Training-Induced Escape from Token Clustering in a Mean-Field Formulation of Transformers

Noboru Isobe[1] **Daisuke Inoue**[2] **Masaaki Imaizumi**[1,3,4]

[1]RIKEN [2]Imperial College London [3]The University of Tokyo [4]Kyoto University

noboru.isobe@riken.jp, dinoue@ic.ac.uk, imaizumi@g.ecc.u-tokyo.ac.jp

Abstract

Transformers perform inference by iteratively transforming token representations across layers. This layerwise computation has been studied empirically, and recent mean-field theories of Transformer dynamics explain how attention can drive token distributions toward clustering. However, existing analyses treat model parameters as prescribed, leaving open how training reshapes this clustering picture. We study this question in a noisy mean-field Transformer in which only a parameter-linear FFN is trained under L^2 regularization. We identify a training-induced phase of the inference dynamics across layers: after clustering in the early layers, the token distribution can leave the clustered regime near the final layers. A single gradient-descent update already produces this escape, and minimizers of the training objective satisfy an exponential turnpike bound under which clustering persists through the interior layers. Our analysis is based on an entropy-regularized interaction energy that captures the clustering bias of attention, for which we establish a double-well landscape and a local Polyak–Łojasiewicz inequality. More broadly, our results point toward a training-aware mean-field theory in which training and inference dynamics of Transformers are treated together.

1 Introduction

Modern language models built on the Transformer architecture (Vaswani et al. 2017) perform inference by iteratively transforming token representations across layers. This naturally suggests a dynamical viewpoint: along the depth axis, a Transformer defines a discrete-time dynamics of tokens. Understanding the internal dynamics of Transformers is important both for post-hoc explanation and for mitigating risks arising from their lack of transparency as discussed in (Zhao et al. 2024). Broad empirical literature has studied aspects of this layerwise computation and its mechanisms, including attention flow (Abnar and Zuidema 2020), mechanistic interpretability (Bereska and Gavves 2024), and causal analysis and model editing (Geiger et al. 2021; Meng et al. 2022). These approaches typically inspect particular trained models and their learned representations at a microscopic level.

Complementary to these empirical approaches, recent mathematical works in (Lu et al. 2020; Sander et al. 2022; Geshkovski et al. 2023) seek macroscopic descriptions of Transformer dynamics by replacing the collection of token representations with their empirical distribution and studying its evolution across layers. Within this line, model weights are typically prescribed rather than selected by training. For prototypical self-attention models, fixed scalar or symmetric weights yield a gradient-flow structure for an attention-induced interaction energy and explain token clustering (Burger et al. 2025; Geshkovski et al. 2025; Kuehn and Yoon 2026). Variants of this prescribed-parameter picture further analyze localization, synchronization, metastability, and multiscale transitions around clustering by adding feedforward blocks with fixed weights or initialization randomness (Geshkovski

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

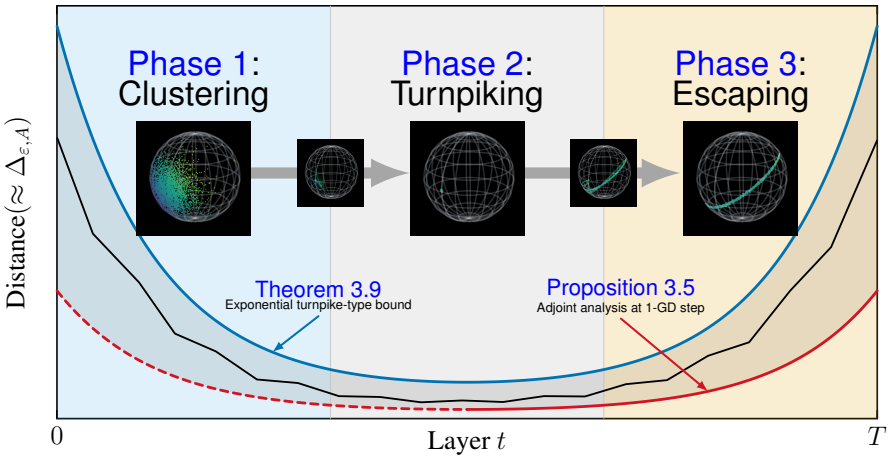


Figure 1: Overview of our results. The y-axis is an informal distance from the token distribution to a clustered state; in the analysis, the corresponding proxy is $\Delta_{\varepsilon,A}$ in Section 3. Training can induce a phase of escaping from the clustered regime. The convergence rate from Phases 1 to 2 and the escape rate from Phases 2 to 3 are approximately given in Remark 3.10 and Proposition 3.5, respectively.

et al. 2024; Bruno et al. 2025a; Bruno et al. 2025b; Polyanskiy et al. 2025; Agazzi et al. 2026; Álvarez-López et al. 2026; Criscitiello et al. 2026).

The missing piece is therefore a macroscopic account of *training-selected* dynamics:

Q. *How does training reshape the mean-field dynamics of token distributions in Transformers?*

As a first step toward this question, we study a simple noisy mean-field Transformer in which each layer t contains a feedforward network (FFN) u_{W_t} with trainable parameters W_t. Our central finding is that the dynamics selected by training—after a single gradient-descent update (Proposition 3.5) and, in the interior-turnpike form, under minimizers of the training objective (Theorem 3.9)—differentiate the clustering picture into the 3 phases as shown in Figure 1:

- **Phase 1 (Initial clustering):** the attention interaction pulls the token distribution toward a clustered state;
- **Phase 2 (Turnpike plateau):** the distribution remains near the clustered state through most intermediate layers, forming a plateau;
- **Phase 3 (Terminal escape):** in the final layers, the trained FFN can drive the distribution away from the plateau in a direction that decreases a loss function ℓ.

Our contribution lies in theoretically elucidating the mechanisms behind these phenomena:

(i) In Section 3.1, we identify the clustering state through an attention-induced interaction energy. We prove that this energy is of the double-well type, providing the background for the persistence of Phase 2 in the intermediate layers.

(ii) Section 3.2 provides a lower bound representing Phase 3 that occurs when the parameter W is updated only once by gradient descent (GD). The proof illustrates that Phase 3 can be caused by a dynamics of gradients between layers through backpropagation.

(iii) Section 3.3 provides an upper bound that supports Phases 1 to 3 globally in layers for the *training-selected inference dynamics*: the depth dynamics generated by a global minimizer of the training objective, as opposed to the trajectory of an optimization algorithm. Parameter optimality constrains the energy dissipation, which can trigger phase transitions.

Rather than following the full optimization trajectory, these results analyze the two tractable ends of the training process—a single gradient-descent step and the minimizers of the objective—and carry complementary statements: a lower bound on the terminal energy lift, and an interior-turnpike upper bound. The finite-step regime in between is probed numerically. Numerical experiments in Section 4

further show that the strength of the turnpike plateau and terminal escape depends quantitatively on the geometric relation between the terminal loss and the interaction energy, as well as on the spectrum of the matrix governing the attention interaction.

Notation. Let $\mathbb{Z}_+$ denote the set of positive integers. For $N \in \mathbb{Z}_+$, $[N] \coloneqq \{1, 2, \dots, N\}$. We denote by $|\bullet|$ and $\langle \bullet, \bullet \rangle$ the Euclidean norm and inner product, respectively, and by $\|\bullet\|_{\mathrm{F}}$ the Frobenius norm. Let $\mathbb{S}^{d-1} \coloneqq \left\{x \in \mathbb{R}^d \;\middle|\; |x| = 1\right\} \subset \mathbb{R}^d$ be the unit sphere with tangent space $T_x\mathbb{S}^{d-1} \coloneqq \left\{v \in \mathbb{R}^d \;\middle|\; \langle v, x\rangle = 0\right\}$ at $x \in \mathbb{S}^{d-1}$. Let ∇, div, and $\triangle$ denote the gradient, divergence, and Laplace–Beltrami operator on $\mathbb{S}^{d-1}$, respectively. For an integer $k \geq 0$, $C^k(\mathbb{S}^{d-1})$ and $H^k(\mathbb{S}^{d-1})$ denote the standard spaces of k-times continuously differentiable and Sobolev functions on $\mathbb{S}^{d-1}$, respectively. For a compact metric space X, $\mathcal{P}(X)$ denotes the space of Borel probability measures on X, endowed with the narrow topology, which is metrized by the (2-)Wasserstein distance W_2. For a Borel map $T\colon X \to X$ and $\mu \in \mathcal{P}(X)$, $T_{\#}\mu \in \mathcal{P}(X)$ denotes the pushforward of μ by T,

2 Framework and problem setting

For theoretical analysis, we model the mean-field dynamics of tokens in a Transformer by a non-local Fokker–Planck equation. Training is then formulated as an optimization problem over the parameters appearing in the drift term, which corresponds to tuning the feedforward network inside each layer.

2.1 Modeling token dynamics by a non-local Fokker–Planck equation

We first recall a discrete-time dynamical system representation of a Transformer and then introduce the mean-field model studied in this paper. According to (Sander et al. 2022; Geshkovski et al. 2025), we can regard the Transformer as mapping an input sequence of (embedded) tokens $(x_i)_{i=1}^N$ to the terminal values $(x_i(L))_{i=1}^N$ of a discrete-time dynamical sequence $((x_i(l))_{i=1}^N)_{l=1}^L$ satisfying $x_i(1) = x_i$ and

$$x_i(l+1) = \mathcal{N}\Big(x_i(l) + u_{W(l)}(x_i(l)) + \sum_{j=1}^{N} \mathrm{e}^{\beta\langle Q(l)x_i(l), K(l)x_j(l)\rangle} V(l)x_j(l)\Big), \quad l \in [L-1],$$

where $\mathcal{N}\colon \mathbb{R}^d \to \mathbb{S}^{d-1}$ is the norm normalization and u_W is an FFN with weight matrices $W(l)$ for each layer l; for example, $u_W(x) = W_1 \tanh(W_2 x)$ where $W = (W_1, W_2)$, $W_1, W_2 \in \mathbb{R}^{d\times d}$. In the theoretical analysis, we specify the parametrization of u_W in Assumption 3.3 below. The coefficient β is a positive constant, and the matrices $Q(l)$, $K(l)$, $V(l)$ are weight matrices that define the query, key, and value, respectively. In the above discrete update the FFN enters additively; in a standard pre-layer-normalization (pre-LN) block it composes with the attention sub-block, and under the residual scaling $h = T/L$ the composed and additive forms coincide to first order—a Lie–Trotter splitting in the sense of (Lu et al. 2020)—with formal accumulated error $O(1/L)$ over L layers. Previous works in, e.g., (Sander et al. 2022; Karagodin et al. 2024; Bruno et al. 2025b; Burger et al. 2025; Geshkovski et al. 2025) have analyzed the case of *symmetric unnormalized self-attention (S-USA)*, in which the weight matrices are shared across layers and satisfy $Q^\top K = V = V^\top$. The S-USA model is introduced as a useful theoretical assumption for analyzing the token clustering phenomenon. Following the literature, we consider a noisy continuous-time counterpart of the model with a (continuous) depth $T > 0$ of layer as follows:

$$\mathrm{d}X_i(t) = P^{\perp}_{X_i(t)}\Big(u_{W(t)}(X_i(t)) + \frac{1}{N}\sum_{j=1}^{N} \mathrm{e}^{\langle X_i(t), AX_j(t)\rangle} AX_j(t)\Big)\,\mathrm{d}t + \sqrt{2\varepsilon}\,\mathrm{d}B_i(t), \quad t \in (0,T). \tag{2.1}$$

Here X_i is a stochastic process that models x_i, $P^{\perp}_x \coloneqq I - xx^\top \in \mathbb{R}^{d\times d}$ is the projection onto $T_x\mathbb{S}^{d-1}$ at $x \in \mathbb{S}^{d-1}$, $A \in \mathbb{R}^{d\times d}$ is a symmetric matrix denoting $Q^\top K$ in S-USA, B_i are independent Brownian motions on $\mathbb{S}^{d-1}$, and $\varepsilon > 0$ controls the strength of the injected noise. We set $\beta = 1$ for simplicity.

By the standard argument of the propagation of chaos as in (Chaintron and Diez 2022), the law μ_t of the tokens $X_i(t)$ at each layer $t \in [0,T]$ satisfies the following non-local Fokker–Planck equation

$$\begin{cases} \partial_t \mu_t(x) + \operatorname{div}\big(\mu_t(x) P^{\perp}_x(\chi_A[\mu_t] + u_{W_t})(x)\big) = \varepsilon \triangle \mu_t(x), \quad t \in (0,T), x \in \mathbb{S}^{d-1}, \\ \mu_t|_{t=0} = \mu_0, \end{cases} \tag{2.2}$$

when the number N of tokens is large and the input tokens (x_i) are sampled i.i.d. from a token distribution $\mu_0 \in \mathcal{P}(\mathbb{S}^{d-1})$. Here $\chi_A[\mu]\colon \mathbb{S}^{d-1} \to \mathbb{R}^d$ is the vector field given by

$$\chi_A[\mu](x) := \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\langle x, Ay\rangle} Ay \,\mathrm{d}\mu(y), \quad x \in \mathbb{S}^{d-1}.$$

Given a fixed input token distribution μ_0, we denote by $\mu^W = (\mu_t^W)_{t\in[0,T]} \in C([0,T]; \mathcal{P}(\mathbb{S}^{d-1}))$ the (weak) solution to (2.2) with initial condition μ_0, corresponding to the parameter $W = (W_t)_{t\in[0,T]}$. We call the terminal distribution $\mu_T^W := \mu_t^W\big|_{t=T}$ the output of the mean-field Transformer. An early measure-level continuity-equation formulation of token dynamics, together with a Wasserstein-gradient-flow interpretation for Sinkhorn-normalized attention, goes back to Sander et al. (2022).

Remark 2.1 (Role of the noise ε). Prior work on noisy Transformer dynamics has focused on simpler settings, such as scalar Q, K, V without an FFN (Balasubramanian et al. 2025; Rigollet 2026). Here, noise is included for two theoretical purposes: it provides parabolic regularization, which facilitates smooth solutions μ_t, and it helps eliminate spurious stationary distributions that can arise in deterministic settings, cf. (S. Chen et al. 2025; Agazzi et al. 2026). At the modeling level, the i.i.d. Brownian motions in (2.1) can be viewed as a simple model of small errors entering each token's computation independently, such as numerical error. Models in which the noise is shared across tokens (Agazzi et al. 2026; Fedorov et al. 2026; Koubbi et al. 2026a) lead instead to stochastic PDE limits and are not covered by the analysis below. Whether the three-phase picture below holds at $\varepsilon = 0$ is theoretically open, although it is observed numerically (Section 4).

2.2 Mean-field optimal control formulation of training

We now formulate training as an optimal control problem for the mean-field dynamics introduced above. Let $\ell\colon \mathcal{P}(\mathbb{S}^{d-1}) \to \mathbb{R}$ be a (narrowly-)continuous loss function with bounded oscillation $\operatorname{osc}\ell := \sup\ell - \inf\ell < +\infty$. We optimize the layer-dependent FFN parameters W_t so as to minimize the terminal loss with an L^2 regularization parameter $\lambda_{\mathrm{reg}} > 0$:

$$\underset{W\in L^2(0,T)}{\text{minimize}}\; J(W) := \ell(\mu_T^W) + \frac{\lambda_{\mathrm{reg}}}{2}\|W\|_{L^2(0,T)}^2, \tag{2.3}$$

where $L^2(0,T)$ is the function space of square integrable functions $t \mapsto W_t$ on $[0,T]$ that take values in the parameters of u_{W_t} and $\|W\|_{L^2(0,T)}^2 := \int_0^T \|W_t\|_{\mathrm{F}}^2 \,\mathrm{d}t$. Under plain gradient descent, the L^2 penalty is equivalent to weight decay, a standard regularizer in Transformer training.

> *Example* 2.2 (Bag-of-words loss). Consider predicting Bag-of-Words proposed in (Ma et al. 2018) using the output token distribution μ_T of a mean-field Transformer. Let $\mathcal{V}$ be a finite set of vocabulary words, and let $R\colon \mathbb{R}^d \to \mathbb{R}^{\mathcal{V}}$ be an affine transformation that maps each output token to word logits; we fix this transformation. A loss function $\ell = \ell_{\mathrm{BoW}}$ with the Bag-of-Words distribution $q \in \mathcal{P}(\mathcal{V})$ as the target can be modeled in the following form
>
> $$\ell_{\mathrm{BoW}}(\mu) := -\sum_{w\in\mathcal{V}} q(w) \log\left(\int \operatorname{softmax}(R(x))_w \,\mathrm{d}\mu(x)\right). \tag{2.4}$$
>
> For finite tokens $(x_i)_{i=1}^N$, this becomes $-\sum_w q(w)\log\big(\frac{1}{N}\sum_i \operatorname{softmax}(R(x_i))_w\big)$. □

The above formulation is closely related to the optimal-control view of ResNets and Neural ODEs (E 2017; E et al. 2019; E et al. 2022).

Remark 2.3 (Other possible formulations of training). More practical learning settings, such as the next-token prediction task, can also be addressed by reformulating (2.2) using a distribution that includes labels (or token positions), following the approach in (E et al. 2019; Bonnet et al. 2023). We adopt the formulation given in (2.3) to align with the literature on mean-field Transformers.

3 Main mathematical results

In this section, we investigate what kind of dynamics, along the depth direction, are exhibited by the solution μ^{W^*} generated by a layer-wise FFN parameter path $W^* = (W_t^*)_{t\in[0,T]} \in L^2(0,T)$ that minimizes the objective in (2.3).

In the same spirit as the analyses of (Sander et al. 2022; Burger et al. 2025; Geshkovski et al. 2025), one observes that (2.2) admits the following gradient-flow-type structure associated with the entropy-regularized interaction energy $E_{\varepsilon,A}$:

$$\partial_t \mu + \operatorname{div}\left(\mu P^{\perp}\left(-\nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu} + u_{W_t}\right)\right) = 0, \quad E_{\varepsilon,A}(\mu) \coloneqq \varepsilon \operatorname{Ent}(\mu) - \frac{1}{2}\iint \mathrm{e}^{\langle x, Ay\rangle} \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y), \tag{3.1}$$

where $\nabla \frac{\delta}{\delta \mu}$ is the (formal) Wasserstein gradient operator, $\operatorname{Ent}\colon \mathcal{P}(\mathbb{S}^{d-1}) \to \mathbb{R} \cup \{+\infty\}$ is the entropy defined by $\operatorname{Ent}(\rho\,\mathrm{d}\omega) = \int \rho \log \rho \,\mathrm{d}\omega$ for a probability density ρ with the uniform distribution ω on $\mathbb{S}^{d-1}$, and by $+\infty$ otherwise. If $W = 0$, the token distribution μ_t moves solely in the direction of decreasing the interaction energy $E_{\varepsilon,A}$, and is expected to concentrate into a small number of clusters.

On the other hand, in the variational problem (2.3), the parameter path W is optimized so that the terminal distribution μ_T^W makes the loss ℓ small. Hence, there are two effects: the attention, which decreases $E_{\varepsilon,A}$ and drives clustering of the token distribution, and the FFN u_{W_t}, which attempts to move μ_T^W in a direction that decreases the loss ℓ. From these two effects, we will find the following three phases to emerge in the trained dynamics:

Phase 1 In the first layers $t \in (0, \tau)$ with $\tau \ll T$, the attention χ_A is dominant, and the token distribution lowers $E_{\varepsilon,A}$ while localizing near a small number of points. This phase helps us to save on the regularization cost $\int_0^\tau \|W_t\|_{\mathrm{F}}^2 \,\mathrm{d}t \approx 0$.

Phase 2 In the middle layers $t \in [\tau, T-\tau]$, the token distribution stays in a plateau state near the set of stationary points along which $E_{\varepsilon,A}$ is almost constant. These dynamics are induced by the tendency to keep the cost $\int_\tau^{T-\tau} \|W_t\|_{\mathrm{F}}^2 \,\mathrm{d}t$ relatively small.

Phase 3 In terminal layers $t \in [T-\tau, T]$, the token distribution can leave the plateau to decrease the loss ℓ. This can be achieved because the FFN u_{W_t} can dominate the attention χ_A over the short layer interval $[T-\tau, T]$.

In this section, we mathematically formulate and verify the above dynamical three-phase structure in the regime of large depth, namely $T \gg 1$. In Section 3.1, we identify the clustered distribution $\overline{\mu}$ that appears in Phase 2. This allows us to formulate the situation in Phases 1 and 3 using the energy gap $\Delta_{\varepsilon,A} \coloneqq E_{\varepsilon,A} - \min E_{\varepsilon,A}$. In Section 3.2, we consider parameters obtained by applying gradient descent (GD) for a single step and demonstrate the mechanism by which Phase 3 can be triggered. Section 3.3 provides an exponential turnpike-type bound for a minimizer of the variational problem (2.3).

3.1 Landscape of the energy $E_{\varepsilon,A}$ in the mean-field Transformer

In this subsection, we study the stationary points of the energy $E_{\varepsilon,A}$ and the landscape around them to formulate Phase 2. In general, not only local minima but also saddle points of $E_{\varepsilon,A}$ exist continuously. For example, considering the case where $\varepsilon \searrow 0$ and $A = I$, it is known that the minimizers of $E_{\varepsilon,A}$ are arbitrary Dirac delta distributions and that other saddle points may also exist continuously (Bruno et al. 2025a; Burger et al. 2025; Geshkovski et al. 2025). To eliminate such spurious situations, we consider the following generic case:

Assumption 3.1 (Non-degenerate positive attention matrix)**.** The symmetric matrix $A \in \mathbb{R}^{d\times d}$ in (2.2) is positive definite and the maximum eigenvalue is simple, i.e., the eigenvalues $(\lambda_i)_{i=1}^d \subset \mathbb{R}$ of A sorted in descending order satisfy $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d > 0$. Without loss of generality, one can assume that the standard orthonormal basis $(e_i)_{i=1}^d \subset \mathbb{R}^d$ diagonalizes A.

Assumption 3.1 can be checked directly from the spectrum of A. Under this assumption, the following statement shows that $E_{\varepsilon,A}$ has exactly two minimizers and that no non-minimizing stationary point is path-connected to a minimizer within the stationary set:

Proposition 3.2 (Landscape analysis of $E_{\varepsilon,A}$)**.** *Under Assumption 3.1, let $e_1 \in \mathbb{S}^{d-1}$ be the principal eigenvector corresponding to λ_1. Then, there exists a constant $\overline{\varepsilon} = \overline{\varepsilon}(A, d)$ such that, for any $\varepsilon \in (0, \overline{\varepsilon})$, there exists a neighborhood $\mathcal{V}_+ \subset \mathcal{P}(\mathbb{S}^{d-1})$ of the Dirac delta $\delta_{e_1} \in \mathcal{P}(\mathbb{S}^{d-1})$ such that the following holds:*

- *$E_{\varepsilon,A}$ has a unique minimizer $\overline{\mu}^+ = \overline{\rho}^+\omega$ in $\mathcal{V}_+$, whose density solves $\overline{\rho}^+(x) \propto \mathrm{e}^{\frac{1}{\varepsilon}\int \mathrm{e}^{\langle x, Ay\rangle}\overline{\rho}^+(y)\mathrm{d}\omega(y)}$. In particular, $\overline{\rho}^+$ is positive and C^∞.*

- *Let $S\colon \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ be the antipodal map $x \mapsto -x$, and let $\overline{\mu}^- = S_{\#}\overline{\mu}^+$. Then, the set of global minimizers of $E_{\varepsilon,A}$ consists only of $\overline{\mu}^+$ and $\overline{\mu}^-$, and the Hessian operator H at $\overline{\mu} \in \{\overline{\mu}^+, \overline{\mu}^-\} =: \{\overline{\mu}^\pm\}$ is bounded for $h \in L^2_0(\overline{\mu}) := \{h \in L^2(\overline{\mu}) \mid \int h \,\mathrm{d}\overline{\mu} = 0\}$ as*

$$\varepsilon \|h\|^2_{L^2(\overline{\mu})} \geq \langle h, \mathsf{H}h \rangle_{L^2(\overline{\mu})} \geq \frac{\varepsilon}{2}\left(1 - \frac{\lambda_2}{\lambda_1}\right)\|h\|^2_{L^2(\overline{\mu})}, \tag{3.2}$$

  where $\mathsf{H}h := \varepsilon h - \Pi_0\big(\int \mathrm{e}^{\langle x, Ay\rangle} h(y)\,\mathrm{d}\overline{\mu}(y)\big)$ with the projection Π_0 onto $L^2_0(\overline{\mu})$.

- *Let $\mathrm{Crit}\, E_{\varepsilon,A}$ be the set of critical points of $E_{\varepsilon,A}$, which consists of solutions of the Gibbs-type equation, and let $\Delta_{\mathrm{crit}}(\varepsilon) := \inf\{\Delta_{\varepsilon,A}(\mu) \mid \mu \in \mathrm{Crit}\, E_{\varepsilon,A} \setminus \{\overline{\mu}^\pm\}\}$. Then*

$$\liminf_{\varepsilon \searrow 0} \Delta_{\mathrm{crit}}(\varepsilon) \geq \Delta_0 := \frac{\min\{\sinh \lambda_1, \mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2}\}}{2} > 0.$$

Informally speaking, Proposition 3.2 means that unless the energy gap of the input token distribution μ_0 is large enough to exceed Δ_0, μ_0 is in a region that is attracted to the set of minimizers $\{\overline{\mu}^\pm\}$. This observation will be useful in proving the key lemma in Section 3.3.

3.2 Adjoint analysis in the first step of gradient descent

In this subsection, we investigate the principles underlying Phase 3. Specifically, we show that even for a model trained for just a single step of GD, an energy gap can grow near the final layer.

For simplicity, consider the following situation: the input token distribution is the minimizer $\overline{\mu} := \overline{\mu}^+$ without loss of generality, and the loss function is sufficiently smooth; specifically, assume that the linear functional derivative $g_\ell := \frac{\delta \ell}{\delta \mu}[\overline{\mu}]\colon \mathbb{S}^{d-1} \to \mathbb{R}$, defined in the sense of (Carmona and Delarue 2018), is C^∞. Suppose the parameters in GD are initialized to $W^{\mathbf{0}} := 0$. Note that in this case $\mu_t^{W^{\mathbf{0}}} \equiv \overline{\mu}$. Furthermore, for the FFN u_W, we assume the parameter-linear parameterization used in (Barboni et al. 2022; Marion 2023; Scagliotti 2023; Isobe and Okumura 2024):

Assumption 3.3 (Linear-in-parameters FFN)**.** The FFN $u_W\colon \mathbb{R}^d \to \mathbb{R}^d$ has the form $u_W(x) = W\sigma(x), W \in \mathbb{R}^{d\times p}$, where $p \in \mathbb{Z}_+$ is the number of hidden features and $\sigma\colon \mathbb{R}^d \to \mathbb{R}^p$ is a smooth function with a constant $C_\sigma := \|\sigma\|^2_{L^\infty(\mathbb{S}^{d-1})}/2 < \infty$. Thus, the parameters to be optimized in (2.3) become an $\mathbb{R}^{d\times p}$-valued L^2 path $(W_t)_{t\in(0,T)} \in L^2(0,T;\mathbb{R}^{d\times p})$.

Let us consider the solution $\mu^{W^{\text{1-GD}}}$ to equation (2.2) under the initial value $\mu_0 = \overline{\mu}$ and the parameters $W^{\text{1-GD}} := W^{\mathbf{0}} - \alpha\nabla J(0)$ updated by one step with a learning rate $\alpha > 0$. Recall that the (Fréchet) gradient $\nabla J(0) \in L^2(0,T)$ of $J(W)$ with respect to $W \in L^2(0,T)$ can be described via the adjoint method, which is the continuous-time version of backpropagation (see (R. T. Q. Chen et al. 2018; Massaroli et al. 2020) and Appendix A.2). We obtain the following expression:

Lemma 3.4. *Under Assumption 3.3 and the regularity of g_ℓ, $W^{\text{1-GD}} \in L^2(0,T;\mathbb{R}^{d\times p})$ is given by*

$$W_t^{\text{1-GD}} = -\alpha \int \nabla\phi_t(x)\sigma(x)^\top \,\mathrm{d}\overline{\mu}(x) \in \mathbb{R}^{d\times p}, \quad t \in (0,T), \tag{3.3}$$

where $(\phi_t)_{t\in[0,T]} \in C^1([0,T]; C(\mathbb{S}^{d-1})) \cap C([0,T]; C^2(\mathbb{S}^{d-1}))$ is the unique (classical) solution of the backward weighted diffusion equation given by

$$\partial_t\phi_t + \mathsf{H}\triangle_{\overline{\rho}}\phi_t = 0, \quad \phi_T = g_\ell, \tag{3.4}$$

with $\triangle_{\overline{\rho}}\phi := \overline{\rho}^{-1}\operatorname{div}(\overline{\rho}\nabla\phi)$ for $\phi \in H^2(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu})$.

