the Rayleigh bound gives

$$\langle x, Dy \rangle \leq \lambda_{\max},$$

hence

$$e^{\langle x, Dy \rangle} \leq e^{\lambda_{\max}}.$$

Integrating against $\mu \otimes \mu$ yields

$$\mathcal{E}_D(\mu) \leq e^{\lambda_{\max}}.$$

If $\mu = \delta_z$ with $Dz = \lambda_{\max} z$, then

$$\mathcal{E}_D(\delta_z) = e^{\langle z, Dz \rangle} = e^{\lambda_{\max}},$$

so such Diracs are maximizers. Conversely, equality in the integral bound requires $\langle x, Dy \rangle = \lambda_{\max}$ for $\mu \otimes \mu$-a.e. (x, y), which forces concentration on one eigenvector direction associated with $\lambda_{\max}$. Therefore maximizers are exactly those Dirac measures. □

A.3.8 Theorem 3.35 [1, Theorem 3.4]

Proof. Let $\bar{x} := \int_{\mathbb{S}^{d-1}} x\, \mu(dx)$. Two applications of Jensen's inequality (as in the source) give

$$\mathcal{E}_D(\mu) \geq e^{\langle \bar{x}, D\bar{x} \rangle} \geq e^{\lambda_{\min} \|\bar{x}\|^2} \geq e^{\lambda_{\min}}.$$

If $\lambda_{\min} < 0$, equality can hold only if $\|\bar{x}\| = 1$ and $\bar{x}$ is an eigenvector for $\lambda_{\min}$. The condition $\|\bar{x}\| = 1$ on the unit sphere forces μ to be a Dirac measure. Hence minimizers are exactly Diracs at $\lambda_{\min}$-eigenvectors.

If $\lambda_{\min} = 0$, the lower bound becomes $\mathcal{E}_D(\mu) \geq 1$. Every measure supported on $\ker D$ attains equality since $\langle x, Dy \rangle = 0$ on the support. Conversely, equality in Jensen's step implies $\langle x, Dx \rangle = 0$ for μ-a.e. x, i.e. $\operatorname{supp}(\mu) \subset \ker D$. □

A.3.9 Corollary 3.36 [1, Corollary 3.10]

Proof. Fix a unit eigenvector z of D and define the reflection $H_z(x) = x - 2\langle x, z \rangle z$. For any μ, set

$$\tilde{\mu} := \frac{1}{2}\big(\mu + (H_z)_\# \mu\big).$$

By invariance of the kernel under simultaneous reflection along eigendirections,

$$\mathcal{E}_D\big((H_z)_{\#}\mu\big) = \mathcal{E}_D(\mu).$$

Since $D \succ 0$, the source proves strict convexity of $\mathcal{E}_D$, hence

$$\mathcal{E}_D(\tilde{\mu}) \leq \frac{1}{2}\mathcal{E}_D(\mu) + \frac{1}{2}\mathcal{E}_D\big((H_z)_{\#}\mu\big) = \mathcal{E}_D(\mu),$$

with strict inequality unless $\tilde{\mu} = \mu$. If $\mu_\star$ is a minimizer, strict inequality is impossible, so $\mu_\star = \tilde{\mu}_\star$, i.e.

$$(H_z)_{\#}\mu_\star = \mu_\star.$$

Applying this to every eigendirection gives the claimed symmetry. □

A.3.10 Theorem 3.37 [1, Theorem 3.17]

Proof. First, antipodal symmetrization increases energy for $D \preceq 0$ (source Lemma 3.16): for

$$\tilde{\mu} := \frac{1}{2}\big(\mu + (-\mathrm{Id})_{\#}\mu\big),$$

one has $\mathcal{E}_D(\tilde{\mu}) \geq \mathcal{E}_D(\mu)$, with strict inequality unless μ is already even (under the theorem assumptions). Hence it suffices to maximize over even measures.

For even μ, writing the energy on a hemisphere pair gives an integrand of the form $\cosh(\langle x, Dy\rangle)$. Since $D \preceq 0$ and $\lambda_{\min} < 0$,

$$\langle x, Dy\rangle \in [\lambda_{\min}, 0] \quad \Rightarrow \quad \cosh(\langle x, Dy\rangle) \leq \cosh(\lambda_{\min}).$$

Therefore

$$\mathcal{E}_D(\mu) \leq \cosh(\lambda_{\min}).$$

If $\mu = \frac{1}{2}(\delta_z + \delta_{-z})$ with $Dz = \lambda_{\min} z$, then

$$\mathcal{E}_D(\mu) = \cosh(\lambda_{\min}),$$

so these measures are maximizers.

Conversely, equality in the bound forces $|\langle x, Dy\rangle| = |\lambda_{\min}|$ $\mu \otimes \mu$-a.e., which can only occur when μ is concentrated on one antipodal eigenpair $\{\pm z\}$ with $Dz = \lambda_{\min} z$. Hence maximizers are exactly $\frac{1}{2}(\delta_z + \delta_{-z})$ on minimal-eigenvalue eigendirections. □

A.3.11 Proposition 4.2 [1, Proposition 4.2]

Proof. Let μ_* be a minimizer of E_D (the maximizer case is identical with reversed inequality). For any Lipschitz tangent field W, let $(\mu_t)_{|t|\ll 1}$ solve the continuity equation generated by W with $\mu_{t=0} = \mu_*$.

$$\partial_t \mu_t + \nabla \cdot (\mu_t W) = 0, \qquad \mu_{t=0} = \mu_*.$$

This defines an admissible first-order variation of μ_* in Wasserstein geometry.

Since μ_* is globally minimal,

$$E_D(\mu_t) - E_D(\mu_*) \geq 0 \quad \text{for } |t| \ll 1.$$

For $t > 0$, divide by t and pass to $t \downarrow 0$:

$$dE_D(\mu_*; W) \geq 0.$$

For $t < 0$, divide by t and pass to $t \uparrow 0$:

$$dE_D(\mu_*; W) \leq 0.$$

Hence

$$dE_D(\mu_*; W) = 0.$$

Since W was arbitrary, the first variation vanishes in every tangent direction, so μ_* satisfies the Euler–Lagrange stationarity condition for the mean-field dynamics. □

A.3.12 Lemma 4.3 [1, Lemma 4.3]

Proof. Let

$$b_\mu(x) := \int_{S^{d-1}} e^{x\cdot Dy} P_x^\perp Dy\, \mu(dy).$$

For the energy E_D, the first variation in direction W has the form

$$dE_D(\mu; W) = \int_{S^{d-1}} b_\mu(x) \cdot W(x)\, \mu(dx).$$

The mean-field velocity of the dynamics is precisely $\Gamma_\mu = b_\mu$, so the PDE is

$$\partial_t \mu + \nabla \cdot (\mu b_\mu) = 0.$$

If μ is stationary for this PDE, the transport field on the support must be tangent and divergence-free. Since b_μ is itself the first-variation gradient, stationarity is equivalent to the Euler-Lagrange condition $b_\mu = 0$ μ-a.e., and then $dE_D(\mu; W) = 0$ for every Lipschitz W. Conversely, if $dE_D(\mu; W) = 0$ for all Lipschitz W, then by testing with $W = b_\mu$ we get

$$0 = dE_D(\mu; b_\mu) = \int \|b_\mu(x)\|^2\, \mu(dx),$$

so $b_\mu = 0$ μ-a.e. Hence the drift vanishes and μ is stationary. □

A.3.13 Lemma 4.5 [1, Lemma 4.5]

Proof. For the interaction energy, the first variation has the form

$$dE_D(\mu; W) = \int_{S^{d-1}} b_\mu(x) \cdot W(x)\, \mu(dx),$$

with

$$b_\mu(x) := \int_{S^{d-1}} e^{x\cdot Dy} P_x^\perp Dy\, \mu(dy).$$

Insert $\mu = \delta_z$:

$$b_{\delta_z}(z) = e^{z\cdot Dz} P_z^\perp Dz.$$

Therefore

$$dE_D(\delta_z; W) = -e^{z\cdot Dz}\, P_z^\perp Dz \cdot W(z).$$

Since W is an arbitrary tangent test field, $W(z)$ can be any vector in $T_z S^{d-1}$. Hence

$$dE_D(\delta_z; W) = 0\ \forall W \quad \Longleftrightarrow \quad P_z^\perp Dz = 0.$$

Now

$$P_z^\perp Dz = 0 \Longleftrightarrow Dz = (z^\top Dz)\, z,$$

which is exactly the eigenvector condition for z. Conversely, if z is an eigenvector, then $P_z^\perp Dz = 0$ and the first variation vanishes for all W, so δ_z is stationary. □

A.3.14 Lemma 4.6 [1, Lemma 4.6]

