CONVERGENCE OF MEAN-FIELD TRANSFORMERS AND BILEVEL TRAINING DYNAMICS IN THE REPULSIVE REGIME

Anonymous authors
Paper under double-blind review

ABSTRACT

We study token convergence and parameter training in continuous-depth, mean-field self-attention on the sphere, in the repulsive regime with parameters shared across depth. For fixed symmetric positive-definite interactions, we investigate a Polyak–Łojasiewicz-type inequality relating energy dissipation to a power of the energy gap. Under uniform regularity and non-degeneracy bounds along the flow, we establish this inequality with exponent $\beta = 2$, yielding an $O(t^{-1})$ energy-gap bound. We then study training as a bilevel optimization problem: the token distribution evolves toward equilibrium, while the parameters are adjusted to minimize a loss defined at that equilibrium. Tokens and parameters evolve simultaneously, without requiring equilibration between parameter updates. Under suitable regularity and gradient-approximation assumptions, a dissipation exponent $\beta \in (1, 2]$ determines a learning-rate schedule yielding an $O(T^{-1/\beta} \log T)$ bound on the time-averaged squared training-loss gradient. In particular, quadratic dissipation gives an $O(\log T/\sqrt{T})$ training bound. Numerical experiments support the predicted convergence rates.

1 INTRODUCTION

Self-attention updates each token by a weighted average of the other tokens. In the continuous-depth, mean-field limit (i.e., when the number of tokens goes to infinity), the token distribution solves a nonlinear transport equation (Sander et al., 2022; Geshkovski et al., 2025). Understanding these dynamics has attracted growing interest, as they describe how token representations evolve with depth. For suitable attention matrices, these dynamics admit an interaction-energy gradient-flow formulation (Geshkovski et al., 2025; Burger et al., 2025). The behavior of these dynamics depends on the attention matrices and the temperature.

Attractive interactions can bring tokens together into clusters, whereas repulsive interactions can spread them over the sphere. Temperature controls how selectively each token x attends to tokens y: for weights proportional to $\exp(x^\top Dy/\tau)$, with $D = Q^\top K$ the query-key interaction matrix, smaller $\tau > 0$ gives greater weight to tokens with larger similarity scores. Chen et al. (2025b) establish exponential convergence to a single cluster for unnormalized attractive attention under suitable assumptions on the initialization and temperature. For normalized attention, Alcalde et al. (2026) derive quantitative concentration bounds over finite time windows in the low-temperature regime. Our work addresses two related but distinct questions. First, how fast does the attention transport equation converge in the repulsive regime? Second, how does the convergence rate of this inner attention flow affect the learning dynamics of the interaction parameters in a transformer ?

In this paper, we consider normalized tokens on the sphere and exponential attention with interaction matrix $D = Q^\top K$. We focus on the symmetric positive-definite case and the repulsive regime, where the value matrix satisfies $V = -D$, so that the dynamics decrease an interaction energy E_D. Our setting models a continuous-depth, parameter-tied attention architecture, with the same query-key interaction shared across depth.

Our analysis proceeds in two steps. At fixed parameters, we combine convexity of the interaction energy with homogeneous Sobolev duality to obtain a quadratic energy-dissipation inequality, under uniform regularity bounds along the flow. We then use this inner-flow estimate to study the training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dynamics of transformer models. We choose the parameter learning rate to balance progress in the training loss against the error caused by the evolving token distribution.

The remainder of this paper is organized as follows. Section 2 introduces the mean-field attention model and the gradient flow viewpoint. Section 3 establishes a conditional $O(t^{-1})$ convergence rate for general symmetric positive-definite interactions. Section 4 turns this fixed-parameter result into a bound for the coupled training dynamics. Section 5 presents our experiments to validate our theoretical findings, Section 6 discusses related work, and Section 7 concludes. The appendix contains the proofs and additional technical details and experiments.

2 Background and problem setting

We use the mean-field formulation of normalized self-attention developed in Geshkovski et al. (2025); Burger et al. (2025). Let $S = \mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, $d \geq 2$, and let ν be its uniform probability measure. We write ∇_S, div_S, and $\Delta_S = \operatorname{div}_S \nabla_S$ for the spherical gradient, divergence, and Laplace-Beltrami operator. For a symmetric positive-definite matrix D, define the kernel, potential, and interaction energy by

$$K_D(x,y) = e^{x^\top Dy}, \qquad m_{D,\mu} = \int_S K_D(\cdot, y)\,\mathrm{d}\mu(y), \qquad E_D(\mu) = \int_S m_{D,\mu}\,\mathrm{d}\mu. \tag{1}$$

For n normalized token vectors $x_i(t) \in S$, we consider a continuous-depth, parameter-tied attention architecture in which the same interaction matrix D and value matrix $V = -D$ are shared across depth. Thus, D is fixed along the depth variable t rather than evolving as $D(t)$. The token dynamics are

$$\dot{x}_i = -P_{x_i} \sum_{j=1}^{n} A_{ij} D x_j, \qquad A_{ij} = \frac{e^{x_i^\top D x_j}}{\sum_{\ell=1}^{n} e^{x_i^\top D x_\ell}}, \qquad P_x = \mathrm{Id} - xx^\top. \tag{2}$$

Here A_{ij} is the softmax attention weight assigned by token i to token j, and P_x projects onto the tangent space of S at x, keeping the tokens on the sphere. In the mean-field limit, as the number of tokens tends to infinity, the token distribution μ_t evolves according to the continuity equation

$$\partial_t \mu_t = \operatorname{div}_S\left(\mu_t \frac{\nabla_S m_{D,\mu_t}}{m_{D,\mu_t}}\right), \qquad v_{D,\mu} = -\frac{\nabla_S m_{D,\mu}}{m_{D,\mu}}. \tag{3}$$

The normalization in attention produces the nonlocal factor $m_{D,\mu}^{-1}$. Accordingly, (3) is the gradient flow of $E_D/2$ in a transport geometry whose kinetic cost is $\int_S m_{D,\mu}|v|^2\,\mathrm{d}\mu$ (Burger et al., 2025, Section 2). This modifies the Benamou–Brenier dynamic formulation (Benamou & Brenier, 2000) of the standard 2-Wasserstein distance by weighting the kinetic cost with $m_{D,\mu}$, which depends on the current distribution μ.

We refer to (3) as the repulsive attention flow. Along this flow,[1] the chain rule of Burger et al. (2025, Lemma 2.6) gives the energy-dissipation identity

$$\frac{\mathrm{d}}{\mathrm{d}t} E_D(\mu_t) = -2 \int_S \frac{|\nabla_S m_{D,\mu_t}|^2}{m_{D,\mu_t}}\,\mathrm{d}\mu_t. \tag{4}$$

We use the energy without the factor $1/2$ throughout, so the factor 2 in this identity is retained.

Infinite-depth equilibrium at fixed parameters. For fixed interaction parameters D, the equilibrium of the attention flow is the infinite-depth limit of the inner dynamics. Equivalently, it is the minimizer of the interaction energy E_D over token distributions. The energy E_D is continuous for weak convergence of probability measures. Since S is compact, $\mathcal{P}(S)$ is compact in this topology, so E_D attains its minimum. For symmetric positive-definite D, strict convexity makes the minimizer $\mu_D^\star$ unique (Burger et al., 2025, Lemma 3.6). For $D = \alpha\mathrm{Id}$, $\alpha > 0$, this minimizer is ν (Burger et al., 2025, Corollary 3.11). For positive diagonal D, the minimizer is invariant under reflections with respect to eigenvectors of D (Burger et al., 2025, Corollary 3.10), although no general explicit formula is available.

[1] We consider weak solutions that are locally absolutely continuous in the transport metric described above. The energy is then locally absolutely continuous, and the dissipation identity holds for almost every t.

Towards studying training dynamics. We view parameter training as a bilevel optimization problem. The inner problem minimizes E_{D_θ} over token distributions at fixed θ; the outer problem adjusts θ to minimize a loss evaluated at the equilibrium $\mu_\theta^\star$. Section 3 studies convergence of the inner flow through its excess energy. Section 4 studies the coupled evolution of the token distribution and the parameters, where $h(\mu, \theta) = E_{D_\theta}(\mu) - E_{D_\theta}(\mu_\theta^\star)$ measures the energy gap relative to the current equilibrium. We allow the outer loss to depend explicitly on θ: when only the isotropic scale α is trained, the equilibrium remains ν, so a loss depending solely on that equilibrium would be constant.

3 Convergence of the inner transformer flow

We now derive a convergence estimate for the fixed-parameter attention flow. The argument relies on convexity of the interaction energy and on regularity estimates along the flow. Together with homogeneous Sobolev duality, these yield a quadratic energy-dissipation inequality.

Throughout this section, let D be a fixed symmetric positive-definite matrix, and let $\mu_D^\star$ denote the unique minimizer of E_D introduced in Section 2. For any probability measure μ on S, we write $e_D(\mu) := E_D(\mu) - E_D(\mu_D^\star)$ for the corresponding energy gap.

The starting point is the following first-order optimality property of the equilibrium potential.

Lemma 3.1 (Equilibrium potential)**.** *Let $\mu_D^\star$ minimize E_D over $\mathcal{P}(S)$. Then*

$$m_{D,\mu_D^\star} \geq E_D(\mu_D^\star) \quad \text{on } S, \qquad m_{D,\mu_D^\star} = E_D(\mu_D^\star) \quad \text{on } \operatorname{supp}(\mu_D^\star).$$

In particular, if $\operatorname{supp}(\mu_D^\star) = S$, then $m_{D,\mu_D^\star} \equiv E_D(\mu_D^\star)$ on S.

The proof is a direct first-variation argument and is given in Appendix A.

By convexity of the interaction energy, for every $\mu \in \mathcal{P}(S)$,

$$e_D(\mu) \leq 2 \int_S m_{D,\mu} \, \mathrm{d}(\mu - \mu_D^\star). \tag{5}$$

The inequality is also proved in Appendix A and relies on the positive definiteness of the kernel. To control the pairing above (i.e. $\int_S m_{D,\mu} \, \mathrm{d}(\mu - \mu_D^\star)$), we use the homogeneous Sobolev seminorm $\|f\|_{\dot{H}^1(\nu)} := \|\nabla_S f\|_{L^2(\nu)}$ for smooth functions f on S. For a finite signed measure σ with $\sigma(S) = 0$, define the dual norm

$$\|\sigma\|_{\dot{H}^{-1}(\nu)} := \sup \left\{ \int_S f \, \mathrm{d}\sigma : f \in C^\infty(S), \ \|\nabla_S f\|_{L^2(\nu)} \leq 1 \right\}. \tag{6}$$

By duality,

$$\left| \int_S f \, \mathrm{d}\sigma \right| \leq \|\sigma\|_{\dot{H}^{-1}(\nu)} \|\nabla_S f\|_{L^2(\nu)}. \tag{7}$$

We make the following uniform regularity assumption along the attention flow.