Roughly speaking, ϕ_t and $\nabla\phi_t$ correspond to the linear functional and Wasserstein gradient of the loss ℓ at layer t, namely $\frac{\delta\ell}{\delta\mu_t}$ and $\nabla\frac{\delta\ell}{\delta\mu_t}$, respectively. Since ϕ_t solves (3.4), the gradient can increase exponentially as we approach the final layer, as shown in Figure 2. However, this does not hold if $g_\ell = 0$, i.e., if the loss ℓ is stationary at the minimizer $\overline{\mu}$ of $E_{\varepsilon,A}$. Based on this analysis, the following estimate shows that the energy gap may increase near the final layer.

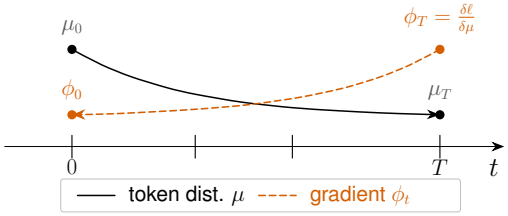


Figure 2: Token distribution μ_t and adjoint gradient ϕ_t.

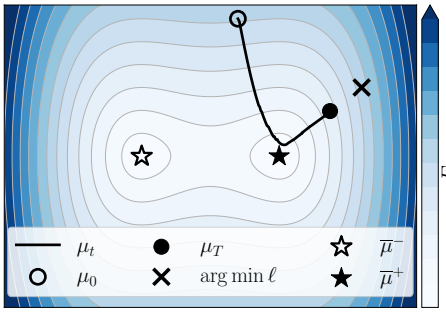


Figure 3: Energy-landscape view of terminal escape.

Proposition 3.5 (Final-layer escape in one-step gradient descent)**.** *Suppose Assumptions 3.1 and 3.3. Let $\mu^{W^{1\text{-GD}}}$ be the solution to equation (2.2) with initial value $\mu_0 = \overline{\mu}$ under the parameter $W^{1\text{-GD}}$ in (3.3). Then, for every $r > 0$ there exist constants $\overline{\alpha}, C_R > 0$, depending on r as in Proposition A.12, such that for any $\alpha \in (0, \overline{\alpha})$ and $t \in (T - r, T)$, it holds that*

$$\Delta_{\varepsilon,A}(\mu_t^{W^{1\text{-GD}}}) \geq \frac{1}{4}\big(\alpha \Omega_{t,T}(g_\ell)\|g_\ell\|_{\mathsf{H}^{-1}} \exp(-(T-t)R_{\overline{\rho}}(g_\ell)) - \alpha^2 C_R\big)_+^2.$$

Here, if $g_\ell = 0$, the right-hand side is understood as 0, $\|g_\ell\|_{\mathsf{H}^{-1}} := \sqrt{\langle g_\ell, \mathsf{H}^{-1} g_\ell \rangle}$, and $\Omega_{t,T}(g_\ell)$ and $R_{\overline{\rho}}(g_\ell)$ are defined for $g_\ell \neq 0$ by

$$\Omega_{t,T}(g_\ell) := \frac{\|\nabla J(0)\|_{L^2(0,t)}^2}{\|\phi_t\|_{\mathsf{H}^{-1}}^2}, \quad R_{\overline{\rho}}(g_\ell) := \frac{\|\nabla g_\ell\|_{L^2(\overline{\mu})}^2}{\|g_\ell\|_{\mathsf{H}^{-1}}^2} = \left\|\nabla \frac{\delta \ell}{\delta \mu}[\overline{\mu}]\right\|_{L^2(\overline{\mu})}^2 \Bigg/ \left\|\frac{\delta \ell}{\delta \mu}[\overline{\mu}]\right\|_{\mathsf{H}^{-1}}^2. \tag{3.5}$$

The proof is given in Appendix A.2, where we use Jensen's inequality as a key step. This result shows that, already in the initial steps of GD, energy can increase exponentially with respect to t near the terminal layer. The Rayleigh quotient $R_{\overline{\rho}}(g_\ell)$ governs the rate of escape.

3.3 Exponential turnpike bound for optimized dynamics

Finally, we study the token distribution $\mu_t^* := \mu_t^{W^*}$ induced by a global minimizer W^* of the variational problem (2.3). We prove an exponential turnpike-type bound showing that $\Delta_{\varepsilon,A}(\mu_t^*)$ becomes exponentially small away from the two boundary layers. Figure 3 schematically shows an optimized trajectory $(\mu_t^*)_t$: after staying near $\{\overline{\mu}^{\pm}\}$, it escapes toward lower ℓ while retaining that clustering bias, and reaches the output μ_T^*. The theorem below formalizes the corresponding plateau–boundary-layer structure. As in (Isobe and Okumura 2024), the variational problem (2.3) admits a minimizer under Assumption 3.3. To state the theorem, we impose that the initial energy gap and the oscillation of the terminal loss are not too large:

Assumption 3.6 (Energy-threshold conditions)**.** The initial energy gap $\Delta_{\varepsilon,A}(\mu_0)$ and $\operatorname{osc} \ell$ satisfy $\mathfrak{E}_{\max} := \Delta_{\varepsilon,A}(\mu_0) + \frac{2C_\sigma}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell < \Delta_0, \operatorname{osc} \ell < \frac{\lambda_{\mathrm{reg}}}{8C_\sigma}\Delta_0$ where Δ_0 is given by Proposition 3.2.

Assumption 3.6 is explicit in the initial energy gap, $\operatorname{osc} \ell$, λ_{reg}, C_σ, and the margin Δ_0. It excludes initial distributions with little bias toward a clustered configuration (e.g., near-uniform ones)—consistent with the biased token representations the model is meant to describe—and it is the only place where closeness to the minimizers enters: convergence toward $\{\overline{\mu}^{\pm}\}$ is not assumed.

The next lemma formulates the main consequence of Assumption 3.6: for sufficiently long depth T, μ_t^* enters the sublevel set $\{\Delta_{\varepsilon,A}(\mu) \leq \eta\}$ at some time. For a small $\eta > 0$, the sublevel set has

exactly two connected components. We denote by $\mathcal{C}_+(\eta), \mathcal{C}_-(\eta)$ the components containing $\overline{\mu}^+$ and $\overline{\mu}^-$, respectively.

Lemma 3.7 (Entry into one low-energy component)**.** *Assume Assumption 3.6. Then, there exists $\overline{\varepsilon} > 0$ such that for every $\varepsilon \in (0, \overline{\varepsilon})$, there exist $\eta, r_* > 0$ and $L_\varepsilon < \infty$ such that the following holds: For every sufficiently large T and every optimal pair (μ^*, W^*), there exist times $s_- \in [0, L_\varepsilon]$, $s_+ \in [T - L_\varepsilon, T]$, and a sign $\varsigma \in \{+, -\}$ such that $\mu^*_{s_-}, \mu^*_{s_+} \in \mathcal{U}_\varsigma \cap \mathcal{C}_\varsigma(\eta)$ where $\mathcal{U}_\varsigma \subset \mathcal{P}(\mathbb{S}^{d-1})$ is defined for each ς by*

$$\mathcal{U}_+ := \left\{ (1+h)\overline{\mu}^+ \in \mathcal{P}(\mathbb{S}^{d-1}) \;\middle|\; h \in L^\infty(\mathbb{S}^{d-1}), \int h \,\mathrm{d}\overline{\mu}^+ = 0, \|h\|_{L^\infty} < r_* \right\}, \tag{3.6}$$

and set $\mathcal{U}_- := S_\# \mathcal{U}_+$.

Proof sketch of Lemma 3.7. By Corollary A.33, the first inequality in Assumption 3.6 gives a positive constant $d_\eta > 0$ such that

$$\Delta_{\varepsilon,A}(\mu) \le \mathfrak{E}_{\max}, \qquad \mu \notin \mathcal{C}_+(\eta) \sqcup \mathcal{C}_-(\eta) \quad \Longrightarrow \quad D(\mu) := \int \left| \nabla \tfrac{\delta E_{\varepsilon,A}}{\delta \mu}[\mu] \right|^2 \mathrm{d}\mu \ge d_\eta.$$

Moreover, $J(W^*) \le J(0)$ gives $\frac{\lambda_{\mathrm{reg}}}{2} \|W^*\|^2_{L^2(0,T)} \le \operatorname{osc} \ell$, and the gradient-flow structure yields $\int_0^T D(\mu^*_t)\,\mathrm{d}t \le 2\Delta_{\varepsilon,A}(\mu_0) + \frac{4C_\sigma}{\lambda_{\mathrm{reg}}} \operatorname{osc} \ell$. Hence Lemma A.39 gives times $s_- \in [0, L_\varepsilon]$ and $s_+ \in [T - L_\varepsilon, T]$ at which the trajectory enters $\mathcal{C}_+(\eta) \sqcup \mathcal{C}_-(\eta)$. The two entries cannot be in different components of $\mathcal{C}_\pm(\eta)$. Indeed, Corollary A.31 gives $\inf_{\int x_1 \mathrm{d}\mu = 0} \Delta_{\varepsilon,A}(\mu) \ge \frac{3}{4}\Delta_0$, while $\eta \le \Delta_0/4$. Therefore, switching components requires an energy increase of at least $\Delta_0/2$, hence an L^2-cost of W^* larger than the terminal-loss improvement allowed by the second inequality in Assumption 3.6. Thus Lemma A.40 places the two entries in the same component. Finally, Lemma A.34 upgrades these entries to the neighborhoods of densities: $\mu^*_{s_-}, \mu^*_{s_+} \in \mathcal{U}_\varsigma \cap \mathcal{C}_\varsigma(\eta)$. On $\mathcal{U}_\varsigma$, the Hessian coercivity in Proposition 3.2 gives the local Polyak–Łojasiewicz (PL) inequality. □

We impose regularity on ℓ as follows to use the optimality condition of (2.3) and its linearization:

Assumption 3.8 (Smooth loss)**.** The loss ℓ is of class C^2 on an $L^2(\omega)$-open set containing all Hölder-continuous probability densities when regarded as a functional of the density via $\ell(\rho\omega)$.

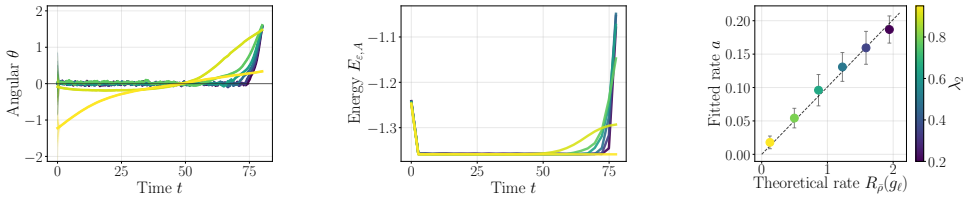
Under Assumption 3.8, every (μ^*, W^*) satisfies the optimality condition $\lambda_{\mathrm{reg}} W^*_t = -\int \nabla \phi_t \sigma^\top \,\mathrm{d}\mu^*_t$—the fixed-point form of the one-step update (3.3)—where the adjoint state ϕ solves the backward equation (3.4) linearized along μ^* rather than at $\overline{\mu}$. The Hessian coercivity of Proposition 3.2 and the weighted Poincaré inequality for $\overline{\mu}$ make this backward dynamics exponentially stable, with rate and constant independent of T; hence ϕ and W^* are exponentially small away from the terminal layer. Substituting this bound on W^* (Corollary A.50) into the local PL inequality on $\mathcal{U}_\pm$ yields the following exponential turnpike bound, whose full proof is given in Appendix A.3. Informally, for every sufficiently deep network, every optimal token flow stays exponentially close in energy to one clustered state, except in two boundary layers of depth-independent width at the input and at the output.

Theorem 3.9. *Fix $\varepsilon \in (0, \overline{\varepsilon})$. Suppose Assumptions 3.1, 3.3, 3.6, and 3.8 hold. Then there exist $T^*, C, a > 0$ such that for every $T > T^*$ and every optimal pair (μ^*, W^*),*

$$\Delta_{\varepsilon,A}(\mu^*_t) \le C\left(\mathrm{e}^{-at} + \mathrm{e}^{-a(T-t)} \right), \quad t \in [0, T].$$

Remark 3.10 (Dependence of a on A). The rate a is the smaller of the local PL rate and the backward decay rate of (3.4) along μ^*. Both depend on A through $1 - \lambda_2/\lambda_1$ and $\lambda_1 \mathrm{e}^{-\lambda_1}$, as made explicit in Remark A.52.

Remark 3.11 (A linearized weak-attention model). To exhibit the three phases in a setting where (2.3) can be solved explicitly, we consider the weak-attention regime. Replacing A by βA and expanding in β, the energy becomes $E_{\varepsilon,A}(\mu) = \varepsilon \operatorname{Ent}(\mu) - \frac{\beta}{2}\langle m_\mu, A m_\mu \rangle + O(\beta^2)$ up to a constant, where $m_\mu := \int x \,\mathrm{d}\mu$, so only m_μ interacts at leading order. For $\sigma \equiv 1$ in Assumption 3.3 and



(a) Angular trajectories. (b) Energy trajectories. (c) Fitted/theoretical rate.

Figure 4: Spectral dependence of the turnpike in the align-target experiment for various λ_2.

$\ell(\mu) = \frac{1}{2}|m_\mu - b|^2$, linearizing around ω gives $\dot{m}_j = -r_j m_j + \frac{d-1}{d} w_j$ with $r_j := (d-1)(\varepsilon - \beta\lambda_j/d)$, and (2.3) becomes a finite-dimensional linear–quadratic problem. For $\beta\lambda_1 < \varepsilon d$, its solution is $m_j^*(t) = m_j(0)\mathrm{e}^{-r_j t} + \theta_j \mathrm{e}^{-r_j(T-t)} \frac{1-\mathrm{e}^{-2r_j t}}{1-\mathrm{e}^{-2r_j T}}$ for some $\theta_j \in \mathbb{R}$ bounded uniformly in T.

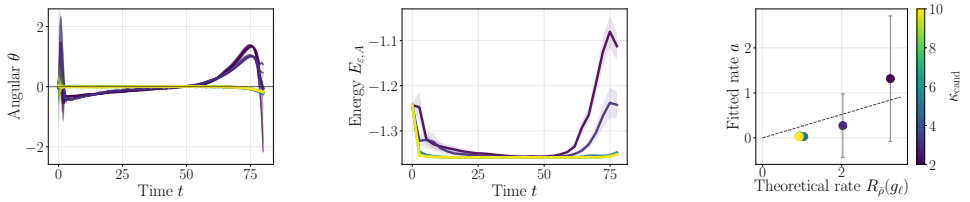
4 Numerical experiments

We validate the turnpike behavior through particle simulations of (2.1). We simulate $N = 64$ particles up to horizon $T = 80$, and work in a projected $d = 2$ setting with $A = \mathrm{diag}(1, \lambda_2)$. The particles are initialized in a cap around $e_1 = (1, 0)$, so the experiments track how the learned control moves an initially clustered token distribution along depth. Implementation details are given in Appendix B.1. We compare two terminal losses: an align-target loss and a Bag-of-Words (BoW) loss as in Example 2.2. The align-target loss is $\ell_{\text{align}}(\mu) = 1 - \left\langle \int x \, \mathrm{d}\mu(x), u_{\text{tar}} \right\rangle$, with $u_{\text{tar}} = e_2 = (0, 1)$; it encourages the first moment to point toward a target direction away from the initial cap center. For the BoW loss, we use the finite-vocabulary version of (2.4) with logits $R(x)_w = \langle x, c_w \rangle$, uniform target q, and candidate vectors $c_w \in \mathbb{S}^1$ sampled from a von Mises–Fisher distribution with concentration κ_{cand}; this tests how the geometry of the readout candidates affects the terminal-layer escape. Before the parameter sweeps below, Appendix B.2 shows that the turnpike profile appears only after optimizing the controls. We report three diagnostics. First, we plot the signed angle θ relative to the reference direction, so $\theta \approx 0$ indicates alignment with the interior plateau. Second, we compute the energy $E_{\varepsilon,A}$ from (3.1) along the simulated trajectories. Third, we fit the terminal energy lift to $C\exp(-a(T-t))$ and compare the fitted rate a with the theoretical rate computed numerically as the Rayleigh quotient $R_{\bar{\rho}}(g_\ell)$ in Proposition 3.5.

Experiment 1: Attention matrix structure induces the turnpike. We fix the align-target loss and vary the eigenvalue λ_2 of $A = \mathrm{diag}(1, \lambda_2)$. In Figure 4(a), the angular trajectories show that smaller λ_2 keeps the particles aligned near $\theta = 0$ through most of the depth and then produces a sharper terminal displacement. The energy curves in Figure 4(b) show the same effect at the level of $E_{\varepsilon,A}$: smaller λ_2 creates a deeper and more persistent interior plateau before the terminal lift. Finally, Figure 4(c) compares the fitted terminal exponent a with the numerically computed theoretical rate $R_{\bar{\rho}}(g_\ell)$. Their shared dependence on λ_2 indicates that the spectral gap of the attention matrix controls the strength of the turnpike and the sharpness of the final-layer escape.

Experiment 2: Target-distribution geometry modulates the turnpike. We fix $\lambda_2 = 0.65$ and vary the concentration κ_{cand} of the von Mises–Fisher distribution used to sample the BoW candidate vectors. In Figure 5(a), smaller κ_{cand} produces a sharper terminal angular dispersion, indicating that a less concentrated readout geometry can destabilize the plateau near the terminal layer. The energy trajectories in Figure 5(b) also show that smaller κ_{cand} produces a stronger terminal escape. Across the different values of κ_{cand}, the fitted terminal exponent a remains consistent with the numerically computed theoretical bound based on $R_{\bar{\rho}}(g_\ell)$, which indicates that the target-distribution geometry modulates the strength of the final-layer escape.

Robustness checks. We complement the sweeps above with a robustness study across seeds, diffusion strengths, horizons, particle numbers, regularization, and targets, using 20 independent runs per configuration and one fixed three-phase criterion applied without retuning. The results are summarized in Table 2, and the protocol and full tables are given in Appendix B.3. The full profile appears in every run at the reference setting, from the deterministic dynamics ($\varepsilon = 0$) through moderate diffusion, and for $T \in \{40, 80, 120\}$, where the plateau grows with T while the boundary layers keep a T-independent width. Across the candidate-bank sweep the interior plateau is always



(a) Angular trajectories. (b) Energy trajectories. (c) Fitted/theoretical rate.

Figure 5: Target-geometry dependence of the turnpike in the BoW experiment for various κ_{cand}.

present, whereas the strength and timing of the terminal escape depend on the target. At $\varepsilon = 10^{-1}$ the clustering and the plateau degrade while the terminal response persists, and under a global random initialization the profile requires early entry into a single well. These experiments are stress tests of the qualitative picture in a controlled particle model. They neither verify the quantitative hypotheses of Theorem 3.9 nor constitute measurements on pretrained Transformers.

5 Related Work, Limitations, and Discussion

Relation to prescribed-weight mean-field Transformer dynamics. Existing mean-field Transformer analyses reviewed in Section 1 enriched the prescribed-weight mean-field picture, but they do not identify the training-induced terminal escape from a clustered regime studied here. The closest work in motivation to ours is that of Koubbi et al. (2026b), which analyzes the effect of LoRA-type attention perturbations on token clustering. However, their theory treats the perturbed parameters as given, whereas our analysis studies how training-selected parameters reshape the dynamics. On the noise side, stationary solutions of the associated McKean–Vlasov equations on spheres are studied in (Shalova and Schlichting 2026), and Mun and Rosenzweig (2026) prove a sharp phase-transition dichotomy for the noisy USA free energy in every dimension $d \geq 2$; both are equilibrium analyses of the same idiosyncratic-noise family, on top of which our results add the training layer.

Relation to turnpike theory. Turnpike phenomena are classical in optimal control, with related estimates for Neural ODEs (Geshkovski and Zuazua 2022; Gugat 2025), mean-field optimal control (Herty and Zhou 2025; Bonnet-Weill et al. 2026), and broader optimal control problems (Trélat and Zuazua 2025). Our estimate is related in spirit, but the plateau regime is generated by the attention-induced interaction rather than imposed as a target or representation regularizer; terminal escape can then be driven by the trained FFN parameters and terminal loss.

Practical implications for Transformer models. The predicted depth profile has qualitative counterparts in pretrained models: large blocks of interior layers can be removed with little damage and interventions harm most at the early and final layers (Gromov et al. 2025; Lad et al. 2025), and the terminal escape in our model echoes reported sharp final-layer representation shifts in pretrained multilingual models (Wendler et al. 2024; Shibata et al. 2026). We do not claim that our FFN-only noisy mean-field model directly explains these phenomena in full-scale Transformers. The practical implications of our model are instead task-dependent: for distributed-information objectives such as ℓ_{BoW}, a final-layer jump may correct attention-induced clustering, whereas for clustering-compatible objectives ℓ, it may signal avoidable over-reliance on the final layers.

6 Conclusion

We studied how training reshapes token dynamics in Transformers by optimizing a parameter-linear FFN in a noisy mean-field model. Training can qualitatively modify the clustering picture of the prescribed-parameter setting: the trained dynamics may exhibit early clustering, a long turnpike plateau, and terminal escape, where the adjoint analysis captures terminal-layer amplification after a gradient-descent update and the exponential turnpike estimate captures the persistence of the clustered plateau for optimized parameters. Our theorems concern the two-well regime; landscapes with additional metastable states, verification of the mechanism in pretrained Transformers, and extensions beyond the noisy S-USA setting with FFN-only training are left for future work.

Acknowledgments and Disclosure of Funding

N.I. was supported by JST ACT-X (JPMJAX25C2) and JSPS KAKENHI (25K24910). M.I. was supported by JSPS KAKENHI (24K02904), JST CREST (JPMJCR21D2), FOREST (JPMJFR216I), ASPIRE (JPMJAP2329), BOOST (JPMJBY24A9), and ERATO (JPMJER2601). The authors declare no competing interests.

References

S. Abnar and W. Zuidema (2020). "Quantifying Attention Flow in Transformers". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. URL: https://aclanthology.org/2020.acl-main.385/ (cit. on p. 1).

A. Agazzi, G. Bruno, E. M. García, S. Saviozzi, and M. Romito (2026). *Stochastic Scaling Limits and Synchronization by Noise in Deep Transformer Models*. arXiv: 2604.26898 [math.PR]. URL: https://arxiv.org/abs/2604.26898v1 (cit. on pp. 2, 4).

A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet (2026). "Perceptrons and Localization of Attention's Mean-Field Landscape". *Proceedings of the 43rd International Conference on Machine Learning*. Vol. 306. Proceedings of Machine Learning Research, pp. 2180–2206. URL: https://proceedings.mlr.press/v306/alvarez-lopez26a.html (cit. on p. 2).

K. Balasubramanian, S. Banerjee, and P. Rigollet (2025). *On the Structure of Stationary Solutions to McKean-Vlasov Equations with Applications to Noisy Transformers*. arXiv: 2510.20094 [math.PR]. URL: https://arxiv.org/abs/2510.20094v2 (cit. on p. 4).

R. Barboni, G. Peyré, and F.-X. Vialard (2022). "On global convergence of ResNets: From finite to infinite width using linear parameterization". *Advances in Neural Information Processing Systems*. Vol. 35, pp. 16385–16397. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/67ea314d1df751bbf99ab664ae3049a5-Paper-Conference.pdf (cit. on p. 6).

L. Bereska and S. Gavves (2024). "Mechanistic Interpretability for AI Safety – A Review". *Transactions on Machine Learning Research*. Survey Certification, Expert Certification. URL: https://openreview.net/forum?id=ePUVetPKu6 (cit. on p. 1).

V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov (2015). *Fokker–Planck–Kolmogorov Equations*. Mathematical Surveys and Monographs 207. URL: https://doi.org/10.1090/surv/207 (cit. on p. 48).

V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov (2019). "Convergence in variation of solutions of nonlinear Fokker–Planck–Kolmogorov equations to stationary measures". *Journal of Functional Analysis* 276.12, pp. 3681–3713. URL: https://www.sciencedirect.com/science/article/pii/S0022123619301041 (cit. on pp. 37, 41).

B. Bonnet, C. Cipriani, M. Fornasier, and H. Huang (2023). "A measure theoretical approach to the mean-field maximum principle for training NeurODEs". *Nonlinear Analysis* 227, p. 113161. URL: https://www.sciencedirect.com/science/article/pii/S0362546X2200236X (cit. on p. 4).

B. Bonnet-Weill, G. Colombo, D. Shishmintsev, and E. Trélat (2026). *Exponential Turnpike Theorems for Nonlinear Deterministic Meanfield Optimal Control Problems*. arXiv: 2603.23303 [math.OC]. URL: https://arxiv.org/abs/2603.23303v1 (cit. on p. 10).

G. Bruno, F. Pasqualotto, and A. Agazzi (2025a). "A multiscale analysis of mean-field transformers in the moderate interaction regime". *Advances in Neural Information Processing Systems*. Vol. 38. URL: https://openreview.net/forum?id=WCRPgBpbcA (cit. on pp. 2, 5).

– (2025b). "Emergence of meta-stable clustering in mean-field transformer models". *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=eBS3dQQ8GV (cit. on pp. 2, 3).

M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand (2025). "Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 383.2298, p. 20240233. URL: https://doi.org/10.1098/rsta.2024.0233 (cit. on pp. 1, 3, 5).

R. Carmona and F. Delarue (2018). "Spaces of Measures and Related Differential Calculus". *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*, pp. 349–512. URL: https://doi.org/10.1007/978-3-319-58920-6_5 (cit. on p. 6).

L.-P. Chaintron and A. Diez (2022). "Propagation of chaos: A review of models, methods and applications. II. Applications". *Kinetic and Related Models* 15.6, pp. 1017–1173. URL: https://www.aimsciences.org/article/id/62948a7a2d80b70dfd2582aa (cit. on pp. 3, 48).
R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). "Neural Ordinary Differential Equations". *Advances in Neural Information Processing Systems*. Vol. 31. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf (cit. on p. 6).
S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet (2025). *Quantitative Clustering in Mean-Field Transformer Models*. arXiv: 2504.14697 [cs.LG]. URL: https://arxiv.org/abs/2504.14697v1 (cit. on p. 4).
C. Criscitiello, Q. Rebjock, A. D. McRae, and N. Boumal (2026). "Synchronization on Circles and Spheres with Nonlinear Interactions". *SIAM Journal on Applied Dynamical Systems* 25.2, pp. 1207–1237. URL: https://doi.org/10.1137/24M1699759 (cit. on p. 2).
W. E (2017). "A Proposal on Machine Learning via Dynamical Systems". *Communications in Mathematics and Statistics* 5.1, pp. 1–11. URL: https://doi.org/10.1007/s40304-017-0103-z (cit. on p. 4).
W. E, J. Han, and Q. Li (2019). "A mean-field optimal control formulation of deep learning". *Research in the Mathematical Sciences* 6.1, p. 10. URL: https://doi.org/10.1007/s40687-018-0172-y (cit. on p. 4).
– (2022). "Dynamical Systems and Optimal Control Approach to Deep Learning". *Mathematical Aspects of Deep Learning*, pp. 422–438. URL: https://doi.org/10.1017/9781009025096.011 (cit. on p. 4).
L. Fedorov, M. E. Sander, R. Elie, P. Marion, and M. Laurière (2026). "Clustering in Deep Stochastic Transformers". *Proceedings of the 43rd International Conference on Machine Learning*. Vol. 306. Proceedings of Machine Learning Research, pp. 29779–29803. URL: https://proceedings.mlr.press/v306/fedorov26a.html (cit. on p. 4).
A. Geiger, H. Lu, T. Icard, and C. Potts (2021). "Causal Abstractions of Neural Networks". *Advances in Neural Information Processing Systems*. Vol. 34, pp. 9574–9586. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf (cit. on p. 1).
B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet (2024). *Dynamic metastability in the self-attention model*. arXiv: 2410.06833 [cs.LG]. URL: https://arxiv.org/abs/2410.06833v1 (cit. on p. 1).
B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet (2023). "The emergence of clusters in self-attention dynamics". *Advances in Neural Information Processing Systems*. Vol. 36, pp. 57026–57037. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/b2b3e1d9840eba17ad9bbf073e009afe-Paper-Conference.pdf (cit. on p. 1).
– (2025). "A mathematical perspective on transformers". *Bulletin of the American Mathematical Society* 62.3, pp. 427–479. URL: https://doi.org/10.1090/bull/1863 (cit. on pp. 1, 3, 5).
B. Geshkovski and E. Zuazua (2022). "Turnpike in optimal control of PDEs, ResNets, and beyond". *Acta Numerica* 31, pp. 135–263. URL: https://doi.org/10.1017/S0962492922000046 (cit. on p. 10).
A. Grigor'yan (1999). "Estimates of heat kernels on Riemannian manifolds". *Spectral Theory and Geometry*. London Mathematical Society Lecture Note Series, pp. 140–225. URL: https://doi.org/10.1017/CBO9780511566165.008 (cit. on p. 41).
A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, and D. A. Roberts (2025). "The Unreasonable Ineffectiveness of the Deeper Layers". *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ngmEcEer8a (cit. on p. 10).
M. Gugat (2025). "The Exponential Turnpike Property for Neural Differential Equations". *29th International Conference on Methods and Models in Automation and Robotics*, pp. 142–147. URL: https://doi.org/10.1109/MMAR65820.2025.11151091 (cit. on p. 10).
M. Herty and Y. Zhou (2025). "Exponential turnpike property for particle systems and mean-field limit". *European Journal of Applied Mathematics* 36.5, pp. 993–1011. URL: https://doi.org/10.1017/S0956792524000871 (cit. on p. 10).
N. Isobe and M. Okumura (2024). "Variational Formulations of ODE-Net as a Mean-Field Optimal Control Problem and Existence Results". *Journal of Machine Learning* 3.4, pp. 413–444. URL: https://www.global-sci.com/jml/article/view/13205 (cit. on pp. 6, 7).
N. Karagodin, Y. Polyanskiy, and P. Rigollet (2024). "Clustering in Causal Attention Masking". *Advances in Neural Information Processing Systems*. Vol. 37, pp. 115652–115681.

URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/d18d208fa9c333483e5724ade7beff0f-Paper-Conference.pdf (cit. on p. 3).
P. Kidger (2022). "On Neural Differential Equations". PhD thesis. University of Oxford. arXiv: 2202.02435 [cs.LG]. URL: https://arxiv.org/abs/2202.02435v1 (cit. on p. 50).
D. P. Kingma and J. Ba (2015). "Adam: A Method for Stochastic Optimization". *The Third International Conference on Learning Representations*. URL: http://arxiv.org/abs/1412.6980 (cit. on p. 50).
H. Koubbi, B. Geshkovski, and P. Rigollet (2026a). *Homogenized Transformers*. arXiv: 2604.01978 [math.PR]. URL: https://arxiv.org/abs/2604.01978v1 (cit. on p. 4).
H. Koubbi, L. Hernandez, and M. Boussard (2026b). "Understanding Catastrophic Forgetting In LoRA via Mean-Field Attention Dynamics". *Proceedings of the 43rd International Conference on Machine Learning*. Vol. 306. Proceedings of Machine Learning Research, pp. 60523–60567. URL: https://proceedings.mlr.press/v306/koubbi26a.html (cit. on p. 10).
C. Kuehn and J. Yoon (2026). *Spectral Selection in Symmetric Self-Attention Dynamics*. arXiv: 2604.26085 [math.DS]. URL: https://arxiv.org/abs/2604.26085v1 (cit. on p. 1).
V. Lad, J. H. Lee, W. Gurnee, and M. Tegmark (2025). "Remarkable Robustness of LLMs: Stages of Inference?" *Advances in Neural Information Processing Systems*. Vol. 38, pp. 144216–144249. URL: https://openreview.net/forum?id=Wxh5Xz7NpJ (cit. on p. 10).
J.-L. Lions and E. Magenes (1972). *Non-Homogeneous Boundary Value Problems and Applications*. Vol. I. Grundlehren der mathematischen Wissenschaften 181. Translated from the French by P. Kenneth. URL: https://doi.org/10.1007/978-3-642-65161-8 (cit. on p. 46).
Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-y. Liu (2020). "Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View". *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*. URL: https://openreview.net/forum?id=pxlqJa21C (cit. on pp. 1, 3).
S. Ma, X. Sun, Y. Wang, and J. Lin (2018). "Bag-of-Words as Target for Neural Machine Translation". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 332–338. URL: https://aclanthology.org/P18-2053/ (cit. on p. 4).
P. Marion (2023). "Generalization bounds for neural ordinary differential equations and deep residual networks". *Advances in Neural Information Processing Systems*. Vol. 36, pp. 48918–48938. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/98ed250b203d1ac6b24bbcf263e3d4a7-Paper-Conference.pdf (cit. on p. 6).
S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama (2020). "Dissecting Neural ODEs". *Advances in Neural Information Processing Systems*. Vol. 33, pp. 3952–3963. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/293835c2cc75b585649498ee74b395f5-Paper.pdf (cit. on p. 6).
K. Meng, D. Bau, A. Andonian, and Y. Belinkov (2022). "Locating and Editing Factual Associations in GPT". *Advances in Neural Information Processing Systems*. Vol. 35, pp. 17359–17372. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf (cit. on p. 1).
K. Mun and M. Rosenzweig (2026). *Phase transitions for the noisy transformer model in arbitrary dimension*. arXiv: 2606.05140 [math.AP]. URL: https://arxiv.org/abs/2606.05140v1 (cit. on p. 10).
Y. Polyanskiy, P. Rigollet, and A. Yao (2025). *Synchronization of mean-field models on the circle*. arXiv: 2507.22857 [math.DS]. URL: https://arxiv.org/abs/2507.22857v1 (cit. on p. 2).
P. Rigollet (2026). "The Mean-Field Dynamics of Transformers". *Proceedings of the International Congress of Mathematicians*. Vol. 7, pp. 389–404. URL: https://doi.org/10.1137/25M1826342 (cit. on p. 4).
M. E. Sander, P. Ablin, M. Blondel, and G. Peyré (2022). "Sinkformers: Transformers with Doubly Stochastic Attention". *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*. Vol. 151. Proceedings of Machine Learning Research, pp. 3515–3530. URL: https://proceedings.mlr.press/v151/sander22a.html (cit. on pp. 1, 3–5).
A. Scagliotti (2023). "A Gradient Flow Equation for Optimal Control Problems With End-point Cost". *Journal of Dynamical and Control Systems* 29.2, pp. 521–568. URL: https://doi.org/10.1007/s10883-022-09604-2 (cit. on p. 6).
A. Shalova and A. Schlichting (2026). "Solutions of stationary McKean–Vlasov equation on a high-dimensional sphere and other Riemannian manifolds". *Advances in Nonlinear Analysis* 15.1, p. 20250141. URL: https://doi.org/10.1515/anona-2025-0141 (cit. on p. 10).

K. Shibata, K. Yano, R. Takahashi, J. Lee, W. Ikeda, and J. Suzuki (2026). "Suppressing Final Layer Hidden State Jumps in Transformer Pretraining". *Findings of the Association for Computational Linguistics: EACL*, pp. 1236–1262. URL: https://aclanthology.org/2026.findings-eacl.64/ (cit. on p. 10).

E. Trélat and E. Zuazua (2025). *Turnpike in optimal control and beyond: a survey*. arXiv: 2503.20342 [math.OC]. URL: https://arxiv.org/abs/2503.20342v1 (cit. on p. 10).

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention is All you Need". *Advances in Neural Information Processing Systems*. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 1).

C. Wendler, V. Veselovsky, G. Monea, and R. West (2024). "Do Llamas Work in English? On the Latent Language of Multilingual Transformers". *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394. URL: https://aclanthology.org/2024.acl-long.820/ (cit. on p. 10).

H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du (2024). "Explainability for Large Language Models: A Survey". *ACM Transactions on Intelligent Systems and Technology* 15.2, pp. 1–38. URL: https://doi.org/10.1145/3639372 (cit. on p. 1).

A Additional proofs

The constants that appear in the main text and in the proofs below are collected in Table 1.

Table 1: Constants used in the main results.

Symbol	Meaning	Defined in
$\lambda_1 > \lambda_2$	two largest eigenvalues of A, spectral gap $\lambda_1 - \lambda_2$	Assumption 3.1
$\bar{\varepsilon}$	small-noise threshold	Proposition 3.2
Δ_0	critical-level gap $\frac{1}{2}\min\{\sinh\lambda_1, \mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2}\}$	Proposition 3.2
C_σ	bound on the feature map σ	Assumption 3.3
λ_{reg}	regularization coefficient	(2.3)
$\mathfrak{E}_{\max}$	initial energy gap plus loss oscillation	Assumption 3.6
η	sublevel threshold defining the low-energy components $\mathcal{C}_\pm(\eta)$	Lemma 3.7
r_*	radius of the density charts $\mathcal{U}_\pm$	(3.6)
L_ε	entry time scale from both ends of $[0,T]$	Lemma 3.7
T^*, C, a	minimal depth, constant, and exponential rate of the turnpike bound	Theorem 3.9

A.1 Proof for Section 3.1

This appendix proves Proposition 3.2. The static small-noise theorem stated first constructs the two entropic wells $\overline{\mu}^{\pm}$ and establishes the gap Δ_0 between the minimal energy and every other critical level. Appendix A.1.1 proves the gap by passing to the small-temperature limit of the critical points, and Appendix A.1.2 completes the proof of the theorem by a Laplace-type expansion around each well that is uniform in ε. Appendices A.1.3 and A.1.4 compute the Hessian of $E_{\varepsilon,A}$ at $\overline{\mu}$ in the density chart and identify its leading scale, which gives the coercivity constant of the local PL inequality. Appendix A.1.5 relates this Hessian to the second variation along transport perturbations. The first static input is the following small-noise theorem. It constructs the two entropic wells and gives the small-temperature critical-level gap used in Proposition 3.2. The chord estimate is kept inside the package because it is the finite-ε mechanism behind one-well uniqueness.

Theorem A.1 (static small-noise package)**.** *Let*

$$m_\varepsilon := \inf_{\nu\in\mathcal{P}(\mathbb{S}^{d-1})} E_{\varepsilon,A}(\nu).$$

Set

$$Q_A := \max\{\mathrm{e}^{\lambda_2}, \cosh\lambda_1\}, \qquad \Delta_0 := \frac{1}{2}(\mathrm{e}^{\lambda_1} - Q_A) = \frac{1}{2}\min\{\mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2}, \sinh\lambda_1\} > 0.$$

There exist $\varepsilon_0 > 0$, disjoint narrow neighborhoods V_+ of δ_{e_1} and V_- of δ_{-e_1}, with $V_- = S_\# V_+$ for $S(x) = -x$, such that the following holds for $0 < \varepsilon < \varepsilon_0$.

(i) Every global minimizer of $E_{\varepsilon,A}$ lies in $V_+ \cup V_-$. There is a global minimizer $\overline{\mu}^+ \in V_+$, and $\overline{\mu}^- := S_\#\overline{\mu}^+$ is a global minimizer in V_-. We write $\overline{\mu}^+ = \overline{\rho}^+\omega$, $\overline{\mu}^- = \overline{\rho}^-\omega$, and, on the selected positive branch, $\overline{\mu} = \overline{\mu}^+$, $\overline{\rho} = \overline{\rho}^+$.

(ii) Let $\eta_\varepsilon \downarrow 0$. Suppose $\mu_\varepsilon = \rho_\varepsilon\omega$ is any family of critical points satisfying

$$E_{\varepsilon,A}(\mu_\varepsilon) \le m_\varepsilon + \eta_\varepsilon, \qquad \mu_\varepsilon \in V_+, \tag{A.1}$$

and hence the Euler–Lagrange equation

$$\varepsilon(\log\rho_\varepsilon + 1) - K_A * \mu_\varepsilon = \Lambda_\varepsilon.$$

Then there are constants $r, c, C > 0$, independent of the particular family, and centers $x_{\mu,\varepsilon} \in \mathbb{S}^{d-1}$ such that

$$x_{\mu,\varepsilon} = e_1 + O(\varepsilon), \tag{A.2}$$
$$\mu_\varepsilon(\mathbb{S}^{d-1} \setminus B_r(x_{\mu,\varepsilon})) \le C\mathrm{e}^{-c/\varepsilon},$$

and, with $u_\mu(x) = \exp^{-1}_{x_{\mu,\varepsilon}}(x)$ inside the cap and $u_\mu = 0$ outside,

$$\sigma_\varepsilon^2 = \frac{\varepsilon}{\mathrm{e}^{\lambda_1}\lambda_1},$$

the rescaled density converges to the standard Gaussian in weighted $L^1(\mathbb{R}^{d-1})$, uniformly over such families. In particular, for every unit tangent vector $\tau \in T_{x_{\mu,\varepsilon}}\mathbb{S}^{d-1}$,

$$\int_{\mathbb{S}^{d-1}} \langle u_\mu(x), \tau\rangle \, \mathrm{d}\mu_\varepsilon(x) = O(\varepsilon), \tag{A.3}$$

$$\int_{\mathbb{S}^{d-1}} \langle u_\mu(x), \tau\rangle^2 \, \mathrm{d}\mu_\varepsilon(x) = \sigma_\varepsilon^2 + o(\varepsilon),$$

and, for every integer $m \geq 0$,

$$\int_{\mathbb{S}^{d-1}} |u_\mu(x)|^m \mathrm{d}\mu_\varepsilon(x) = O(\varepsilon^{m/2}). \tag{A.4}$$

The same statement holds in V_- after applying the antipodal map.

For the selected branch $\overline{\mu} = \overline{\mu}^+$ we write $x_\varepsilon = x_{\overline{\mu},\varepsilon}$ and $u(x) = \exp_{x_\varepsilon}^{-1}(x)$ on the cap, extended by zero outside. Thus

$$\overline{\mu}(\mathbb{S}^{d-1} \setminus B_r(x_\varepsilon)) \leq C\mathrm{e}^{-c/\varepsilon}, \tag{A.5}$$

$$\int_{\mathbb{S}^{d-1}} \langle u(x), \tau\rangle \, \mathrm{d}\overline{\mu}(x) = O(\varepsilon), \tag{A.6}$$

$$\int_{\mathbb{S}^{d-1}} \langle u(x), \tau\rangle^2 \, \mathrm{d}\overline{\mu}(x) = \sigma_\varepsilon^2 + o(\varepsilon),$$

and

$$\int_{\mathbb{S}^{d-1}} |u(x)|^m \mathrm{d}\overline{\mu}(x) = O(\varepsilon^{m/2}). \tag{A.7}$$

(iii) Let $\eta_\varepsilon \downarrow 0$, and let $\mu_\varepsilon^0, \mu_\varepsilon^1 \in V_+$ be two low-energy critical point families satisfying (A.1), with μ_ε replaced by each μ_ε^i. Set

$$\mu_\varepsilon^s = (1-s)\mu_\varepsilon^0 + s\mu_\varepsilon^1, \qquad s \in [0,1].$$

For $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ define $\Pi_\mu f := f - \int_{\mathbb{S}^{d-1}} f \, \mathrm{d}\mu$ and

$$\mathsf{K}_\mu h := \Pi_\mu \left[\int_{\mathbb{S}^{d-1}} K_A(\cdot, y) h(y) \, \mathrm{d}\mu(y) \right], \qquad h \in L_0^2(\mu).$$

Then, uniformly in $s \in [0,1]$ and in the pair $\mu_\varepsilon^0, \mu_\varepsilon^1$,

$$\left\| \mathsf{K}_{\mu_\varepsilon^s} \right\|_{\mathrm{op}, L_0^2(\mu_\varepsilon^s)} \leq \left(\frac{\lambda_2}{\lambda_1} + o(1) \right) \varepsilon. \tag{A.8}$$

The same estimate holds for chords of low-energy critical points in V_-.

(iv) For every sufficiently small fixed $\varepsilon > 0$, there exists $\eta_{\mathrm{crit}} = \eta_{\mathrm{crit}}(\varepsilon) > 0$ such that each well contains at most one critical point μ satisfying

$$E_{\varepsilon,A}(\mu) \leq m_\varepsilon + \eta_{\mathrm{crit}}.$$

In particular, the global minimizer set consists of exactly two antipodal branches:

$$\{\overline{\mu}^\pm\} := \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{S}^{d-1})} E_{\varepsilon,A}(\mu) = \{\overline{\mu}^+, \overline{\mu}^-\}, \qquad \overline{\mu}^- = S_\# \overline{\mu}^+.$$

A.1.1 Gibbs-limit proof of the critical-level gap

We record the Gibbs-limit facts needed for the non-minimal critical-level gap in Proposition 3.2. The point is that limits of positive-temperature Gibbs critical points satisfy a global max-support condition coming from the Laplace principle.

Throughout this appendix, we write $K_A(x,y) := \mathrm{e}^{\langle x, Ay\rangle}$ and

$$I_A(\mu) := \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} K_A(x,y) \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y), \qquad W_\mu(x) := (K_A * \mu)(x) = \int_{\mathbb{S}^{d-1}} K_A(x,y) \, \mathrm{d}\mu(y).$$

Definition A.2 (positive-temperature critical set and its zero-temperature limits)**.** For the fixed noise level ε, define

$$\operatorname{Crit} E_{\varepsilon,A} \coloneqq \left\{ \mu = \rho\omega \in \mathcal{P}(\mathbb{S}^{d-1}) \;\middle|\; \rho > 0, \rho(x) = \frac{\exp(W_\mu(x)/\varepsilon)}{\int_{\mathbb{S}^{d-1}} \exp(W_\mu(z)/\varepsilon)\,\mathrm{d}\omega(z)} \right\}.$$

For a varying noise level ε_n, the corresponding critical set is denoted by $\operatorname{Crit} E_{\varepsilon_n,A}$. Equivalently,

$$\varepsilon(\log\rho + 1) - K_A * \mu \quad \text{is constant on } \mathbb{S}^{d-1}.$$

We define the zero-temperature Gibbs-limit set by

$$\mathfrak{G}_A^{\lim} \coloneqq \left\{ \nu \in \mathcal{P}(\mathbb{S}^{d-1}) \;\middle|\; \exists\, \varepsilon_n \downarrow 0,\ \exists\, \mu_n \in \operatorname{Crit} E_{\varepsilon_n,A} \text{ such that } \mu_n \rightharpoonup \nu \right\}.$$

We also define the explicit max-support outer class

$$\mathfrak{G}_A^{\max} \coloneqq \left\{ \nu \in \mathcal{P}(\mathbb{S}^{d-1}) \;\middle|\; \operatorname{supp}\nu \subset \operatorname*{arg\,max}_{x\in\mathbb{S}^{d-1}} W_\nu(x) \right\}.$$

Thus $\mathfrak{G}_A^{\lim}$ is the intrinsic object, while $\mathfrak{G}_A^{\max}$ is a tractable necessary outer condition.

If $\nu \in \mathfrak{G}_A^{\max}$, then

$$W_\nu(x) = \max_{\mathbb{S}^{d-1}} W_\nu \qquad \nu\text{-a.e.}$$

and hence

$$I_A(\nu) = \int_{\mathbb{S}^{d-1}} W_\nu \,\mathrm{d}\nu = \max_{\mathbb{S}^{d-1}} W_\nu.$$

Equivalently,

$$W_\nu \le I_A(\nu) \quad \text{on } \mathbb{S}^{d-1}, \qquad W_\nu = I_A(\nu) \quad \nu\text{-a.e.}$$

The converse also holds by continuity of W_ν.

Lemma A.3 (Gibbs limits are max-support states)**.** *One has*

$$\mathfrak{G}_A^{\lim} \subset \mathfrak{G}_A^{\max}.$$

Equivalently, if $\varepsilon_n \downarrow 0$, $\mu_n \in \operatorname{Crit} E_{\varepsilon_n,A}$, and $\mu_n \rightharpoonup \nu$, then

$$\operatorname{supp}\nu \subset \operatorname*{arg\,max}_{x\in\mathbb{S}^{d-1}} W_\nu(x).$$

Proof. Set

$$W_n \coloneqq W_{\mu_n}, \qquad W \coloneqq W_\nu.$$

Since $K_A \in C(\mathbb{S}^{d-1}\times\mathbb{S}^{d-1})$ and $\mathbb{S}^{d-1}$ is compact, narrow convergence gives

$$\|W_n - W\|_{L^\infty(\mathbb{S}^{d-1})} \to 0. \tag{A.9}$$

Indeed, uniform continuity of K_A in the first variable reduces the claim to finitely many scalar narrow-convergence limits.

Let $W_{\max} \coloneqq \max_{\mathbb{S}^{d-1}} W$. Fix a point $x_0 \notin \arg\max_{\mathbb{S}^{d-1}} W$. Choose an open neighborhood O of x_0 and a number $\delta > 0$ such that

$$\sup_{\overline{O}} W \le W_{\max} - 3\delta.$$

Choose also an open set $U \subset \mathbb{S}^{d-1}$ with $\omega(U) > 0$ and

$$\inf_{U} W \ge W_{\max} - \delta.$$

For all sufficiently large n, (A.9) gives

$$\sup_{\overline{O}} W_n \le W_{\max} - 2\delta, \qquad \inf_{U} W_n \ge W_{\max} - \frac{3}{2}\delta.$$

Using the Gibbs representation of μ_n,

$$\mu_n(O) \le \mu_n(\overline{O}) \le \frac{\omega(\mathbb{S}^{d-1})\exp((W_{\max}-2\delta)/\varepsilon_n)}{\omega(U)\exp((W_{\max}-\frac{3}{2}\delta)/\varepsilon_n)} \le C_U \exp\left(-\frac{\delta}{2\varepsilon_n}\right) \to 0.$$

By the Portmanteau theorem for the open set O,

$$\nu(O) \le \liminf_{n\to\infty} \mu_n(O) = 0.$$

Thus every point outside $\arg\max_{\mathbb{S}^{d-1}} W$ has a neighborhood of zero ν-mass. This proves $\operatorname{supp}\nu \subset \arg\max_{\mathbb{S}^{d-1}} W$. ∎

Lemma A.4 (support-minimality inside the max-support class)**.** *Let $\nu \in \mathfrak{G}_A^{\max}$. Then, for every $\eta \in \mathcal{P}(\operatorname{supp}\nu)$,*

$$I_A(\nu) \le I_A(\eta). \tag{A.10}$$

Proof. Since $\nu \in \mathfrak{G}_A^{\max}$, we have $W_\nu = I_A(\nu)$ on $\operatorname{supp}\nu$. Set $\sigma := \eta - \nu$. Then $\sigma(\mathbb{S}^{d-1}) = 0$ and

$$\int_{\mathbb{S}^{d-1}} W_\nu \,\mathrm{d}\sigma = 0.$$

Therefore

$$I_A(\eta) = I_A(\nu+\sigma) = I_A(\nu) + 2\int_{\mathbb{S}^{d-1}} W_\nu \,\mathrm{d}\sigma + I_A(\sigma) = I_A(\nu) + I_A(\sigma).$$

It remains only to recall that K_A is positive definite. Indeed, for every finite signed measure σ, the uniformly absolutely convergent expansion on the compact sphere gives

$$I_A(\sigma) = \sum_{n=0}^{\infty} \frac{1}{n!} \iint (x^\top A y)^n \,\mathrm{d}\sigma(x)\,\mathrm{d}\sigma(y),$$

and

$$\iint (x^\top A y)^n \,\mathrm{d}\sigma(x)\,\mathrm{d}\sigma(y) = \left| \int_{\mathbb{S}^{d-1}} (A^{1/2}x)^{\otimes n} \,\mathrm{d}\sigma(x) \right|^2 \ge 0.$$

Thus $I_A(\sigma) \ge 0$, proving (A.10). ∎

Lemma A.5 (two-point bound across the equator)**.** *Let $x, y \in \mathbb{S}^{d-1}$ satisfy $x_1 y_1 \le 0$, where $x_1 = x \cdot e_1$ and $y_1 = y \cdot e_1$. Then*

$$\min_{t\in[0,1]} I_A(t\delta_x + (1-t)\delta_y) \le \max\{\mathrm{e}^{\lambda_2}, \cosh\lambda_1\}.$$

Proof. After interchanging x and y, we may write

$$x_1 = \alpha \ge 0, \qquad y_1 = -\beta \le 0, \qquad \alpha, \beta \in [0,1].$$

Set

$$s := \sqrt{1-\alpha^2}, \qquad r := \sqrt{1-\beta^2}.$$

Then

$$x^\top A x \le \lambda_2 + (\lambda_1 - \lambda_2)\alpha^2 =: a, \qquad y^\top A y \le \lambda_2 + (\lambda_1 - \lambda_2)\beta^2 =: c,$$

and, by Cauchy–Schwarz in the orthogonal complement of e_1,

$$x^\top A y \le -\lambda_1 \alpha\beta + \lambda_2 sr =: b.$$

Hence

$$I_A(t\delta_x + (1-t)\delta_y) \le t^2 \mathrm{e}^a + 2t(1-t)\mathrm{e}^b + (1-t)^2 \mathrm{e}^c.$$

If $\alpha = \beta = 0$, then $a, c, b \le \lambda_2$, so the claim follows. Assume $\alpha + \beta > 0$, and choose the trial weight

$$t_0 := \frac{\beta}{\alpha+\beta}, \qquad 1 - t_0 = \frac{\alpha}{\alpha+\beta}.$$

This weight cancels the top coordinate. Thus it is enough to show

$$\frac{\beta^2 \mathrm{e}^a + \alpha^2 \mathrm{e}^c + 2\alpha\beta \mathrm{e}^b}{(\alpha+\beta)^2} \le (1-\alpha\beta)\mathrm{e}^{\lambda_2} + \alpha\beta\cosh\lambda_1. \tag{A.11}$$

By convexity of the exponential,

$$\mathrm{e}^a \le (1-\alpha^2)\mathrm{e}^{\lambda_2} + \alpha^2 \mathrm{e}^{\lambda_1}, \qquad \mathrm{e}^c \le (1-\beta^2)\mathrm{e}^{\lambda_2} + \beta^2 \mathrm{e}^{\lambda_1}.$$

Moreover $sr \le (s^2 + r^2)/2$, so

$$b \le \bar{b} := -\lambda_1 \alpha\beta + \frac{\lambda_2}{2}(s^2 + r^2).$$

With

$$\omega_0 := \frac{s^2+r^2}{2}, \qquad \omega_+ := \frac{(\alpha-\beta)^2}{4}, \qquad \omega_- := \frac{(\alpha+\beta)^2}{4},$$

one has $\omega_0+\omega_++\omega_-=1$ and

$$\bar{b} = \omega_0\lambda_2 + \omega_+\lambda_1 + \omega_-(-\lambda_1).$$

Hence

$$\mathrm{e}^{b} \le \omega_0 \mathrm{e}^{\lambda_2} + \omega_+ \mathrm{e}^{\lambda_1} + \omega_- \mathrm{e}^{-\lambda_1}.$$

Substituting these three estimates into the numerator in (A.11), the coefficient of e^{λ_2} is

$$\beta^2(1-\alpha^2) + \alpha^2(1-\beta^2) + \alpha\beta(s^2+r^2) = (\alpha+\beta)^2(1-\alpha\beta),$$

the coefficient of e^{λ_1} is

$$2\alpha^2\beta^2 + 2\alpha\beta\omega_+ = \frac{\alpha\beta}{2}(\alpha+\beta)^2,$$

and the coefficient of $\mathrm{e}^{-\lambda_1}$ is

$$2\alpha\beta\omega_- = \frac{\alpha\beta}{2}(\alpha+\beta)^2.$$

This proves (A.11). Since the right-hand side of (A.11) is a convex combination of e^{λ_2} and $\cosh\lambda_1$, it is bounded by $\max\{\mathrm{e}^{\lambda_2}, \cosh\lambda_1\}$. This proves the lemma. ■

Proposition A.6 (sharp first-excited max-support bound)**.** *One has*

$$\sup_{\nu \in \mathfrak{G}_A^{\max} \setminus \{\delta_{e_1}, \delta_{-e_1}\}} I_A(\nu) = Q_A := \max\{\mathrm{e}^{\lambda_2}, \cosh\lambda_1\}.$$

In particular $Q_A < \mathrm{e}^{\lambda_1}$.

Proof. We first prove the upper bound. Let $\nu \in \mathfrak{G}_A^{\max} \setminus \{\delta_{e_1}, \delta_{-e_1}\}$.