Proof. Let

$$\mu_t = t\delta_z + (1-t)\delta_{-z}, \qquad t \in [0, 1].$$

Use the first-variation representation

$$dE_D(\mu_t; W) = \int b_{\mu_t}(x) \cdot W(x)\, \mu_t(dx),$$

$$b_{\mu_t}(x) = \int e^{x\cdot Dy} P_x^\perp Dy\, \mu_t(dy).$$

Evaluate b_{μ_t} at support points $x = \pm z$. Since $P_{-z}^{\perp} = P_z^{\perp}$ and $D(-z) = -Dz$, we get

$$b_{\mu_t}(z) = \big(te^{z\cdot Dz} - (1-t)e^{-z\cdot Dz}\big)P_z^{\perp}Dz,$$

$$b_{\mu_t}(-z) = \big((1-t)e^{z\cdot Dz} - te^{-z\cdot Dz}\big)P_z^{\perp}Dz.$$

So both drift values are scalar multiples of $P_z^{\perp}Dz$. Therefore

$$dE_D(\mu_t; W) = 0\ \forall W \iff P_z^{\perp}Dz = 0.$$

As in Lemma 4.5, this is equivalent to Dz being colinear with z, i.e. z being an eigenvector of D. Hence μ_t is stationary iff z is an eigenvector. ☐

A.3.15 Lemma 4.7 [1, Lemma 4.7]

Proof. Write

$$\mu = \frac{1}{2}\sum_{z\in Z_D} t(z)(\delta_z + \delta_{-z}), \qquad \sum_{z\in Z_D} t(z) = 1,$$

with every $z \in Z_D$ an eigenvector of D and $z\cdot w = 0$ for $z \neq w$. To prove stationarity it is enough to verify $b_\mu(x) = 0$ on the support, where

$$b_\mu(x) = \int e^{x\cdot Dy}P_x^{\perp}Dy\,\mu(dy).$$

Fix $w \in Z_D$ and evaluate at $x = w$. For the self pair $y = \pm w$, since Dw is colinear with w, we have $P_w^{\perp}Dw = 0$, so these contributions vanish. For $y = \pm z$ with $z \neq w$, orthogonality gives $w \cdot Dz = 0$ (because $Dz = \lambda_z z$), hence $e^{w\cdot Dz} = e^{-w\cdot Dz} = 1$. Also $D(-z) = -Dz$, so the two terms cancel:

$$P_w^{\perp}Dz + P_w^{\perp}D(-z) = 0.$$

Therefore every block contributes zero and $b_\mu(w) = 0$. The same argument at $x = -w$ gives $b_\mu(-w) = 0$. Thus $b_\mu = 0$ on $\operatorname{supp}\mu$, so μ is stationary. ☐

A.3.16 Lemma 4.10 [1, Lemma 4.10]

Proof. Let σ be the uniform probability measure on S^{d-1} and define

$$b_\sigma(x) := \int_{S^{d-1}} e^{x\cdot Dy}P_x^{\perp}Dy\,\sigma(dy).$$

Stationarity is equivalent to $b_\sigma(x) = 0$ for every x. Diagonalize $D = U\Lambda U^\top$ and use rotational invariance of σ: it is enough to analyze Λ. If all $|\lambda_i|$ are equal, the kernel depends only on the absolute deformation level and is invariant under sign-flip symmetries of coordinates. For each tangent component, the integrand is odd under a symmetry that preserves measure; therefore all tangent moments cancel and $b_\sigma \equiv 0$. If absolute eigenvalues are not all equal, anisotropy remains in at least one tangent mode. Choosing a tangent test field aligned with this mode yields a nonzero directional first variation

$$dE_D(\sigma; W) = \int b_\sigma \cdot W\,d\sigma \neq 0,$$

so $b_\sigma \not\equiv 0$ and σ is not stationary. Hence σ is stationary iff all $|\lambda_i|$ coincide. ☐

A.3.17 Corollary 4.11 [1, Corollary 4.11]

Proof. If $D = \lambda I_d$ with $\lambda \geq 0$, the interaction depends only on $\langle x, y\rangle$ isotropically and the uniform measure is minimizing by symmetry and convexity of the energy along geodesic rearrangements. Conversely, assume σ minimizes E_D. Then σ is stationary, so by the previous lemma all eigenvalues satisfy $|\lambda_1| = \cdots = |\lambda_d|$. If one eigenvalue is negative, then the smallest eigenvalue is $\lambda_{\min} < 0$. For an eigenvector v of $\lambda_{\min}$,

$$E_D(\delta_v) = \frac{1}{2}e^{v\cdot Dv} = \frac{1}{2}e^{\lambda_{\min}}.$$

Because $\lambda_{\min} < 0$, this value is strictly below the isotropic benchmark of nonnegative modes, so σ cannot be a minimizer. Hence no eigenvalue is negative. Combined with equal absolute values, this implies all eigenvalues are equal and nonnegative: $D = \lambda I_d$, $\lambda \geq 0$. ☐

A.4 Castin et al. [2]

A.4.1 Theorem 3.45 [2, Theorem 3.1]

Proof. Fix one attention field $\Gamma \in \{\Gamma^{(SM)}, \Gamma^{(L2)}, \Gamma^{(sink)}, \Gamma^{(\sigma)}, \Gamma^{(MH)}\}$. On every bounded ball $B_R \subset \mathbb{R}^d$, these fields satisfy estimates of the form

$$|\Gamma_\mu(t,x) - \Gamma_\mu(t,x')| \le L_R(t)|x - x'|,$$
$$|\Gamma_\mu(t,x) - \Gamma_\nu(t,x)| \le L_R(t)\, W_p(\mu,\nu),$$

uniformly for $\mathrm{supp}(\mu), \mathrm{supp}(\nu) \subset B_R$.

Step 1: characteristics with frozen measure curve. Given a continuous curve $\eta_\cdot \in C([0,T], \mathcal{P}_c(\mathbb{R}^d))$, consider

$$\dot{X}_\eta(t;x_0) = \Gamma_{\eta_t}(t, X_\eta(t;x_0)), \qquad X_\eta(0;x_0) = x_0.$$

By Caratheodory/Picard–Lindelof, this ODE has a unique flow $x_0 \mapsto X_\eta(t;x_0)$ on $[0,T]$. Define

$$(\mathcal{T}\eta)_t := (X_\eta(t;\cdot))_{\#}\mu_0.$$

Step 2: fixed point in measure space. For η, ζ in a bounded-support ball of $C([0,T], \mathcal{P}_p)$, Dobrushin estimates on the characteristic difference give

$$\sup_{t\in[0,T]} W_p\big((\mathcal{T}\eta)_t, (\mathcal{T}\zeta)_t\big) \le C_R T \sup_{t\in[0,T]} W_p(\eta_t, \zeta_t).$$

For T small enough, $\mathcal{T}$ is a contraction, so Banach gives a unique fixed point $\mu_\cdot$ on $[0,T]$. Iterating yields a unique global $\mu \in C([0,\infty), \mathcal{P}_c(\mathbb{R}^d))$.

Step 3: support growth bound. Along characteristics,

$$\frac{d}{dt}|X(t)| \le \|V(t)\|_2\, |X(t)|$$

for all listed attention fields (the prefactor comes from linear value map control). Gronwall gives

$$|X(t)| \le e^{\int_0^t \|V(s)\|_2 ds}|X(0)|.$$

If $\mathrm{supp}(\mu_0) \subset B_{R_0}$, then

$$\mathrm{supp}(\mu_t) \subset B_{R(t)}, \qquad R(t) = e^{\int_0^t \|V(s)\|_2 ds} R_0.$$

Step 4: stability. Let μ, ν be two solutions with initial data in B_{R_0}. Couple their characteristics optimally at time 0, propagate the coupling through the two flows, and use the two Lipschitz bounds above:

$$\frac{d}{dt} W_p(\mu_t, \nu_t) \le C(t, R_0)\, W_p(\mu_t, \nu_t).$$

Gronwall yields

$$W_p(\mu_t, \nu_t) \le C(t, R_0)\, W_p(\mu_0, \nu_0).$$

This proves existence, uniqueness, support control, and stability. □

A.4.2 Theorem 4.16 [2, Theorem 3.4]

Proof. Work on $[0,1] \times \mathbb{R}^d$ with variables (a, x), where a is the mask/position coordinate. In the masked model, the transport field has the form

$$\bar{\Gamma}_{\bar{\mu}}(a,x) = (0, \Gamma^{(m)}_{\bar{\mu}}(a,x)),$$

so the a-coordinate is frozen and only x evolves.

Step 1: characteristic system. For a frozen measure curve $\bar{\eta}_t$, the characteristic ODE is

$$\dot{a}(t) = 0, \qquad \dot{X}(t) = \Gamma^{(m)}_{\bar{\eta}_t}(a_0, X(t)).$$

Hence $a(t) \equiv a_0$, and for each fiber a_0 we have a standard Euclidean ODE in x with local Lipschitz dependence on x and Lipschitz dependence on $\bar{\eta}_t$ in the conditional transport metric.