Assumption 3.2 (Regularity along the attention flow)**.** *The flow is defined for all $t \geq 0$, with $\mu_t = \rho_t \nu$, and there exist constants $c_0, B > 0$ such that, uniformly in t,*

$$\rho_t \geq c_0 \quad \text{a.e. on } S, \qquad \|\mu_t - \mu_D^\star\|_{\dot{H}^{-1}(\nu)}^2 \leq B. \tag{8}$$

Under this assumption, homogeneous Sobolev duality directly yields a quadratic energy-dissipation inequality and the corresponding $O(t^{-1})$ convergence rate.

Proposition 3.3 (Quadratic energy dissipation)**.** *Let D be a fixed symmetric positive-definite matrix, and let $(\mu_t)_{t \geq 0}$ be a sufficiently regular solution of* (3) *satisfying the dissipation identity* (4)*. Suppose that Assumption 3.2 holds. Writing $e_t := e_D(\mu_t)$ and $\Lambda := \|D\|_{\mathrm{op}}$, the energy gap satisfies*

$$-\dot{e}_t \geq \kappa_{\mathrm{in}} e_t^2, \qquad \kappa_{\mathrm{in}} := \frac{c_0 e^{-\Lambda}}{2B}, \tag{9}$$

almost everywhere, and therefore

$$e_t \leq \frac{e_0}{1 + \kappa_{\mathrm{in}} e_0 t}, \qquad t \geq 0. \tag{10}$$

By compactness of $\mathcal{P}(S)$, weak continuity of E_D, and uniqueness of the minimizer, $e_t \to 0$ also implies $\mu_t \rightharpoonup \mu_D^\star$.

Proof. Set $\sigma_t := \mu_t - \mu_D^\star$. By (5), homogeneous Sobolev duality, and Assumption 3.2,

$$e_t^2 \le 4\|\sigma_t\|^2_{\dot{H}^{-1}(\nu)} \|\nabla_S m_{D,\mu_t}\|^2_{L^2(\nu)} \le 4B \int_S |\nabla_S m_{D,\mu_t}|^2 \,\mathrm{d}\nu.$$

Since $|x^\top D y| \le \Lambda$ on $S \times S$, we have $m_{D,\mu_t} \le e^\Lambda$. Together with $\rho_t \ge c_0$, the dissipation identity (4) gives

$$-\dot{e}_t = 2 \int_S \frac{\rho_t}{m_{D,\mu_t}} |\nabla_S m_{D,\mu_t}|^2 \,\mathrm{d}\nu \ge 2c_0 e^{-\Lambda} \int_S |\nabla_S m_{D,\mu_t}|^2 \,\mathrm{d}\nu.$$

Combining the two estimates proves (9). The rate (10) then follows from the standard nonlinear Grönwall inequality. □

Discussion of the assumptions. The first part of Assumption 3.2 is a lower bound on the attention flow density. Under Assumption 3.2, Proposition 3.3 gives $\mu_t \rightharpoonup \mu_D^\star$, and the bound $\mu_t \ge c_0\nu$ passes to the limit, giving $\mu_D^\star \ge c_0\nu$, hence full support of the equilibrium measure. In the isotropic case $D = \alpha \mathrm{Id}$, $\alpha > 0$, the unique minimizer is the uniform measure $\mu_D^\star = \nu$ (Burger et al., 2025, Proposition 3.11), so full support holds. For general positive-definite D, Burger et al. (2025, Section 1) only conjecture that the minimizer has full support. The second part of Assumption 3.2 requires the $\dot{H}^{-1}$ discrepancy to remain uniformly bounded in time. This condition is examined numerically in Section 5.2.

4 Convergence rates for training dynamics of transformers: coupling the outer loop

We now turn to parameter training and study how convergence of the attention flow affects the training dynamics. Let $\theta \in \Theta \subset \mathbb{R}^r$ denote the trainable transformer parameters specifying D_θ. We consider the bilevel problem

$$\min_\theta \mathcal{L}(\theta) = F(\mu_\theta^\star, \theta), \qquad \mu_\theta^\star \in \arg\min_{\mu \in \mathcal{P}(S)} E_{D_\theta}(\mu).$$

The inner problem optimizes the token distribution at fixed parameters; the outer problem optimizes the parameters using the resulting equilibrium. We assume that $\mathcal{L}$ is differentiable and bounded below by a finite constant $\mathcal{L}_{\inf} \in \mathbb{R}$, and that there exists a gradient representation $\mathcal{I}$ such that $\mathcal{I}(\mu_\theta^\star, \theta) = \nabla \mathcal{L}(\theta)$.

Above, for a given θ, the outer objective F is evaluated at the inner equilibrium. Under the convergence assumptions of section 3, this equilibrium is the infinite-depth limit of the attention flow with fixed parameters D_θ, shared across depth. However, at finite depth, the token distribution generally differs from this equilibrium. We therefore study coupled dynamics that update the transformer parameters while the token distribution is still evolving:

$$\begin{aligned} \dot{\theta}_t &= -\eta_t \mathcal{I}(\mu_t, \theta_t), \\ \partial_t \mu_t &= \mathrm{div}_S \left(\mu_t \frac{\nabla_S m_{D_{\theta_t},\mu_t}}{m_{D_{\theta_t},\mu_t}} \right), \qquad \eta_t = \eta_0 (1+t)^{-\delta}, \quad 0 \le \delta < 1. \end{aligned} \tag{11}$$

with initial conditions $\mu_0 \in \mathcal{P}(S)$ and $\theta_0 \in \Theta$. Here, μ_t is the token distribution at finite depth t along the coupled dynamics, in which the parameters also evolve. Here η_t sets the time scale of the parameter dynamics relative to the inner flow, while δ controls how this relative time scale decays over time.

For the attention model, define the inner energy gap at parameter θ by

$$h(\mu, \theta) := E_{D_\theta}(\mu) - E_{D_\theta}(\mu_\theta^\star) = e_{D_\theta}(\mu), \qquad h_t := h(\mu_t, \theta_t) \ge 0. \tag{12}$$

Thus h_t measures the excess inner energy relative to the equilibrium at the current parameter. The following assumption relates the decrease of the inner energy gap at fixed parameters to two sources of error during training: approximating the outer gradient using μ_t, and changing the parameters before the token distribution reaches equilibrium.

Assumption 4.1. *Let $\mathcal{A} \subset \mathcal{P}(S) \times \Theta$ be a set of pairs (μ, θ) containing the coupled trajectory defined by* (11)*. The gap h is differentiable in θ and obeys the chain rule along regular coupled curves in $\mathcal{A}$. There exist constants $a > 0$, $C_{\mathcal{I}}, L \geq 0$, and $\beta \in [1, 2]$ such that, for every $(\mu, \theta) \in \mathcal{A}$,*

$$-\frac{\mathrm{d}}{\mathrm{d}s} h(\mu_s, \theta) \geq a\, h(\mu_s, \theta)^{\beta}, \tag{13}$$

where μ_s is the inner flow started from μ with θ fixed, for all s such that $(\mu_s, \theta) \in \mathcal{A}$. Moreover,

$$\begin{aligned} \|\mathcal{I}(\mu, \theta) - \mathcal{I}(\mu^{\star}_{\theta}, \theta)\| &\leq C_{\mathcal{I}} \sqrt{h(\mu, \theta)}, \\ \|\partial_{\theta} h(\mu, \theta)\| &\leq L \sqrt{h(\mu, \theta)}. \end{aligned} \tag{14}$$

The bounds are stated uniformly on $\mathcal{A}$, although the proof only uses them along the coupled trajectory. For $\beta = 2$, (13) is precisely the quadratic dissipation estimate (9). If the assumptions of Section 3 hold uniformly on $\mathcal{A}$, with $c_0 \geq \underline{c} > 0$, $B \leq \overline{B}$, and $\|D_{\theta}\|_{\mathrm{op}} \leq \overline{\Lambda}$, we can take $a = \frac{\underline{c}\, e^{-\overline{\Lambda}}}{2\overline{B}} > 0$. Section 3 establishes the case $\beta = 2$. We also consider $\beta \in [1, 2]$ to show how the convergence rate for parameter training depends on the dissipation inequality satisfied by the token dynamics. The two bounds in (14), which control the outer-gradient error and the sensitivity of h to θ, require separate verification.

The two approximation bounds lead to a closed pair of differential inequalities:

$$\dot{\mathcal{L}}(\theta_t) \leq -\tfrac{1}{2}\eta_t \|\mathcal{I}(\mu^{\star}_{\theta_t}, \theta_t)\|^2 + \tfrac{1}{2} C_{\mathcal{I}}^2 \eta_t h_t, \tag{15}$$

$$\dot{h}_t \leq -a h_t^{\beta} + (L C_{\mathcal{I}} + L^2)\eta_t h_t + \tfrac{1}{4}\eta_t \|\mathcal{I}(\mu^{\star}_{\theta_t}, \theta_t)\|^2. \tag{16}$$

Consequently, $V_t = \mathcal{L}(\theta_t) - \mathcal{L}_{\inf} + h_t \geq 0$ satisfies

$$\dot{V}_t \leq -a h_t^{\beta} + A \eta_t h_t - \tfrac{1}{4}\eta_t \|\mathcal{I}(\mu^{\star}_{\theta_t}, \theta_t)\|^2, \qquad A = \tfrac{1}{2} C_{\mathcal{I}}^2 + L C_{\mathcal{I}} + L^2. \tag{17}$$

The competition between h_t^{β} and $\eta_t h_t$ determines the schedule. In the following theorem, we provide the convergence rates for the training dynamics according to the latter, as well as the parameter balancing the speed of outer and inner loops respectively.

Theorem 4.2 (Outer stationarity)**.** *Under Assumption 4.1, let $1 < \beta \leq 2$ and $p = \beta/(\beta - 1)$. For $T > 0$,*

$$\frac{1}{T}\int_0^T \|\nabla \mathcal{L}(\theta_t)\|^2 \,\mathrm{d}t \leq \frac{4}{T\eta_T}\left[V_0 + C_{\beta}\int_0^T \eta_t^p \,\mathrm{d}t\right], \tag{18}$$

where C_{β} depends only on a, A, β. If $0 < \delta < 1$, the resulting rates are

$$\frac{1}{T}\int_0^T \|\nabla \mathcal{L}(\theta_t)\|^2 \,\mathrm{d}t = \begin{cases} O(T^{-\delta/(\beta-1)}), & \delta < (\beta-1)/\beta, \\ O(T^{-1/\beta}\log T), & \delta = (\beta-1)/\beta, \\ O(T^{-(1-\delta)}), & \delta > (\beta-1)/\beta. \end{cases} \tag{19}$$

For $\beta = 1$, $0 \leq \delta < 1$, and $A\eta_0 \leq a$, the rate is $O(T^{-(1-\delta)})$. In particular, $\delta = 0$ gives a constant learning rate and the rate $O(T^{-1})$.