Suppose first that $\operatorname{supp}\nu \subset \{x_1 \ge 0\}$. Let $x \in \operatorname{supp}\nu$ minimize x_1 on the support and set $c := x_1$. If $0 < c < 1$, define

$$v := \frac{e_1 - cx}{\sqrt{1-c^2}} \in T_x\mathbb{S}^{d-1}.$$

For every $y \in \operatorname{supp}\nu$, one has $y_1 \ge c$. Also

$$x^\top A y \le \lambda_1 c y_1 + \lambda_2\sqrt{1-c^2}\sqrt{1-y_1^2}.$$

Consequently

$$(e_1 - cx)^\top A y \ge \lambda_1(1-c^2)y_1 - c\lambda_2\sqrt{1-c^2}\sqrt{1-y_1^2} > 0,$$

where the strict inequality follows from $y_1 \ge c > 0$ and $\lambda_1 > \lambda_2$. Hence

$$\partial_v W_\nu(x) = \int_{\mathbb{S}^{d-1}} \mathrm{e}^{x^\top A y} v^\top A y \,\mathrm{d}\nu(y) > 0,$$

contradicting $x \in \arg\max_{\mathbb{S}^{d-1}} W_\nu$. Therefore either $c = 1$, in which case $\nu = \delta_{e_1}$, or $c = 0$. Since $\nu \ne \delta_{e_1}$, we must have $c = 0$. Thus there is $x \in \operatorname{supp}\nu$ with $x_1 = 0$. By Lemma A.4,

$$I_A(\nu) \le I_A(\delta_x) = \mathrm{e}^{x^\top A x} \le \mathrm{e}^{\lambda_2}.$$

The case $\operatorname{supp}\nu \subset \{x_1 \le 0\}$ is identical and gives $I_A(\nu) \le \mathrm{e}^{\lambda_2}$, unless $\nu = \delta_{-e_1}$, which is excluded.

It remains to consider the case where the support meets both closed hemispheres. Then there are $x, y \in \operatorname{supp}\nu$ such that $x_1 y_1 \le 0$. By Lemma A.4, for every $t \in [0,1]$,

$$I_A(\nu) \le I_A(t\delta_x + (1-t)\delta_y).$$

Taking the minimum over t and applying Lemma A.5 yields

$$I_A(\nu) \le Q_A.$$

This proves the upper bound.

For sharpness, δ_{e_2} and δ_{-e_2} belong to $\mathfrak{G}_A^{\max}$, and

$$I_A(\delta_{\pm e_2}) = \mathrm{e}^{\lambda_2}.$$

Also

$$\sigma_1 := \frac{1}{2}(\delta_{e_1} + \delta_{-e_1}) \in \mathfrak{G}_A^{\max}, \qquad I_A(\sigma_1) = \cosh \lambda_1.$$

Thus the supremum is exactly Q_A. Finally, $\mathrm{e}^{\lambda_2} < \mathrm{e}^{\lambda_1}$ and $\cosh \lambda_1 < \mathrm{e}^{\lambda_1}$, so $Q_A < \mathrm{e}^{\lambda_1}$. ■

Proposition A.7 (explicit small-temperature excited critical gap)**.** *Let*

$$\Delta_{\mathrm{crit}}(\varepsilon) := \inf\left\{\Delta_{\varepsilon,A}(\mu) : \mu \in \mathrm{Crit}\, E_{\varepsilon,A} \setminus \left\{\overline{\mu}^{\pm}\right\}\right\}.$$

Then

$$\liminf_{\varepsilon \downarrow 0} \Delta_{\mathrm{crit}}(\varepsilon) \geq \Delta_0 := \frac{1}{2}(\mathrm{e}^{\lambda_1} - Q_A). \tag{A.12}$$

Equivalently, for every $\gamma > 0$, there exists $\varepsilon_\gamma > 0$ such that, for $0 < \varepsilon < \varepsilon_\gamma$,

$$\mu \in \mathrm{Crit}\, E_{\varepsilon,A} \setminus \left\{\overline{\mu}^{\pm}\right\} \quad \Longrightarrow \quad \Delta_{\varepsilon,A}(\mu) \geq \Delta_0 - \gamma.$$

Proof. Let $\varepsilon_n \downarrow 0$ and let $\mu_n \in \mathrm{Crit}\, E_{\varepsilon_n,A}$ be critical points that are not global minimizers of $E_{\varepsilon_n,A}$. Passing to a subsequence, assume $\mu_n \rightharpoonup \nu$. By Lemma A.3, $\nu \in \mathfrak{G}_A^{\max}$.

If $\nu = \delta_{e_1}$ or $\nu = \delta_{-e_1}$, then the Gibbs representation and standard Laplace estimates imply $E_{\varepsilon_n,A}(\mu_n) \to -\frac{1}{2}\mathrm{e}^{\lambda_1}$. Since $m_{\varepsilon_n} \to -\frac{1}{2}\mathrm{e}^{\lambda_1}$, we get $\Delta_{\varepsilon_n,A}(\mu_n) \to 0$. The one-well low-energy critical uniqueness in Theorem A.1 then forces μ_n to be the corresponding minimizer for all large n, contrary to the choice of μ_n. Thus a non-minimal critical sequence cannot converge to either ground Dirac.

Therefore $\nu \notin \{\delta_{e_1}, \delta_{-e_1}\}$. By Proposition A.6, $I_A(\nu) \leq Q_A$. The test-cap construction in the proof of Theorem A.1 gives

$$m_{\varepsilon_n} \leq -\frac{1}{2}\mathrm{e}^{\lambda_1} + C\varepsilon_n|\log \varepsilon_n|.$$

Using $\varepsilon_n \operatorname{Ent}(\mu_n) \geq 0$, we obtain

$$\Delta_{\varepsilon_n,A}(\mu_n) = E_{\varepsilon_n,A}(\mu_n) - m_{\varepsilon_n} \geq \frac{1}{2}(\mathrm{e}^{\lambda_1} - I_A(\mu_n)) - C\varepsilon_n|\log \varepsilon_n|.$$

Since $I_A(\mu_n) \to I_A(\nu)$,

$$\liminf_{n\to\infty} \Delta_{\varepsilon_n,A}(\mu_n) \geq \frac{1}{2}(\mathrm{e}^{\lambda_1} - I_A(\nu)) \geq \frac{1}{2}(\mathrm{e}^{\lambda_1} - Q_A) = \Delta_0.$$

Taking the infimum over all such sequences proves (A.12). ■

A.1.2 Proof of the static small-noise package

We give the finite-ε argument behind Theorem A.1. The zero-temperature interaction only supplies a separation estimate; the actual one-well uniqueness comes from an Euler–Lagrange difference identity and the same-well chord estimate (A.8).

Proof of Theorem A.1. **1. Direct method and Euler–Lagrange equation.** Since $\mathbb{S}^{d-1}$ is compact, $\mathcal{P}(\mathbb{S}^{d-1})$ is narrowly compact. The entropy is narrowly lower semicontinuous and I_A is narrowly continuous because K_A is continuous and bounded. Hence $E_{\varepsilon,A}$ admits a global minimizer. Every finite-energy minimizer, and more generally every interior critical point, has a smooth positive density ρ and satisfies

$$\varepsilon(\log \rho(x) + 1) - W_\mu(x) = \Lambda_\mu.$$

Equivalently,

$$\rho(x) = \frac{\exp(W_\mu(x)/\varepsilon)}{\int_{\mathbb{S}^{d-1}} \exp(W_\mu(z)/\varepsilon)\, \mathrm{d}\omega(z)}.$$

Smoothness and positivity follow immediately from the smooth positive kernel in this fixed-point formula.

2. Zero-temperature separation. For $x, y \in \mathbb{S}^{d-1}$,

$$x^\top A y \le \lambda_1.$$

Because the top eigenvalue is simple, equality holds only for $(x, y) = (e_1, e_1)$ or $(x, y) = (-e_1, -e_1)$. Thus

$$I_A(\mu) \le \mathrm{e}^{\lambda_1},$$

and equality holds only for $\mu = \delta_{e_1}$ or $\mu = \delta_{-e_1}$.

Let $r_\varepsilon = \sqrt{\varepsilon|\log\varepsilon|}$ and let ν_ε be the normalized surface measure on $B_{r_\varepsilon}(e_1)$. Then

$$\mathrm{Ent}(\nu_\varepsilon) = O(|\log\varepsilon|), \qquad I_A(\nu_\varepsilon) = \mathrm{e}^{\lambda_1} + O(\varepsilon|\log\varepsilon|),$$

and hence

$$m_\varepsilon \le -\frac{1}{2}\mathrm{e}^{\lambda_1} + C\varepsilon|\log\varepsilon|. \tag{A.13}$$

If a family μ_ε satisfies $E_{\varepsilon,A}(\mu_\varepsilon) \le m_\varepsilon + \eta_\varepsilon$ with $\eta_\varepsilon \to 0$, then $\mathrm{Ent} \ge 0$ and (A.13) give

$$I_A(\mu_\varepsilon) \ge \mathrm{e}^{\lambda_1} - C\varepsilon|\log\varepsilon| - 2\eta_\varepsilon.$$

Every narrow limit point of μ_ε must therefore be one of $\delta_{e_1}, \delta_{-e_1}$. Fix disjoint narrow neighborhoods V_+ and V_- of these two measures, with $V_- = S_\# V_+$. Then, for small ε, every global minimizer lies in $V_+ \cup V_-$. By the direct method there is at least one minimizer; reflecting by S if necessary gives a minimizer $\overline{\mu}^+ \in V_+$, and $\overline{\mu}^- := S_\#\overline{\mu}^+ \in V_-$ is also a minimizer.

The same argument applies to any low-energy critical family satisfying (A.1): after restricting to V_+, it converges narrowly to δ_{e_1}. This convergence is uniform over all such families, in the usual contradiction sense.

3. Uniform one-well Laplace estimates for critical points. The last paragraph of Step 2 gives uniform narrow convergence to δ_{e_1} for all low-energy critical families in V_+. Fix such a family $\mu_\varepsilon = \rho_\varepsilon\omega$ satisfying (A.1). Since all x-derivatives of $K_A(x, y)$ are continuous and bounded on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, the uniform narrow convergence to δ_{e_1} gives

$$W_{\mu_\varepsilon} \to W_0, \qquad W_0(x) = \mathrm{e}^{\lambda_1 x_1},$$

in $C^4(\mathbb{S}^{d-1})$, uniformly over the family. The function W_0 has a unique strict global maximum at e_1, and

$$-\operatorname{Hess}_{\mathbb{S}^{d-1}} W_0(e_1) = a_0 I_{T_{e_1}\mathbb{S}^{d-1}}, \qquad a_0 := \mathrm{e}^{\lambda_1}\lambda_1.$$

Therefore, for a fixed small cap and all small ε, W_{μ_ε} has a unique maximizer $x_{\mu,\varepsilon}$ in that cap, this maximizer is global, and $x_{\mu,\varepsilon} \to e_1$. Moreover, with $u \in T_{x_{\mu,\varepsilon}}\mathbb{S}^{d-1}$,

$$W_{\mu_\varepsilon}(x_{\mu,\varepsilon}) - W_{\mu_\varepsilon}(\exp_{x_{\mu,\varepsilon}} u) = \frac{1}{2}\langle A_{\mu,\varepsilon}u, u\rangle + R_{\mu,\varepsilon}(u), \tag{A.14}$$

where

$$A_{\mu,\varepsilon} = a_0 I + o(1), \qquad |R_{\mu,\varepsilon}(u)| \le C|u|^3$$

uniformly. In particular, for some fixed $c_1, c_2 > 0$,

$$c_1|u|^2 \le W_{\mu_\varepsilon}(x_{\mu,\varepsilon}) - W_{\mu_\varepsilon}(\exp_{x_{\mu,\varepsilon}} u) \le c_2|u|^2$$

inside the cap, and there is a fixed gap outside the cap.

Subtracting the Euler–Lagrange equation at the maximizer and at $\exp_{x_{\mu,\varepsilon}} u$ gives the centered Gibbs representation

$$\rho_\varepsilon(\exp_{x_{\mu,\varepsilon}} u) = \rho_\varepsilon(x_{\mu,\varepsilon}) \exp\left(-\frac{W_{\mu_\varepsilon}(x_{\mu,\varepsilon}) - W_{\mu_\varepsilon}(\exp_{x_{\mu,\varepsilon}} u)}{\varepsilon}\right). \tag{A.15}$$

The quadratic bounds imply the peak-height estimate

$$C^{-1}\varepsilon^{-(d-1)/2} \le \rho_\varepsilon(x_{\mu,\varepsilon}) \le C\varepsilon^{-(d-1)/2},$$

and the fixed outside gap gives

$$\mu_\varepsilon(\mathbb{S}^{d-1} \setminus B_r(x_{\mu,\varepsilon})) \le C\mathrm{e}^{-c/\varepsilon}.$$

After the scaling $u = \sigma_\varepsilon \xi$, $\sigma_\varepsilon^2 = \varepsilon/a_0$, (A.14) and (A.15) give convergence of the scaled density to the standard Gaussian in weighted $L^1(\mathbb{R}^{d-1})$. The usual Laplace expansion also yields the first-moment refinement

$$\int u \, \mathrm{d}\mu_\varepsilon = O(\varepsilon),$$

because the leading Gaussian is even and the first non-even contribution is of order $\sqrt{\varepsilon}$ in the scaled variables. This proves the moment estimates (A.3)–(A.4).

It remains to prove the location (A.2) of the center. Since $x_{\mu,\varepsilon}$ maximizes W_{μ_ε},

$$0 = P^{\perp}_{x_{\mu,\varepsilon}} \int_{\mathbb{S}^{d-1}} \mathrm{e}^{x_{\mu,\varepsilon}^\top Ay} Ay \, \mathrm{d}\mu_\varepsilon(y). \tag{A.16}$$

Using the just-proved moments in coordinates centered at $x_{\mu,\varepsilon}$, the integral in (A.16) equals

$$\mathrm{e}^{x_{\mu,\varepsilon}^\top Ax_{\mu,\varepsilon}} Ax_{\mu,\varepsilon} + O(\varepsilon),$$

after tangent projection. Hence

$$P^{\perp}_{x_{\mu,\varepsilon}} Ax_{\mu,\varepsilon} = O(\varepsilon). \tag{A.17}$$

Write $x_{\mu,\varepsilon} = (\sqrt{1-|z|^2}, z)$ in the eigenbasis of A, with $z \in \mathbb{R}^{d-1}$. The tangent components of $P^{\perp}_x Ax$ near e_1 are

$$((\lambda_2 - \lambda_1)z_2, \dots, (\lambda_d - \lambda_1)z_d) + O(|z|^2).$$

Since $\lambda_1 > \lambda_2$, (A.17) gives $z = O(\varepsilon)$. Thus $x_{\mu,\varepsilon} = e_1 + O(\varepsilon)$. This proves Theorem A.1(ii) in V_+, and the statement in V_- follows by applying the antipodal map.

4. Chord operator estimate. Let $\mu_\varepsilon^0, \mu_\varepsilon^1 \in V_+$ be two low-energy critical point families satisfying (A.1), with μ_ε replaced by each μ_ε^i, and set $\mu_\varepsilon^s = (1-s)\mu_\varepsilon^0 + s\mu_\varepsilon^1$. By the previous step, the centers of μ_ε^0 and μ_ε^1 are both $e_1 + O(\varepsilon)$. Consequently, if $z(x) = \exp_{e_1}^{-1}(x)$ on a fixed cap and $z = 0$ outside, then uniformly in $s \in [0,1]$,

$$\mu_\varepsilon^s(\mathbb{S}^{d-1} \setminus B_r(e_1)) \le C\mathrm{e}^{-c/\varepsilon}, \qquad \int_{\mathbb{S}^{d-1}} z_i z_j \, \mathrm{d}\mu_\varepsilon^s = \sigma_\varepsilon^2 \delta_{ij} + o(\varepsilon), \tag{A.18}$$

$$\int_{\mathbb{S}^{d-1}} z_i \, \mathrm{d}\mu_\varepsilon^s = O(\varepsilon), \qquad \int_{\mathbb{S}^{d-1}} |z|^m \mathrm{d}\mu_\varepsilon^s = O(\varepsilon^{m/2}).$$

Here i, j range over the tangent coordinates corresponding to $e_2, \dots, e_d$.

Let $h \in L_0^2(\mu_\varepsilon^s)$. Since $\Pi_{\mu_\varepsilon^s}$ only removes constants,

$$\langle \mathsf{K}_{\mu_\varepsilon^s} h, h \rangle_{L^2(\mu_\varepsilon^s)} = \iint K_A(x,y) h(x) h(y) \, \mathrm{d}\mu_\varepsilon^s(x) \, \mathrm{d}\mu_\varepsilon^s(y).$$

On the fixed cap, with $x = \exp_{e_1} z$, $y = \exp_{e_1} w$, Taylor expansion at (e_1, e_1) gives

$$K_A(x,y) = \mathrm{e}^{\lambda_1} + \mathrm{e}^{\lambda_1} \sum_{j=2}^{d} \lambda_j z_j w_j + q_1(z) + q_2(w) + R(z,w), \tag{A.19}$$

where q_1 and q_2 depend on only one variable, and

$$|R(z,w)| \le C((|z|+|w|)^3 + |z|^2|w|^2)$$

for small z, w. The constant term and the one-variable terms vanish in the quadratic form because $\int h \, \mathrm{d}\mu_\varepsilon^s = 0$. The tail contribution and the remainder in (A.19) are $o(\varepsilon)\|h\|^2_{L^2(\mu_\varepsilon^s)}$ by (A.18) and Cauchy–Schwarz. Therefore

$$\langle \mathsf{K}_{\mu_\varepsilon^s} h, h \rangle_{L^2(\mu_\varepsilon^s)} = \mathrm{e}^{\lambda_1} \sum_{j=2}^{d} \lambda_j \left(\int_{\mathbb{S}^{d-1}} z_j h \, \mathrm{d}\mu_\varepsilon^s \right)^2 + o(\varepsilon)\|h\|^2_{L^2(\mu_\varepsilon^s)}. \tag{A.20}$$

By (A.18), the Gram matrix of $\sigma_\varepsilon^{-1} z_j$ in $L^2(\mu_\varepsilon^s)$ is $I + o(1)$. Hence

$$\sum_{j=2}^{d} \left(\sigma_\varepsilon^{-1} \int_{\mathbb{S}^{d-1}} z_j h \, \mathrm{d}\mu_\varepsilon^s \right)^2 \le (1 + o(1))\|h\|^2_{L^2(\mu_\varepsilon^s)}. \tag{A.21}$$

Combining (A.20) and (A.21), and using $\sigma_\varepsilon^2 = \varepsilon/(\mathrm{e}^{\lambda_1}\lambda_1)$, gives

$$\langle \mathsf{K}_{\mu_\varepsilon^s} h, h\rangle \le \left(\frac{\lambda_2}{\lambda_1} + o(1)\right) \varepsilon \|h\|^2_{L^2(\mu_\varepsilon^s)}.$$

The kernel $\mathrm{e}^{x^\top Ay}$ is positive definite because it is $\mathrm{e}^{\langle A^{1/2}x, A^{1/2}y\rangle}$. Thus $\mathsf{K}_{\mu_\varepsilon^s}$ is nonnegative self-adjoint on $L^2_0(\mu_\varepsilon^s)$, and the last Rayleigh bound is exactly (A.8). The negative well is obtained by applying the antipodal map.

5. One-well uniqueness of critical points. Let $\mu^0 = \rho^0\omega$ and $\mu^1 = \rho^1\omega$ be two low-energy critical points in the same positive well. Their Euler–Lagrange equations imply

$$\varepsilon(\log\rho^1 - \log\rho^0) - K_A * (\mu^1 - \mu^0) = c$$

for a constant c. Put $\nu = \mu^1 - \mu^0$, $\mu^s = (1-s)\mu^0 + s\mu^1$, and $h_s = \mathrm{d}\nu/\mathrm{d}\mu^s$. Since $\nu(\mathbb{S}^{d-1}) = 0$, multiplying the last equation by ν gives

$$0 = \varepsilon \int_{\mathbb{S}^{d-1}} (\log\rho^1 - \log\rho^0)\,\mathrm{d}\nu - \iint K_A(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y).$$

Moreover,

$$\int_{\mathbb{S}^{d-1}} (\log\rho^1 - \log\rho^0)\,\mathrm{d}\nu = \int_0^1 \int_{\mathbb{S}^{d-1}} h_s^2\,\mathrm{d}\mu^s\,\mathrm{d}s,$$

and

$$\iint K_A\,\mathrm{d}\nu\,\mathrm{d}\nu = \langle \mathsf{K}_{\mu^s} h_s, h_s\rangle_{L^2(\mu^s)} \qquad (0 \le s \le 1).$$

Integrating the latter identity in s and using the chord estimate yields, for small ε,

$$0 \ge \left(1 - \frac{\lambda_2}{\lambda_1} + o(1)\right) \varepsilon \int_0^1 \|h_s\|^2_{L^2(\mu^s)}\,\mathrm{d}s.$$

Since $\lambda_2 < \lambda_1$, this forces $h_s = 0$ for a.e. s, hence $\nu = 0$ and $\mu^0 = \mu^1$. The same proof applies in V_-.

The preceding argument is an asymptotic uniqueness statement for critical families with $E_{\varepsilon,A} \le m_\varepsilon + o(1)$. Equivalently, after fixing ε sufficiently small, there is a positive threshold $\eta_{\mathrm{crit}}(\varepsilon)$ such that each static well contains at most one critical point satisfying

$$E_{\varepsilon,A}(\mu) \le m_\varepsilon + \eta_{\mathrm{crit}}(\varepsilon).$$

Indeed, otherwise one could choose a sequence $\varepsilon_n \downarrow 0$ and two distinct same-well critical points with excess energy tending to zero, and the argument above would force them to coincide for large n.

Finally, Step 2 gives at least one minimizer in one of the wells and symmetry gives its antipodal partner. Since minimizers have zero excess energy, the one-well low-energy critical uniqueness gives at most one minimizer in each well. Therefore the global minimizer set is exactly $\{\overline{\mu}^+, \overline{\mu}^-\}$, as claimed.

■

A.1.3 Hessian operator

For the selected branch define the mean-zero projection $\Pi_0^{\overline{\mu}} f := f - \int_{\mathbb{S}^{d-1}} f\,\mathrm{d}\overline{\mu}$ and

$$(\mathsf{K}h)(x) := \Pi_0^{\overline{\mu}}\left[\int_{\mathbb{S}^{d-1}} \mathrm{e}^{x^\top Ay} h(y)\,\mathrm{d}\overline{\mu}(y)\right], \qquad \mathsf{H} = \varepsilon I - \mathsf{K}. \tag{A.22}$$

A.1.4 Hessian operator-norm asymptotics

The next lemma is the exact reduction behind the top scale.

Lemma A.8 (first reduction of the kernel form)**.** *Assume Theorem A.1. Define the tangent compression of A at x_ε by*

$$B_\varepsilon := P_{x_\varepsilon} A P_{x_\varepsilon}|_{T_{x_\varepsilon}\mathbb{S}^{d-1}},$$

and write its eigenpairs as

$$\nu_{\varepsilon,1} \geq \cdots \geq \nu_{\varepsilon,d-1} > 0, \qquad B_\varepsilon \tau_{\varepsilon,m} = \nu_{\varepsilon,m}\tau_{\varepsilon,m}.$$

Then $\nu_{\varepsilon,m} \to \lambda_{m+1}$ for $m = 1, \dots, d-1$. Moreover, if

$$\mathfrak{m}_{\varepsilon,m}(h) := \sigma_\varepsilon^{-1} \int_{\mathbb{S}^{d-1}} \langle u(x), \tau_{\varepsilon,m}\rangle h(x)\, \mathrm{d}\overline{\mu}(x), \qquad h \in L^2_0(\overline{\mu}),$$

then, uniformly for $\|h\|_{L^2(\overline{\mu})} \leq 1$,

$$\frac{1}{\varepsilon}\langle \mathsf{K}h, h\rangle_{L^2(\overline{\mu})} = \frac{1}{\lambda_1}\sum_{m=1}^{d-1} \nu_{\varepsilon,m}\, \mathfrak{m}_{\varepsilon,m}(h)^2 + o(1). \tag{A.23}$$

Proof. Since $x_\varepsilon \to e_1$, the compressed operators B_ε converge in operator norm to $A_\perp := P_{e_1} A P_{e_1}|_{T_{e_1}\mathbb{S}^{d-1}}$, whose eigenvalues are $\lambda_2, \dots, \lambda_d$. Hence $\nu_{\varepsilon,m} \to \lambda_{m+1}$.

For $h \in L^2_0(\overline{\mu})$, the projection disappears inside the quadratic form because $\int h\, \mathrm{d}\overline{\mu} = 0$, so

$$\langle \mathsf{K}h, h\rangle_{L^2(\overline{\mu})} = \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \mathrm{e}^{x^\top A y} h(x) h(y)\, \mathrm{d}\overline{\mu}(x)\, \mathrm{d}\overline{\mu}(y). \tag{A.24}$$

By (A.5), the part of the integral where at least one of (x, y) lies outside $B_r(x_\varepsilon)$ is $o(\varepsilon)\|h\|^2_{L^2(\overline{\mu})}$. On the cap, write

$$x = \exp_{x_\varepsilon}(u), \qquad y = \exp_{x_\varepsilon}(v), \qquad u, v \in T_{x_\varepsilon}\mathbb{S}^{d-1}.$$

A Taylor expansion at $(u, v) = (0, 0)$ gives

$$x^\top A y = x_\varepsilon^\top A x_\varepsilon + u^\top B_\varepsilon v + R_\varepsilon^{(1)}(u, v),$$

where $R_\varepsilon^{(1)}$ is a sum of terms depending only on u or only on v, a term of size $|P_{x_\varepsilon} A x_\varepsilon|\,|u|\,|v|$, and terms of total degree at least 3 in (u, v). Since $x_\varepsilon \to e_1$, one has $P_{x_\varepsilon} A x_\varepsilon \to 0$. Exponentiating once more yields

$$\mathrm{e}^{x^\top A y} = \mathrm{e}^{x_\varepsilon^\top A x_\varepsilon}\Big(1 + u^\top B_\varepsilon v\Big) + R_\varepsilon(u, v),$$

with

$$|R_\varepsilon(u, v)| \leq C\Big(|P_{x_\varepsilon} A x_\varepsilon|\,|u|\,|v| + (|u|+|v|)^3 + |u|^2|v|^2\Big).$$

The constant term disappears in (A.24) because h has mean zero; the terms depending only on one variable disappear for the same reason. Therefore

$$\langle \mathsf{K}h, h\rangle_{L^2(\overline{\mu})} = \mathrm{e}^{x_\varepsilon^\top A x_\varepsilon} \sum_{m=1}^{d-1} \nu_{\varepsilon,m} \left(\int_{\mathbb{S}^{d-1}} \langle u(x), \tau_{\varepsilon,m}\rangle h(x)\, \mathrm{d}\overline{\mu}(x)\right)^2 + o(\varepsilon)\|h\|^2_{L^2(\overline{\mu})}.$$

Finally, $\mathrm{e}^{x_\varepsilon^\top A x_\varepsilon} = \mathrm{e}^{\lambda_1} + o(1)$ and $\sigma_\varepsilon^2 = \varepsilon/(\mathrm{e}^{\lambda_1}\lambda_1)$, so dividing by ε gives (A.23). ∎

Theorem A.9 (exact Hessian top scale)**.** *Assume Theorem A.1. Then*

$$\|\mathsf{K}\|_{\mathrm{op}} = \frac{\lambda_2}{\lambda_1}\varepsilon + o(\varepsilon), \qquad \varepsilon \downarrow 0, \tag{A.25}$$

as an operator on $L^2_0(\overline{\mu})$. Consequently,

$$\langle \mathsf{H}h, h\rangle_{L^2(\overline{\mu})} \geq \left(\frac{\lambda_1 - \lambda_2}{\lambda_1}\varepsilon + o(\varepsilon)\right)\|h\|^2_{L^2(\overline{\mu})}, \qquad h \in L^2_0(\overline{\mu}). \tag{A.26}$$

Proof. Define the normalized transverse linear modes

$$\psi_{\varepsilon,m}(x) \coloneqq \sigma_\varepsilon^{-1}\langle u(x), \tau_{\varepsilon,m}\rangle, \qquad m = 1, \dots, d-1.$$

By (A.6)–(A.7),

$$\langle \psi_{\varepsilon,m}, \psi_{\varepsilon,n}\rangle_{L^2(\overline{\mu})} = \delta_{mn} + o(1).$$

Hence the Gram matrix of $\{\psi_{\varepsilon,1}, \dots, \psi_{\varepsilon,d-1}\}$ is $I + o(1)$, and therefore

$$\sum_{m=1}^{d-1} \mathfrak{m}_{\varepsilon,m}(h)^2 \le (1 + o(1))\|h\|^2_{L^2(\overline{\mu})}$$

uniformly for all $h \in L^2_0(\overline{\mu})$. Applying Lemma A.8 gives

$$\frac{1}{\varepsilon}\langle \mathsf{K}h, h\rangle_{L^2(\overline{\mu})} \le \frac{\nu_{\varepsilon,1}}{\lambda_1}(1 + o(1))\|h\|^2_{L^2(\overline{\mu})},$$

so

$$\|\mathsf{K}\|_{\mathrm{op}} \le \frac{\nu_{\varepsilon,1}}{\lambda_1}\varepsilon + o(\varepsilon) = \frac{\lambda_2}{\lambda_1}\varepsilon + o(\varepsilon).$$

For the matching lower bound, choose

$$h_{\varepsilon,1} \coloneqq \psi_{\varepsilon,1} - \int_{\mathbb{S}^{d-1}} \psi_{\varepsilon,1}\, \mathrm{d}\overline{\mu} \in L^2_0(\overline{\mu}).$$

Then $\|h_{\varepsilon,1}\|_{L^2(\overline{\mu})} = 1 + o(1)$, $\mathfrak{m}_{\varepsilon,1}(h_{\varepsilon,1}) = 1 + o(1)$, and $\mathfrak{m}_{\varepsilon,m}(h_{\varepsilon,1}) = o(1)$ for $m \neq 1$. Plugging this test function into (A.23) yields the reverse bound and proves (A.25). Since $\mathsf{H} = \varepsilon I - \mathsf{K}$ and $\mathsf{K} \ge 0$, (A.26) follows immediately. ∎

A.1.5 Relation to transport second variations

Theorem A.9 perturbs the minimizer through the density ratio, $\mu_s = (1 + sh)\overline{\mu}$, which is adapted to the $L^2(\overline{\mu})$ inner product: the second variation of the entropy is exactly the flat term $\varepsilon\|h\|^2_{L^2(\overline{\mu})}$, without curvature corrections. A transport perturbation $\mu_t = (\mathrm{id} + t\xi)_{\#}\overline{\mu}$ corresponds to $h = -\operatorname{div}(\overline{\rho}\xi)/\overline{\rho}$; the associated displacement second variation carries additional curvature and shape terms, while its coercivity content coincides with that of Theorem A.9.