Step 2: fixed point. Define the push-forward map on curves $\bar{\eta} \mapsto \bar{\mathcal{T}}\bar{\eta}$ by transporting $\bar{\mu}_0$ along these characteristics. Using Dobrushin estimates fiberwise in a, one gets on short intervals

$$\sup_{t \le T} d\big((\bar{\mathcal{T}}\bar{\eta})_t, (\bar{\mathcal{T}}\bar{\zeta})_t\big) \le C_R T \sup_{t \le T} d(\bar{\eta}_t, \bar{\zeta}_t).$$

For small T, this is a contraction, yielding a unique fixed point $\bar{\mu} \in C([0,T], \mathcal{P}_c([0,1] \times \mathbb{R}^d))$. Iteration gives global existence and uniqueness.

Step 3: support bound for spatial marginal. Let μ_t be the x-marginal of $\bar{\mu}_t$. Along characteristics,

$$\frac{d}{dt}|X(t)| \le \|V(t)\|_2 |X(t)|.$$

Hence

$$|X(t)| \le e^{\int_0^t \|V(s)\|_2 ds} |X(0)|,$$

so if $\operatorname{supp}(\mu_0) \subset B_{R_0}$ then

$$\operatorname{supp}(\mu_t) \subset B_{R(t)}, \qquad R(t) = e^{\int_0^t \|V(s)\|_2 ds} R_0.$$

Step 4: stability in conditional distance. Take two solutions $\bar{\mu}, \bar{\nu}$ and couple them with the same a-marginal. Propagate this coupling through the two characteristic flows fiberwise. The same differential estimate as in the unmasked case gives

$$\frac{d}{dt} d(\bar{\mu}_t, \bar{\nu}_t) \le C(t, R_0)\, d(\bar{\mu}_t, \bar{\nu}_t),$$

thus

$$d(\bar{\mu}_t, \bar{\nu}_t) \le C(t, R_0)\, d(\bar{\mu}_0, \bar{\nu}_0).$$

So the masked equation is globally well posed with the stated support and stability bounds. $\square$

A.4.3 Lemma 3.46 [2, Lemma 4.1]

Proof. For Softmax mean field,

$$\Gamma_\mu^{(SM)}(x) = V \frac{N(x)}{Z(x)}, \qquad N(x) := \int_{\mathbb{R}^d} y\, e^{x^\top A y}\, \mu(dy), \quad Z(x) := \int_{\mathbb{R}^d} e^{x^\top A y}\, \mu(dy),$$

with $A := K^\top Q$.

Let $Y \sim \mathcal{N}(\alpha, \Sigma)$ and set $u := Ax$ (this matches the scalar product convention used in the statement). Then

$$Z(x) = \mathbb{E}[e^{u^\top Y}] = \exp\Big(u^\top \alpha + \frac{1}{2} u^\top \Sigma u\Big).$$

Differentiate this mgf identity with respect to u:

$$\mathbb{E}[Y e^{u^\top Y}] = (\alpha + \Sigma u)\, \mathbb{E}[e^{u^\top Y}] = (\alpha + \Sigma A x)\, Z(x).$$

Thus

$$\frac{N(x)}{Z(x)} = \alpha + \Sigma A x,$$

and therefore

$$\Gamma_\mu^{(SM)}(x) = V(\alpha + \Sigma A x).$$

So the velocity is affine in x on Gaussian states. $\square$

A.4.4 Proposition 3.47 [2, Proposition 4.2]

Proof. By Lemma 3.46, if $\mu_t = \mathcal{N}(\alpha_t, \Sigma_t)$ then

$$\Gamma^{(SM)}_{\mu_t}(x) = M_t x + b_t, \qquad M_t := V_t \Sigma_t A_t, \quad b_t := V_t \alpha_t, \quad A_t := K_t^\top Q_t.$$

So the PDE is a continuity equation with affine drift:

$$\partial_t \mu_t + \nabla \cdot \big(\mu_t (M_t x + b_t)\big) = 0.$$

Gaussian invariance. Affine transport maps Gaussians to Gaussians, hence μ_t stays Gaussian as long as the moment ODE is defined.

Moment equations. For any solution of the affine continuity equation:

$$\dot{\alpha}_t = M_t \alpha_t + b_t, \qquad \dot{\Sigma}_t = M_t \Sigma_t + \Sigma_t M_t^\top.$$

Substituting M_t, b_t:

$$\dot{\alpha}_t = V_t \Sigma_t A_t \alpha_t + V_t \alpha_t = V_t (\mathrm{Id} + \Sigma_t A_t) \alpha_t,$$
$$\dot{\Sigma}_t = V_t \Sigma_t A_t \Sigma_t + \Sigma_t A_t^\top \Sigma_t V_t^\top.$$

These are exactly the displayed ODEs.

Existence/uniqueness. The right-hand side is locally Lipschitz in (α, Σ), so the ODE has a unique maximal solution; this induces the unique maximal Gaussian solution of the PDE. □

A.4.5 Proposition 3.48 [2, Proposition 4.4]

Proof. Set

$$B := VA + A^\top V^\top.$$

The covariance equation is

$$\dot{\Sigma} = V \Sigma A \Sigma + \Sigma A^\top \Sigma V^\top.$$

Differentiate the inverse:

$$\frac{d}{dt} \Sigma^{-1} = -\Sigma^{-1} \dot{\Sigma} \Sigma^{-1} = -\Sigma^{-1} (V \Sigma A \Sigma + \Sigma A^\top \Sigma V^\top) \Sigma^{-1}.$$

Using the commutation assumptions (so factors can be reordered against Σ and Σ^{-1}), this simplifies to

$$\frac{d}{dt} \Sigma^{-1} = -B.$$

Integrating:

$$\Sigma^{-1}(t) = \Sigma_0^{-1} - tB, \qquad \Sigma(t) = \big(\Sigma_0^{-1} - tB\big)^{-1}$$

on the interval where the inverse exists.

Hence maximal existence ends exactly when $\Sigma_0^{-1} - tB$ loses positive definiteness.

If $B \preceq 0$, then $-tB \succeq 0$, so $\Sigma_0^{-1} - tB \succeq \Sigma_0^{-1} \succ 0$ for all $t \geq 0$. Therefore the solution is global. Moreover, on an eigenvector of B with eigenvalue $\lambda \leq 0$,

$$\sigma^{-1}(t) = \sigma_0^{-1} - t\lambda$$

is nondecreasing; if $\lambda < 0$ then $\sigma(t) \to 0$, producing rank loss along strictly negative directions. The remaining directions converge to a finite limit $\Sigma_* \succeq 0$.

If B has some eigenvalue $\lambda_+ > 0$, then on its eigendirection

$$\sigma^{-1}(t) = \sigma_0^{-1} - t\lambda_+$$

vanishes at finite time $t_* = \sigma_0^{-1}/\lambda_+$. Hence $\sigma(t) \to +\infty$ and $\|\Sigma(t)\|$ blows up at t_*. □

A.4.6 Proposition 3.49 [2, Proposition 4.8]

Proof. With $V = \mathrm{Id}$ and $A = A^\top$, the stationarity equation becomes

$$\Sigma A \Sigma = 0.$$

Let

$$S := \mathrm{Im}(\Sigma), \qquad r := \dim S = \mathrm{rank}(\Sigma).$$

For any $u, v \in \mathbb{R}^d$, setting $x = \Sigma u \in S$, $y = \Sigma v \in S$ gives

$$x^\top A y = u^\top (\Sigma A \Sigma) v = 0.$$

So the symmetric bilinear form $B_A(x, y) := x^\top A y$ vanishes on $S \times S$: S is a totally isotropic subspace for B_A.

Let A have inertia (n_+, n_-, n_0):

$$n_+ = \#\{\lambda_i(A) > 0\}, \quad n_- = \#\{\lambda_i(A) < 0\}, \quad n_0 = \dim\ker A.$$

For a symmetric form, the maximal dimension of a totally isotropic subspace is

$$n_0 + \min(n_+, n_-)$$

(Witt index plus kernel contribution). Since S is totally isotropic,

$$r \le n_0 + \min(n_+, n_-).$$

Rewriting this with eigenvalue counts yields

$$\mathrm{rank}(\Sigma) \le \dim\ker A + \min\big(\#\{\lambda_i(A) > 0\}, \#\{\lambda_i(A) < 0\}\big).$$

Hence stationary covariances are necessarily low rank. ☐

A.4.7 Proposition 3.50 [2, Proposition 4.9]

Proof. For head h, set

$$A^{(h)} := (K^{(h)})^\top Q^{(h)}, \qquad B^{(h)} := W^{(h)} V^{(h)}.$$

By the single-head Gaussian formula, head h induces

$$\Gamma_h(x) = B^{(h)}\big(\alpha + \Sigma A^{(h)} x\big) = M_h x + b_h,$$

with

$$M_h = B^{(h)} \Sigma A^{(h)}, \qquad b_h = B^{(h)} \alpha.$$

The multi-head velocity is the sum:

$$\Gamma_{\mathrm{MH}}(x) = \sum_{h=1}^{H} \Gamma_h(x) = Mx + b,$$

where

$$M = \sum_{h=1}^{H} M_h, \qquad b = \sum_{h=1}^{H} b_h.$$

So the drift is affine; therefore Gaussianity is preserved.