A detailed proof of Theorem 4.2 is given in Appendix B. Inspired by the analysis of Marion et al. (2025), the proof uses a Lyapunov function V_t that combines the outer loss and the inner energy gap, here allowing polynomial inner dissipation h_t^{β}. Integrating the differential inequality (17) for V_t, after absorbing the coupling term into the inner dissipation, yields (18). The three rates in (19) then follow by evaluating the integral of $\eta_t^{\beta/(\beta-1)}$ for each learning-rate schedule. The balanced exponent $\delta = (\beta - 1)/\beta$ maximizes the polynomial decay certified by this estimate within the family (11).

In particular, if Assumption 4.1 holds with $\beta = 2$, the choice $\delta = 1/2$ gives

$$\eta_t = \eta_0 (1 + t)^{-1/2}, \qquad \frac{1}{T}\int_0^T \|\nabla \mathcal{L}(\theta_t)\|^2 \,\mathrm{d}t = O\left(\frac{\log T}{\sqrt{T}}\right). \tag{20}$$

Thus, under the stated assumptions, the time-averaged squared norm of the training-loss gradient tends to zero even though the token distribution need not reach equilibrium before the parameters are updated. The learning-rate schedule depends on the dissipation exponent of the inner flow.

5 NUMERICAL EXPERIMENTS

We first test the quadratic energy decay predicted by Proposition 3.3, and then examine the $\dot{H}^{-1}$ regularity condition in Assumption 3.2. We finally illustrate the convergence rate of the time-averaged squared training-loss gradient established in Theorem 4.2.

5.1 INNER ENERGY DECAY

We first study the attention flow (2) at fixed parameters, comparing energy decay for isotropic and anisotropic interactions.

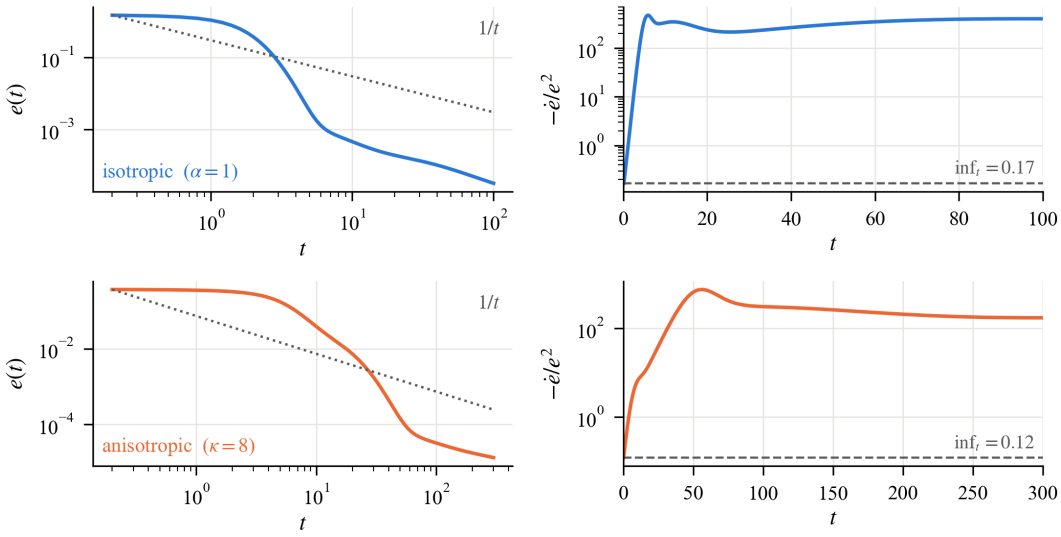

Figure 1: Energy decay along the minimizing flow (2). *Left:* energy gap $e(t)$ in log-log scale, compared with the $1/t$ rate predicted by (10). *Right:* the ratio $-\dot{e}/e^2$, whose lower bound is the quadratic dissipation constant in (9).

We integrate (2) with explicit Euler, using step size $\tau = 10^{-2}$ and re-projecting onto S after each step. Initial conditions are spherical caps around $(1, 0, \ldots, 0)$, with angular spread $\sigma_d = 0.15\sqrt{2/(d-1)}$ to keep the initial gap comparable across dimensions. We consider the isotropic case $D = \alpha \mathrm{Id}$, with $\alpha = 1$, and the anisotropic case $D = \mathrm{diag}(\kappa^{-1/2}, 1, \ldots, 1, \kappa^{1/2})$, with $\kappa = 8$. We use $d = 10$ and $N = 500$.

To compute the energy gap $e(t) = E_D(\mu_t) - E_D(\mu_D^\star)$, we need the equilibrium energy. For isotropic D, the equilibrium is the uniform measure ν, so $E_D(\nu)$ is known explicitly. For anisotropic D, the equilibrium is not known explicitly, and we approximate its energy using a longer reference run; details are given in Appendix C.

Under its assumptions, Proposition 3.3 gives $-\dot{e}(t) \geq \kappa_{\mathrm{in}} e(t)^2$, and hence $e(t) = O(t^{-1})$. We examine whether the finite-token dynamics exhibit similar behavior. The left panels compare the observed energy decay with a $1/t$ reference, while the right panels plot $-\dot{e}/e^2$.

At the beginning of the flow, the energy gap decays slowly and the ratio $-\dot{e}/e^2$ is smallest. This initial phase starts from a concentrated configuration, with tokens occupying only a small region of the sphere. As repulsion spreads the tokens across the sphere, the energy gap decays faster than $1/t$. At later times, the decay gradually approaches the predicted $1/t$ scaling, while $-\dot{e}/e^2$ remains bounded away from zero and becomes approximately constant over the observed time interval. This is consistent with the quadratic dissipation underlying the $O(t^{-1})$ estimate. Anisotropy changes the relaxation time and the observed constant, but not this overall behavior. The same qualitative behavior persists up to $d = 100$ and down to $N = 25$ (Appendix C).

5.2 Homogeneous Sobolev discrepancy along the inner flow

The previous experiment examined the energy decay predicted by Proposition 3.3. We now turn to one of its assumptions: the uniform bound on $\|\mu_t - \mu_D^\star\|^2_{\dot{H}^{-1}(\nu)}$ in Assumption 3.2.

On $\mathbb{S}^2$, this squared norm has the spectral representation

$$\|\mu - \mu^\star\|^2_{\dot{H}^{-1}(\nu)} = \sum_{k=1}^{\infty} \frac{\|c_k(\mu) - c_k(\mu^\star)\|_2^2}{k(k+1)}, \quad \text{where } c_k(\mu) = \left(\int_{\mathbb{S}^2} Y_{k,j}(x)\, \mathrm{d}\mu(x) \right)_{j=1}^{2k+1}$$

collects the degree-k spherical-harmonic coefficients of μ. The spherical harmonics form an orthonormal basis of $L^2(\nu)$, and those of degree k are eigenfunctions of $-\Delta_S$ with eigenvalue $k(k+1)$. The degree-zero term vanishes because both measures have unit mass.

For numerical evaluation, we consider an approximation of the previous quantity with a cutoff M:

$$H_M(\mu, \mu^\star) = \sum_{k=1}^{M} \frac{\|c_k(\mu) - c_k(\mu^\star)\|_2^2}{k(k+1)}.$$

Thus, H_M increases to the squared Sobolev norm as $M \to \infty$. Appendix D derives this representation and the particle estimator used below.

We use $N = 400$ particles on $\mathbb{S}^2$, time step $\tau = 0.01$, and $D = \mathrm{diag}(\kappa^{-1/2}, 1, \kappa^{1/2})$, $\kappa \in \{1, 2, 4, 8\}$. For each κ, we perform 20 independent runs until $T = 100$, using the same spherical-cap initialization as in the previous subsection, with $\sigma_3 = 0.15$. We approximate the reference equilibrium by pooling three independent reference simulations of length $T_{\mathrm{ref}} = 200$. We evaluate the discrepancy for $M \in \{5, 10, 20, 40\}$. The shaded bands show 95% bootstrap intervals at $M = 40$, computed across the independent runs.

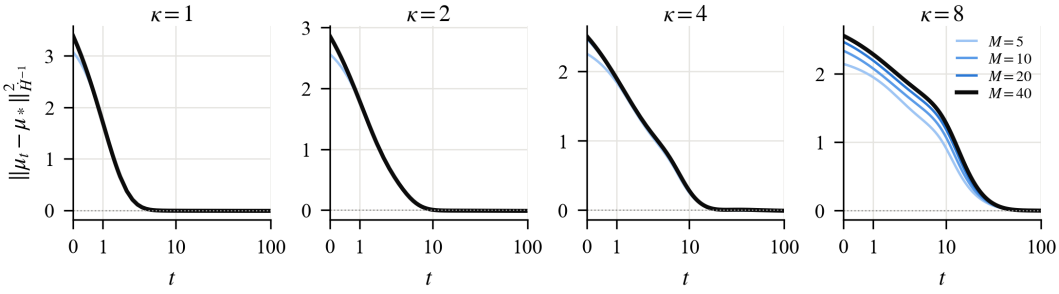

Figure 2: Estimated truncated homogeneous $\dot{H}^{-1}$ discrepancy $H_M(t)$ for $\kappa \in \{1, 2, 4, 8\}$ and harmonic cutoffs $M \in \{5, 10, 20, 40\}$, on a $\log(1 + t)$ time axis. The shaded band is the 95% bootstrap interval across the 20 independent runs at $M = 40$.

Figure 2 shows that the estimated truncated discrepancy decreases toward zero for all tested anisotropies, as the flow converges to equilibrium. Differences between cutoffs are most visible at early times, particularly for $\kappa = 8$, and become small as the flow evolves. Increasing anisotropy slows this decay. The decrease of the estimated discrepancy and its similar behavior across the tested cutoffs provide numerical support for the boundedness assumption along the flow, although they do not establish a uniform bound on the full Sobolev norm.

5.3 Coupled inner–outer training dynamics

We now allow the interaction parameters to evolve alongside the tokens and examine how the learning-rate schedule affects the time-averaged squared training-loss gradient. We consider two training objectives for which Appendix E verifies the coupling estimates required by Theorem 4.2. If the inner-flow assumptions of Section 3 hold uniformly along the coupled dynamics, these examples satisfy the theorem's hypotheses with $\beta = 2$.

In the isotropic example, we train the scale $D_\alpha = \alpha \mathrm{Id}$ using the objective

$$z_\alpha(\mu) = \int_S e^{\alpha b^\top x}\, \mathrm{d}\mu(x), \qquad F(\mu, \alpha) = \frac{1}{2}\big(z_\alpha(\mu) - y\big)^2,$$

where b is a fixed vector and y is a prescribed scalar target. The training loss is $\mathcal{L}(\alpha) = F(\nu, \alpha)$. Its explicit dependence on α is necessary: since the equilibrium is ν for every α, a loss depending only on the equilibrium measure would be constant.