Proof of Proposition 3.2. The Gibbs fixed-point equation and smooth positivity of the one-well minimizer follow from the Euler–Lagrange equation in the direct-method part of the proof of Theorem A.1. The one-well uniqueness and antipodal symmetry give

$$\{\overline{\mu}^{\pm}\} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{S}^{d-1})} E_{\varepsilon,A}(\mu) = \{\overline{\mu}^{+}, \overline{\mu}^{-}\}, \qquad \overline{\mu}^{-} = S_{\#}\overline{\mu}^{+}.$$

The Hessian at either branch is $\mathsf{H} = \varepsilon I - \mathsf{K}$ on the mean-zero space. By Theorem A.9,

$$\|\mathsf{K}\|_{\mathrm{op}} = \frac{\lambda_2}{\lambda_1}\varepsilon + o(\varepsilon).$$

After decreasing the small-noise threshold, this gives

$$\varepsilon\|h\|^2_{L^2(\overline{\mu})} \ge \langle h, \mathsf{H}h\rangle_{L^2(\overline{\mu})} \ge \frac{\varepsilon}{2}\left(1 - \frac{\lambda_2}{\lambda_1}\right)\|h\|^2_{L^2(\overline{\mu})}, \qquad h \in L^2_0(\overline{\mu}).$$

Finally, Proposition A.7 gives

$$\liminf_{\varepsilon \downarrow 0} \Delta_{\mathrm{crit}}(\varepsilon) \ge \frac{1}{2}(\mathrm{e}^{\lambda_1} - Q_A) = \frac{1}{2}\min\{\mathrm{e}^{\lambda_1} - \mathrm{e}^{\lambda_2}, \sinh\lambda_1\}.$$

This is exactly the asserted landscape statement. ∎

A.2 Proof for Section 3.2

This appendix proves Lemma 3.4 and Proposition 3.5. Appendix A.2.1 fixes the operator notation and the sign convention of the adjoint equation, and Appendix A.2.2 derives the one-step gradient-descent update of Lemma 3.4. Appendix A.2.3 introduces the finite-time Gramian and the visibility factor, which measure how much of the terminal gradient reaches layer t through the backward dynamics. Appendices A.2.4 and A.2.5 prove local Taylor expansions of the energy gap and of the first variation in the density chart and a small-learning-rate estimate for the nonlinear response of the state. Appendix A.2.6 combines these ingredients with a Rayleigh–Jensen lower bound for the costate to prove Proposition 3.5.

A.2.1 Operator notation and sign convention

Throughout this appendix we work on the positive minimizer branch $\overline{\mu} = \overline{\rho}\,\omega$ supplied by Proposition 3.2. We write

$$\Pi_0^{\overline{\mu}} f := f - \int_{\mathbb{S}^{d-1}} f \,\mathrm{d}\overline{\mu}$$

for the orthogonal projection of $L^2(\overline{\mu})$ onto the mean-zero subspace $L^2_0(\overline{\mu})$. When no confusion is possible we write Π_0 for $\Pi_0^{\overline{\mu}}$.

The Hessian kernel and the reduced Hessian at $\overline{\mu}$ are

$$\begin{aligned}(\mathsf{K}h)(x) &:= \Pi_0\left(\int_{\mathbb{S}^{d-1}} \mathrm{e}^{\langle x, Ay\rangle} h(y)\,\mathrm{d}\overline{\mu}(y)\right),\\ \mathsf{H}h &:= \varepsilon h - \mathsf{K}h, \qquad h \in L^2_0(\overline{\mu}).\end{aligned} \tag{A.27}$$

By (3.2) in Proposition 3.2, after decreasing the small-noise threshold if necessary,

$$\frac{\varepsilon}{2}\left(1 - \frac{\lambda_2}{\lambda_1}\right)\|h\|^2_{L^2(\overline{\mu})} \le \langle h, \mathsf{H}h\rangle_{L^2(\overline{\mu})} \le \varepsilon \|h\|^2_{L^2(\overline{\mu})}, \qquad h \in L^2_0(\overline{\mu}). \tag{A.28}$$

We use the equivalent Hilbert norms

$$\|h\|^2_{\mathsf{H}} := \langle h, \mathsf{H}h\rangle_{L^2(\overline{\mu})}, \qquad \|z\|^2_{\mathsf{H}^{-1}} := \left\langle \mathsf{H}^{-1} z, z\right\rangle_{L^2(\overline{\mu})}.$$

We use the sign convention fixed in the main text:

$$\triangle_{\overline{\rho}}\phi := \overline{\rho}^{-1}\,\mathrm{div}(\overline{\rho}\nabla\phi).$$

Thus $\triangle_{\overline{\rho}}$ agrees with the Laplace–Beltrami operator when $\overline{\rho} \equiv 1$, and it is non-positive on $L^2_0(\overline{\mu})$:

$$\langle \triangle_{\overline{\rho}} f, g\rangle_{L^2(\overline{\mu})} = -\int_{\mathbb{S}^{d-1}} \nabla f \cdot \nabla g \,\mathrm{d}\overline{\mu}. \tag{A.29}$$

The forward linearized generator and its $L^2(\overline{\mu})$-adjoint are

$$\mathsf{A} := \triangle_{\overline{\rho}}\mathsf{H}, \qquad \mathsf{A}^\dagger := \mathsf{H}\triangle_{\overline{\rho}}. \tag{A.30}$$

The positive backward generator is therefore

$$\mathsf{L} := -\mathsf{A}^\dagger = -\mathsf{H}\triangle_{\overline{\rho}}.$$

Viewed on the Hilbert space $(L^2_0(\overline{\mu}), \langle\cdot,\cdot\rangle_{\mathsf{H}^{-1}})$, L is self-adjoint and non-negative, with closed quadratic form

$$q_{\mathsf{L}}[z,w] := \int_{\mathbb{S}^{d-1}} \nabla z \cdot \nabla w \,\mathrm{d}\overline{\mu}, \qquad D(q_{\mathsf{L}}) = H^1(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu}).$$

Indeed, for smooth z, w,

$$\begin{aligned}\langle \mathsf{L}z, w\rangle_{\mathsf{H}^{-1}} &= \left\langle \mathsf{H}^{-1}(-\mathsf{H}\triangle_{\overline{\rho}} z), w\right\rangle_{L^2(\overline{\mu})}\\ &= -\langle \triangle_{\overline{\rho}} z, w\rangle_{L^2(\overline{\mu})} = \int_{\mathbb{S}^{d-1}} \nabla z \cdot \nabla w \,\mathrm{d}\overline{\mu}.\end{aligned}$$

For nonzero $g_\ell \in H^1(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu})$, this gives the Rayleigh quotient

$$R_{\overline{\rho}}(g_\ell) := \frac{q_{\mathsf{L}}[g_\ell, g_\ell]}{\|g_\ell\|^2_{\mathsf{H}^{-1}}} = \frac{\int_{\mathbb{S}^{d-1}} |\nabla g_\ell|^2 \, \mathrm{d}\overline{\mu}}{\|g_\ell\|^2_{\mathsf{H}^{-1}}}.$$

The control operator generated by the parameter-linear FFN is

$$(\mathsf{B}W)(x) := -\overline{\rho}(x)^{-1} \operatorname{div}\big(\overline{\rho}(x) P_x^{\perp}(W\sigma(x))\big), \qquad W \in \mathbb{R}^{d\times p}. \tag{A.31}$$

Its adjoint is taken with respect to the $L^2(\overline{\mu})$-inner product on $L^2_0(\overline{\mu})$ and the Frobenius inner product on $\mathbb{R}^{d\times p}$. For smooth z, integration by parts gives

$$\mathsf{B}^* z = \int_{\mathbb{S}^{d-1}} \nabla z(x)\sigma(x)^\top \, \mathrm{d}\overline{\mu}(x).$$

A.2.2 Proof of Lemma 3.4

The controlled gradient-flow form of (2.2) is, in density variables,

$$\partial_t \rho_t = \operatorname{div}\left(\rho_t \nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu}(x; \rho_t\omega)\right) - \operatorname{div}\big(\rho_t P_x^{\perp}(W_t\sigma(x))\big). \tag{A.32}$$

At $W \equiv 0$, the initial datum $\mu_0 = \overline{\mu}$ is stationary because $\delta E_{\varepsilon,A}/\delta\mu(\overline{\mu})$ is constant. If $\rho_t = \overline{\rho}(1+h_t)$, then the first variation of (A.32) at $(h, W) = (0, 0)$ is

$$\partial_t h_t = \mathsf{A}h_t + \mathsf{B}W_t, \qquad h_0 = 0, \tag{A.33}$$

with $\mathsf{A} = \triangle_{\overline{\rho}}\mathsf{H}$.

The functional derivative $g_\ell = \frac{\delta \ell}{\delta \mu}[\overline{\mu}]$ is defined only up to an additive constant. We take its $\overline{\mu}$-mean-zero representative, still denoted g_ℓ. Let ϕ_t solve the adjoint equation

$$\partial_t \phi_t + \mathsf{A}^\dagger \phi_t = 0, \qquad \phi_T = g_\ell.$$

Since $\mathsf{A}^\dagger = \mathsf{H}\triangle_{\overline{\rho}}$, this is exactly (3.4). Equivalently,

$$\phi_t = \mathrm{e}^{(T-t)\mathsf{A}^\dagger} g_\ell = \mathrm{e}^{-(T-t)\mathsf{L}} g_\ell.$$

The coefficients are smooth and $g_\ell \in C^\infty(\mathbb{S}^{d-1})$, so the solution is classical.

Let $V \in L^2(0, T; \mathbb{R}^{d\times p})$, and let h^V be the linearized state response from (A.33). Then

$$\begin{aligned} D\ell(\overline{\mu})[h^V_T \overline{\mu}] &= \big\langle g_\ell, h^V_T \big\rangle_{L^2(\overline{\mu})} \\ &= \int_0^T \langle \mathsf{B}^*\phi_t, V_t\rangle_{\mathrm{F}} \, \mathrm{d}t. \end{aligned}$$

Thus the unregularized terminal-loss gradient at the zero initialization is

$$G_t := \mathsf{B}^*\phi_t = \int_{\mathbb{S}^{d-1}} \nabla\phi_t(x)\sigma(x)^\top \, \mathrm{d}\overline{\mu}(x).$$

The quadratic regularizer in (2.3) has zero derivative at $W \equiv 0$, so the one-step update with learning rate α is

$$W_t^{\text{1-GD}} = -\alpha G_t = -\alpha \int_{\mathbb{S}^{d-1}} \nabla\phi_t(x)\sigma(x)^\top \, \mathrm{d}\overline{\mu}(x),$$

which is (3.3).

A.2.3 Finite-time Gramian and the visibility factor

For $0 \le t \le T$, define the finite-time control Gramian

$$\mathsf{W}^{\mathsf{c}}_t := \int_0^t \mathrm{e}^{u\mathsf{A}} \mathsf{B}\mathsf{B}^* \mathrm{e}^{u\mathsf{A}^\dagger} \, \mathrm{d}u. \tag{A.34}$$

The linearized state generated by the one-step parameter $W^{\text{1-GD}}$ satisfies

$$\partial_t h_t^{\mathrm{lin},\alpha} = \mathsf{A} h_t^{\mathrm{lin},\alpha} - \alpha \mathsf{B}\mathsf{B}^* \phi_t, \qquad h_0^{\mathrm{lin},\alpha} = 0.$$

Using $\phi_s = \mathrm{e}^{(t-s)\mathsf{A}^\dagger}\phi_t$ for $0 \le s \le t$, Duhamel's formula gives

$$h_t^{\mathrm{lin},\alpha} = -\alpha \mathsf{W}_t^{\mathsf{c}} \phi_t.$$

For $z \in L_0^2(\overline{\mu})$, define

$$\mathcal{O}_t(z) := \int_0^t \left\| \mathsf{B}^* \mathrm{e}^{u\mathsf{A}^\dagger} z \right\|_{\mathrm{F}}^2 \mathrm{d}u.$$

Then

$$\mathcal{O}_t(z) = \left\langle z, \mathsf{W}_t^{\mathsf{c}} z \right\rangle_{L^2(\overline{\mu})}.$$

For $z = \phi_t$, this is precisely the squared norm of the unregularized loss-gradient signal on $[0,t]$:

$$\mathcal{O}_t(\phi_t) = \int_0^t \|G_s\|_{\mathrm{F}}^2 \,\mathrm{d}s = \|\nabla J(0)\|_{L^2(0,t)}^2,$$

where the last equality uses the notation of Section 3.2. Hence the visibility factor in (3.5) is

$$\Omega_{t,T}(g_\ell) = \frac{\mathcal{O}_t(\phi_t)}{\|\phi_t\|_{\mathsf{H}^{-1}}^2}. \tag{A.35}$$

By Cauchy–Schwarz in the H^{-1}-geometry,

$$\begin{aligned} \mathcal{O}_t(z) &= \left\langle \mathsf{H}^{-1/2} z, \mathsf{H}^{1/2} \mathsf{W}_t^{\mathsf{c}} z \right\rangle_{L^2(\overline{\mu})} \\ &\le \|z\|_{\mathsf{H}^{-1}} \left\| \mathsf{H}^{1/2} \mathsf{W}_t^{\mathsf{c}} z \right\|_{L^2(\overline{\mu})}. \end{aligned}$$

Consequently,

$$\left\| \mathsf{H}^{1/2} \mathsf{W}_t^{\mathsf{c}} z \right\|_{L^2(\overline{\mu})} \ge \frac{\mathcal{O}_t(z)}{\|z\|_{\mathsf{H}^{-1}}} \quad (z \neq 0). \tag{A.36}$$

A.2.4 Local Taylor expansions in the density chart

Lemma A.10 (local energy expansion in the density chart)**.** *Fix $r_0 \in (0, 1/2)$. For $\mu = (1+h)\overline{\mu}$ with $h \in L_0^2(\overline{\mu})$ and $\|h\|_{L^\infty(\mathbb{S}^{d-1})} \le r_0$,*

$$\Delta_{\varepsilon,A}(\mu) = \frac{1}{2}\|h\|_{\mathsf{H}}^2 + R_E(h), \qquad |R_E(h)| \le C_E \varepsilon \|h\|_{L^\infty(\mathbb{S}^{d-1})} \|h\|_{L^2(\overline{\mu})}^2. \tag{A.37}$$

In particular, there exists a chart radius $r_{\mathrm{ch}} \in (0, r_0]$ such that, whenever $\|h\|_{L^\infty(\mathbb{S}^{d-1})} \le r_{\mathrm{ch}}$, one has

$$\Delta_{\varepsilon,A}(\mu) \ge \frac{1}{4}\|h\|_{\mathsf{H}}^2. \tag{A.38}$$

Proof. Since $\overline{\mu}$ is a global minimizer, $\Delta_{\varepsilon,A}(\mu) = E_{\varepsilon,A}((1+h)\overline{\mu}) - E_{\varepsilon,A}(\overline{\mu})$. The Euler–Lagrange equation at $\overline{\mu}$ and $\int h \,\mathrm{d}\overline{\mu} = 0$ give the exact expansion

$$\begin{aligned} E_{\varepsilon,A}((1+h)\overline{\mu}) - E_{\varepsilon,A}(\overline{\mu}) = \varepsilon \int_{\mathbb{S}^{d-1}} &((1+h)\log(1+h) - h)\,\mathrm{d}\overline{\mu} \\ &- \frac{1}{2} \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \mathrm{e}^{\langle x, Ay\rangle} h(x) h(y)\,\mathrm{d}\overline{\mu}(x)\,\mathrm{d}\overline{\mu}(y). \end{aligned}$$

Write $(1+z)\log(1+z) - z = \frac{1}{2}z^2 + r(z)$ with $|r(z)| \le C|z|^3$ for $|z| \le r_0$. The interaction term is exactly quadratic in h, so (A.37) holds with $R_E(h) = \varepsilon \int_{\mathbb{S}^{d-1}} r(h)\,\mathrm{d}\overline{\mu}$. Using the lower bound in (A.28), we choose r_{ch} so small that the remainder is bounded by $\frac{1}{4}\|h\|_{\mathsf{H}}^2$, and (A.38) follows. ∎

Lemma A.11 (centered first-variation Taylor expansion)**.** *For $\mu = (1+h)\overline{\mu}$ with $h \in L^2_0(\overline{\mu})$ and $\|h\|_{L^\infty(\mathbb{S}^{d-1})} \le r_0$, set*

$$\Phi(h) := \Pi_0\left(\frac{\delta E_{\varepsilon,A}}{\delta \mu}((1+h)\overline{\mu}) - \frac{\delta E_{\varepsilon,A}}{\delta \mu}(\overline{\mu})\right). \tag{A.39}$$

Then

$$\Phi(h) = \mathsf{H}h + R_F(h), \qquad R_F(h) = \varepsilon \Pi_0(\log(1+h) - h), \tag{A.40}$$

where

$$\|R_F(h)\|_{L^2(\overline{\mu})} \le C_F \varepsilon \|h\|_{L^\infty(\mathbb{S}^{d-1})} \|h\|_{L^2(\overline{\mu})}. \tag{A.41}$$

Moreover, for every integer $s > \frac{d-1}{2} + 1$ there exists $C_{F,s}$ such that

$$\|R_F(h)\|_{H^s(\mathbb{S}^{d-1})} \le C_{F,s} \varepsilon \|h\|^2_{H^s(\mathbb{S}^{d-1})}. \tag{A.42}$$

Proof. From the explicit formula for $\frac{\delta E_{\varepsilon,A}}{\delta \mu}$,

$$\frac{\delta E_{\varepsilon,A}}{\delta \mu}((1+h)\overline{\mu}) - \frac{\delta E_{\varepsilon,A}}{\delta \mu}(\overline{\mu}) = \varepsilon \log(1+h) - \int_{\mathbb{S}^{d-1}} \mathrm{e}^{\langle \cdot, Ay\rangle} h(y)\, \mathrm{d}\overline{\mu}(y).$$

Projecting onto mean-zero functions and using (A.27) gives (A.40). Since $|\log(1+z) - z| \le C|z|^2$ for $|z| \le r_0$, (A.41) follows. The Sobolev bound (A.42) is the standard composition estimate for C^∞ functions on $\mathbb{S}^{d-1}$. ■

A.2.5 Small-learning-rate nonlinear response

Proposition A.12 (small-learning-rate nonlinear response)**.** *Fix a terminal-layer thickness $r > 0$ and an integer $s > \dim(\mathbb{S}^{d-1})/2 + 2$. Assume $g_\ell \in H^{s+1}(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu})$. Then there exist constants $\overline{\alpha} > 0$ and $C_R < \infty$ such that, for every $0 < \alpha \le \overline{\alpha}$, the one-step-GD trajectory $\mu_t^{W^{1\text{-GD}}}$ remains in the selected local density chart on $[0,T]$. Moreover, for all $t \in [T-r, T]$, if*

$$\mu_t^{W^{1\text{-GD}}} = (1 + h_t^\alpha)\overline{\mu},$$

then

$$h_t^\alpha = -\alpha \mathsf{W}_t^{\mathsf{c}} \phi_t + R_t^\alpha, \qquad \|R_t^\alpha\|_{\mathsf{H}} \le C_R \alpha^2. \tag{A.43}$$

Proof. Let r_{ch} be a chart radius for which Lemma A.10 holds. Inside the chart write

$$\mu_t^{W^{1\text{-GD}}} = (1 + h_t^\alpha)\overline{\mu}, \qquad h_t^\alpha \in L^2_0(\overline{\mu}).$$

Using (A.32), (A.39), and (A.40), the chart equation is

$$\partial_t h_t = \mathsf{A} h_t + \mathsf{B} W_t + \mathcal{R}(h_t, W_t), \tag{A.44}$$

where

$$\mathcal{R}(h,W) := \triangle_{\overline{\rho}} R_F(h) + \overline{\rho}^{-1} \operatorname{div}(h\overline{\rho} \nabla \Phi(h)) - \overline{\rho}^{-1} \operatorname{div}\big(h\overline{\rho} P_x^\perp (W\sigma(x))\big).$$

The Sobolev algebra property and smoothness of $\overline{\rho}$ and σ imply

$$\|\mathcal{R}(h,W)\|_{H^{s-2}(\mathbb{S}^{d-1})} \le C\Big(\|h\|^2_{H^s(\mathbb{S}^{d-1})} + \|h\|_{H^s(\mathbb{S}^{d-1})} \|W\|_{\mathrm{F}}\Big) \tag{A.45}$$

for $\|h\|_{L^\infty} < r_{\mathrm{ch}}$.

Let

$$\zeta_t := -\mathsf{W}_t^{\mathsf{c}} \phi_t.$$

A direct differentiation using (A.34) and $\partial_t \phi_t + \mathsf{A}^\dagger \phi_t = 0$ gives

$$\partial_t \zeta_t = \mathsf{A}\zeta_t - \mathsf{B}\mathsf{B}^* \phi_t, \qquad \zeta_0 = 0. \tag{A.46}$$

Thus $\alpha\zeta_t$ is the first-order response forced by $W_t^{1\text{-GD}} = -\alpha \mathsf{B}^* \phi_t$. By standard parabolic estimates for smooth coefficients and $g_\ell \in H^{s+1}$,

$$\sup_{0 \le t \le T} \|\zeta_t\|_{H^s(\mathbb{S}^{d-1})} \le C_\zeta.$$

Moreover,

$$\sup_{0\le t\le T}\left\|W_t^{\text{1-GD}}\right\|_{\mathsf{F}} \le C_W\alpha. \tag{A.47}$$

On the maximal chart interval, set

$$Z_t^\alpha := h_t^\alpha - \alpha\zeta_t = h_t^\alpha + \alpha \mathsf{W}_t^{\mathsf{C}}\phi_t.$$

Subtracting α times (A.46) from (A.44) gives

$$\partial_t Z_t^\alpha = \mathsf{A}Z_t^\alpha + \mathcal{R}(h_t^\alpha, W_t^{\text{1-GD}}), \qquad Z_0^\alpha = 0.$$

Local parabolic stability gives, for every T' contained in the chart interval,

$$\sup_{0\le t\le T'}\|Z_t^\alpha\|_{H^s(\mathbb{S}^{d-1})} \le C_{\text{par}}\int_0^{T'}\left\|\mathcal{R}(h_\tau^\alpha, W_\tau^{\text{1-GD}})\right\|_{H^{s-2}(\mathbb{S}^{d-1})}\mathrm{d}\tau. \tag{A.48}$$

Bootstrap $\sup_{0\le t\le T'}\|h_t^\alpha\|_{H^s} \le M_{\text{boot}}\alpha$, with $M_{\text{boot}} > 2C_\zeta$. Then (A.45) and (A.47) give

$$\sup_{0\le t\le T'}\left\|\mathcal{R}(h_t^\alpha, W_t^{\text{1-GD}})\right\|_{H^{s-2}} \le C\alpha^2.$$

Therefore (A.48) yields

$$\sup_{0\le t\le T'}\|Z_t^\alpha\|_{H^s(\mathbb{S}^{d-1})} \le C_Z\alpha^2.$$

For $\alpha \le \overline{\alpha}$ small, this improves the bootstrap bound and, by Sobolev embedding, keeps $\|h_t^\alpha\|_{L^\infty} < r_{\text{ch}}$. The standard continuation criterion therefore gives chart retention on the whole interval $[0, T]$. Finally, (A.28) gives $\|f\|_{\mathsf{H}} \lesssim \|f\|_{L^2(\overline{\mu})} \lesssim \|f\|_{H^s(\mathbb{S}^{d-1})}$, so

$$\|Z_t^\alpha\|_{\mathsf{H}} \le C_R\alpha^2.$$

Since $Z_t^\alpha = h_t^\alpha + \alpha\mathsf{W}_t^{\mathsf{C}}\phi_t$, this is (A.43). ∎

A.2.6 Proof of Proposition 3.5

Lemma A.13 (Rayleigh–Jensen lower bound for the costate)**.** *Assume $g_\ell \neq 0$ and $g_\ell \in H^1(\mathbb{S}^{d-1}) \cap L_0^2(\overline{\mu})$. Then, for every $r \ge 0$,*

$$\left\|\mathrm{e}^{r\mathsf{A}^\dagger}g_\ell\right\|_{\mathsf{H}^{-1}} \ge \|g_\ell\|_{\mathsf{H}^{-1}}\exp(-rR_{\overline{\rho}}(g_\ell)).$$

Proof. Since $\mathsf{L} = -\mathsf{A}^\dagger$ is self-adjoint and non-negative on $(L_0^2(\overline{\mu}), \langle\cdot,\cdot\rangle_{\mathsf{H}^{-1}})$, the spectral theorem gives a probability measure ν_{g_ℓ}, normalized by $\|g_\ell\|_{\mathsf{H}^{-1}}^2$, such that

$$\left\|\mathrm{e}^{r\mathsf{A}^\dagger}g_\ell\right\|_{\mathsf{H}^{-1}}^2 = \|g_\ell\|_{\mathsf{H}^{-1}}^2\int_0^\infty \mathrm{e}^{-2r\lambda}\,\mathrm{d}\nu_{g_\ell}(\lambda).$$

Moreover,

$$\int_0^\infty \lambda\,\mathrm{d}\nu_{g_\ell}(\lambda) = R_{\overline{\rho}}(g_\ell).$$

The map $\lambda \mapsto \mathrm{e}^{-2r\lambda}$ is convex, so Jensen's inequality gives

$$\int_0^\infty \mathrm{e}^{-2r\lambda}\,\mathrm{d}\nu_{g_\ell}(\lambda) \ge \mathrm{e}^{-2rR_{\overline{\rho}}(g_\ell)}.$$

Taking square roots proves the claim. ∎

Proof of Proposition 3.5. If $g_\ell = 0$, then the one-step parameter vanishes and the statement is understood with zero right-hand side. Assume $g_\ell \neq 0$. Let $t \in [T-r, T]$ and set $z_t := \phi_t = \mathrm{e}^{(T-t)\mathsf{A}^\dagger}g_\ell$. By Proposition A.12,

$$h_t^\alpha = -\alpha\mathsf{W}_t^{\mathsf{C}}z_t + R_t^\alpha, \qquad \|R_t^\alpha\|_{\mathsf{H}} \le C_R\alpha^2.$$

Therefore

$$\|h_t^\alpha\|_{\mathsf{H}} \ge \alpha\left\|\mathsf{H}^{1/2}\mathsf{W}_t^{\mathsf{C}}z_t\right\|_{L^2(\overline{\mu})} - C_R\alpha^2. \tag{A.49}$$

Using (A.36) and the definition (A.35),

$$\left\| \mathsf{H}^{1/2} \mathsf{W}_t^{\mathsf{c}} z_t \right\|_{L^2(\overline{\mu})} \geq \Omega_{t,T}(g_\ell) \|z_t\|_{\mathsf{H}^{-1}}.$$

Finally, Lemma A.13 gives

$$\|z_t\|_{\mathsf{H}^{-1}} \geq \|g_\ell\|_{\mathsf{H}^{-1}} \exp(-(T-t) R_{\overline{\rho}}(g_\ell)). \tag{A.50}$$

Combining (A.49)–(A.50) and then taking positive parts yields

$$\|h_t^\alpha\|_{\mathsf{H}} \geq \left(\alpha \Omega_{t,T}(g_\ell) \|g_\ell\|_{\mathsf{H}^{-1}} \exp(-(T-t) R_{\overline{\rho}}(g_\ell)) - C_R \alpha^2 \right)_+.$$

The trajectory remains in the density chart by Proposition A.12; hence Lemma A.10 applies and gives

$$\Delta_{\varepsilon,A}(\mu_t^{W^{\text{1-GD}}}) \geq \frac{1}{4} \|h_t^\alpha\|_{\mathsf{H}}^2.$$

This proves the asserted lower bound. ∎

A.3 Proof for Section 3.3

This appendix proves Theorem 3.9 in three stages. Appendices A.3.1 to A.3.3 collect the analytic inputs around a single well, namely the weighted Poincaré estimates, the local quadratic equivalence in the density chart, and the local PL inequality with a controlled strict dissipativity estimate, which localize the state near a well without yet giving the turnpike. Appendices A.3.4 to A.3.8 treat the double-well structure under Assumption 3.6. The energy budget forces entry into a low-energy component, the sign barrier excludes switching between the two branches, and the parabolic regularization upgrades this entry to the density chart, which proves Lemma 3.7. Appendices A.3.9 to A.3.11 control the optimal parameters under Assumption 3.8. A T-uniform stability estimate for the linearized evolution along the optimal trajectory, the optimality system with the stationarity condition $\lambda_{\text{reg}} W_t^* = -\int \nabla \phi_t \sigma^\top \, \mathrm{d}\mu_t^*$, and the backward decay of the adjoint state make W^* exponentially small away from the terminal layer, which completes the proof.