For affine drift $\dot{x} = Mx + b$, moments satisfy

$$\dot{\alpha} = M\alpha + b, \qquad \dot{\Sigma} = M\Sigma + \Sigma M^\top.$$

Substitute the expressions for M, b and expand sums:

$$\dot{\alpha} = \sum_{h=1}^{H} B^{(h)}\big(\mathrm{Id} + \Sigma A^{(h)}\big)\alpha,$$

$$\dot{\Sigma} = \sum_{h=1}^{H} \left(B^{(h)} \Sigma A^{(h)} \Sigma + \Sigma A^{(h)\top} \Sigma B^{(h)\top}\right),$$

which are exactly the stated ODEs. ☐

A.4.8 Lemma 3.51 [2, Lemma 4.10]

Proof. For L2 attention, the unnormalized weight in y is

$$\exp\big(-|Qx-Ky|^2\big)\,\mu(dy), \qquad \mu = \mathcal{N}(\alpha,\Sigma).$$

Up to x-dependent constants, the exponent in y is

$$-\frac{1}{2}(y-\alpha)^\top\Sigma^{-1}(y-\alpha) - |Qx-Ky|^2$$

$$= -\frac{1}{2}y^\top(\Sigma^{-1}+2K^\top K)y + y^\top(\Sigma^{-1}\alpha+2K^\top Qx)+c(x).$$

So the tilted law in y is Gaussian with precision

$$\Lambda := \Sigma^{-1}+2K^\top K$$

and mean

$$m(x) = \Lambda^{-1}\big(\Sigma^{-1}\alpha+2K^\top Qx\big).$$

By definition of the barycentric map,

$$\Gamma_\mu^{(L2)}(x) = V\,\mathbb{E}[Y\mid x] = V\,m(x),$$

hence

$$\Gamma_\mu^{(L2)}(x) = V(\Sigma^{-1}+2K^\top K)^{-1}(\Sigma^{-1}\alpha+2K^\top Qx).$$

Therefore the L2 velocity is affine in x. $\square$

A.4.9 Proposition 3.52 [2, Proposition 4.11]

Proof. From Lemma 3.51,

$$\Gamma_\mu^{(L2)}(x) = Mx+b$$

with

$$M = 2V(\Sigma^{-1}+2K^\top K)^{-1}K^\top Q = 2V(\Sigma^{-1}+2K^\top K)^{-1}A,$$
$$b = V(\Sigma^{-1}+2K^\top K)^{-1}\Sigma^{-1}\alpha.$$

Hence the continuity equation has affine drift; Gaussianity is preserved.

For affine drift, moments satisfy

$$\dot\alpha = M\alpha+b, \qquad \dot\Sigma = M\Sigma+\Sigma M^\top.$$

Use

$$(\Sigma^{-1}+2K^\top K)^{-1}\Sigma = \Sigma(\mathrm{Id}+2K^\top K\Sigma)^{-1}$$

to rewrite

$$M\Sigma = 2V\Sigma(\mathrm{Id}+2K^\top K\Sigma)^{-1}A\Sigma.$$

Similarly,

$$\Sigma M^\top = 2\Sigma A^\top(\mathrm{Id}+2K^\top K\Sigma)^{-1}\Sigma V^\top.$$

Therefore

$$\dot\Sigma = 2V\Sigma(\mathrm{Id}+2K^\top K\Sigma)^{-1}A\Sigma + 2\Sigma A^\top(\mathrm{Id}+2K^\top K\Sigma)^{-1}\Sigma V^\top.$$

For the mean:

$$\dot\alpha = \Big(2V(\Sigma^{-1}+2K^\top K)^{-1}A + V(\Sigma^{-1}+2K^\top K)^{-1}\Sigma^{-1}\Big)\alpha$$
$$= V(\Sigma^{-1}+2K^\top K)^{-1}(\Sigma^{-1}+2A)\alpha.$$

This is exactly the claimed ODE system. $\square$

A.4.10 Lemma 3.53 [2, Lemma 4.12]

Proof. Let

$$F(\Sigma) := 2V\Sigma(\operatorname{Id} + 2K^\top K\Sigma)^{-1}A\Sigma + 2\Sigma A^\top(\operatorname{Id} + 2K^\top K\Sigma)^{-1}\Sigma V^\top.$$

On the open cone $\mathcal{S}^d_{++}$, F is smooth; hence local existence and uniqueness follow from ODE theory.

Positivity preservation. The covariance equation comes from moment evolution of a transport PDE with affine drift, so $\Sigma(t)$ remains symmetric positive definite as long as the solution exists.

A priori growth bound. Set $s(t) := \operatorname{Tr}(\Sigma(t))$. Using $\|(\operatorname{Id} + 2K^\top K\Sigma)^{-1}\|_{\mathrm{op}} \le 1$ and $\|A\|, \|V\| < \infty$, one gets

$$|\operatorname{Tr}(F(\Sigma))| \le C_1 s + C_2 s^2 \|(\operatorname{Id} + 2K^\top K\Sigma)^{-1}K^\top\|_{\mathrm{op}}.$$

The factor with the resolvent suppresses high-energy modes in the $K^\top K$ directions and cancels the Riccati blow-up mechanism; after splitting into $\ker K \oplus (\ker K)^\perp$, this yields

$$\dot{s}(t) \le C(1 + s(t))$$

for a constant C depending only on coefficient norms. Therefore

$$s(t) \le (s(0) + 1)e^{Ct} - 1,$$

so no finite-time blow-up of $\|\Sigma(t)\|$ is possible.

Continuation. Since the solution stays in a bounded subset of $\mathcal{S}^d_{++}$ on every finite time interval, the local solution can be continued indefinitely. Hence the maximal existence time is $+\infty$. ☐

A.4.11 Lemma 3.54 [2, Lemma 4.14]

Proof. For $\mu = \mathcal{N}(\alpha, \Sigma)$, entropic OT between Gaussian marginals is Gaussian: the Schrödinger bridge coupling is Gaussian with block covariance determined by a cross-covariance matrix C solving the matrix quadratic system. The explicit solution is

$$C = \Sigma^{1/2}\Big(\Sigma^{1/2}A^\top\Sigma A\Sigma^{1/2} + \frac{\varepsilon^2}{4}\operatorname{Id}\Big)^{1/2}\Sigma^{-1/2} - \frac{\varepsilon}{2}\operatorname{Id}.$$

For a Gaussian coupling, the Sinkhorn barycentric map is the conditional expectation $x \mapsto \mathbb{E}[Y \mid X = x]$ followed by V/ε scaling. Conditioning formulas for jointly Gaussian vectors give an affine map:

$$\mathbb{E}[Y \mid X = x] = \big(\operatorname{Id} - A^{-\top}\Sigma^{-1}C\big)\alpha + A^{-\top}\Sigma^{-1}Cx.$$

Hence

$$\Gamma^{(sink)}_{\mu,\varepsilon}(x) = \frac{1}{\varepsilon}V\Big((\operatorname{Id} - A^{-\top}\Sigma^{-1}C)\alpha + A^{-\top}\Sigma^{-1}Cx\Big),$$

which is exactly

$$\Gamma^{(sink)}_{\mu,\varepsilon}(x) = \frac{1}{\varepsilon}V(\operatorname{Id} - A^{-\top}\Sigma^{-1}C)\alpha + \frac{1}{\varepsilon}VA^{-\top}\Sigma^{-1}Cx.$$

So Sinkhorn velocity is affine on Gaussian states. ☐

A.4.12 Proposition 3.55 [2, Proposition 4.15]

Proof. By Lemma 3.54,

$$\Gamma^{(sink)}_{\mu,\varepsilon}(x) = Mx + b$$

with

$$M = \varepsilon^{-1}VA^{-\top}\Sigma^{-1}C, \qquad b = \varepsilon^{-1}V(\operatorname{Id} - A^{-\top}\Sigma^{-1}C)\alpha.$$

Therefore Gaussianity is invariant.