In the anisotropic example, we train the orientation $D_\theta = R_\theta D_0 R_\theta^\top$, where D_0 is a fixed positive diagonal matrix and R_θ rotates by angle θ in the (e_1, e_d) plane. We use the objective

$$M_\mu = \int_S xx^\top \, \mathrm{d}\mu(x), \qquad q_W(\mu) = \mathrm{Tr}(W M_\mu), \qquad F(\mu) = \frac{1}{2}\big(q_W(\mu) - y\big)^2,$$

where W is a fixed symmetric matrix and y is a prescribed scalar target. Here, the equilibrium rotates with the parameter, $\mu_\theta^\star = (R_\theta)_\# \mu_0^\star$, so the training loss $\mathcal{L}(\theta) = F(\mu_\theta^\star)$ depends on θ through the equilibrium measure.

We discretize the coupled dynamics (11) by explicit Euler. Writing θ for either trainable parameter, $\mu_k = N^{-1}\sum_{i=1}^N \delta_{x_i^k}$, and $t_k = k\tau$, one iteration is

$$\widetilde{x}_i^{k+1} = x_i^k - \tau P_{x_i^k} \sum_{j=1}^N A_{ij}^k D_{\theta_k} x_j^k, \qquad x_i^{k+1} = \frac{\widetilde{x}_i^{k+1}}{\|\widetilde{x}_i^{k+1}\|}, \qquad \theta_{k+1} = \theta_k - \tau\eta_{t_k}\mathcal{I}(\mu_k, \theta_k), \quad (21)$$

where A_{ij}^k is the softmax weight in (2) evaluated at (x^k, θ_k). The learning rate η_t controls the relative timescales of token evolution and parameter training. Smaller η_t lets the attention flow evolve over a greater continuous depth before the parameters change appreciably, allowing the token distribution to track its evolving equilibrium more closely.

We use $N = 128$ particles, step size $\tau = 0.03$, and $d \in \{3, 5, 10\}$. For $\beta = 2$, the theorem suggests the learning-rate schedule $\eta_t = \eta_0(1+t)^{-1/2}$. We compare this schedule with exponents $\delta = 1/4$ and $\delta = 3/4$ in $\eta_t = \eta_0(1+t)^{-\delta}$, corresponding to more slowly and more rapidly decreasing learning rates. We report the time-averaged squared gradient

$$G(T) = \frac{1}{T}\int_0^T \|\nabla\mathcal{L}(\theta_t)\|^2 \, \mathrm{d}t.$$

Curves show the median over five independent runs, with 20–80% percentile bands.

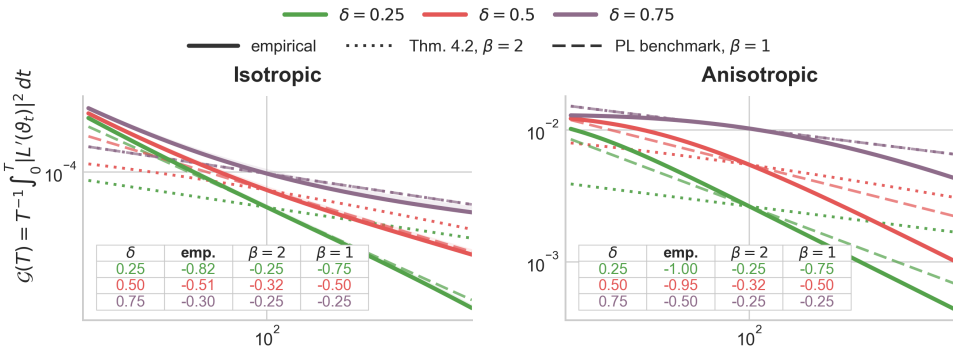

Figure 3: **Coupled parameter training at $d = 10$.** Stationarity metric $G(T)$ for isotropic kernel regression (left) and rotating anisotropy (right). Solid curves show empirical results. Dotted guides show the rates of Theorem 4.2 with $\beta = 2$, while dashed guides show the stronger $\beta = 1$ linear-dissipation benchmark. All guides are normalized at $T = 100$. The tables report fitted empirical and reference slopes.

For both objectives and all three learning-rate schedules, the observed decay of $G(T)$ over the simulated interval is at least as fast as the corresponding rate predicted by Theorem 4.2 with $\beta = 2$. For comparison, we also plot the rates obtained under the stronger, linear dissipation assumption $\beta = 1$, without claiming that this assumption holds for these dynamics. The observed slopes are close to these reference rates in the isotropic example and often steeper in the anisotropic one. All reference curves are normalized at $T = 100$, so the comparison concerns decay rates rather than multiplicative constants. These experiments therefore support the convergence-rate predictions for coupled token

and parameter dynamics. The faster decay observed in these examples also suggests that the general bounds can be conservative. Further experimental details and results for the other dimensions are reported in Appendix F.

6 Related work

We build on the mean-field and gradient-flow formulations of Geshkovski et al. (2025); Burger et al. (2025). Below, we discuss relevant work.

Clustering and other attention regimes. Without the spherical constraint, Geshkovski et al. (2023) relate clustering to the value matrix spectrum. For attractive spherical attention, Geshkovski et al. (2024) establish long-lived clusters at low temperature, while Bruno et al. (2025a) study initial pattern growth for unnormalized attention. Chen et al. (2025a) prove exponential convergence in W_2 to a single point for unnormalized attractive attention under suitable initialization or low-temperature assumptions. In joint large-token and low-temperature regimes, Bruno et al. (2025b) study concentration onto subspaces followed by spreading or clustering, while Alcalde et al. (2026) bound concentration for normalized spherical attention under matrix and initial-density assumptions. Peletier & Shalova (2025) derive a porous-medium limit for localized repulsion without softmax normalization. For random layers, Fedorov et al. (2026) derive stochastic spherical dynamics; Agazzi et al. (2026) include MLP blocks and prove synchronization when shared noise dominates attention drift. Beyond the sphere, Castin et al. (2025) establish well-posedness and mean-field limits, including multi-head and masked models. Boufadène & Vialard (2025a) study the cost and approximation power of sliced ReLU attention.

Convergence of kernel-discrepancy flows. Our work also connects to the literature on maximum mean discrepancy (MMD) flows, which evolve a distribution to reduce its discrepancy from a target. Our analysis shares similarities with that of Arbel et al. (2019), who prove an $O(t^{-1})$ energy bound for standard Wasserstein MMD flow under a bounded weighted H^{-1} discrepancy. For standard Wasserstein Coulomb flow on compact manifolds without boundary, Boufadène & Vialard (2025b) prove persistent density bounds and exponential convergence when initial and target densities are positive with bounded, Hölder-continuous logarithms. Chizat et al. (2026) treat less regular densities for Coulomb flow on the torus and prove local polynomial convergence for smoother Sobolev/Riesz kernels under positivity and Sobolev regularity assumptions.

Training attention parameters. In an infinite-depth, infinite-head limit, Barboni et al. (2026) establish well-posedness of the parameter gradient flow and exponential loss decay under small initial loss and positivity of the value-parameter neural tangent kernel at initialization. Their output is evaluated at finite rescaled depth with depth-dependent parameters. Gao et al. (2024) consider this limit with weight decay; their small-loss bound includes regularization and finite-size errors and assumes convergence, eventual boundedness, and conditions on the support of the limiting parameter distribution.

7 Conclusion

We established an $O(t^{-1})$ energy-gap bound for repulsive mean-field attention under uniform regularity assumptions. We then showed how the dissipation rate of the token dynamics guides the learning-rate schedule for simultaneous parameter training. In particular, under the stated coupling assumptions, quadratic dissipation yields an $O(\log T/\sqrt{T})$ bound on the time-averaged squared training-loss gradient. The experiments support the predicted decay rates and provide numerical evidence for the assumed Sobolev bound. Together, these results connect convergence of the attention flow with the timescale of parameter learning in the bilevel model.

Limitations. The model uses continuous depth, infinitely many tokens, and shared parameters with $V = -D$; arbitrary value matrices and full multi-head transformers are outside its scope. Establishing the assumed uniform trajectory bounds for general positive-definite interactions remains open. Convergence guarantees for a finite number of particles also remain to be established.

AI USE STATEMENT

Generative AI tools assisted with parts of the proofs, experimental design, and code development. ChatGPT and Codex were also used for literature searches, drafting and revising text, mathematical exposition and proof checking, and preparation of the LaTeX manuscript. The authors are responsible for the final text, mathematical claims, references, code, and experimental results.

ETHICS STATEMENT

This work studies an idealized mathematical model of self-attention. The reported experiments use synthetic token configurations and do not involve human participants or personal data. The results describe convergence in this model; they do not establish the safety, fairness, or reliability of deployed transformer systems.

REPRODUCIBILITY STATEMENT

The assumptions and convergence arguments are given in Sections 3 and 4, with supporting proofs in Appendices A and B. Appendix E verifies the coupling bounds for the two training objectives used in the experiments. Section 5 and Appendices C and F describe the numerical settings, while Appendix D defines the truncated Sobolev estimator.

REFERENCES

Andrea Agazzi, Giuseppe Bruno, Eloy Mosig García, Samuele Saviozzi, and Marco Romito. Stochastic scaling limits and synchronization by noise in deep transformer models. *arXiv preprint arXiv:2604.26898*, 2026. URL https://arxiv.org/abs/2604.26898.

Albert Alcalde, Leon Bungert, Konstantin Riedl, and Tim Roith. Quantifying concentration phenomena of mean-field transformers in the low-temperature regime. *arXiv preprint arXiv:2605.10931*, 2026. URL https://arxiv.org/abs/2605.10931.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.04370.

Raphaël Barboni, Maarten V. de Hoop, Takashi Furuya, and Gabriel Peyré. Training infinitely deep and wide transformers. *arXiv preprint arXiv:2605.17660*, 2026. URL https://arxiv.org/abs/2605.17660.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002.

Siwan Boufadène and François-Xavier Vialard. Sliced ReLU attention: Quasi-linear contextual expressivity via sorting. *arXiv preprint arXiv:2512.11411*, 2025a. URL https://arxiv.org/abs/2512.11411.

Siwan Boufadène and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4):4556–4587, 2025b. doi: 10.1137/24M1647631. URL https://arxiv.org/abs/2312.00800.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.23228.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://arxiv.org/abs/2509.25040.

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. *Philosophical Transactions of the Royal Society A*, 383(2298):20240233, 2025. doi: 10.1098/rsta.2024.0233. URL https://arxiv.org/abs/2501.03096.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025. URL https://arxiv.org/abs/2501.18322.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025a. URL https://arxiv.org/abs/2504.14697.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025b.

Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative convergence of Wasserstein gradient flows of kernel mean discrepancies. *arXiv preprint arXiv:2603.01977*, 2026. URL https://arxiv.org/abs/2603.01977.

Lev Fedorov, Michaël E. Sander, Romuald Elie, Pierre Marion, and Mathieu Laurière. Clustering in deep stochastic transformers. *arXiv preprint arXiv:2601.21942*, 2026. URL https://arxiv.org/abs/2601.21942.

Cheng Gao, Yuan Cao, Zihao Li, Yihan He, Mengdi Wang, Han Liu, Jason Klusowski, and Jianqing Fan. Global convergence in training large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29213–29284, 2024. URL https://arxiv.org/abs/2410.23610.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57026–57037, 2023. URL https://arxiv.org/abs/2305.05465.

Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024. URL https://arxiv.org/abs/2410.06833.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62:427–479, 2025. URL https://arxiv.org/abs/2312.10794.

Pierre Marion, Anna Korba, Peter Bartlett, Mathieu Blondel, Valentin De Bortoli, Arnaud Doucet, Felipe Llinares-López, Courtney Paquette, and Quentin Berthet. Implicit diffusion: Efficient optimization through stochastic sampling. In *International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1999–2007, 2025. URL https://proceedings.mlr.press/v258/marion25a.html.

C. Müller. *Analysis of Spherical Symmetries in Euclidean Spaces*. Springer, 1998.

Mark A. Peletier and Anna Shalova. Nonlinear diffusion limit of non-local interactions on a sphere. *arXiv preprint arXiv:2512.03185*, 2025. URL https://arxiv.org/abs/2512.03185.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v151/sander22a.html.

A EQUILIBRIUM IDENTITIES FOR SECTION 3

As recalled in Section 2, continuity of E_D and compactness of $\mathcal{P}(S)$ give existence of a minimizer, which is unique for $D \succ 0$ (Burger et al., 2025, Lemma 3.6). We denote it by $\mu_D^\star$.

Proof of Lemma 3.1. Write $\mu^\star = \mu_D^\star$. For any $x \in S$, consider $\mu_s = (1-s)\mu^\star + s\delta_x$, $s \in [0,1]$. Since $\mu^\star$ minimizes E_D,

$$0 \le \frac{\mathrm{d}}{\mathrm{d}s} E_D(\mu_s)\Big|_{s=0^+} = 2\big(m_{D,\mu^\star}(x) - E_D(\mu^\star)\big).$$

Thus $m_{D,\mu^\star} \ge E_D(\mu^\star)$ on S. Moreover,

$$\int_S \big(m_{D,\mu^\star} - E_D(\mu^\star)\big)\,\mathrm{d}\mu^\star = 0.$$

The integrand is nonnegative, hence equality holds $\mu^\star$-almost everywhere. By continuity of $m_{D,\mu^\star}$, it holds on $\mathrm{supp}(\mu^\star)$. □

For $\sigma := \mu - \mu_D^\star$, symmetry and positive definiteness of K_D give

$$\begin{aligned} e_D(\mu) &= 2\int_S m_{D,\mu}\,\mathrm{d}\sigma - \iint_{S\times S} K_D(x,y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y) \\ &\le 2\int_S m_{D,\mu}\,\mathrm{d}\sigma. \end{aligned}$$

This proves (5) without a support assumption.

Assume now that $m_{D,\mu_D^\star}$ is constant on S. Since μ and $\mu_D^\star$ have unit mass,

$$\int_S m_{D,\mu_D^\star}\,\mathrm{d}\mu = E_D(\mu_D^\star).$$

It follows that

$$\mathrm{MMD}^2_{K_D}(\mu,\mu_D^\star) = E_D(\mu) + E_D(\mu_D^\star) - 2\int_S m_{D,\mu_D^\star}\,\mathrm{d}\mu = e_D(\mu).$$

Moreover, symmetry of K_D gives $\int_S m_{D,\mu}\,\mathrm{d}\mu_D^\star = \int_S m_{D,\mu_D^\star}\,\mathrm{d}\mu$, and therefore

$$e_D(\mu) = \int_S m_{D,\mu}\,\mathrm{d}(\mu - \mu_D^\star). \tag{22}$$

B PROOF OF THEOREM 4.2

B.1 COUPLED LOSS AND LAG INEQUALITIES

Set $q_t := \mathcal{I}(\mu_t,\theta_t) - \mathcal{I}(\mu_{\theta_t}^\star,\theta_t)$, so that $\|q_t\| \le C_{\mathcal{I}}\sqrt{h_t}$. Then

$$\begin{aligned} \dot{\mathcal{L}}(\theta_t) &= \langle \nabla\mathcal{L}(\theta_t), \dot\theta_t\rangle \\ &= -\eta_t \langle \mathcal{I}(\mu_{\theta_t}^\star,\theta_t), \mathcal{I}(\mu_{\theta_t}^\star,\theta_t) + q_t\rangle \\ &= -\eta_t \|\mathcal{I}(\mu_{\theta_t}^\star,\theta_t)\|^2 - \eta_t\langle \mathcal{I}(\mu_{\theta_t}^\star,\theta_t), q_t\rangle \\ &\le -\frac{\eta_t}{2}\|\mathcal{I}(\mu_{\theta_t}^\star,\theta_t)\|^2 + \frac{\eta_t}{2}\|q_t\|^2 \\ &\le -\frac{\eta_t}{2}\|\mathcal{I}(\mu_{\theta_t}^\star,\theta_t)\|^2 + \frac{C_{\mathcal{I}}^2\eta_t}{2}h_t. \end{aligned} \tag{23}$$

Here the first two equalities use the chain rule, the outer dynamics (11), and $\nabla\mathcal{L}(\theta_t) = \mathcal{I}(\mu_{\theta_t}^\star,\theta_t)$. The penultimate line uses Young's inequality, and the last uses (14).

For the inner gap, the same chain rule gives

$$\begin{aligned}\dot{h}_t &\le -a h_t^\beta + \langle \partial_\theta h(\mu_t, \theta_t), \dot{\theta}_t \rangle \\ &= -a h_t^\beta - \eta_t \langle \partial_\theta h(\mu_t, \theta_t), \mathcal{I}(\mu^\star_{\theta_t}, \theta_t) + q_t \rangle \\ &\le -a h_t^\beta + \eta_t L \sqrt{h_t} \big(\|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\| + \|q_t\| \big) \\ &\le -a h_t^\beta + (L^2 + L C_{\mathcal{I}}) \eta_t h_t + \frac{\eta_t}{4} \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2. \end{aligned} \tag{24}$$

The first line uses the fixed-parameter decay bound (13); the equality substitutes (11). The next line uses Cauchy–Schwarz and $\|\partial_\theta h\| \le L\sqrt{h_t}$. The last uses $L\sqrt{h_t}\, u \le L^2 h_t + u^2/4$ and $\|q_t\| \le C_{\mathcal{I}} \sqrt{h_t}$.

B.2 Construction of the joint Lyapunov function

With $V_t := \mathcal{L}(\theta_t) - \mathcal{L}_{\inf} + h_t \ge 0$ and $A := C_{\mathcal{I}}^2/2 + L C_{\mathcal{I}} + L^2$, we have

$$\begin{aligned}\dot{V}_t &= \dot{\mathcal{L}}(\theta_t) + \dot{h}_t \\ &\le -a h_t^\beta + \left(\frac{C_{\mathcal{I}}^2}{2} + L C_{\mathcal{I}} + L^2 \right) \eta_t h_t - \left(\frac{1}{2} - \frac{1}{4} \right) \eta_t \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2 \\ &= -a h_t^\beta + A \eta_t h_t - \frac{\eta_t}{4} \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2. \end{aligned} \tag{25}$$

The inequality adds (23) and (24). It remains to control the coupling term $A\eta_t h_t$ using the inner dissipation $a h_t^\beta$.

B.3 Absorption and integration

Case $\beta > 1$. Put $p = \beta/(\beta - 1)$ and

$$C_\beta = \frac{\beta - 1}{\beta} A^p \left(\frac{a\beta}{2} \right)^{-1/(\beta - 1)}.$$

Then

$$\begin{aligned}\dot{V}_t &\le -a h_t^\beta + A \eta_t h_t - \frac{\eta_t}{4} \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2 \\ &\le -\frac{a}{2} h_t^\beta + C_\beta \eta_t^p - \frac{\eta_t}{4} \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2 \\ &\le C_\beta \eta_t^p - \frac{\eta_t}{4} \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2. \end{aligned}$$

We used (25), then Young's inequality $Axy \le (a/2) x^\beta + C_\beta y^p$ with $x = h_t$, $y = \eta_t$, and finally $h_t \ge 0$. Integrating gives

$$\begin{aligned}\frac{\eta_T}{4} \int_0^T \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2 \, \mathrm{d}t &\le \frac{1}{4} \int_0^T \eta_t \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2 \, \mathrm{d}t \\ &\le V_0 - V_T + C_\beta \int_0^T \eta_t^p \, \mathrm{d}t \\ &\le V_0 + C_\beta \int_0^T \eta_t^p \, \mathrm{d}t. \end{aligned}$$

The first inequality uses $\eta_t \ge \eta_T$ on $[0, T]$, the second integrates the preceding bound, and the last uses $V_T \ge 0$. Dividing by $T\eta_T/4$ proves

$$\frac{1}{T} \int_0^T \|\mathcal{I}(\mu^\star_{\theta_t}, \theta_t)\|^2 \, \mathrm{d}t \le \frac{4}{T\eta_T} \left[V_0 + C_\beta \int_0^T \eta_t^p \, \mathrm{d}t \right]. \tag{26}$$

For $\eta_t = \eta_0(1+t)^{-\delta}$, direct integration yields

$$\int_0^T \eta_t^p \,\mathrm{d}t = \eta_0^p \int_0^T (1+t)^{-p\delta}\,\mathrm{d}t$$
$$= \eta_0^p \begin{cases} \big((1+T)^{1-p\delta}-1\big)/(1-p\delta), & p\delta \neq 1, \\ \log(1+T), & p\delta = 1. \end{cases}$$

Using $T\eta_T \asymp T^{1-\delta}$ in (26) therefore gives

$$\frac{1}{T}\int_0^T \|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2 \,\mathrm{d}t = \begin{cases} O(T^{-\delta/(\beta-1)}), & \delta < (\beta-1)/\beta, \\ O(T^{-1/\beta}\log T), & \delta = (\beta-1)/\beta, \\ O(T^{-(1-\delta)}), & \delta > (\beta-1)/\beta. \end{cases} \tag{27}$$

The three cases correspond to an integral growing as $T^{1-p\delta}$, growing logarithmically, or remaining bounded; in the first case, $\delta - p\delta = -\delta/(\beta-1)$.