A.3.1 Weighted Laplacian and one-well Poincaré estimates

Recall the weighted Laplacian $\triangle_{\overline{\rho}} \phi = \overline{\rho}^{-1} \operatorname{div}(\overline{\rho} \nabla \phi)$ from Lemma 3.4.

Both K and $\triangle_{\overline{\rho}}$ preserve $L_0^2(\overline{\mu})$, since the projection is built into K and $\int_{\mathbb{S}^{d-1}} \triangle_{\overline{\rho}} \phi \, \mathrm{d}\overline{\mu} = \int_{\mathbb{S}^{d-1}} \operatorname{div}(\overline{\rho} \nabla \phi) \, \mathrm{d}\omega = 0$ on the closed sphere.

Proposition A.14 (weighted Poincaré gap)**.** *For every fixed $\varepsilon > 0$, there exists $\kappa_P(\varepsilon) > 0$ such that*

$$-\langle \triangle_{\overline{\rho}} z, z \rangle_{L^2(\overline{\mu})} \geq \kappa_P(\varepsilon) \|z\|_{L^2(\overline{\mu})}^2, \qquad z \in D(\triangle_{\overline{\rho}}) \cap L_0^2(\overline{\mu}). \tag{A.51}$$

Proof. Since $\overline{\rho} \in C^\infty(\mathbb{S}^{d-1})$ and $\inf_{\mathbb{S}^{d-1}} \overline{\rho} > 0$, the operator $\triangle_{\overline{\rho}}$ is uniformly elliptic and self-adjoint on the compact sphere $\mathbb{S}^{d-1}$. Its kernel consists of the constants, so the first positive eigenvalue of $-\triangle_{\overline{\rho}}$ gives (A.51) on the mean-zero subspace. ∎

For the selected branch define

$$W_{\overline{\mu}}(x) := \int_{\mathbb{S}^{d-1}} K_A(x,y) \, \mathrm{d}\overline{\mu}(y), \qquad U_\varepsilon(x) := -W_{\overline{\mu}}(x), \qquad U_0(x) := -\mathrm{e}^{\lambda_1 x_1}.$$

The Euler–Lagrange equation gives

$$\overline{\rho}(x) = Z_\varepsilon^{-1} \exp\left(-\frac{U_\varepsilon(x)}{\varepsilon} \right).$$

Lemma A.15 (C^3-one-well convergence of the branch potential)**.** *Assume Theorem A.1. Then there exists $\delta_3(\varepsilon; A, d) \downarrow 0$ such that*

$$\|W_{\overline{\mu}} - K_A(\cdot, e_1)\|_{C^3(\mathbb{S}^{d-1})} + \|U_\varepsilon - U_0\|_{C^3(\mathbb{S}^{d-1})} \leq \delta_3(\varepsilon; A, d).$$

More concretely, after decreasing the small-noise threshold if necessary,

$$\delta_3(\varepsilon; A, d) \leq C_{3,A,d} \varepsilon^{1/2}.$$

Proof. For $0 \le j \le 3$, the map $(x,y) \mapsto \nabla_x^j K_A(x,y)$ is smooth on the compact set $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. Hence it is uniformly Lipschitz in the second variable. By Theorem A.1, the selected branch satisfies $x_\varepsilon = e_1 + O(\varepsilon)$, the tail estimate (A.5), and the moment bounds (A.7). Since $u = \exp_{x_\varepsilon}^{-1}(y)$ on the good cap and is set to zero outside,

$$\int_{\mathbb{S}^{d-1}} \operatorname{dist}_{\mathbb{S}^{d-1}}(y,e_1)\,\mathrm{d}\overline{\mu}(y) \le C\left(\int_{\mathbb{S}^{d-1}} |u(y)|^2 \mathrm{d}\overline{\mu}(y)\right)^{1/2} + C\overline{\mu}(\mathbb{S}^{d-1}\setminus B_r(x_\varepsilon)) + C\operatorname{dist}_{\mathbb{S}^{d-1}}(x_\varepsilon,e_1) \le C_{A,d}\varepsilon^{1/2}.$$

Therefore, for $0 \le j \le 3$,

$$\sup_{x\in\mathbb{S}^{d-1}} \left|\nabla_x^j W_{\overline{\mu}}(x) - \nabla_x^j K_A(x,e_1)\right| \le C_{j,A,d} \int_{\mathbb{S}^{d-1}} \operatorname{dist}_{\mathbb{S}^{d-1}}(y,e_1)\,\mathrm{d}\overline{\mu}(y) \le C_{j,A,d}\varepsilon^{1/2}.$$

Summing over j proves the claim. Since $K_A(x,e_1) = \mathrm{e}^{\lambda_1 x_1}$, the statement for U_ε follows by changing sign. ∎

Proposition A.16 (one-well scale of the weighted Poincaré gap)**.** *Assume Theorem A.1 and $d \ge 2$. There exist $\bar{\varepsilon}_{1\mathrm{w}}(A,d) > 0$ and $c_{1\mathrm{w}}(A,d) > 0$ such that, for every $0 < \varepsilon < \bar{\varepsilon}_{1\mathrm{w}}(A,d)$, the weighted Poincaré constant in Proposition A.14 satisfies*

$$\kappa_P(\varepsilon) \ge \frac{c_{1\mathrm{w}}(A,d)}{\varepsilon}. \tag{A.52}$$

Moreover, one may choose the constant so that, with

$$a_+ := \lambda_1 \mathrm{e}^{\lambda_1}, \qquad a_- := \lambda_1 \mathrm{e}^{-\lambda_1}, \qquad m := d-1,$$

there is a dimensional constant $c_d > 0$ for which

$$c_{1\mathrm{w}}(A,d) \ge c_d \min\{a_+, m a_-\} \ge c_d \lambda_1 \mathrm{e}^{-\lambda_1}. \tag{A.53}$$

Proof. We use a standard Lyapunov–Poincaré argument combined with the Bakry–Émery local convexity criterion. The estimates required for the present one-well Gibbs measure are recorded explicitly below.

Set

$$G_\varepsilon := U_\varepsilon - \min_{\mathbb{S}^{d-1}} U_\varepsilon, \qquad \mathcal{L}_\varepsilon := \varepsilon\triangle - \nabla G_\varepsilon \cdot \nabla.$$

Then $\overline{\mu} = Z_\varepsilon^{-1} \exp(-G_\varepsilon/\varepsilon)\omega$, and $\triangle_{\overline{\rho}} = \varepsilon^{-1}\mathcal{L}_\varepsilon$. Thus $\kappa_P(\varepsilon)$ is the usual Poincaré constant for the Gibbs measure $\exp(-G_\varepsilon/\varepsilon)\omega$, with Dirichlet form $\int_{\mathbb{S}^{d-1}} |\nabla f|^2 \mathrm{d}\overline{\mu}$.

For the limiting potential $U_0 = -\mathrm{e}^{\lambda_1 x_1}$, the only critical points on $\mathbb{S}^{d-1}$ are e_1 and $-e_1$. The point e_1 is a nondegenerate minimum and $-e_1$ is a nondegenerate maximum, with

$$\operatorname{Hess}_{\mathbb{S}^{d-1}} U_0(e_1) = a_+ I_{T_{e_1}\mathbb{S}^{d-1}}, \qquad \operatorname{Hess}_{\mathbb{S}^{d-1}} U_0(-e_1) = -a_- I_{T_{-e_1}\mathbb{S}^{d-1}}.$$

By Lemma A.15 and Morse stability, after decreasing $\bar{\varepsilon}_{1\mathrm{w}}$, the potential U_ε has exactly two critical points: a minimum $x_{\varepsilon,+}$ near e_1 and a maximum $x_{\varepsilon,-}$ near $-e_1$. Moreover,

$$\operatorname{Hess}_{\mathbb{S}^{d-1}} G_\varepsilon(x_{\varepsilon,+}) \ge \frac{1}{2} a_+ I, \qquad \triangle G_\varepsilon(x_{\varepsilon,-}) \le -\frac{1}{2} m a_-, \tag{A.54}$$

and, outside fixed small neighborhoods of these two points,

$$|\nabla G_\varepsilon| \ge g_A > 0. \tag{A.55}$$

Choose $R = R(A,d)$ large and define

$$\mathcal{U}_\varepsilon := B_{R\sqrt{\varepsilon}}(x_{\varepsilon,+}).$$

For ε small, this ball is geodesically convex and contained in a region where $\operatorname{Hess}_{\mathbb{S}^{d-1}} G_\varepsilon \ge a_+ I/4$. The Bakry–Émery criterion applied to the restricted Gibbs measure on $\mathcal{U}_\varepsilon$, with Neumann boundary condition, gives the local Poincaré estimate

$$\operatorname{Var}_{\overline{\mu}^{\mathcal{U}_\varepsilon}}(f) \le \frac{4\varepsilon}{a_+} \int_{\mathcal{U}_\varepsilon} |\nabla f|^2 \mathrm{d}\overline{\mu}^{\mathcal{U}_\varepsilon},$$

where $\overline{\mu}^{\mathcal{U}_\varepsilon}$ is $\overline{\mu}$ conditioned on $\mathcal{U}_\varepsilon$. Hence the restricted Poincaré constant satisfies

$$\rho_{\mathcal{U}_\varepsilon} \geq \frac{a_+}{4\varepsilon}. \tag{A.56}$$

Now take the Lyapunov function

$$V_\varepsilon := \exp\left(\frac{G_\varepsilon}{2\varepsilon}\right) \geq 1.$$

A direct computation gives

$$\frac{\mathcal{L}_\varepsilon V_\varepsilon}{\varepsilon V_\varepsilon} = \frac{\triangle G_\varepsilon}{2\varepsilon} - \frac{|\nabla G_\varepsilon|^2}{4\varepsilon^2}. \tag{A.57}$$

In the annulus around $x_{\varepsilon,+}$ outside $\mathcal{U}_\varepsilon$, the quadratic lower bound on $|\nabla G_\varepsilon|$ and the choice of R make the negative term in (A.57) dominate the Laplacian term. Near $x_{\varepsilon,-}$, the second estimate in (A.54) gives a negative contribution of order ma_-/ε. Away from the two critical neighborhoods, (A.55) gives a stronger negative bound for small ε. Consequently there exist constants $\Lambda_A, B_A > 0$, independent of ε, such that

$$\frac{\mathcal{L}_\varepsilon V_\varepsilon}{\varepsilon V_\varepsilon} \leq -\frac{\Lambda_A}{\varepsilon} + \frac{B_A}{\varepsilon}\mathbf{1}_{\mathcal{U}_\varepsilon},$$

with

$$\Lambda_A \geq c_d \min\{a_+, ma_-\}, \qquad B_A \leq C_d a_+. \tag{A.58}$$

The Lyapunov–Poincaré criterion now yields

$$\kappa_P(\varepsilon) \geq \frac{(\Lambda_A/\varepsilon)\rho_{\mathcal{U}_\varepsilon}}{B_A/\varepsilon + \rho_{\mathcal{U}_\varepsilon}}.$$

Using (A.56) and (A.58), we obtain

$$\kappa_P(\varepsilon) \geq \frac{\Lambda_A a_+}{4B_A + a_+}\frac{1}{\varepsilon} \geq \frac{c_d}{\varepsilon}\min\{a_+, ma_-\}.$$

This proves (A.52)–(A.53). ∎

A.3.2 Local quadratic equivalence in the density chart

We use the mean-zero projection $\Pi_0^{\overline{\mu}} f = f - \int_{\mathbb{S}^{d-1}} f \, \mathrm{d}\overline{\mu}$ from (A.22) and the density chart of Lemma A.10, that is, measures $\mu = (1+h)\overline{\mu}$ with $h \in L_0^2(\overline{\mu})$ and $\|h\|_{L^\infty(\mathbb{S}^{d-1})} \leq r_0$. The bound on $\|h\|_{L^\infty}$ is only the definition of the local chart; it is not a model assumption. For such μ, we write $\Phi_\mu := \Phi(h)$ for the centered first-variation increment of Lemma A.11. Because $\frac{\delta E_{\varepsilon,A}}{\delta \mu}(\overline{\mu})$ is constant on $\mathbb{S}^{d-1}$, this is the same as $\Pi_0^{\overline{\mu}}(\frac{\delta E_{\varepsilon,A}}{\delta \mu}(\mu))$.

Corollary A.17 (local quadratic equivalence)**.** *There exists $r_1 \in (0, r_0]$ such that, whenever $\|h\|_{L^\infty} \leq r_1$,*

$$c_1\|h\|^2_{L^2(\overline{\mu})} \leq \Delta_{\varepsilon,A}(\mu) \leq c_2\|h\|^2_{L^2(\overline{\mu})}, \tag{A.59}$$

and

$$\|\Phi_\mu\|_{L^2(\overline{\mu})} \geq c_3\|h\|_{L^2(\overline{\mu})}. \tag{A.60}$$

After decreasing the small-noise threshold and the chart radius if necessary, one may take

$$c_3 = \frac{\gamma_A}{4}\varepsilon, \qquad \gamma_A := 1 - \frac{\lambda_2}{\lambda_1}, \tag{A.61}$$

and there is a constant $C_{\mathrm{up}} = C_{\mathrm{up}}(A, d, r_1)$, independent of ε in the small-noise range, such that

$$c_2 \leq C_{\mathrm{up}}\varepsilon. \tag{A.62}$$

Proof. By (A.26), after decreasing the small-noise threshold if necessary,

$$\langle \mathsf{H}h, h\rangle_{L^2(\overline{\mu})} \geq \frac{\gamma_A}{2}\varepsilon\|h\|^2_{L^2(\overline{\mu})}, \qquad h \in L_0^2(\overline{\mu}).$$

Since H is self-adjoint and nonnegative on $L^2_0(\overline{\mu})$, the same spectral lower bound gives

$$\|\mathsf{H}h\|_{L^2(\overline{\mu})} \geq \frac{\gamma_A}{2}\varepsilon\|h\|_{L^2(\overline{\mu})}.$$

Apply Lemmas A.10 and A.11, and choose r_1 so that their remainder terms are absorbed by $\gamma_A\varepsilon/8$. This gives the lower energy bound and (A.61). The upper bound in (A.59) follows from the Taylor expansion, the positivity of K, and $\mathsf{H} = \varepsilon I - \mathsf{K}$, together with the remainder estimate (A.37); after possibly decreasing r_1, this yields (A.62). ■

A.3.3 Local PL, strict dissipativity, and forced localization

Define the local dissipation by

$$D(\mu) := \int_{\mathbb{S}^{d-1}} |\nabla\Phi_\mu|^2 \mathrm{d}\mu.$$

Lemma A.18 (dissipation controls the L^2 distance)**.** *There exist $r_2 \in (0, r_1]$ and $c_D > 0$ such that, whenever $\mu = (1+h)\overline{\mu}$ and $\|h\|_{L^\infty} \leq r_2$,*

$$D(\mu) \geq c_D\|h\|^2_{L^2(\overline{\mu})}.$$

Proof. If $\|h\|_{L^\infty} \leq r_2 < 1$, then

$$(1-r_2)\,\mathrm{d}\overline{\mu} \leq \mathrm{d}\mu \leq (1+r_2)\,\mathrm{d}\overline{\mu}.$$

Hence

$$D(\mu) = \int_{\mathbb{S}^{d-1}} |\nabla\Phi_\mu|^2 \mathrm{d}\mu \geq (1-r_2)\int_{\mathbb{S}^{d-1}} |\nabla\Phi_\mu|^2 \mathrm{d}\overline{\mu}.$$

By Proposition A.14,

$$\int_{\mathbb{S}^{d-1}} |\nabla\Phi_\mu|^2 \mathrm{d}\overline{\mu} = -\langle \triangle_{\overline{\rho}}\Phi_\mu, \Phi_\mu\rangle_{L^2(\overline{\mu})} \geq \kappa_P(\varepsilon)\|\Phi_\mu\|^2_{L^2(\overline{\mu})}.$$

Now use (A.60) from Corollary A.17. ■

Theorem A.19 (local PL / local entropy-dissipation)**.** *There exist $r_2 \in (0, r_1]$ and $c_{\mathrm{PL}} > 0$ such that, whenever $\mu = (1+h)\overline{\mu}$ and $\|h\|_{L^\infty} \leq r_2$,*

$$D(\mu) \geq c_{\mathrm{PL}}\,\Delta_{\varepsilon,A}(\mu). \tag{A.63}$$

Proof. Combine Lemma A.18 with the upper bound in (A.59). ■

Proposition A.20 (controlled strict dissipativity)**.** *Let (μ_t, W_t) solve (2.2), and assume μ_t remains in the neighborhood from Theorem A.19. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta_{\varepsilon,A}(\mu_t) = -D(\mu_t) + \int_{\mathbb{S}^{d-1}} \nabla\frac{\delta E_{\varepsilon,A}}{\delta\mu}(\mu_t)\cdot u_{W_t}\,\mathrm{d}\mu_t, \tag{A.64}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta_{\varepsilon,A}(\mu_t) \leq -\frac{1}{2}D(\mu_t) + \frac{1}{2}\|\sigma\|^2_{L^\infty(\mathbb{S}^{d-1})}\|W_t\|^2_{\mathrm{F}} \leq -a\,\Delta_{\varepsilon,A}(\mu_t) + C_\sigma\|W_t\|^2_{\mathrm{F}}, \tag{A.65}$$

where $a := c_{\mathrm{PL}}/2$ and C_σ is the constant in Assumption 3.3.

Proof. Differentiate $E_{\varepsilon,A}(\mu_t)$ along (3.1). Since $\partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0$ with

$$v_t = -\nabla\frac{\delta E_{\varepsilon,A}}{\delta\mu}(\mu_t) + u_{W_t},$$

the chain rule and integration by parts give

$$\frac{\mathrm{d}}{\mathrm{d}t}E_{\varepsilon,A}(\mu_t) = \int_{\mathbb{S}^{d-1}} \nabla\frac{\delta E_{\varepsilon,A}}{\delta\mu}(\mu_t)\cdot v_t\,\mathrm{d}\mu_t,$$

which is exactly (A.64). Apply Young's inequality to the mixed term:

$$\int_{\mathbb{S}^{d-1}} \nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu}(\mu_t) \cdot u_{W_t} \, \mathrm{d}\mu_t \le \frac{1}{2} D(\mu_t) + \frac{1}{2} \int_{\mathbb{S}^{d-1}} |u_{W_t}|^2 \mathrm{d}\mu_t.$$

Since $|u_W(x)| = |P_x^{\perp} W \sigma(x)| \le \|W\|_{\mathrm{F}} \|\sigma(x)\|$, one has

$$\int_{\mathbb{S}^{d-1}} |u_{W_t}|^2 \mathrm{d}\mu_t \le \|\sigma\|^2_{L^\infty(\mathbb{S}^{d-1})} \|W_t\|^2_{\mathrm{F}}.$$

Insert (A.63). ■

Corollary A.21 (forced state localization, not yet turnpike)**.** *Under the assumptions of Proposition A.20, every* $0 \le s \le t \le T$ *satisfies*

$$\Delta_{\varepsilon,A}(\mu_t) \le \mathrm{e}^{-a(t-s)} \Delta_{\varepsilon,A}(\mu_s) + C_\sigma \int_s^t \mathrm{e}^{-a(t-\tau)} \|W_\tau\|^2_{\mathrm{F}} \, \mathrm{d}\tau.$$

Proof. Apply Gronwall to (A.65). ■

Remark A.22 (explicit safe scale of the localization exponent). Tracing the proof of Theorem A.19 gives

$$c_{\mathrm{PL}} \ge \frac{(1-r_2)\kappa_P(\varepsilon) c_3^2}{c_2}, \qquad a = \frac{1}{2} c_{\mathrm{PL}}.$$

The parameter r_2 is only the radius of the relative-density chart:

$$\mu = (1+h)\overline{\mu}, \qquad \int_{\mathbb{S}^{d-1}} h \, \mathrm{d}\overline{\mu} = 0, \qquad \|h\|_{L^\infty(\overline{\mu})} \le r_2.$$

It enters the estimate through the measure comparison

$$(1-r_2) \, \mathrm{d}\overline{\mu} \le \mathrm{d}\mu \le (1+r_2) \, \mathrm{d}\overline{\mu}.$$

Thus $1 - r_2$ is a chart-safety factor rather than a spectral or dynamical parameter. By shrinking the local neighborhood one may make r_2 smaller, at the cost of requiring a more localized initial state.

With the choice $c_3 = \gamma_A \varepsilon / 4$ from (A.61), Proposition A.16 gives

$$a \ge \frac{(1-r_2) c_{1\mathrm{w}}(A,d) \gamma_A^2 \varepsilon}{32 c_2}.$$

Using the upper quadratic scale (A.62), we get

$$a \ge a_{\mathrm{safe}} := \frac{(1-r_2) c_{1\mathrm{w}}(A,d) \gamma_A^2}{32 C_{\mathrm{up}}} \ge \frac{(1-r_2) c_d}{32 C_{\mathrm{up}}} \left(1 - \frac{\lambda_2}{\lambda_1}\right)^2 \lambda_1 \mathrm{e}^{-\lambda_1}. \tag{A.66}$$

This is a lower bound on the uniform global state-localization exponent in the relative-density chart. It should not be read as a sharp equivalence for the bulk relaxation rate near the bottom of the selected well.

Remark A.23 (uniform localization time from the energy gap). Assume first that the control is absent, $W_t \equiv 0$, and that the trajectory remains in the local chart of Theorem A.19. Set

$$\Delta := \Delta_{\varepsilon,A}(\mu_0) = E_{\varepsilon,A}(\mu_0) - E_{\varepsilon,A}(\overline{\mu}).$$

Then Corollary A.21 gives

$$\Delta_{\varepsilon,A}(\mu_t) \le \mathrm{e}^{-at} \Delta.$$

Consequently, for any target level $0 < \eta < \Delta$, the sufficient uniform localization time

$$T_{\mathrm{uniform}}(\eta) := \inf\{t \ge 0 : \ \Delta_{\varepsilon,A}(\mu_s) \le \eta \text{ for all } s \ge t\}$$

satisfies

$$T_{\mathrm{uniform}}(\eta) \le \frac{1}{a} \log \frac{\Delta}{\eta} \le \frac{1}{a_{\mathrm{safe}}} \log \frac{\Delta}{\eta}.$$

Using (A.66), this yields the rough scale

$$T_{\mathrm{uniform}}(\eta) \lesssim \frac{\mathrm{e}^{\lambda_1}}{\left(1 - \frac{\lambda_2}{\lambda_1}\right)^2 \lambda_1} \log \frac{\Delta}{\eta},$$

where the suppressed constant depends on the chart constants and on the one-well Poincaré comparison constants. This is a worst-case global relative-density estimate. It may overestimate the effective relaxation time for perturbations supported in the bulk of the selected well.

At this stage the local one-well theory above is a toolkit around one fixed minimizer branch. The global problem is different: the free energy has an antipodal symmetry, and a branch-specific Wasserstein statement is not stable under this symmetry. The robust target is therefore the scalar free-energy gap.

Globally, $\Delta_{\varepsilon,A}$ measures distance to the minimizer set in energy and not distance to a chosen element of that set.

A.3.4 Static double-well landscape

Theorem A.1 now supplies the finite-ε double-well minimizer statement. The zero-temperature equality case gives only separation into the two caps; the branch count itself is obtained in the proof of Theorem A.1 from the Euler–Lagrange difference identity and the chord estimate (A.8). The next proposition records the static double-well structure. The local basins below are density-chart basins around the two smooth minimizers.

Proposition A.24 (static double-well local package)**.** *For every sufficiently small fixed $\varepsilon > 0$, the following hold.*

(i) The global minimizer set consists of exactly two antipodal branches:

$$\{\overline{\mu}^{\pm}\} = \{\overline{\mu}^{+}, \overline{\mu}^{-}\}, \qquad \overline{\mu}^{-} = S_{\#}\overline{\mu}^{+}, \qquad \overline{\mu}^{+} = \overline{\mu}.$$

(ii) There exist disjoint local density-chart basins $\mathcal{U}_+$ and $\mathcal{U}_-$ of $\overline{\mu}^+$ and $\overline{\mu}^-$, and constants $a_{\mathrm{loc}} > 0$, $b_{\mathrm{loc}} > 0$, such that the local PL/strict-dissipativity estimate above holds on both basins:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta_{\varepsilon,A}(\mu_t) \le -a_{\mathrm{loc}}\,\Delta_{\varepsilon,A}(\mu_t) + b_{\mathrm{loc}}\|W_t\|_{\mathrm{F}}^2 \tag{A.67}$$

whenever the controlled solution remains in one of the two basins. By symmetry one may take the constants from Proposition A.20 on the selected branch, namely $a_{\mathrm{loc}} = a$ and $b_{\mathrm{loc}} = C_\sigma$.

(iii) On each basin there is a branchwise metric-energy comparison

$$W_2(\mu, \overline{\mu}^{\pm})^2 \le C_{W_2,\Delta_{\varepsilon,A}}\,\Delta_{\varepsilon,A}(\mu) \qquad \forall \mu \in \mathcal{U}_{\pm}. \tag{A.68}$$

(iv) There exists $\eta_{\mathrm{loc}} > 0$ such that, for every $0 < \eta < \eta_{\mathrm{loc}}$, the branchwise chart sublevels

$$\{\Delta_{\varepsilon,A} \le \eta\} \cap \mathcal{U}_+, \qquad \{\Delta_{\varepsilon,A} \le \eta\} \cap \mathcal{U}_-$$

are connected.

Proof. Item (i) is exactly Theorem A.1(iv). Choose $r_* > 0$ so small that

$$r_* \le r_2, \qquad \frac{1}{1+r_*} > \frac{\lambda_2}{\lambda_1},$$

where r_2 is the radius in Theorem A.19, and then decrease r_* if necessary so that the two relative-density charts below are disjoint. Define

$$\mathcal{U}_+ := \left\{ (1+h)\overline{\mu}^+ : h \in L^\infty(\overline{\mu}^+), \int_{\mathbb{S}^{d-1}} h\,\mathrm{d}\overline{\mu}^+ = 0, \|h\|_{L^\infty} < r_* \right\},$$

and set $\mathcal{U}_- := S_{\#}\mathcal{U}_+$. By Theorem A.19 and Proposition A.20, the dissipativity estimate (A.67) holds on $\mathcal{U}_+$. The antipodal symmetry transfers the same estimate, with the same constants, to $\mathcal{U}_-$.