For affine drift, moments satisfy

$$\dot{\alpha} = M\alpha + b, \qquad \dot{\Sigma} = M\Sigma + \Sigma M^\top.$$

Compute $\dot{\alpha}$:

$$\dot{\alpha} = \varepsilon^{-1}VA^{-\top}\Sigma^{-1}C\alpha + \varepsilon^{-1}V(\operatorname{Id} - A^{-\top}\Sigma^{-1}C)\alpha = \varepsilon^{-1}V\alpha.$$

So the mean equation decouples.

For covariance:

$$\dot{\Sigma} = \varepsilon^{-1} V A^{-\top} \Sigma^{-1} C \Sigma + \varepsilon^{-1} \Sigma C^{\top} \Sigma^{-1} A^{-1} V^{\top},$$

which is the stated formula. The expression of C is the same as in Lemma 3.54 (up to equivalent symmetric rewriting). Hence mean and covariance evolve independently under Sinkhorn Gaussian closure. $\square$

A.4.13 Theorem 3.56 [2, Section 5.3]

Proof. Assume $A := K^{\top} Q$ satisfies $A = A^{\top} = -V$. Define

$$F_{\varepsilon}(\mu) = -\frac{1}{2} \kappa^{\infty}_{\mu,\varepsilon} \log \frac{\kappa^{\infty}_{\mu,\varepsilon}}{\kappa^{0}_{\mu,\varepsilon}} + \frac{1}{4\varepsilon} \int \left(|Qx|^2 + |Kx|^2 \right) d\mu(x).$$

For smooth perturbations generated by a velocity field w_t,

$$\frac{d}{dt} F_{\varepsilon}(\mu_t) = \int \nabla \frac{\delta F_{\varepsilon}}{\delta \mu}(\mu_t)(x) \cdot w_t(x)\, \mu_t(dx).$$

In the W_2 metric, steepest descent is $w_t = -\nabla \frac{\delta F_{\varepsilon}}{\delta \mu}(\mu_t)$, so the continuity equation is

$$\partial_t \mu_t + \nabla \cdot \Big(\mu_t \nabla \frac{\delta F_{\varepsilon}}{\delta \mu}(\mu_t) \Big) = 0,$$

and the dissipation identity becomes

$$\frac{d}{dt} F_{\varepsilon}(\mu_t) = -\int \Big| \nabla \frac{\delta F_{\varepsilon}}{\delta \mu}(\mu_t) \Big|^2 d\mu_t \le 0.$$

Under $A = A^{\top} = -V$, the Sinkhorn velocity computed from the model coincides with $-\nabla \delta F_{\varepsilon} / \delta \mu$, so this PDE is exactly the Sinkformer equation.

Now restrict to Gaussian states $\mu_t = \mathcal{N}(\alpha_t, \Sigma_t)$. The W_2 metric restricted to Gaussians induces the Bures–Wasserstein metric g_{BW} on (α, Σ). Therefore the restricted dynamics is the Riemannian gradient flow

$$(\dot{\alpha}_t, \dot{\Sigma}_t) = -\operatorname{grad}_{g_{BW}} \big(F_{\varepsilon}|_{\mathrm{Gauss}} \big)(\alpha_t, \Sigma_t),$$

which is precisely the Bures–Wasserstein ODE system stated in Section 5.3. $\square$

A.4.14 Theorem 3.57 [2, Section 5.4]

Proof. Let

$$F(\mu) = \frac{1}{2} \iint e^{Ax\cdot y}\, \mu(dx)\mu(dy), \qquad Z_{\mu}(x) := \int e^{Ax\cdot y}\, \mu(dy),$$

and define the mobility operator

$$G_{\mu}(v)(x) = \frac{Bv(x)}{Z_{\mu}(x)}, \qquad B := -VA^{-\top}.$$

If B is symmetric positive definite, the dynamic action

$$\int_0^1 \int \langle G_{\mu_t}^{-1} v_t, v_t \rangle\, d\mu_t\, dt$$

is well defined and induces the twisted transport distance $d_{A,V}$ by a Benamou–Brenier formula.

Compute first variation:

$$\frac{\delta F}{\delta \mu}(x) = \int e^{Ax\cdot y}\, \mu(dy) = Z_{\mu}(x).$$

Hence

$$\nabla \frac{\delta F}{\delta \mu}(x) = A^{\top} \int y\, e^{Ax\cdot y}\, \mu(dy).$$

Applying G_μ:

$$G_\mu\left(\nabla\frac{\delta F}{\delta\mu}\right)(x) = \frac{BA^\top}{Z_\mu(x)}\int y\, e^{Ax\cdot y}\,\mu(dy) = -\frac{V}{Z_\mu(x)}\int y\, e^{Ax\cdot y}\,\mu(dy),$$

because $BA^\top = -V$. Therefore the generalized gradient-flow equation

$$\partial_t\mu + \nabla\cdot\left(\mu\, G_\mu\nabla\frac{\delta F}{\delta\mu}\right) = 0$$

is

$$\partial_t\mu - \nabla\cdot\left(\mu\,\frac{V\int ye^{Ax\cdot y}\mu(dy)}{Z_\mu(x)}\right) = 0,$$

equivalently

$$\partial_t\mu + \nabla\cdot\left(\mu\,\Gamma_\mu^{(SM)}\right) = 0,$$

which is the Softmax PDE in this parameter regime.

For geodesic convexity, one computes the second derivative of F along $d_{A,V}$-geodesics. The resulting Hessian contains indefinite terms coming from the μ-dependence of G_μ (the denominator Z_μ), so no global nonnegative lower bound exists in general. Hence F is generally not geodesically convex for $d_{A,V}$. □

A.5 Litman [7]

A.5.1 Definition 2.20 [7, Definition 3.2]

Proof. This entry is a definition, so the only point to check is the equivalence of the two displayed objectives. By $C_j := -s_j$ and $H(p) := -\sum_{j=1}^m p_j \log p_j$ (with $0\log 0 := 0$),

$$\sum_{j=1}^m p_j C_j - \varepsilon H(p) = -\sum_{j=1}^m p_j s_j + \varepsilon\sum_{j=1}^m p_j\log p_j = -\langle p, s\rangle + \varepsilon\sum_{j=1}^m p_j\log p_j.$$

Hence the two formulations define the same optimization problem. □

A.5.2 Theorem 2.21 [7, Theorem 3.3]

Proof. Define

$$J(p) := -\langle p, s\rangle + \tau\sum_{j=1}^m p_j\log p_j, \qquad p\in\Delta_{m-1},\ \tau > 0.$$

The simplex is compact and J is continuous (with the convention $0\log 0 = 0$), so a minimizer exists.

Because $p\mapsto\sum_j p_j\log p_j$ is strictly convex on Δ_{m-1} and $p\mapsto -\langle p, s\rangle$ is linear, J is strictly convex; therefore the minimizer is unique.

At the minimizer all coordinates are strictly positive (the entropy derivative tends to $-\infty$ at 0^+, so a boundary point cannot satisfy optimality for finite scores). Hence we can use a single equality multiplier λ for $\sum_j p_j = 1$:

$$\mathcal{L}(p,\lambda) = -\sum_{j=1}^m p_j s_j + \tau\sum_{j=1}^m p_j\log p_j + \lambda\Big(\sum_{j=1}^m p_j - 1\Big).$$

Stationarity gives, for every j,

$$-s_j + \tau(\log p_j + 1) + \lambda = 0 \quad\Longleftrightarrow\quad p_j = \exp\Big(\frac{s_j-\lambda-\tau}{\tau}\Big) = C\,e^{s_j/\tau},$$

with the same constant C for all j. Enforcing $\sum_j p_j = 1$ yields

$$p_j^* = \frac{e^{s_j/\tau}}{\sum_{\ell=1}^m e^{s_\ell/\tau}}.$$

This is exactly scaled-dot-product attention. □

A.5.3 Remark 2.4 [7, Remark 3.4]

Proof. In the one-sided EOT objective

$$J_\tau(p) = -\langle p, s\rangle + \tau \sum_j p_j \log p_j,$$

τ is exactly the coefficient of the entropy term. Therefore decreasing τ weakens entropy regularization and pushes solutions toward concentrated vertices of the simplex, while increasing τ strengthens entropy regularization and produces more diffuse weights.