Case $\beta = 1$. If $A\eta_0 \le a$, then

$$\dot{V}_t \le -(a - A\eta_t)h_t - \frac{\eta_t}{4}\|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2$$
$$\le -\frac{\eta_t}{4}\|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2,$$

where the last inequality uses $\eta_t \le \eta_0$ and $h_t \ge 0$. Consequently,

$$\frac{\eta_T}{4}\int_0^T \|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2 \,\mathrm{d}t \le \frac{1}{4}\int_0^T \eta_t \|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2 \,\mathrm{d}t$$
$$\le V_0 - V_T \le V_0,$$

using monotonicity of η_t, integration, and nonnegativity of V_T, respectively. Thus the averaged squared gradient is at most $4V_0/(T\eta_T) = O(T^{-(1-\delta)})$, including $O(T^{-1})$ for a constant learning rate.

Conclusion. Since

$$\min_{0\le t\le T} \|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2 \le \frac{1}{T}\int_0^T \|\mathcal{I}(\mu_{\theta_t}^\star, \theta_t)\|^2 \,\mathrm{d}t,$$

the same bounds control the smallest squared outer-gradient norm attained up to time T. The choice $\delta = (\beta-1)/\beta$ maximizes the polynomial decay certified by (27), giving $O(T^{-1/\beta}\log T)$. In particular, $\beta = 2$ gives $\eta_t = \eta_0(1+t)^{-1/2}$ and the rate $O(\log T/\sqrt{T})$. The lower bound on $\mathcal{L}$ is needed to ensure $V_T \ge 0$ in the integration step.

C Additional experiments

This appendix collects the dimension and particle-number sweeps behind Figure 1, in the same two-panel format: $e(t)$ in log-log scale with the $1/t$ reference, and the Łojasiewicz ratio $-\dot{e}/e^2$ in semilog scale. All runs use the discretisation of Section 5,explicit Euler on (2) with $\tau = 10^{-2}$ and reprojection, constant-angular cap initial condition, and $\mathcal{E}^\star$ from the tail of a decoupled run four times longer than the analysis window.

Figures 4 and 5 show that dimension changes the time scale of the flow but not its shape. The infimum of the ratio, i.e. the largest empirical κ_{in} for which (9) holds over the window, is essentially dimension-free: it lies between 0.163 and 0.169 for d from 3 to 100 isotropically, and between 0.09 and 0.19 at $\kappa = 8$. Every curve also exhibits the plateau discussed in Section 5, at a level that does depend on d.

Figures 6 and 7 separate what is intrinsic to the flow from what is an artefact of the discretisation. The curves coincide over the transient and the intermediate regime for every N, and differ only in the final approach, where the smaller systems leave the common curve earlier and settle at a larger gap. The certified constant is likewise insensitive to N: at $\alpha = 1$ the infimum of the ratio stays within $[0.12, 0.14]$ over $N \in \{25, \dots, 400\}$, and at $\kappa = 8$ within $[0.07, 0.09]$. The height of the late-time plateau, by contrast, grows with N, since it measures how close to equilibrium a finite system can get.

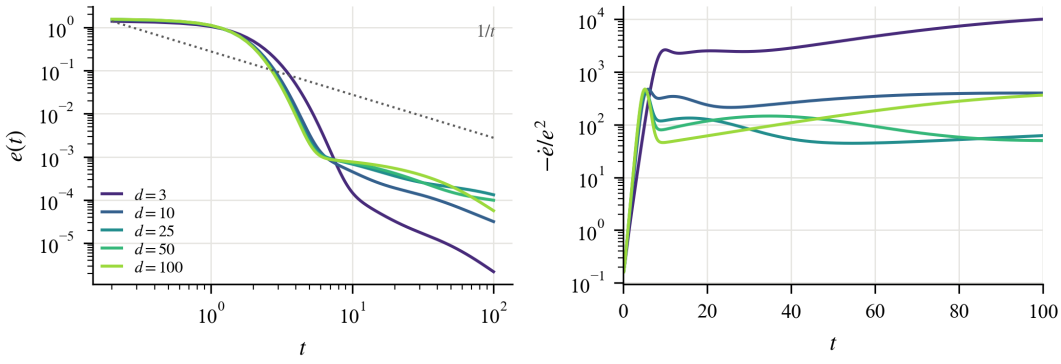

Figure 4: Isotropic interactions, $\alpha = 1$, for $d \in \{3, 10, 25, 50, 100\}$ at $N = 500$, analysis window $T = 100$.

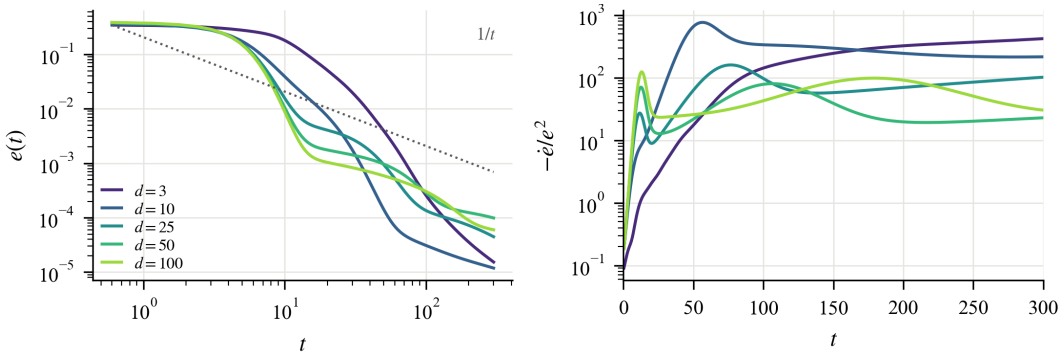

Figure 5: The same sweep for $D = \mathrm{diag}(\kappa^{-1/2}, 1, \ldots, 1, \kappa^{1/2})$ with $\kappa = 8$, $\det D = 1$; $T = 300$, three times longer than isotropically because the flow is slower.

D Estimating the truncated $\dot{H}^{-1}$ discrepancy with spherical harmonics (Section 5.2)

Spherical harmonics are used here only to evaluate the Sobolev diagnostic of Section 5.2. We follow the standard conventions of Müller (1998). Let $\{Y_{k,j}\}_{k\geq 0,\, 1\leq j\leq N(d,k)}$ be a real orthonormal spherical-harmonic basis of $L^2(S, \nu)$, where $N(d, k)$ is the dimension of the space of degree-k harmonics. These functions satisfy

$$-\Delta_S Y_{k,j} = \omega_k Y_{k,j}, \qquad \omega_k = k(k + d - 2).$$

For a finite signed measure σ with $\sigma(S) = 0$, write $c_{k,j}(\sigma) := \int_S Y_{k,j}\, \mathrm{d}\sigma$. Expanding the dual definition (6) in the spherical-harmonic basis gives

$$\|\sigma\|^2_{\dot{H}^{-1}(\nu)} = \sum_{k\geq 1} \frac{1}{\omega_k} \sum_{j=1}^{N(d,k)} |c_{k,j}(\sigma)|^2, \tag{28}$$

with the value $+\infty$ when the series diverges. The degree-zero term is absent because $\sigma(S) = 0$.

For the experiments, we truncate (28). On $\mathbb{S}^2$, for two probability measures μ and $\mu^\star$, we define

$$H_M(\mu, \mu^\star) := \sum_{k=1}^{M} \frac{\|c_k(\mu) - c_k(\mu^\star)\|_2^2}{k(k + 1)}, \tag{29}$$

where $c_k(\mu) = (c_{k,j}(\mu))_{j=1}^{2k+1}$. For fixed μ and $\mu^\star$, H_M is nondecreasing in M and converges to $\|\mu - \mu^\star\|^2_{\dot{H}^{-1}(\nu)}$, possibly with infinite limit.

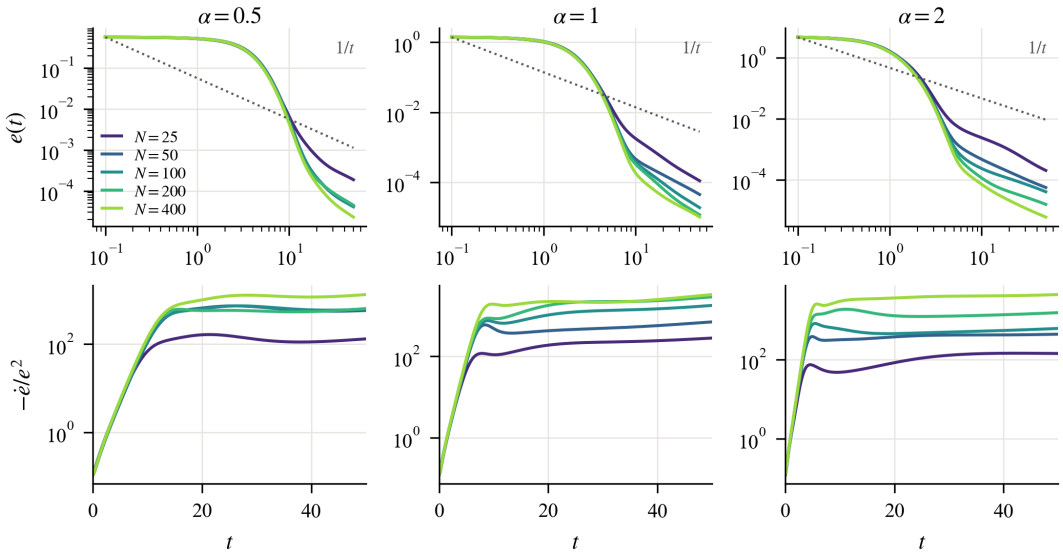

Figure 6: Finite-size effect, isotropic case: $N \in \{25, 50, 100, 200, 400\}$ overlaid, one column per $\alpha \in \{0.5, 1, 2\}$, $d = 3$, $T = 50$.