For the metric-energy comparison on $\mathcal{U}_+$, write $\mu = (1+h)\overline{\mu}^+$ with $\|h\|_{L^\infty} < r_*$, and connect $\overline{\mu}^+$ to μ by the linear density path $\mu_s = (1+sh)\overline{\mu}^+$. Since μ_s is uniformly equivalent to the smooth positive measure $\overline{\mu}^+$, the elliptic problem

$$-\operatorname{div}(\mu_s \nabla \phi_s) = h\overline{\mu}^+, \qquad \int_{\mathbb{S}^{d-1}} \phi_s\,\mathrm{d}\mu_s = 0,$$

has the standard estimate

$$\int_{\mathbb{S}^{d-1}} |\nabla \phi_s|^2 \mathrm{d}\mu_s \le C\|h\|^2_{L^2(\overline{\mu}^+)},$$

with C uniform for $0 \le s \le 1$. The Benamou–Brenier formula along this path gives

$$W_2(\mu, \overline{\mu}^+)^2 \le C\|h\|^2_{L^2(\overline{\mu}^+)}.$$

Combining this with the local quadratic lower bound (A.59) gives (A.68). The negative branch is again identical by symmetry.

It remains to prove the connectedness of small branchwise chart sublevels. Let $\mu_i = (1+h_i)\overline{\mu}^+ \in \mathcal{U}_+$, $i = 0, 1$, and set $h_t = (1-t)h_0 + th_1$, $g = h_1 - h_0$. The chart is convex, so $(1+h_t)\overline{\mu}^+ \in \mathcal{U}_+$. For $f(t) := E_{\varepsilon,A}((1+h_t)\overline{\mu}^+)$, the exact second derivative is

$$f''(t) = \varepsilon \int_{\mathbb{S}^{d-1}} \frac{g^2}{1+h_t}\, \mathrm{d}\overline{\mu}^+ - \iint K_A(x,y)g(x)g(y)\, \mathrm{d}\overline{\mu}^+(x)\, \mathrm{d}\overline{\mu}^+(y).$$

Since $\int g\, \mathrm{d}\overline{\mu}^+ = 0$, Theorem A.9 yields

$$f''(t) \ge \left(\frac{\varepsilon}{1+r_*} - \|\mathsf{K}\|_{\mathrm{op}} \right) \|g\|^2_{L^2(\overline{\mu}^+)} \ge c\varepsilon \|g\|^2_{L^2(\overline{\mu}^+)}$$

for small ε. Hence $E_{\varepsilon,A}$ is convex along every density chord in $\mathcal{U}_+$, and every sublevel intersection $\{\Delta_{\varepsilon,A} \le \eta\} \cap \mathcal{U}_+$ is connected. The proof in $\mathcal{U}_-$ is the antipodal image of this one. Taking $\eta_{\mathrm{loc}} > 0$ arbitrary, or smaller if desired for later uses, finishes the proof. ■

Proposition A.25 (weak double-well barrier)**.** *Fix a sufficiently small $\varepsilon > 0$. Let V_+^{stat} and V_-^{stat} denote the two static wells from Theorem A.1, and choose disjoint narrow open neighborhoods $\mathcal{V}_+$, $\mathcal{V}_-$ of $\overline{\mu}^+$, $\overline{\mu}^-$ such that*

$$S_{\#}\mathcal{V}_+ = \mathcal{V}_-, \qquad \mathcal{U}_\pm \subset \mathcal{V}_\pm, \qquad \overline{\mathcal{V}_+} \subset V_+^{\mathrm{stat}}, \qquad \overline{\mathcal{V}_-} \subset V_-^{\mathrm{stat}}.$$

This choice is possible after decreasing the chart radius in Proposition A.24, because the relative-density charts then lie in an arbitrarily small total-variation, hence W_2, neighborhood of the corresponding minimizer. Then

$$\beta_{\mathrm{w}} := \inf_{\mathcal{P}(\mathbb{S}^{d-1}) \setminus (\mathcal{V}_+ \cup \mathcal{V}_-)} \Delta_{\varepsilon,A} > 0.$$

Proof. The functional $E_{\varepsilon,A}$ is narrowly lower semicontinuous: the entropy is narrowly lower semicontinuous and the interaction term is narrowly continuous because K_A is continuous and bounded. Since $\mathbb{S}^{d-1}$ is compact, $\mathcal{P}(\mathbb{S}^{d-1})$ is narrowly compact. If $\beta_{\mathrm{w}} = 0$, there is a sequence $\mu_n \notin \mathcal{V}_+ \cup \mathcal{V}_-$ such that $\Delta_{\varepsilon,A}(\mu_n) \to 0$. Passing to a subsequence, $\mu_n \rightharpoonup \mu_*$. Lower semicontinuity gives $\Delta_{\varepsilon,A}(\mu_*) = 0$, so Proposition A.24(i) gives $\mu_* \in \{\overline{\mu}^+, \overline{\mu}^-\}$. Since $\mathcal{V}_\pm$ are open, this contradicts $\mu_n \notin \mathcal{V}_+ \cup \mathcal{V}_-$ for all large n. ■

Lemma A.26 (free-energy gradient-flow deformation inside one weak well)**.** *Let $0 < \eta < \min\{\beta_{\mathrm{w}}, \eta_{\mathrm{crit}}\}$, where η_{crit} is the one-well critical-point uniqueness threshold in Theorem A.1(iv). Define*

$$\mathcal{C}_\pm(\eta) := \{\mu \in \mathcal{P}(\mathbb{S}^{d-1}) : \Delta_{\varepsilon,A}(\mu) \le \eta\} \cap \mathcal{V}_\pm.$$

Then each $\mathcal{C}_\pm(\eta)$ is path connected. More precisely, every $\mu \in \mathcal{C}_\pm(\eta)$ can be joined to $\overline{\mu}^\pm$ by a continuous path contained in $\mathcal{C}_\pm(\eta)$.

Proof. We prove the statement in the positive well. Let $G_t\mu$ be the uncontrolled Wasserstein gradient flow of $E_{\varepsilon,A}$, equivalently the smooth solution for $t > 0$ of

$$\partial_t \nu_t + \mathrm{div}(\nu_t P_x^\perp \chi_A[\nu_t]) = \varepsilon \triangle \nu_t, \qquad \nu_{t=0} = \mu.$$

For finite-energy initial data this uniformly parabolic equation is well posed for positive times, is continuous in the narrow topology down to $t = 0$, and satisfies the energy identity

$$\frac{\mathrm{d}}{\mathrm{d}t} \Delta_{\varepsilon,A}(G_t\mu) = -D(G_t\mu) \tag{A.69}$$

for $t > 0$. These are the standard parabolic-gradient-flow facts for entropy plus a smooth interaction on a compact sphere; see, for example, (Bogachev et al. 2019) for related Fokker–Planck–Kolmogorov stability and convergence arguments.

Since energy is nonincreasing, $G_t\mu \in \{\Delta_{\varepsilon,A} \le \eta\}$ for all $t \ge 0$. Because $\eta < \beta_{\mathrm{w}}$, Proposition A.25 gives

$$G_t\mu \in \mathcal{V}_+ \cup \mathcal{V}_- \qquad \forall t \ge 0.$$

The path $t \mapsto G_t\mu$ is continuous and starts in $\mathcal{V}_+$; the sets $\mathcal{V}_+$ and $\mathcal{V}_-$ are disjoint open sets whose union contains the path. Hence connectedness of $[0,\infty)$ implies

$$G_t\mu \in \mathcal{V}_+ \qquad \forall t \ge 0.$$

It remains to identify the limit as $t \to \infty$. From (A.69), $\int_0^\infty D(G_t\mu)\,\mathrm{d}t < \infty$. Parabolic regularity gives relative compactness of $\{G_t\mu : t \ge t_0\}$ in smooth topologies for every $t_0 > 0$. Thus every ω-limit point is a critical point of $E_{\varepsilon,A}$, has energy at most η, and belongs to $\overline{\mathcal{V}_+} \subset V_+^{\mathrm{stat}}$. Since $\eta < \eta_{\mathrm{crit}}$, Theorem A.1(iv) implies that the only such critical point is $\overline{\mu}^+$. Therefore

$$G_t\mu \rightharpoonup \overline{\mu}^+ \qquad \text{as } t \to \infty.$$

On compact $\mathbb{S}^{d-1}$, narrow convergence is equivalent to W_2-convergence. Thus

$$\gamma(s) := \begin{cases} G_{s/(1-s)}\mu, & 0 \le s < 1, \\ \overline{\mu}^+, & s = 1, \end{cases}$$

is a continuous path in $\mathcal{C}_+(\eta)$ joining μ to $\overline{\mu}^+$. The negative well follows by applying the antipodal map. ∎

Proposition A.27 (low-energy sublevel has exactly two weak components)**.** *For every sufficiently small fixed $\varepsilon > 0$, there exists $\eta_{\mathrm{sep}} > 0$ such that, for every $0 < \eta < \eta_{\mathrm{sep}}$,*

$$\{\mu : \Delta_{\varepsilon,A}(\mu) \le \eta\} = \mathcal{C}_+(\eta) \sqcup \mathcal{C}_-(\eta), \tag{A.70}$$

where $\mathcal{C}_\pm(\eta)$ are the sets defined in Lemma A.26. Moreover $\mathcal{C}_+(\eta)$ and $\mathcal{C}_-(\eta)$ are precisely the two connected components of the low-energy sublevel.

Proof. Let $\eta_{\mathrm{crit}} = \eta_{\mathrm{crit}}(\varepsilon)$ be the critical-point uniqueness threshold from Theorem A.1(iv). Choose η_{sep} with

$$0 < \eta_{\mathrm{sep}} < \min\{\beta_{\mathrm{w}}, \eta_{\mathrm{crit}}\}.$$

Then any μ with $\Delta_{\varepsilon,A}(\mu) \le \eta < \eta_{\mathrm{sep}}$ lies in $\mathcal{V}_+ \cup \mathcal{V}_-$ by Proposition A.25. This proves the disjoint decomposition (A.70). The two sets are nonempty because they contain $\overline{\mu}^+$ and $\overline{\mu}^-$. They are path connected by Lemma A.26. Since they are separated inside the sublevel by the disjoint open sets $\mathcal{V}_+$, $\mathcal{V}_-$, they are exactly the two connected components. ∎

Lemma A.28 (well-distance modulus on weak low-energy components)**.** *For $0 < \eta < \eta_{\mathrm{sep}}$, define*

$$\Theta_{\mathrm{well}}(\eta) := \max_{\varsigma \in \{+,-\}} \sup_{\mu \in \mathcal{C}_\varsigma(\eta)} W_2(\mu, \overline{\mu}^\varsigma).$$

Then

$$\Theta_{\mathrm{well}}(\eta) \to 0 \qquad \text{as } \eta \downarrow 0.$$

Proof. If not, then for some $\delta > 0$ there are $\eta_n \downarrow 0$, signs ς_n, and $\mu_n \in \mathcal{C}_{\varsigma_n}(\eta_n)$ such that $W_2(\mu_n, \overline{\mu}^{\varsigma_n}) \ge \delta$. Passing to a subsequence fixes the sign and gives $\mu_n \rightharpoonup \mu_*$. Lower semicontinuity implies $\Delta_{\varepsilon,A}(\mu_*) = 0$, hence $\mu_* = \overline{\mu}^\varsigma$. Since $\mathbb{S}^{d-1}$ is compact, narrow convergence implies W_2-convergence, contradicting the lower bound by δ. ∎

A.3.5 Global energy budget and set-valued entry

The part of the local dissipativity argument that does not use the local PL inequality is global. Define the Wasserstein slope dissipation

$$D(\mu) := \int_{\mathbb{S}^{d-1}} \left| \nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu}(\mu) \right|^2 \mathrm{d}\mu.$$

For every smooth controlled solution of (2.2), the chain rule and Young's inequality give

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta_{\varepsilon,A}(\mu_t) = -D(\mu_t) + \int_{\mathbb{S}^{d-1}} \nabla \frac{\delta E_{\varepsilon,A}}{\delta \mu}(\mu_t) \cdot u_{W_t}\,\mathrm{d}\mu_t \le -\frac{1}{2}D(\mu_t) + C_\sigma \|W_t\|_{\mathrm{F}}^2. \tag{A.71}$$

Since $\operatorname{osc}\ell < \infty$ in the standing problem formulation, comparing the optimal control with the zero control gives the T-uniform bound

$$\frac{\lambda_{\mathrm{reg}}}{2}\int_0^T \left\|W_t^{*,T}\right\|_{\mathrm{F}}^2 \mathrm{d}t \le \operatorname{osc}\ell. \tag{A.72}$$

For later use set

$$R_{\mathrm{opt}} := \frac{2}{\lambda_{\mathrm{reg}}}\operatorname{osc}\ell, \tag{A.73}$$

so that every subinterval of an optimal trajectory has control energy at most R_{opt}. Combining (A.71) and (A.72) yields

$$\sup_{t\in[0,T]} \Delta_{\varepsilon,A}(\mu_t^{*,T}) \le \mathfrak{E}_{\max} := \Delta_{\varepsilon,A}(\mu_0) + \frac{2C_\sigma}{\lambda_{\mathrm{reg}}}\operatorname{osc}\ell, \tag{A.74}$$

$$\int_0^T D(\mu_t^{*,T})\,\mathrm{d}t \le \mathfrak{D}_{\max} := 2\Delta_{\varepsilon,A}(\mu_0) + \frac{4C_\sigma}{\lambda_{\mathrm{reg}}}\operatorname{osc}\ell. \tag{A.75}$$

These constants are independent of the horizon.

For $0 < \eta < \eta_{\mathrm{sep}}$, with the components $\mathcal{C}_\pm(\eta)$ of Proposition A.27, define the low-energy neighborhood of the minimizer set

$$\mathcal{N}_\eta := \mathcal{C}_+(\eta)\cup\mathcal{C}_-(\eta).$$

The correct global slope gap is a gap outside this set:

$$d_{\mathrm{set}}(\eta) := \inf\Big\{D(\mu):\ \Delta_{\varepsilon,A}(\mu)\le\mathfrak{E}_{\max}, \mu\notin\mathcal{N}_\eta\Big\}. \tag{A.76}$$

We use the first-coordinate order parameter and the finite-temperature sign-changing barrier

$$m_1(\mu) := \int_{\mathbb{S}^{d-1}} x_1\,\mathrm{d}\mu(x), \qquad \eta_{\mathrm{w},\varepsilon} := \inf\{\Delta_{\varepsilon,A}(\mu): m_1(\mu)=0\}. \tag{A.77}$$

Remark A.29 (how the slope gap and no-switching are obtained). The entry argument below uses a threshold $\eta\in(0,\eta_{\mathrm{sep}})$ for which

$$d_{\mathrm{set}}(\eta) > 0. \tag{A.78}$$

This follows from Assumption 3.6. Propositions A.7 and A.32 and Corollary A.33 show that its first inequality excludes all non-minimal Gibbs critical points with $0 < \Delta_{\varepsilon,A} \le \mathfrak{E}_{\max}$ for all sufficiently small fixed ε, and a fixed-temperature compactness argument then gives (A.78) for every $0<\eta<\eta_{\mathrm{sep}}$. Its second inequality, together with the sign-changing barrier in Corollary A.31, excludes switching between the two branches.

A.3.6 Sign barrier and slope-gap inputs

Lemma A.30 (zero-temperature sign-changing barrier)**.** *Let*

$$\mathcal{P}_0 := \{\mu\in\mathcal{P}(\mathbb{S}^{d-1}): m_1(\mu)=0\}.$$

Then

$$\sup_{\mu\in\mathcal{P}_0} I_A(\mu) = Q_A.$$

Proof. The lower bound follows from the two admissible measures δ_{e_2} and $\frac{1}{2}(\delta_{e_1}+\delta_{-e_1})$, whose interaction energies are e^{λ_2} and $\cosh\lambda_1$, respectively. For the upper bound, use that K_A is positive definite, hence I_A is convex in μ. The set $\mathcal{P}_0$ is convex and compact. The maximum of a continuous convex functional over this moment set is attained at an extreme point. Since $\mathcal{P}_0$ is described by the two affine moment constraints $\int 1\,\mathrm{d}\mu = 1$ and $\int x_1\,\mathrm{d}\mu = 0$, its extreme points are supported on at most two points. Hence it suffices to consider $\mu = t\delta_x + (1-t)\delta_y$ with $tx_1 + (1-t)y_1 = 0$. If this measure is a single Dirac mass, then its support point satisfies $x_1 = 0$, and $I_A(\delta_x)\le \mathrm{e}^{\lambda_2}$. Otherwise $x_1y_1\le 0$, and the value of t is the top-coordinate-cancelling weight used in Lemma A.5. Therefore

$$I_A(t\delta_x+(1-t)\delta_y)\le Q_A.$$

This proves the upper bound. ∎

Corollary A.31 (finite-temperature sign-changing barrier)**.** *Let $\eta_{\mathrm{w},\varepsilon}$ be defined by (A.77). Then there exists $C_{\mathrm{w}} < \infty$, depending only on the fixed model data, such that, for all sufficiently small $\varepsilon > 0$,*

$$\eta_{\mathrm{w},\varepsilon} \geq \Delta_0 - C_{\mathrm{w}}\varepsilon|\log\varepsilon|.$$

In particular, after decreasing ε_0,

$$\eta_{\mathrm{w},\varepsilon} \geq \frac{3}{4}\Delta_0.$$

Proof. Let μ satisfy $m_1(\mu) = 0$. By Lemma A.30, $I_A(\mu) \leq Q_A$. Since $\varepsilon\,\mathrm{Ent}(\mu) \geq 0$,

$$E_{\varepsilon,A}(\mu) \geq -\frac{1}{2}Q_A.$$

On the other hand, the test-cap construction in the proof of Theorem A.1 gives

$$m_\varepsilon \leq -\frac{1}{2}\mathrm{e}^{\lambda_1} + C_{\mathrm{w}}\varepsilon|\log\varepsilon|.$$

Therefore

$$\Delta_{\varepsilon,A}(\mu) = E_{\varepsilon,A}(\mu) - m_\varepsilon \geq \frac{1}{2}(\mathrm{e}^{\lambda_1} - Q_A) - C_{\mathrm{w}}\varepsilon|\log\varepsilon| = \Delta_0 - C_{\mathrm{w}}\varepsilon|\log\varepsilon|.$$

Taking the infimum over $m_1(\mu) = 0$ proves the first claim. The second claim follows by decreasing ε_0. ∎

Proposition A.32 (fixed-temperature slope gap from critical-point exclusion)**.** *Fix $\varepsilon > 0$ and $0 < \eta < \eta_{\mathrm{sep}}$. Suppose that*

$$\mathrm{Crit}\,E_{\varepsilon,A} \cap \{\mu : 0 < \Delta_{\varepsilon,A}(\mu) \leq \mathfrak{E}_{\max}\} \subset \{\overline{\mu}^{\pm}\}. \tag{A.79}$$

Equivalently, there is no non-minimal stationary Gibbs state with $0 < \Delta_{\varepsilon,A} \leq \mathfrak{E}_{\max}$. Then

$$d_{\mathrm{set}}(\eta) > 0.$$

Proof. Suppose not. Then there are $\mu_n = \rho_n\omega$ such that

$$\Delta_{\varepsilon,A}(\mu_n) \leq \mathfrak{E}_{\max}, \qquad \mu_n \notin \mathcal{N}_\eta, \qquad D(\mu_n) \to 0.$$

Since $0 < \eta < \eta_{\mathrm{sep}}$, Proposition A.27 implies that $\mu_n \notin \mathcal{N}_\eta$ forces $\Delta_{\varepsilon,A}(\mu_n) > \eta$. Passing to a subsequence, $\Delta_{\varepsilon,A}(\mu_n) \to L \in [\eta, \mathfrak{E}_{\max}]$.

Set $b_n := \nabla(K_A * \mu_n)$. Smoothness of K_A on the compact set $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ gives $\|b_n\|_{L^\infty} \leq C_A$. From the definition of D,

$$\varepsilon^2 \int_{\mathbb{S}^{d-1}} \frac{|\nabla\rho_n|^2}{\rho_n}\,\mathrm{d}\omega = \int_{\mathbb{S}^{d-1}} |\varepsilon\nabla\log\rho_n|^2 \mathrm{d}\mu_n \leq 2D(\mu_n) + 2C_A^2.$$

Thus $\sqrt{\rho_n}$ is bounded in $H^1(\mathbb{S}^{d-1})$. After passing to a subsequence, $\sqrt{\rho_n} \to g$ strongly in $L^2(\mathbb{S}^{d-1})$, and hence $\rho_n \to \rho_* := g^2$ strongly in $L^1(\mathbb{S}^{d-1})$. Put $\mu_* := \rho_*\omega$. The entropy is continuous along this sequence by uniform integrability, and the interaction is continuous under narrow convergence; hence $\Delta_{\varepsilon,A}(\mu_*) = L \in [\eta, \mathfrak{E}_{\max}]$.

For every $\phi \in C^\infty(\mathbb{S}^{d-1})$,

$$\left|\int_{\mathbb{S}^{d-1}} \nabla\phi \cdot (\varepsilon\nabla\log\rho_n - b_n)\,\mathrm{d}\mu_n\right| \leq \|\nabla\phi\|_{L^\infty} D(\mu_n)^{1/2} \to 0.$$

Passing to the limit, using $b_n \to b_* := \nabla(K_A * \mu_*)$ uniformly, gives

$$-\varepsilon\int_{\mathbb{S}^{d-1}} \triangle\phi\,\mathrm{d}\mu_* - \int_{\mathbb{S}^{d-1}} \nabla\phi \cdot b_*\,\mathrm{d}\mu_* = 0.$$

Thus

$$\mathrm{div}(\varepsilon\nabla\rho_* - \rho_*\nabla(K_A * \mu_*)) = 0$$

in distributions. Elliptic regularity and the strong maximum principle imply $\rho_* \in C^\infty(\mathbb{S}^{d-1})$ and $\rho_* > 0$. Therefore

$$\varepsilon(\log\rho_* + 1) - K_A * \mu_*$$

is constant on $\mathbb{S}^{d-1}$, i.e. $\mu_* \in \mathrm{Crit}\,E_{\varepsilon,A}$. Since $\Delta_{\varepsilon,A}(\mu_*) \in [\eta, \mathfrak{E}_{\max}]$, this contradicts (A.79). Hence $d_{\mathrm{set}}(\eta) > 0$. ∎

Corollary A.33 (slope gap from the explicit barrier)**.** *Assume that there exists* $\gamma_{\mathrm{bar}} > 0$ *such that*

$$\mathfrak{E}_{\max} \le \Delta_0 - \gamma_{\mathrm{bar}}.$$

Then, after decreasing ε_0 *depending on* γ_{bar}*, for every* $0 < \varepsilon < \varepsilon_0$ *and every* $0 < \eta < \eta_{\mathrm{sep}}$*,*

$$d_{\mathrm{set}}(\eta) > 0.$$

Proof. Apply Proposition A.7 with $\gamma = \gamma_{\mathrm{bar}}/2$. For all sufficiently small ε, every non-minimal Gibbs critical point has excess energy at least $\Delta_0 - \gamma_{\mathrm{bar}}/2 > \mathfrak{E}_{\max}$. Hence the critical-point exclusion hypothesis in Proposition A.32 holds, and the slope gap follows. ■

A.3.7 Parabolic chart entry and exclusion of branch switching

Lemma A.34 (parabolic weak-to-chart upgrade for fixed ε)**.** *Fix the local density-chart basins* $\mathcal{U}_\pm$ *from Proposition A.24. Let* $\tau_{\mathrm{reg}} > 0$ *and* $R_{\mathrm{reg}} < \infty$ *be fixed. Then there exists*

$$0 < \eta_{\mathrm{chart}} = \eta_{\mathrm{chart}}(\varepsilon, \tau_{\mathrm{reg}}, R_{\mathrm{reg}}, \mathcal{U}_\pm) < \eta_{\mathrm{sep}},$$

possibly very small as $\varepsilon \downarrow 0$*, such that the following holds. Let* $\mu_t = \rho_t \omega$ *be a controlled solution of* (2.2) *on* $[t - \tau_{\mathrm{reg}}, t]$*. If, for some* $\varsigma \in \{+, -\}$*,*

$$\int_{t-\tau_{\mathrm{reg}}}^{t} \|W_s\|_{\mathrm{F}}^2 \, \mathrm{d}s \le R_{\mathrm{reg}}, \qquad \mu_t \in \mathcal{C}_\varsigma(\eta_{\mathrm{chart}}), \tag{A.80}$$

then

$$\mu_t \in \mathcal{U}_\varsigma.$$

A.3.8 Parabolic regularization and same-branch entry

This section proves the regularization input used in Lemma A.34. The heat-kernel input is recalled in Remark A.35, following Grigor'yan (Grigor'yan 1999). For related nonlinear Fokker–Planck–Kolmogorov stability and convergence results, see (Bogachev et al. 2019). The drift is treated by an explicit parametrix expansion. Throughout this section

$$m := d - 1,$$

and $\mathbb{S}^{d-1}$ is equipped with its standard Riemannian metric and Riemannian volume ω. All constants below may depend on the fixed value of $\varepsilon > 0$, but not on the horizon T.

Remark A.35 (heat-kernel input from Grigor'yan). Let $p_\varepsilon(t, x, y)$ be the heat kernel of $\varepsilon\triangle$, so that

$$P_t^\varepsilon f(x) := \int_{\mathbb{S}^{d-1}} p_\varepsilon(t, x, y) f(y) \, \mathrm{d}\omega(y)$$

is the heat semigroup. Grigor'yan constructs the heat kernel on Riemannian manifolds as the positive fundamental solution of the heat equation and proves its basic semigroup, positivity, and Gaussian-estimate properties (Grigor'yan 1999). On the compact sphere, the same construction, together with local parabolic regularity in finitely many coordinate charts, yields the following fixed-time derivative and Hölder bounds. For every $\alpha \in (0, 1)$ and every $\tau_0 > 0$, there are constants $G_{0,\alpha}, G_{1,\alpha} < \infty$, depending only on $\mathbb{S}^{d-1}, \alpha, \tau_0$, such that for $0 < s \le \tau_0$,

$$\sup_{x \in \mathbb{S}^{d-1}} \|p_\varepsilon(s, x, \cdot)\|_{C^\alpha(\mathbb{S}^{d-1})} \le G_{0,\alpha} (\varepsilon s)^{-(m+\alpha)/2}, \tag{A.81}$$

$$\sup_{x \in \mathbb{S}^{d-1}} \|\nabla_y p_\varepsilon(s, x, \cdot)\|_{C^\alpha(\mathbb{S}^{d-1})} \le G_{1,\alpha} (\varepsilon s)^{-(m+1+\alpha)/2}. \tag{A.82}$$

These estimates are not meant to be sharp. They are the only heat-kernel input used below. The drift-dependent estimates are derived by the parametrix series.