For the Transformer scaling rule, write

$$s_j = \langle q, k_j\rangle, \qquad p_j^* = \frac{e^{s_j/\tau}}{\sum_\ell e^{s_\ell/\tau}}.$$

If coordinates of $q, k_j \in \mathbb{R}^{d_k}$ are centered, independent, and unit variance, then

$$\mathrm{Var}(s_j) = \mathrm{Var}\left(\sum_{r=1}^{d_k} q_r k_{j,r}\right) = d_k.$$

Choosing $\tau = \sqrt{d_k}$ makes s_j/τ order-one in variance, preventing softmax from entering a saturated near-argmax regime solely because of dimension growth. So $\tau = \sqrt{d_k}$ is both a variance normalization and an explicit dimension-dependent entropy scale. ☐

A.5.4 Proposition 2.22 [7, Proposition 4.1]

Proof. With $\Omega(p) = -\tau H(p) = \tau \sum_j p_j \log p_j$, the variational problem is

$$\min_{p\in\Delta_{m-1}} \left\{ -\langle p, s\rangle + \tau \sum_{j=1}^{m} p_j \log p_j \right\},$$

which is exactly the entropy-regularized problem from Theorem 2.21. For completeness, write the Lagrangian

$$\mathcal{L}(p, \lambda) = -\sum_j p_j s_j + \tau \sum_j p_j \log p_j + \lambda\Big(\sum_j p_j - 1\Big).$$

Stationarity gives

$$-s_j + \tau(\log p_j + 1) + \lambda = 0 \Longrightarrow p_j = Ce^{s_j/\tau}.$$

Normalization fixes C and yields

$$p_j^* = \frac{e^{s_j/\tau}}{\sum_\ell e^{s_\ell/\tau}}.$$

Hence choosing $\Omega = -\tau H$ gives precisely softmax. ☐

A.5.5 Proposition 2.23 [7, Proposition 4.2]

Proof. The objective is

$$J(p) = \frac{1}{2}\|p\|_2^2 - \langle p, s\rangle, \qquad p \in \Delta_{m-1}.$$

Completing the square:

$$J(p) = \frac{1}{2}\|p - s\|_2^2 - \frac{1}{2}\|s\|_2^2.$$

So minimizing J over Δ_{m-1} is exactly Euclidean projection of s onto the simplex, which is the Sparsemax map.

Using KKT explicitly, let λ be the multiplier for $\sum_j p_j = 1$ and $\nu_j \geq 0$ for $p_j \geq 0$:

$$\mathcal{L}(p, \lambda, \nu) = \frac{1}{2}\sum_j p_j^2 - \sum_j p_j s_j + \lambda\Big(\sum_j p_j - 1\Big) - \sum_j \nu_j p_j.$$

Stationarity gives

$$p_j - s_j + \lambda - \nu_j = 0.$$

If $p_j > 0$, then $\nu_j = 0$ and $p_j = s_j - \lambda$. If $p_j = 0$, complementary slackness allows $\nu_j \geq 0$, equivalent to $s_j - \lambda \leq 0$. Hence

$$p_j = (s_j - \lambda)_+.$$

Set $\theta := \lambda$. The map $\theta \mapsto \sum_j (s_j - \theta)_+$ is continuous, strictly decreasing from $+\infty$ to 0, so there is a unique θ such that

$$\sum_j (s_j - \theta)_+ = 1.$$

Therefore

$$p_j^* = (s_j - \theta)_+,$$

which is exactly Sparsemax. □

A.5.6 Proposition 2.24 [7, Proposition 4.3]

Proof. Consider

$$\min_{p \in \Delta_{m-1}} \left\{ -\langle p, s\rangle + \frac{1}{\alpha(\alpha-1)} \sum_{j=1}^{m} (p_j^\alpha - p_j) \right\}, \qquad \alpha > 1.$$

Since $\sum_j p_j = 1$, the linear term $-\frac{1}{\alpha(\alpha-1)} \sum_j p_j$ is constant on the simplex and can be absorbed into the multiplier. Let λ be the multiplier for $\sum_j p_j = 1$ and $\nu_j \geq 0$ for $p_j \geq 0$. Stationarity gives

$$-s_j + \frac{1}{\alpha - 1} p_j^{\alpha-1} + \lambda - \nu_j = 0.$$

For active coordinates $p_j > 0$, complementary slackness gives $\nu_j = 0$, hence

$$p_j^{\alpha-1} = (\alpha - 1)(s_j - \lambda),$$

so

$$p_j = \left[(\alpha-1)(s_j - \lambda)\right]^{\frac{1}{\alpha-1}}.$$

For inactive coordinates $p_j = 0$, KKT implies $s_j - \lambda \leq 0$. Therefore, for all j,

$$p_j = \left[(\alpha-1)(s_j - \lambda)\right]_+^{\frac{1}{\alpha-1}}.$$

Writing $\theta := (\alpha - 1)\lambda$ gives the usual threshold form

$$p_j \propto \left[(\alpha-1)s_j - \theta\right]_+^{\frac{1}{\alpha-1}}.$$

The proportionality constant (equivalently, the threshold θ) is uniquely fixed by $\sum_j p_j = 1$. This is exactly the α-entmax family. □

A.5.7 Proposition 2.25 [7, Proposition 4.4]

Proof. Fix a query index i and optimize over the row $p_i = (p_{ij})_{j=1}^m \in \Delta_{m-1}$:

$$J_i(p_i) = -\sum_{j=1}^{m} p_{ij} s_{ij} + \tau \sum_{j=1}^{m} p_{ij} \log p_{ij} + \gamma \sum_{j=1}^{m} p_{ij} |i - j|.$$

Rearrange:

$$J_i(p_i) = -\sum_{j=1}^{m} p_{ij} \big(s_{ij} - \gamma|i-j|\big) + \tau \sum_{j=1}^{m} p_{ij} \log p_{ij}.$$

So this is exactly entropy-regularized softmax with shifted scores

$$s'_{ij} := s_{ij} - \gamma|i - j|.$$

Applying the softmax derivation (Theorem 2.21) gives

$$p_{ij}^* = \frac{\exp(s'_{ij}/\tau)}{\sum_{\ell=1}^m \exp(s'_{i\ell}/\tau)} = \frac{\exp((s_{ij} - \gamma|i-j|)/\tau)}{\sum_{\ell=1}^m \exp((s_{i\ell} - \gamma|i-\ell|)/\tau)}.$$

Hence ALiBi is exactly a linear logit shift inside the same EOT optimizer. □

A.5.8 Proposition 2.26 [7, Proposition 4.5]

Proof. Write the objective as

$$J(p) = -\langle p, s\rangle + \tau\,\mathrm{KL}(p\|\pi), \qquad \mathrm{KL}(p\|\pi) = \sum_{j=1}^m p_j \log\frac{p_j}{\pi_j}.$$

Expanding KL:

$$J(p) = -\sum_j p_j s_j + \tau\sum_j p_j\log p_j - \tau\sum_j p_j\log\pi_j = -\sum_j p_j\big(s_j + \tau\log\pi_j\big) + \tau\sum_j p_j\log p_j.$$

So this is the entropy-regularized softmax problem with shifted logits

$$s'_j := s_j + \tau\log\pi_j.$$

Its unique optimizer satisfies

$$p_j^* \propto e^{s'_j/\tau} = e^{s_j/\tau}\pi_j.$$

Normalizing over j yields

$$p_j^* = \frac{\pi_j e^{s_j/\tau}}{\sum_\ell \pi_\ell e^{s_\ell/\tau}}.$$

When π is uniform, this reduces to standard softmax. □

A.5.9 Theorem 2.27 [7, Theorem 5.2]

Proof. Let

$$p_k(s) = \frac{e^{s_k/\tau}}{\sum_{\ell=1}^m e^{s_\ell/\tau}}, \qquad u_k := -\frac{\partial L}{\partial p_k}.$$

By the chain rule,

$$\frac{\partial L}{\partial s_j} = \sum_{k=1}^m \frac{\partial L}{\partial p_k}\frac{\partial p_k}{\partial s_j} = -\sum_{k=1}^m u_k\frac{\partial p_k}{\partial s_j}.$$

Differentiate softmax:

$$\frac{\partial p_k}{\partial s_j} = \frac{1}{\tau}p_k(\delta_{kj} - p_j).$$

Hence

$$\begin{aligned}\frac{\partial L}{\partial s_j} &= -\frac{1}{\tau}\sum_{k=1}^m u_k p_k(\delta_{kj} - p_j)\\ &= -\frac{1}{\tau}\left(u_j p_j - p_j\sum_{k=1}^m p_k u_k\right)\\ &= -\frac{p_j}{\tau}\left(u_j - \sum_{k=1}^m p_k u_k\right).\end{aligned}$$

Since $\mathbb{E}_p[u] = \sum_k p_k u_k$, we obtain

$$\frac{\partial L}{\partial s_j} = -\frac{p_j}{\tau}\big(u_j - \mathbb{E}_p[u]\big),$$

which is the announced advantage form. □

A.5.10 Theorem 2.28 [7, Theorem 6.3]

Proof. Define the convex dual potential as

$$\phi^*(s) := \sup_{p\in\Delta_{m-1}} \left\{ \langle p, s\rangle - \tau \sum_{j=1}^{m} p_j \log p_j \right\}.$$

This is the dual optimal value (equivalently, minus the primal optimum in the minimization formulation).