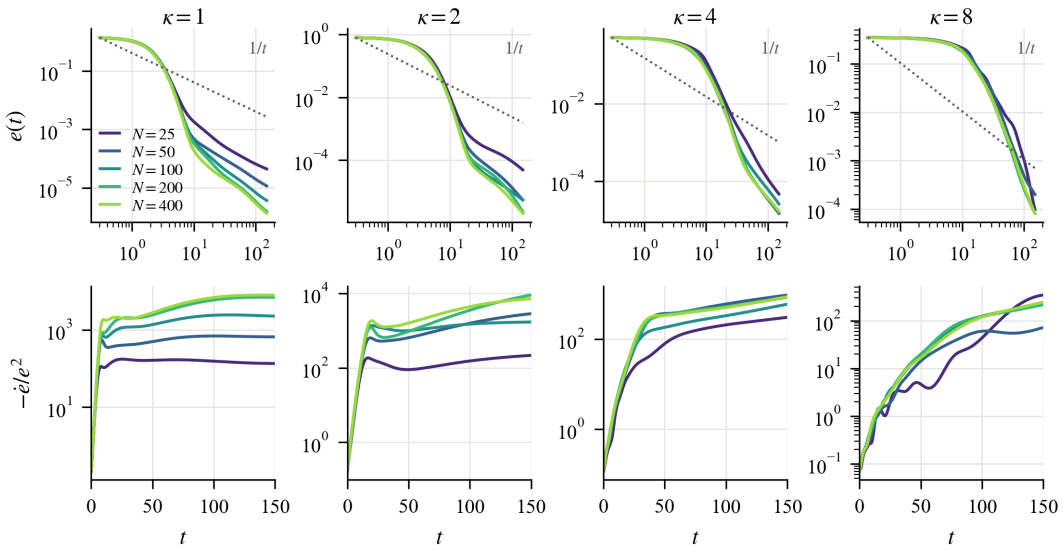

Figure 7: The same particle-number sweep for anisotropic D, one column per $\kappa \in \{1, 2, 4, 8\}$, $d = 3$, $T = 150$.

To evaluate (29) without computing the individual harmonic coefficients, we use the spherical-harmonic addition theorem (Müller, 1998, Thm. 2, §2). With harmonics orthonormal in $L^2(S, \nu)$,

$$\sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(y) = N(d,k) P_k(x \cdot y), \tag{30}$$

where P_k is the normalized degree-k zonal polynomial, $P_k(1) = 1$. On $\mathbb{S}^2$, $N(3,k) = 2k+1$ and P_k is the usual Legendre polynomial. Hence, defining

$$\mathcal{K}_M(x,y) := \sum_{k=1}^{M} \frac{2k+1}{k(k+1)} P_k(x \cdot y), \tag{31}$$

we obtain

$$H_M(\mu,\mu^\star) = \mathbb{E}_{X,X'\sim\mu}[\mathcal{K}_M(X,X')] + \mathbb{E}_{Y,Y'\sim\mu^\star}[\mathcal{K}_M(Y,Y')] - 2\mathbb{E}_{X\sim\mu,\,Y\sim\mu^\star}[\mathcal{K}_M(X,Y)], \tag{32}$$

where all expectations are taken over independent draws.

Given particles $x_1,\dots,x_N$ from an analysis run and reference particles $y_1,\dots,y_{N_{\mathrm{ref}}}$, we estimate the two self-interaction terms by their off-diagonal U-statistics and the cross term by the empirical average:

$$\widehat{H}_M = \frac{1}{N(N-1)}\sum_{i\neq i'}\mathcal{K}_M(x_i,x_{i'}) + \frac{1}{N_{\mathrm{ref}}(N_{\mathrm{ref}}-1)}\sum_{\ell\neq\ell'}\mathcal{K}_M(y_\ell,y_{\ell'}) - \frac{2}{NN_{\mathrm{ref}}}\sum_{i,\ell}\mathcal{K}_M(x_i,y_\ell). \tag{33}$$

Removing the diagonal self-pairs avoids the finite-sample self-interaction bias. As a consequence, $\widehat{H}_M$ need not be nonnegative for every finite sample, even though the population quantity H_M is nonnegative. Since particles within a run interact, the independent statistical units in Section 5.2 are the simulation runs rather than the individual particles.

E Verification of the assumptions of Theorem 4.2 for the transformer architectures

We give two training objectives for which the coupling bounds (14) can be verified directly from the exponential kernel. In both cases, the frozen-parameter estimate of Section 3 gives $\beta = 2$, provided the assumptions of that section hold uniformly along the coupled trajectory.

We use the following tensor expansion of the exponential kernel. For any finite signed measure σ on S,

$$\iint_{S\times S} K_D(x,y)\,\mathrm{d}\sigma(x)\mathrm{d}\sigma(y) = \sum_{n=0}^{\infty}\frac{1}{n!}\left\|\int_S (D^{1/2}x)^{\otimes n}\,\mathrm{d}\sigma(x)\right\|^2, \tag{34}$$

where $\|\cdot\|$ denotes the canonical Euclidean norm on tensor powers. This follows directly by expanding $e^{x^\top Dy}$ into its power series.

Isotropic kernel regression. Let $D_\alpha = \alpha\mathrm{Id}$, with $\alpha \in [\alpha_{\min},\alpha_{\max}] \subset (0,\infty)$. Since $\mu^\star_\alpha = \nu$ for every α, a loss depending only on the equilibrium measure would be independent of α. We therefore introduce an explicit parameter dependence. Fix $b \in S$ and $y \in \mathbb{R}$, and define

$$z_\alpha(\mu) := \int_S e^{\alpha b^\top x}\,\mathrm{d}\mu(x), \qquad F(\mu,\alpha) := \frac{1}{2}\big(z_\alpha(\mu)-y\big)^2.$$

The outer objective and gradient representation are

$$\mathcal{L}(\alpha) := F(\nu,\alpha), \qquad \mathcal{I}(\mu,\alpha) := \partial_\alpha F(\mu,\alpha),$$

so that $\mathcal{I}(\nu,\alpha) = \partial_\alpha\mathcal{L}(\alpha)$.

By the constant-potential identity in Appendix A,

$$h(\mu,\alpha) = E_{D_\alpha}(\mu) - E_{D_\alpha}(\nu) = \mathrm{MMD}^2_{K_\alpha}(\mu,\nu), \qquad K_\alpha(x,y) = e^{\alpha x^\top y}.$$

Let $\mathcal{H}_\alpha$ denote the RKHS of K_α. By the reproducing property,

$$|z_\alpha(\mu) - z_\alpha(\nu)| \le e^{\alpha/2}\sqrt{h(\mu,\alpha)}.$$

Moreover,

$$\|\partial_\alpha K_\alpha(b,\cdot)\|^2_{\mathcal{H}_\alpha} = \sum_{n\ge1}\frac{n^2\alpha^{n-2}}{n!} = e^\alpha(1+\alpha^{-1}),$$

and hence

$$|\partial_\alpha z_\alpha(\mu) - \partial_\alpha z_\alpha(\nu)| \le e^{\alpha/2}\sqrt{1+\alpha^{-1}}\,\sqrt{h(\mu,\alpha)}.$$

Since $z_\alpha(\mu)$ and $\partial_\alpha z_\alpha(\mu)$ are uniformly bounded for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, it follows that

$$|\mathcal{I}(\mu, \alpha) - \partial_\alpha \mathcal{L}(\alpha)| \le C_{\mathcal{I}} \sqrt{h(\mu, \alpha)}.$$

For the second coupling bound, define

$$T_n(\mu) := \int_S x^{\otimes n} \, \mathrm{d}(\mu - \nu)(x).$$

Since $\mu - \nu$ has zero mass, (34) gives

$$h(\mu, \alpha) = \sum_{n \ge 1} \frac{\alpha^n}{n!} \|T_n(\mu)\|^2.$$

Differentiating at fixed μ and applying Cauchy–Schwarz yields

$$|\partial_\alpha h(\mu, \alpha)| \le \sqrt{h(\mu, \alpha)} \left(\sum_{n \ge 1} \frac{n^2 \alpha^{n-2}}{n!} \|T_n(\mu)\|^2 \right)^{1/2}.$$

Since $\|T_n(\mu)\| \le 2$,

$$|\partial_\alpha h(\mu, \alpha)| \le 2 e^{\alpha_{\max}/2} \sqrt{1 + \alpha_{\min}^{-1}} \, \sqrt{h(\mu, \alpha)}.$$

Thus both bounds in (14) hold uniformly on $[\alpha_{\min}, \alpha_{\max}]$.

Rotating anisotropy and moment regression. Let θ belong to an interval $I \subset \mathbb{R}$, let $\theta \mapsto R_\theta \in \mathrm{SO}(d)$ be a C^1 rotation path, and set

$$D_\theta := R_\theta D_0 R_\theta^\top,$$

where $D_0 \succ 0$ is diagonal and anisotropic. The interaction energy is rotation equivariant:

$$E_{D_\theta}(\mu) = E_{D_0}\big((R_\theta^\top)_{\#}\mu\big).$$

Hence, by uniqueness of the minimizer,

$$\mu_\theta^\star = (R_\theta)_{\#} \mu_0^\star.$$

Full support and the regularity quantities of Assumption 3.2 are invariant under rotations. Thus the frozen inner estimate of Section 3 holds with the same constants along this family whenever it holds for D_0.

Define the second-moment matrix

$$M_\mu := \int_S x x^\top \, \mathrm{d}\mu(x), \qquad M_\theta^\star := M_{\mu_\theta^\star} = R_\theta M_0^\star R_\theta^\top.$$

For a fixed symmetric matrix W, let

$$q_W(\mu) := \mathrm{Tr}(W M_\mu), \qquad F(\mu) := \frac{1}{2}\big(q_W(\mu) - y\big)^2, \qquad \mathcal{L}(\theta) := F(\mu_\theta^\star).$$

Set

$$\Omega_\theta := (\partial_\theta R_\theta) R_\theta^\top.$$

Since R_θ is orthogonal, Ω_θ is skew-symmetric, and

$$\partial_\theta M_\theta^\star = \Omega_\theta M_\theta^\star - M_\theta^\star \Omega_\theta.$$

Therefore

$$\partial_\theta \mathcal{L}(\theta) = \big(q_W(\mu_\theta^\star) - y\big) \, \mathrm{Tr}\big(W\big(\Omega_\theta M_\theta^\star - M_\theta^\star \Omega_\theta\big)\big).$$

We take

$$\mathcal{I}(\mu, \theta) := \big(q_W(\mu) - y\big) \, \mathrm{Tr}\big(W\big(\Omega_\theta M_\mu - M_\mu \Omega_\theta\big)\big),$$

so that $\mathcal{I}(\mu_\theta^\star, \theta) = \partial_\theta \mathcal{L}(\theta)$.

Lemma 3.1 gives $\int_S m_{D_\theta,\mu_\theta^\star}\,\mathrm{d}\mu \geq E_{D_\theta}(\mu_\theta^\star)$, and hence

$$h(\mu,\theta) \geq \mathrm{MMD}^2_{K_{D_\theta}}(\mu,\mu_\theta^\star).$$

Applying (34) to $\sigma = \mu - \mu_\theta^\star$ and retaining only its degree-two contribution gives

$$h(\mu,\theta) \geq \frac{1}{2}\left\| D_\theta^{1/2}(M_\mu - M_\theta^\star)D_\theta^{1/2}\right\|_F^2.$$

If $d_{\min} := \lambda_{\min}(D_0)$, then

$$\|M_\mu - M_\theta^\star\|_F \leq \frac{\sqrt{2}}{d_{\min}}\sqrt{h(\mu,\theta)}. \tag{35}$$

Since $\|M_\mu\|_F \leq 1$ for probability measures on S, the map

$$M \longmapsto \big(\mathrm{Tr}(WM) - y\big)\,\mathrm{Tr}(W(\Omega_\theta M - M\Omega_\theta))$$

is uniformly Lipschitz on the relevant set whenever $\|\Omega_\theta\|_{\mathrm{op}}$ is uniformly bounded. Hence

$$|\mathcal{I}(\mu,\theta) - \partial_\theta\mathcal{L}(\theta)| \leq C_{\mathcal{I}}\sqrt{h(\mu,\theta)}.$$

For the second coupling bound, fix (μ,θ) and set $\Omega := \Omega_\theta$, $\Lambda := \|D_0\|_{\mathrm{op}}$. Consider the auxiliary rotation

$$D(s) := e^{s\Omega}D_\theta e^{-s\Omega}, \qquad f(s) := E_{D(s)}(\mu) - E_{D_0}(\mu_0^\star).$$

Rotational equivariance makes the minimum energy independent of s, so $f(s) \geq 0$ for every $s \in \mathbb{R}$. Moreover, $f(0) = h(\mu,\theta)$ and $f'(0) = \partial_\theta h(\mu,\theta)$, since $D'(0) = \Omega D_\theta - D_\theta\Omega = \partial_\theta D_\theta$. Differentiating under the integral gives

$$\begin{aligned}
|f''(s)| &= \left|\iint e^{x^\top D(s)y}\big((x^\top D'(s)y)^2 + x^\top D''(s)y\big)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y)\right| \\
&\leq e^{\Lambda}\big(\|D'(s)\|_{\mathrm{op}}^2 + \|D''(s)\|_{\mathrm{op}}\big) \\
&\leq 4\Lambda(1+\Lambda)e^{\Lambda}\|\Omega\|_{\mathrm{op}}^2 =: C_\theta.
\end{aligned}$$

The last inequality uses $\|D'(s)\|_{\mathrm{op}} \leq 2\Lambda\|\Omega\|_{\mathrm{op}}$ and $\|D''(s)\|_{\mathrm{op}} \leq 4\Lambda\|\Omega\|_{\mathrm{op}}^2$. If $C_\theta > 0$, Taylor's inequality at $s = -f'(0)/C_\theta$ gives

$$0 \leq f\left(-\frac{f'(0)}{C_\theta}\right) \leq f(0) - \frac{|f'(0)|^2}{2C_\theta},$$

and hence

$$|\partial_\theta h(\mu,\theta)| \leq 2\sqrt{2}\,e^{\Lambda/2}\sqrt{\Lambda(1+\Lambda)}\|\Omega_\theta\|_{\mathrm{op}}\sqrt{h(\mu,\theta)}.$$

If $C_\theta = 0$, then $\Omega_\theta = 0$ and the derivative vanishes. Thus both coupling bounds in (14) hold uniformly whenever $\|\Omega_\theta\|_{\mathrm{op}}$ is uniformly bounded on I, without assuming full support of the equilibrium.

F Details on the Coupled-training experiments in 5.3

Numerical scheme. We use the coupled explicit-Euler discretization described in Section 5.3. At $t_k = k\tau$, both the particle velocity and the outer gradient oracle are evaluated at (μ_k, θ_k) before applying the two updates. We then take one projected particle step for (2) and one parameter update

$$\theta_{k+1} = \theta_k - \tau\eta_{t_k}\mathcal{I}(\mu_k, \theta_k), \qquad \eta_t = \eta_0(1+t)^{-\delta}.$$

Thus each outer update is paired with a single inner-flow step; no inner loop is run to convergence between parameter updates. Unless otherwise stated, we use $\tau = 0.03$, $N = 128$, $T = 600$, $d \in \{3, 5, 10\}$, $\delta \in \{0.25, 0.5, 0.75\}$, and five independent random seeds.

Two training objectives. For the isotropic example, we use $D_\alpha = \alpha\mathrm{Id}$ and the scalar head

$$z_\alpha(\mu) = \int_S e^{\alpha b^\top x}\,\mathrm{d}\mu(x), \qquad F(\mu, \alpha) = \frac{1}{2}\big(z_\alpha(\mu) - y\big)^2,$$

as in Appendix E. The outer objective is $\mathcal{L}(\alpha) = F(\nu, \alpha)$. We initialize $\alpha_0 = 2.3$ and set

$$y = z_{\alpha_\star}(\nu), \qquad \alpha_\star = 0.9,$$

so that $\mathcal{L}(\alpha_\star) = 0$. Since $\mu^\star_\alpha = \nu$ is known explicitly, $\partial_\alpha\mathcal{L}(\alpha)$ is evaluated exactly.

For the rotating-anisotropy example, we take

$$D_\theta = R_\theta D_0 R_\theta^\top,$$

where D_0 has anisotropy $\kappa = 8$ and R_θ is the rotation of angle θ in the (e_1, e_d) plane. We use the second-moment head

$$q_W(\mu) = \mathrm{Tr}(WM_\mu), \qquad M_\mu = \int_S xx^\top\,\mathrm{d}\mu(x),$$

with outer loss

$$\mathcal{L}(\theta) = \frac{1}{2}\big(q_W(\mu^\star_\theta) - y\big)^2.$$

We initialize $\theta_0 = 0.4$ and choose

$$y = q_W(\mu^\star_{\theta_\star}), \qquad \theta_\star = 1.1,$$

so that the reference outer objective vanishes at $\theta_\star$. A reference approximation of $\mu^\star_0$ is obtained from a fixed-parameter simulation with 256 particles up to $T_{\mathrm{ref}} = 360$, and reference equilibria at other angles are obtained by rotation.

Stationarity metric. For every run we compute

$$G(T) = \frac{1}{T}\int_0^T \|\nabla\mathcal{L}(\theta_t)\|^2\,\mathrm{d}t.$$

Curves show the median over seeds, with 20–80% percentile bands. Empirical log–log slopes are fitted over $T \in [100, 600]$. The theoretical curves are normalized to the empirical curves at $T = 100$, so that the comparison concerns decay rates rather than multiplicative constants. Figure 8 reports the complete dimension sweep corresponding to Figure 3.

Tracking and robustness. To complement the stationarity metric, we monitor how accurately the inner particle distribution tracks the equilibrium associated with the current parameter. In the isotropic example this is measured by the exact energy gap $h(\mu_t, \alpha_t)$. In the anisotropic example we use the kernel MMD^2 relative to the rotated numerical equilibrium. We also repeat the balanced-schedule experiment for several initial token distributions to assess sensitivity to initialization.

Figure 9 shows decreasing inner tracking error across the tested dimensions and initializations. In the anisotropic experiment, the late-time plateau is consistent with the finite-particle approximation used for the reference equilibrium.

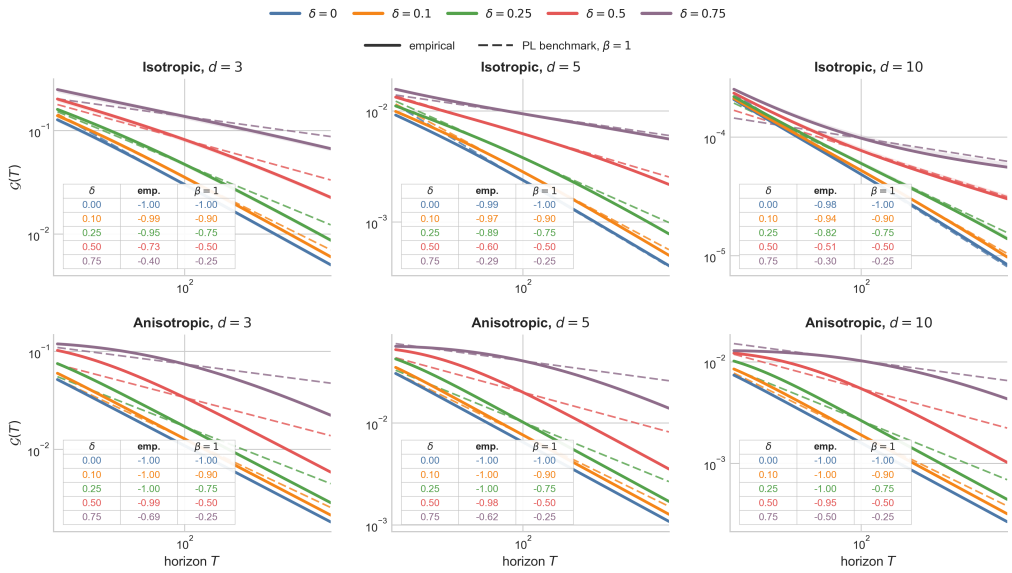

Figure 8: **Dimension sweep for coupled training.** Stationarity metric $G(T)$ for $d \in \{3, 5, 10\}$, for isotropic kernel regression (top) and rotating anisotropy (bottom). Solid curves show empirical results, while dashed guides show the $\beta = 1$ linear-dissipation rates of Theorem 4.2. These provide a stronger comparison benchmark than the $\beta = 2$ rates established from the general inner-flow estimate of Section 3.

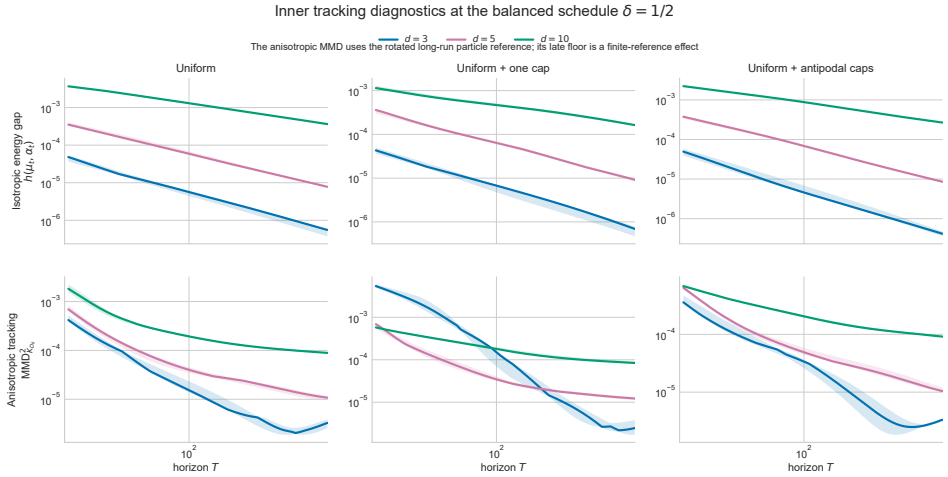

Figure 9: **Inner tracking at the balanced schedule** $\delta = 1/2$**.** Top: isotropic energy gap $h(\mu_t, \alpha_t)$. Bottom: anisotropic kernel MMD^2 relative to the rotated numerical equilibrium. Columns correspond to different initial token distributions and curves to $d \in \{3, 5, 10\}$. Bands show the 20–80% range across seeds.