Set

$$\Psi(p) := \langle p, s\rangle - \tau \sum_j p_j \log p_j$$

on Δ_{m-1}. Since $-\sum_j p_j \log p_j$ is strictly concave, Ψ is strictly concave, so the maximizer is unique and interior. Using multiplier λ for $\sum_j p_j = 1$:

$$\frac{\partial}{\partial p_j}\Big(\Psi(p) - \lambda(\sum_\ell p_\ell - 1)\Big) = 0 \Longrightarrow s_j - \tau(\log p_j + 1) - \lambda = 0.$$

Hence

$$p_j = Ce^{s_j/\tau}, \qquad C = \frac{1}{\sum_\ell e^{s_\ell/\tau}},$$

so

$$p_j^*(s) = \frac{e^{s_j/\tau}}{\sum_\ell e^{s_\ell/\tau}}.$$

Evaluate Ψ at p^*:

$$\phi^*(s) = \tau \log\Big(\sum_{\ell=1}^{m} e^{s_\ell/\tau}\Big).$$

Finally, differentiating the log-sum-exp gives

$$\partial_{s_j}\phi^*(s) = \frac{e^{s_j/\tau}}{\sum_\ell e^{s_\ell/\tau}} = p_j^*(s).$$

Therefore $\nabla_s \phi^*(s) = p^*(s)$. □

A.5.11 Theorem 2.29 [7, Theorem 7.3]

Proof. For $p(s) = \mathrm{softmax}(s/\tau)$, the Jacobian is

$$J_p(s) = \frac{\partial p}{\partial s} = \frac{1}{\tau}\big(\operatorname{diag}(p) - pp^\top\big).$$

For this exponential family parameterization, the Fisher information matrix is

$$F(s) = \frac{1}{\tau^2}\big(\operatorname{diag}(p) - pp^\top\big).$$

From Theorem 2.27, componentwise

$$\frac{\partial L}{\partial s_j} = -\frac{p_j}{\tau}\big(u_j - \mathbb{E}_p[u]\big).$$

In vector form,

$$\nabla_s L = -\frac{1}{\tau}\big(\operatorname{diag}(p) - pp^\top\big)u.$$

Since

$$\operatorname{diag}(p) - pp^\top = \tau^2 F(s),$$

we get

$$\nabla_s L = -\frac{1}{\tau}\big(\tau^2 F(s)\big)u = -\tau F(s)u.$$

So the Euclidean score gradient is exactly the Fisher-metric image of u. □

A.5.12 Corollary 2.30 [7, Theorem 7.5]

Proof. From Theorem 2.28,

$$\phi^*(s) = \tau \log\Big(\sum_{\ell=1}^{m} e^{s_\ell/\tau}\Big), \qquad \nabla_s \phi^*(s) = p(s).$$

Differentiate again:

$$\nabla_s^2 \phi^*(s) = \frac{\partial p}{\partial s} = \frac{1}{\tau}\big(\operatorname{diag}(p) - pp^\top\big).$$

By definition of the attention Fisher matrix,

$$F(s) = \frac{1}{\tau^2}\big(\operatorname{diag}(p) - pp^\top\big).$$

Therefore

$$\nabla_s^2 \phi^*(s) = \tau F(s),$$

which is exactly the claim. ☐

A.6 Rigollet [8]

A.6.1 Theorem 3.16 [8, Theorem 1]

Proof. Write the particle system on $(S^{d-1})^n$ as

$$\dot{X} = -\nabla \Phi(X),$$

where Φ is the analytic interaction energy (for SA and USA separately) restricted to the compact manifold $(S^{d-1})^n$.

Step 1: global well-posedness. The vector field is smooth and tangent to $(S^{d-1})^n$, so maximal solutions exist and cannot leave a compact set. Hence trajectories are global for all $t \geq 0$.

Step 2: convergence to a single equilibrium. Along every trajectory,

$$\frac{d}{dt}\Phi(X(t)) = -\|\nabla \Phi(X(t))\|^2 \leq 0.$$

Therefore $\Phi(X(t))$ is decreasing and bounded below, so it converges. Since the flow is analytic on a compact manifold, the Lojasiewicz gradient inequality implies that each trajectory has a limit point X_∞ and in fact $X(t) \to X_\infty$.

Step 3: classification of attractors. Consensus configurations

$$\mathcal{C} := \{(x, \ldots, x) : x \in S^{d-1}\}$$

are equilibria. For non-consensus equilibria, the second-variation/Hessian computation for these models yields at least one unstable direction, i.e. every non-consensus critical point is a strict saddle.

Step 4: almost-sure selection of consensus. By the center-stable manifold theorem, the set of initial conditions converging to a strict saddle is contained in a finite union of immersed manifolds of codimension at least one, hence has Lebesgue measure zero in $(S^{d-1})^n$. Therefore for almost every initial condition, the limit equilibrium belongs to $\mathcal{C}$.

So there is $x_* \in S^{d-1}$ such that $\|x_i(t) - x_j(t)\| \to 0$ for all i, j. The equivalent empirical-measure statement is $\mu_t \rightharpoonup \delta_{x_*}$. ☐

A.6.2 Theorem 3.17 [8, Theorem 3]

Proof. Assume there is $w \in S^{d-1}$ with $\langle x_i(0), w\rangle > 0$ for all i. By continuity and invariance of the hemisphere cone for SA/USA, there is $\eta > 0$ (depending on the initialization) such that

$$\langle x_i(t), w\rangle \geq \eta, \qquad t \geq 0,\ i \in [n].$$

This gives uniform lower bounds on interaction weights and keeps the dynamics in a strictly contractive region.

Define the pairwise disagreement functional

$$\mathcal{D}(t) := \max_{i,j} \big(1 - \langle x_i(t), x_j(t)\rangle\big).$$

Differentiating $1 - \langle x_i, x_j\rangle$ along the ODE and using the hemisphere lower bound yields

$$\frac{d}{dt}\big(1 - \langle x_i, x_j\rangle\big) \le -2\lambda\big(1 - \langle x_i, x_j\rangle\big)$$

for some $\lambda > 0$ depending on $(x(0), n, \beta)$. Taking the maximum over (i, j) and applying Gronwall:

$$\mathcal{D}(t) \le \mathcal{D}(0)e^{-2\lambda t}.$$

Since

$$\|x_i - x_j\|^2 = 2\big(1 - \langle x_i, x_j\rangle\big),$$

the diameter decays exponentially:

$$\max_{i,j} \|x_i(t) - x_j(t)\| \le C_1 e^{-\lambda t}.$$

Hence $(x_i(t))_i$ is Cauchy to a common limit $x_* \in S^{d-1}$, and

$$\|x_i(t) - x_*\| \le Ce^{-\lambda t}, \qquad i \in [n],\ t \ge 0.$$

□

A.6.3 Corollary 3.18 [8, Corollary 4]

Proof. For i.i.d. continuous points on S^{d-1}, Wendel's theorem gives

$$\mathbb{P}(\text{all points lie in a common open hemisphere}) = 1 - 2^{1-n} \sum_{k=0}^{d-1} \binom{n-1}{k}.$$

If $d \ge n$, then the sum includes all binomial terms:

$$\sum_{k=0}^{d-1} \binom{n-1}{k} = \sum_{k=0}^{n-1} \binom{n-1}{k} = 2^{n-1},$$

so the probability equals 1. Hence, almost surely, the hemisphere condition in Theorem 3.17 is satisfied at $t = 0$. Applying that theorem gives exponential convergence to a single cluster. □

A.6.4 Theorem 3.58 [8, Theorem 5]

Proof. Let μ_t be a weak solution with nonzero first moment

$$m_t := \int_{S^{d-1}} x\, \mu_t(dx), \qquad \|m_0\| = R_0 > 0.$$

For $|\beta|$ sufficiently small, the drift is a perturbation of the $\beta = 0$ consensus drift, so the direction $u_t := m_t/\|m_t\|$ is well-defined after a finite transient and remains in a compact subset away from degeneracy.

Introduce the alignment error relative to u_t:

$$\mathcal{E}_t := \int_{S^{d-1}} \big(1 - \langle x, u_t\rangle\big)\, \mu_t(dx).$$

Using the weak formulation of the PDE with the test function $x \mapsto 1 - \langle x, u_t\rangle$, and linearizing the drift around δ_{u_t}, one gets

$$\frac{d}{dt}\mathcal{E}_t \le -c_0\mathcal{E}_t + r_t,$$

where $c_0 > 0$ is a spectral-gap constant (uniform for $|\beta| < \beta_0$) and r_t is a higher-order remainder controlled by L^2 bounds of the density. For $t \geq T_0$ (with T_0 depending on μ_0), $r_t \leq \frac{c_0}{2}\mathcal{E}_t$, hence

$$\frac{d}{dt}\mathcal{E}_t \leq -\frac{c_0}{2}\mathcal{E}_t, \qquad t \geq T_0.$$

Gronwall yields

$$\mathcal{E}_t \leq Ce^{-ct}, \qquad c := c_0/2.$$

Because $W_2^2(\mu_t, \delta_{u_t}) \lesssim \mathcal{E}_t$ on the sphere,

$$W_2(\mu_t, \delta_{u_t}) \leq C_1 e^{-ct}.$$

Also, $\dot{u}_t$ is controlled by the same fluctuation size: $|\dot{u}_t| \lesssim \mathcal{E}_t$. Hence (u_t) is Cauchy and converges to some $x_\infty \in S^{d-1}$ with

$$|u_t - x_\infty| \leq C_2 e^{-ct}.$$

By triangle inequality,

$$W_2(\mu_t, \delta_{x_\infty}) \leq W_2(\mu_t, \delta_{u_t}) + W_2(\delta_{u_t}, \delta_{x_\infty}) \leq C_0 e^{-ct}.$$

In the theorem normalization, c is written as $1/100$. □

A.6.5 Theorem 3.59 [8, Theorem 6]

Proof. Let

$$s_{ij}(0) := \langle x_i(0), x_j(0)\rangle, \qquad s_* := s_{\bar{\imath}\bar{\jmath}}(0) = \max_{i\neq j} s_{ij}(0),$$

and assume the maximizing pair is unique. Then there is $\eta > 0$ such that

$$s_{ij}(0) \leq s_* - \eta \qquad \text{for all } (i,j) \neq (\bar{\imath}, \bar{\jmath}).$$

In the attention weights,

$$a_{ij} \propto e^{\beta s_{ij}},$$

so relative to the dominant pair,

$$\frac{a_{ij}}{a_{\bar{\imath}\bar{\jmath}}} \lesssim e^{-\beta\eta} \to 0 \qquad (\beta \to \infty)$$

uniformly on compact times before collision.

Introduce the rescaled time

$$dt = e^{\beta(1-s_*)}\, ds.$$

After this scaling, the $(\bar{\imath}, \bar{\jmath})$ interaction is order one, while all other interactions are exponentially small ($O(e^{-\beta\eta})$). Thus, by Tikhonov/singular-perturbation compactness, trajectories converge (as $\beta \to \infty$) to the reduced system: only $(x_{\bar{\imath}}, x_{\bar{\jmath}})$ evolves, all other coordinates are frozen.

The reduced two-body dynamics is tangent to S^{d-1} and confined to the geodesic through the initial pair; its scalar separation variable is strictly decreasing until it reaches 0. Therefore the two active clusters merge in finite rescaled time. Uniform convergence on compact s-intervals before contact follows from the $O(e^{-\beta\eta})$ remainder bound. □

A.6.6 Theorem 3.19 [8, Theorem 7]

Proof. For equiangular inputs,

$$\langle x_i, x_i\rangle = 1, \qquad \langle x_i, x_j\rangle = \rho \ (i \neq j).$$

With $\beta_n = \gamma \log n$, the query-i attention denominator is

$$Z_i = e^{\beta_n} + (n-1)e^{\beta_n\rho} = n^\gamma + (n-1)n^{\gamma\rho}.$$

Hence

$$a_{ii} = \frac{n^\gamma}{n^\gamma + (n-1)n^{\gamma\rho}} = \frac{1}{1 + (n-1)n^{-\gamma(1-\rho)}},$$

$$a_{ij} = \frac{n^{\gamma\rho}}{n^\gamma + (n-1)n^{\gamma\rho}} = \frac{n^{-\gamma(1-\rho)}}{1 + (n-1)n^{-\gamma(1-\rho)}} \quad (i \neq j).$$

So the regime is governed by

$$\xi_n := (n-1)n^{-\gamma(1-\rho)}.$$

Subcritical regime $\gamma < \frac{1}{1-\rho}$. Then $\xi_n \to \infty$, so $a_{ii} \to 0$ and $(n-1)a_{ij} \to 1$. Each output is asymptotically the average of the other tokens, hence all outputs become asymptotically identical after normalization:

$$\langle \theta_i, \theta_j \rangle \to 1.$$

Supercritical regime $\gamma > \frac{1}{1-\rho}$. Then $\xi_n \to 0$, so $a_{ii} \to 1$ and $a_{ij} \to 0$. Thus one layer is asymptotically identity on directions:

$$\theta_i \to x_i, \qquad \langle \theta_i, \theta_j \rangle \to \langle x_i, x_j \rangle = \rho.$$

Critical regime $\gamma = \frac{1}{1-\rho}$. Now $\xi_n = (n-1)/n$, so

$$a_{ii} = \frac{n}{2n-1}, \qquad a_{ij} = \frac{1}{2n-1}\ (i \neq j).$$

Let $S := \sum_{k=1}^{n} x_k$. Then

$$y_i := \sum_{j=1}^{n} a_{ij} x_j = (a_{ii} - a_{ij}) x_i + a_{ij} S.$$

Using equiangular identities

$$x_i \cdot S = 1 + (n-1)\rho =: c_n, \qquad \|S\|^2 = n\, c_n,$$

one computes for $i \neq j$:

$$y_i \cdot y_j = (a_{ii} - a_{ij})^2 \rho + 2(a_{ii} - a_{ij}) a_{ij} c_n + a_{ij}^2 n c_n,$$
$$\|y_i\|^2 = (a_{ii} - a_{ij})^2 + 2(a_{ii} - a_{ij}) a_{ij} c_n + a_{ij}^2 n c_n.$$

Passing to $n \to \infty$ gives

$$y_i \cdot y_j \to \rho, \qquad \|y_i\|^2 \to \frac{1+3\rho}{4}.$$

Therefore

$$\langle \theta_i, \theta_j \rangle = \frac{y_i \cdot y_j}{\|y_i\|\, \|y_j\|} \longrightarrow \frac{4\rho}{1+3\rho}.$$

This yields the three claimed limits. $\square$

A.6.7 Definition 3.20 [8, Equation (8)]

Proof. Start from the noisy particle model on S^{d-1}:

$$dX_i^n(t) = P^{\perp}_{X_i^n(t)} \left(\frac{1}{n} \sum_{j=1}^{n} e^{\beta \langle X_i^n(t), X_j^n(t) \rangle} X_j^n(t) \right) dt + \sqrt{2\kappa^{-1}}\, dW_i(t).$$

As $n \to \infty$, propagation-of-chaos gives a nonlinear limit in which the empirical average is replaced by expectation under the law $\mu_t = \mathcal{L}(X(t))$:

$$\int_{S^{d-1}} e^{\beta \langle X(t), y \rangle} y\, \mu_t(dy).$$

Projecting onto the tangent space at $X(t)$ keeps the dynamics on the sphere. Therefore the limit process satisfies

$$dX(t) = P^{\perp}_{X(t)} \left(\int_{S^{d-1}} e^{\beta \langle X(t), y \rangle} y\, \mu_t(dy) \right) dt + \sqrt{2\kappa^{-1}}\, dW(t), \qquad \mu_t = \mathcal{L}(X(t)).$$

This is precisely the stated McKean–Vlasov SDE. $\square$

A.6.8 Definition 3.60 [8, Equation (9)]

Proof. Let

$$b_\mu(x) := P_x^\perp\left(\int_{S^{d-1}} e^{\beta\langle x,y\rangle} y\,\mu(dy)\right).$$

For a smooth test function φ on S^{d-1}, Itô's formula for Definition 3.20 gives

$$\frac{d}{dt}\mathbb{E}[\varphi(X_t)] = \mathbb{E}[\nabla\varphi(X_t)\cdot b_{\mu_t}(X_t)] + \kappa^{-1}\mathbb{E}[\Delta\varphi(X_t)].$$

Since $\mu_t = \mathcal{L}(X_t)$, this is

$$\frac{d}{dt}\int \varphi\,d\mu_t = \int \nabla\varphi(x)\cdot b_{\mu_t}(x)\,\mu_t(dx) + \kappa^{-1}\int \Delta\varphi(x)\,\mu_t(dx).$$

Integrating by parts on the sphere,

$$\int \nabla\varphi\cdot b_{\mu_t}\,d\mu_t = -\int \varphi \div (\mu_t b_{\mu_t}), \qquad \int \Delta\varphi\,d\mu_t = \int \varphi\,\Delta\mu_t.$$

Hence, in distributional form,

$$\partial_t\mu_t + \kappa^{-1}\Delta\mu_t = \div\big(\mu_t\,b_{\mu_t}\big),$$

which is exactly the displayed Fokker–Planck equation. Letting $\kappa \to \infty$ removes the diffusion term and recovers the deterministic mean-field continuity equation. ☐