Lemma A.36 (controlled drift size)**.** *Let*

$$b_s(x) := P_x^\perp (\chi_A[\mu_s](x) + W_s \sigma(x))$$

be the drift in the controlled Fokker–Planck equation

$$\partial_s \rho_s = \varepsilon \triangle \rho_s - \mathrm{div}(\rho_s b_s).$$

There are model constants $B_0, B_1 < \infty$, depending only on A, σ and $\mathbb{S}^{d-1}$, such that for every controlled trajectory,

$$\big\|x \mapsto P_x^{\perp}\chi_A[\mu_s](x)\big\|_{C^2(\mathbb{S}^{d-1})} \le B_0, \qquad \big\|x \mapsto P_x^{\perp}(W_s\sigma(x))\big\|_{C^2(\mathbb{S}^{d-1})} \le B_1\|W_s\|_{\mathrm{F}}, \tag{A.83}$$

and hence, if

$$\int_{t-\tau}^{t} \|W_s\|_{\mathrm{F}}^2 \,\mathrm{d}s \le R, \tag{A.84}$$

then

$$\Lambda_1(\tau, R) := \int_{t-\tau}^{t} \|b_s\|_{C^2(\mathbb{S}^{d-1})} \,\mathrm{d}s \le B_0\tau + B_1\sqrt{R\tau}. \tag{A.85}$$

Proof. The map $x \mapsto K_A(x, y)$ is smooth on the compact set $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, so all spatial derivatives up to order three of K_A are uniformly bounded. Therefore $x \mapsto P_x^{\perp}\chi_A[\mu](x)$, which is obtained by tangent projection of the gradient of a smooth interaction potential, has a uniform C^2-bound independent of μ. The feature map σ is smooth on compact $\mathbb{S}^{d-1}$, and $x \mapsto P_x^{\perp}(W\sigma(x))$ is linear in W, so the second estimate in (A.83) follows. The integral bound is Cauchy–Schwarz. ∎

Lemma A.37 (parametrix Hölder bound with explicit dependence)**.** *Fix $\varepsilon, \tau > 0$, $R < \infty$, and $\alpha \in (0, 1)$. Let $p^b(r, x; s, y)$ be the transition density associated with*

$$\partial_s\rho_s = \varepsilon\triangle\rho_s - \operatorname{div}(\rho_s b_s)$$

on $[t - \tau, t]$, where b satisfies (A.85). Then there is a constant $C_{P,\alpha} = C_{P,\alpha}(\mathbb{S}^{d-1}, m, \alpha, \tau) > 0$ such that

$$\sup_{x\in\mathbb{S}^{d-1}} \big\|p^b(t - \tau, x; t, \cdot)\big\|_{C^{\alpha}(\mathbb{S}^{d-1})} \le C_{\mathrm{reg}}(\varepsilon, \tau, R), \tag{A.86}$$

where one may take

$$C_{\mathrm{reg}}(\varepsilon, \tau, R) := C_{P,\alpha}(\varepsilon\tau)^{-(m+\alpha)/2} \exp\left\{C_{P,\alpha}\left(1 + \frac{\Lambda_1(\tau, R)}{\sqrt{\varepsilon\tau}} + \frac{\Lambda_1(\tau, R)^2}{\varepsilon}\right)\right\}. \tag{A.87}$$

Proof. We spell out the parametrix estimate because this is the point where we want to use only the unperturbed heat-kernel theory. Write

$$H(r, x; s, y) := p_{\varepsilon}(s - r, x, y), \qquad r < s.$$

The kernel of the drifted equation is represented by the series

$$p^b(r, x; s, y) = H(r, x; s, y) + \sum_{n\ge1} P_n(r, x; s, y),$$

where

$$P_{n+1}(r, x; s, y) := \int_r^s \int_{\mathbb{S}^{d-1}} P_n(r, x; u, z)\, \mathcal{B}_u^{(z)} H(u, z; s, y)\, \mathrm{d}\omega(z)\, \mathrm{d}u,$$

and

$$\mathcal{B}_u^{(z)} H(u, z; s, y) := -\operatorname{div}_z(b_u(z)H(u, z; s, y)).$$

Equivalently, after integrating by parts in z, the derivative may be placed on P_n. The two forms are identical on the compact boundaryless manifold $\mathbb{S}^{d-1}$. Using (A.81)–(A.82), the semigroup property, and the compactness of $\mathbb{S}^{d-1}$, there is a constant $C = C(\mathbb{S}^{d-1}, m, \alpha, \tau)$ such that the standard parametrix estimate gives

$$\sup_{x\in\mathbb{S}^{d-1}} \|P_n(t - \tau, x; t, \cdot)\|_{C^{\alpha}(\mathbb{S}^{d-1})} \le C(\varepsilon\tau)^{-(m+\alpha)/2} \frac{\Xi(\varepsilon, \tau, R)^n}{\Gamma(1 + n/2)}, \tag{A.88}$$

where

$$\Xi(\varepsilon, \tau, R) := C\left(\frac{\Lambda_1(\tau, R)}{\sqrt{\varepsilon\tau}} + \frac{\Lambda_1(\tau, R)^2}{\varepsilon}\right). \tag{A.89}$$

Here the factor $\Gamma(1 + n/2)^{-1}$ is the usual Beta-integral gain from the successive heat-kernel convolutions. More explicitly, each insertion of the drift contributes one spatial derivative of the heat kernel and one factor $\|b_u\|_{C^1}$. The singular factors in time are integrable in the iterated simplex, and

the total contribution is bounded by the Beta-integral factor in (A.88). The additional quadratic term in (A.89) is a convenient way to dominate the commutator terms where derivatives hit the spatially varying drift.

Summing (A.88) over $n \geq 0$ and using

$$\sum_{n\geq 0} \frac{\Xi^n}{\Gamma(1+n/2)} \leq C_\alpha \exp\{C_\alpha(1+\Xi^2)\}$$

yields (A.86)–(A.87), after changing $C_{P,\alpha}$. The construction is local in coordinate charts, but no boundary terms appear because $\mathbb{S}^{d-1}$ is compact and without boundary; a partition of unity patches the local estimates. ■

Lemma A.38 (positive-time Hölder regularization). *Fix $\varepsilon > 0$, $\tau > 0$, and $R < \infty$. Let $s \mapsto \mu_s = \rho_s \omega$ be a controlled solution of (2.2) on $[t-\tau, t]$, started from an arbitrary probability measure at time $t-\tau$, and assume (A.84). Then μ_t has a density $\rho_t \in C^\alpha(\mathbb{S}^{d-1})$, with α as in Lemma A.37, and*

$$\|\rho_t\|_{C^\alpha(\mathbb{S}^{d-1})} \leq C_{\text{reg}}(\varepsilon, \tau, R). \tag{A.90}$$

Proof. By the transition-kernel representation for uniformly parabolic equations with bounded smooth drift,

$$\rho_t(y) = \int_{\mathbb{S}^{d-1}} p^b(t-\tau, x; t, y)\, \mathrm{d}\mu_{t-\tau}(x).$$

Taking the C^α_y-norm and using that $\mu_{t-\tau}$ is a probability measure gives (A.90) directly from (A.86). ■

Proof of Lemma A.34. We prove the positive branch. Fix $\tau_{\text{reg}} > 0$ and $R_{\text{reg}} < \infty$. Lemma A.38 gives a uniform bound

$$\|\rho_t\|_{C^\alpha(\mathbb{S}^{d-1})} \leq C_{\text{reg}}(\varepsilon, \tau_{\text{reg}}, R_{\text{reg}}) \tag{A.91}$$

for every endpoint of a controlled window satisfying (A.80).

Assume, for contradiction, that no positive η_{chart} exists. Then there are $\eta_n \downarrow 0$ and endpoint measures $\mu_n = \rho_n \omega \in \mathcal{C}_+(\eta_n)$ satisfying the window control bound, but $\mu_n \notin \mathcal{U}_+$. By (A.91), ρ_n is bounded in $C^\alpha(\mathbb{S}^{d-1})$. Passing to a subsequence, ρ_n converges uniformly to some continuous density ρ_∞. On the other hand, Lemma A.28 gives

$$W_2(\mu_n, \overline{\mu}^+) \leq \Theta_{\text{well}}(\eta_n) \to 0.$$

Hence $\mu_n \rightharpoonup \overline{\mu}^+ = \overline{\rho}^+ \omega$, and the uniform limit must satisfy $\rho_\infty = \overline{\rho}^+$. Therefore

$$\left\|\rho_n - \overline{\rho}^+\right\|_{L^\infty(\mathbb{S}^{d-1})} \to 0.$$

For fixed $\varepsilon > 0$, $\overline{\rho}^+$ is smooth and strictly positive on compact $\mathbb{S}^{d-1}$. If

$$m_\varepsilon^+ := \min_{\mathbb{S}^{d-1}} \overline{\rho}^+ > 0,$$

and if r_* is the density-chart radius in the definition of $\mathcal{U}_+$, then for all large n,

$$\left\|\rho_n - \overline{\rho}^+\right\|_{L^\infty(\mathbb{S}^{d-1})} < r_* m_\varepsilon^+.$$

Thus $h_n := \rho_n/\overline{\rho}^+ - 1$ satisfies $\|h_n\|_{L^\infty} < r_*$ and $\int h_n \,\mathrm{d}\overline{\mu}^+ = 0$, so $\mu_n = (1+h_n)\overline{\mu}^+ \in \mathcal{U}_+$, a contradiction. The negative branch follows by applying the antipodal map. ■

For definiteness, and to make all constants functions of the fixed data, we set

$$\tau_{\text{reg}} := 1, \qquad R_{\text{reg}} := R_{\text{opt}} + 1. \tag{A.92}$$

Let η_{chart} be the threshold supplied by Lemma A.34 with the choices in (A.92), and set

$$\eta_{\text{eff}} := \min\{\eta_{\text{sep}}, \eta_{\text{chart}}\}.$$

Lemma A.39 (regularized entry into the weak wells). *Fix $0 < \eta < \eta_{\text{sep}}$ satisfying (A.78). If*

$$L > \tau_{\text{reg}} + \frac{\mathfrak{D}_{\max}}{d_{\text{set}}(\eta)}, \tag{A.93}$$

and $T > L + \tau_{\text{reg}}$, then every global minimizer satisfies

$$\exists s_0^- \in [\tau_{\text{reg}}, L], \qquad \exists s_0^+ \in [T-L, T] \quad \textit{such that} \quad \mu_{s_0^-}^{*,T}, \mu_{s_0^+}^{*,T} \in \mathcal{N}_\eta.$$

Moreover, if $\eta \le \eta_{\text{chart}}$ and one of the selected times belongs to $\mathcal{C}_\varsigma(\eta)$, then the corresponding state belongs to $\mathcal{U}_\varsigma$.

Proof. If $\mu_t^{*,T} \notin \mathcal{N}_\eta$ for every $t \in [\tau_{\text{reg}}, L]$, then (A.74) and (A.76) imply

$$\int_0^T D(\mu_t^{*,T})\,\mathrm{d}t \ge (L - \tau_{\text{reg}}) d_{\text{set}}(\eta),$$

contradicting (A.75) and (A.93). The proof on $[T-L, T]$ is the same averaging argument. Since $T > L + \tau_{\text{reg}}$, every $s_0^+ \in [T-L, T]$ has a past regularizing window of length τ_{reg}. The final assertion follows from Lemma A.34, because the global control bound (A.73) gives

$$\int_{t-\tau_{\text{reg}}}^t \left\|W_s^{*,T}\right\|_{\text{F}}^2 \mathrm{d}s \le R_{\text{opt}} \le R_{\text{reg}}.$$

■

Lemma A.40 (same-branch entry from the sign-changing barrier). *Assume Assumption 3.6. Suppose that*

$$\eta_{\text{w},\varepsilon} \ge \frac{3}{4}\Delta_0, \tag{A.94}$$

which holds for all sufficiently small ε by Corollary A.31, and let $0 < \eta \le \Delta_0/4$. Let $(\mu^{*,T}, W^{*,T})$ be a global minimizer of (2.3). If

$$s_- < s_+, \qquad \mu_{s_-}^{*,T}, \mu_{s_+}^{*,T} \in \mathcal{N}_\eta,$$

then $\mu_{s_-}^{*,T}$ and $\mu_{s_+}^{*,T}$ belong to the same connected component of $\mathcal{N}_\eta$.

Proof. Assume, to get a contradiction, that $\mu_{s_-}^{*,T} \in \mathcal{C}_+(\eta)$ and $\mu_{s_+}^{*,T} \in \mathcal{C}_-(\eta)$, up to exchanging the two signs. By choosing the weak neighborhoods $\mathcal{V}_\pm$ inside the sign regions $\{m_1 > 0\}$ and $\{m_1 < 0\}$, respectively, we have

$$m_1(\mu_{s_-}^{*,T}) > 0, \qquad m_1(\mu_{s_+}^{*,T}) < 0.$$

Since $t \mapsto m_1(\mu_t^{*,T})$ is continuous, there is $r \in (s_-, s_+)$ such that $m_1(\mu_r^{*,T}) = 0$. Hence $\Delta_{\varepsilon,A}(\mu_r^{*,T}) \ge \eta_{\text{w},\varepsilon}$. Using (A.71) without the dissipative term gives

$$\eta_{\text{w},\varepsilon} - \eta \le \Delta_{\varepsilon,A}(\mu_r^{*,T}) - \Delta_{\varepsilon,A}(\mu_{s_-}^{*,T}) \le C_\sigma \int_{s_-}^r \left\|W_t^{*,T}\right\|_{\text{F}}^2 \mathrm{d}t.$$

By (A.94) and $\eta \le \Delta_0/4$, the control cost on $[s_-, r]$ is at least

$$\frac{\lambda_{\text{reg}}}{2} \int_{s_-}^r \left\|W_t^{*,T}\right\|_{\text{F}}^2 \mathrm{d}t \ge \frac{\lambda_{\text{reg}}}{2C_\sigma}\left(\frac{3}{4}\Delta_0 - \frac{1}{4}\Delta_0\right) = \frac{\lambda_{\text{reg}}\Delta_0}{4C_\sigma}.$$

Now keep the optimal trajectory on $[0, s_-]$ and set the control to zero on $[s_-, T]$. This gives an admissible competitor whose terminal cost can be worse by at most $\operatorname{osc}\ell$, while its control cost is smaller by at least $\lambda_{\text{reg}}\Delta_0/(4C_\sigma)$. This contradicts optimality because the second inequality in Assumption 3.6 is stronger than $\operatorname{osc}\ell < \lambda_{\text{reg}}\Delta_0/(4C_\sigma)$. Hence the two entry states must lie in the same component. ■

Proof of Lemma 3.7. Fix $\varepsilon \in (0, \varepsilon_0)$, where ε_0 is taken small enough for Corollaries A.31 and A.33 and Proposition A.24. The first inequality in Assumption 3.6 gives $\mathfrak{E}_{\max} \le \Delta_0 - \gamma_{\text{bar}}$ with $\gamma_{\text{bar}} := \Delta_0 - \mathfrak{E}_{\max} > 0$, so Corollary A.33 yields $d_{\text{set}}(\eta) > 0$ for every $0 < \eta < \eta_{\text{sep}}$. Choose $\eta \le \min\{\eta_{\text{chart}}, \Delta_0/4\}$ with η_{chart} from Lemma A.34, and let L satisfy (A.93). For $T > L + \tau_{\text{reg}}$, Lemma A.39 gives times $s_0^- \in [\tau_{\text{reg}}, L]$ and $s_0^+ \in [T-L, T]$ with $\mu^{*,T}_{s_0^\pm} \in \mathcal{N}_\eta$. By Corollary A.31, the hypothesis (A.94) holds, so Lemma A.40 places both states in the same component $\mathcal{C}_\varsigma(\eta)$ for one sign $\varsigma \in \{+,-\}$. Since $\eta \le \eta_{\text{chart}}$, the last assertion of Lemma A.39 gives $\mu^{*,T}_{s_0^\pm} \in \mathcal{U}_\varsigma$. This proves the lemma with $L_\varepsilon := L$, $s_\pm := s_0^\pm$, and r_* the radius of the density charts $\mathcal{U}_\pm$ in Proposition A.24. ∎

A.3.9 Adjoint stability along optimal trajectories

The remaining steps prove Theorem 3.9 by a direct analysis of the first-order optimality system of (2.3). The argument has four parts. First, the Hessian coercivity of Proposition 3.2 and the one-well Poincaré estimate of Proposition A.16 make the linearized generator $\mathsf{A} = \triangle_{\overline{\rho}}\mathsf{H}$ of (A.30) exponentially stable. Second, an exact divergence-form identity extends this stability to the linearization along an arbitrary controlled trajectory, with a perturbation measured only by the L^2-distance of the density to the selected well and by the instantaneous control cost; combined with the global budgets (A.73)–(A.75), the slope gap (A.78), and Lemma A.40, this yields exponential stability of the linearized evolution along every optimal trajectory, uniformly in the horizon. Third, under the smooth loss of Assumption 3.8, every optimal pair satisfies the first-order optimality system; by duality, the adjoint state ϕ inherits backward exponential decay, and the stationarity condition makes the optimal control exponentially small away from the terminal layer. Fourth, substituting this bound on $W^{*,T}$ into the strict dissipativity estimate (A.65) closes the turnpike bound.

Throughout this part we fix the sign $\varsigma \in \{+,-\}$ selected by Lemma A.39 and assume $\varsigma = +$ without loss of generality, by applying the antipodal map $S_\#$; we write $\overline{\mu} = \overline{\mu}^+ = \overline{\rho}\,\omega$, and we freely use the operators H, $\triangle_{\overline{\rho}}$, $\mathsf{A} = \triangle_{\overline{\rho}}\mathsf{H}$, and B from (A.30) and (A.31). By the Hessian coercivity (3.2) of Proposition 3.2,

$$c_H \|v\|^2_{L^2(\overline{\mu})} \le \langle \mathsf{H}v, v\rangle_{L^2(\overline{\mu})} \le \varepsilon \|v\|^2_{L^2(\overline{\mu})}, \qquad c_H := \frac{\varepsilon}{2}\Big(1 - \frac{\lambda_2}{\lambda_1}\Big), \qquad v \in L^2_0(\overline{\mu}), \tag{A.95}$$

and we use the norms $\|v\|^2_{\mathsf{H}} := \langle \mathsf{H}v, v\rangle_{L^2(\overline{\mu})}$ and $\|\phi\|^2_{\mathsf{H}^{-1}} := \langle \mathsf{H}^{-1}\phi, \phi\rangle_{L^2(\overline{\mu})}$, as in Proposition 3.5; by (A.95) both are equivalent to $\|\cdot\|_{L^2(\overline{\mu})}$. With the notation $(K_A * \nu)(x) = \int_{\mathbb{S}^{d-1}} K_A(x,y)\,\mathrm{d}\nu(y)$ of the landscape analysis, $\mathsf{H}v = \varepsilon v - \Pi_0[K_A * (v\overline{\mu})]$ by Proposition 3.2; since Π_0 changes a function only by an additive constant,

$$\nabla \mathsf{H}v = \varepsilon \nabla v - \nabla K_A * (v\overline{\mu}). \tag{A.96}$$

From $\nabla_x K_A(x,y) = \mathrm{e}^{\langle x, Ay\rangle} P^\perp_x(Ay)$ and the Cauchy–Schwarz inequality,

$$\|\nabla K_A * (v\overline{\mu})\|_{L^\infty(\mathbb{S}^{d-1})} \le C_A \|v\|_{L^2(\overline{\mu})}, \qquad C_A := \|A\|_{\text{op}} \mathrm{e}^{\|A\|_{\text{op}}}. \tag{A.97}$$

For an interval I, we say that v belongs to the *energy class* on I if $v \in C(\overline{I}; L^2_0(\overline{\mu})) \cap L^2(I; H^1(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu}))$ and $\partial_t v \in L^2(I; (H^1(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu}))')$; the analogous class over $L^2_0(\omega)$ is used for additive densities.

Lemma A.41 (linear well-posedness in the energy class)**.** *Let $I = [\tau, t_1]$, let $t \mapsto b_t$ be a tangential vector field with $\|b_\cdot\|_{C^1(\mathbb{S}^{d-1})} \in L^2(I)$, let $m \in L^\infty(I; L^2(\omega))$, $q_\tau \in L^2_0(\omega)$, and $f \in L^2(I; H^{-1}(\mathbb{S}^{d-1}))$. Then the linear equation*

$$\partial_t q = \varepsilon \triangle q - \operatorname{div}[q\, b_t] - \operatorname{div}\big[m_t\, P^\perp_x \chi_A[q\omega]\big] + f, \qquad q(\tau) = q_\tau, \tag{A.98}$$

has a unique solution in the energy class on I, and

$$\sup_{t\in I} \|q_t\|_{L^2(\omega)} + \|q\|_{L^2(I;H^1)} + \|\partial_t q\|_{L^2(I;H^{-1})} \le C\Big(\|q_\tau\|_{L^2(\omega)} + \|f\|_{L^2(I;H^{-1})}\Big), \tag{A.99}$$

where C depends only on ε, $t_1 - \tau$, A, $\|b_\cdot\|_{L^2(I;C^1)}$, and $\|m\|_{L^\infty(I;L^2)}$.

Proof. The smooth kernel gives $\big\|P^\perp_x \chi_A[q\omega]\big\|_{W^{1,\infty}(\mathbb{S}^{d-1})} \le C\|q\|_{L^2(\omega)}$, so testing (A.98) by q yields the Gårding-type inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\|q_t\|^2_{L^2(\omega)} + \varepsilon\|\nabla q_t\|^2_{L^2(\omega)} \le C\Big(1 + \|b_t\|^2_{C^1} + \|m_t\|^2_{L^2}\Big)\|q_t\|^2_{L^2(\omega)} + C\|f_t\|^2_{H^{-1}},$$

with a time-integrable coefficient. Existence, uniqueness, continuity in time, and (A.99) then follow from the standard Galerkin construction in the Gelfand triple $H^1 \subset L^2 \subset H^{-1}$; see (Lions and Magenes 1972). ∎

We next extend the stationary system (A.33) to a non-autonomous one. For a chart displacement $h \in L^2_0(\overline{\mu})$ with $\overline{\rho}(1+h) \geq 0$ and a parameter $W \in \mathbb{R}^{d\times p}$, define on $L^2_0(\overline{\mu})$

$$\mathsf{A}_{h,W} v \coloneqq \mathsf{A} v - \frac{1}{\overline{\rho}} \operatorname{div}\big\{\overline{\rho}\big(v\,\nabla K_A * (h\overline{\mu}) + h\,\nabla K_A * (v\overline{\mu}) + v\,P_x^{\perp}(W\sigma)\big)\big\}, \tag{A.100}$$

$$\mathsf{B}_h V \coloneqq -\frac{1}{\overline{\rho}} \operatorname{div}\big\{\overline{\rho}(1+h)\,P_x^{\perp}(V\sigma)\big\},$$

so that $\mathsf{A}_{0,0} = \mathsf{A}$ and $\mathsf{B}_0 = \mathsf{B}$.

Lemma A.42 (first variation along a controlled trajectory)**.** *Let $\rho_t = \overline{\rho}(1+h_t)$ solve (2.2) with control $W \in L^2$, and let $V \in L^2$. The first variation of the relative density in the direction V solves*

$$\partial_t v = \mathsf{A}_{h_t,W_t} v + \mathsf{B}_{h_t} V,$$

the non-autonomous form of (A.33). Equivalently, for every smooth $v \in L^2_0(\overline{\mu})$, the linearization $\mathcal{A}_{\rho,W}$ of the right-hand side of (2.2) at $(\rho\omega, W)$ satisfies the exact divergence-form identity

$$\frac{1}{\overline{\rho}}\,\mathcal{A}_{\rho,W}(\overline{\rho}v) = \mathsf{A}_{h,W} v. \tag{A.101}$$

In particular, the perturbation $\mathsf{A}_{h,W} - \mathsf{A}$ involves no derivative of the displacement h.

Proof. The linearization of the right-hand side of (2.2) at (μ, W), $\mu = \rho\,\omega$, acting on an additive density q, is

$$\mathcal{A}_{\rho,W} q = \varepsilon\triangle q - \operatorname{div}\big[q\,P_x^{\perp}(\chi_A[\mu] + W\sigma)\big] - \operatorname{div}\big[\rho\,P_x^{\perp}\chi_A[q\omega]\big],$$

and the control direction contributes $-\operatorname{div}(\rho\,P_x^{\perp}(V\sigma)) = \overline{\rho}\,\mathsf{B}_h V$. It therefore suffices to prove (A.101) for $q = \overline{\rho}v$. The minimizer satisfies the Euler–Lagrange equation $\overline{\rho} \propto \exp((K_A * \overline{\mu})/\varepsilon)$ (Proposition 3.2), hence

$$\varepsilon\nabla \log \overline{\rho} = \nabla K_A * \overline{\mu}. \tag{A.102}$$

Moreover, $P_x^{\perp}\chi_A[\mu] = \nabla K_A * [(1+h)\overline{\mu}]$ and $P_x^{\perp}\chi_A[\overline{\rho}v\,\omega] = \nabla K_A * (v\overline{\mu})$. Using (A.102),

$$\varepsilon\triangle(\overline{\rho}v) = \operatorname{div}[\varepsilon\nabla(\overline{\rho}v)] = \operatorname{div}\{\overline{\rho}(\varepsilon\nabla v + v\,\nabla K_A * \overline{\mu})\}.$$

Substituting these identities into $\mathcal{A}_{\rho,W}(\overline{\rho}v)$, the two terms containing $\nabla K_A * \overline{\mu}$ cancel, and

$$\mathcal{A}_{\rho,W}(\overline{\rho}v) = \operatorname{div}\{\overline{\rho}(\varepsilon\nabla v - \nabla K_A * (v\overline{\mu}))\} - \operatorname{div}\big\{\overline{\rho}\big(v\nabla K_A * (h\overline{\mu}) + h\nabla K_A * (v\overline{\mu}) + v\,P_x^{\perp}(W\sigma)\big)\big\}.$$

By (A.96), the first term equals $\operatorname{div}(\overline{\rho}\,\nabla \mathsf{H} v) = \overline{\rho}\,\triangle_{\overline{\rho}}\mathsf{H} v = \overline{\rho}\,\mathsf{A} v$. Dividing by $\overline{\rho}$ proves (A.101). ∎

Proposition A.43 (non-autonomous energy inequality)**.** *Let $I \subset [\tau_{\mathrm{reg}}, T]$, let $h \in L^\infty(I; L^2_0(\overline{\mu}))$ and $W \in L^2(I)$, and let v belong to the energy class on I and solve $\partial_t v = \mathsf{A}_{h_t,W_t} v$. Then, for almost every $t \in I$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\|v_t\|_{\mathsf{H}}^2 \leq \left[-\kappa_{\mathrm{lin}} + \frac{8C_A^2}{c_H}\|h_t\|_{L^2(\overline{\mu})}^2 + \frac{2\|\sigma\|_{L^\infty(\mathbb{S}^{d-1})}^2}{c_H}\|W_t\|_{\mathsf{F}}^2\right]\|v_t\|_{\mathsf{H}}^2, \qquad \kappa_{\mathrm{lin}} \coloneqq c_H\,\kappa_P(\varepsilon), \tag{A.103}$$

where $\kappa_P(\varepsilon)$ is the weighted Poincaré constant of Proposition A.14. By Proposition A.16 and (A.95),

$$\kappa_{\mathrm{lin}} \geq \frac{c_{\mathrm{lw}}(A,d)}{2}\left(1 - \frac{\lambda_2}{\lambda_1}\right) > 0, \tag{A.104}$$

uniformly as $\varepsilon \downarrow 0$.

Proof. Since H is bounded, symmetric, and maps $H^1(\mathbb{S}^{d-1}) \cap L^2_0(\overline{\mu})$ into itself, the Lions–Magenes product rule gives, with $z := \mathsf{H}v$,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_t\|_{\mathsf{H}}^2 = -\int_{\mathbb{S}^{d-1}} |\nabla z_t|^2 \,\mathrm{d}\overline{\mu} + \int_{\mathbb{S}^{d-1}} \nabla z_t \cdot \left(v_t \nabla K_A * (h_t\overline{\mu}) + h_t \nabla K_A * (v_t\overline{\mu}) + v_t\, u_{W_t}\right) \mathrm{d}\overline{\mu},$$

where the identity follows by testing (A.100) with z_t, using (A.29) for the A-part and integrating the divergence term by parts. By (A.97) and $|u_W| \le \|W\|_{\mathrm{F}}\|\sigma\|_{L^\infty}$,

$$\|v\nabla K_A * (h\overline{\mu}) + h\nabla K_A * (v\overline{\mu}) + v\, u_W\|_{L^2(\overline{\mu})} \le \Big(2C_A\|h\|_{L^2(\overline{\mu})} + \|\sigma\|_{L^\infty}\|W\|_{\mathrm{F}}\Big)\|v\|_{L^2(\overline{\mu})}.$$

Young's inequality absorbs half of the dissipation:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|v_t\|_{\mathsf{H}}^2 \le -\frac{1}{2}\int_{\mathbb{S}^{d-1}} |\nabla z_t|^2 \,\mathrm{d}\overline{\mu} + \Big(4C_A^2\|h_t\|_{L^2(\overline{\mu})}^2 + \|\sigma\|_{L^\infty}^2\|W_t\|_{\mathrm{F}}^2\Big)\|v_t\|_{L^2(\overline{\mu})}^2.$$

Since $z_t \in L^2_0(\overline{\mu})$, the weighted Poincaré inequality of Proposition A.14 and (A.95) give $\int|\nabla z_t|^2\,\mathrm{d}\overline{\mu} \ge \kappa_P(\varepsilon)\|z_t\|_{L^2(\overline{\mu})}^2 \ge \kappa_P(\varepsilon)\, c_H\|v_t\|_{\mathsf{H}}^2$ and $\|v_t\|_{L^2(\overline{\mu})}^2 \le c_H^{-1}\|v_t\|_{\mathsf{H}}^2$. Combining the three displays proves (A.103). ∎

We now apply Proposition A.43 along an optimal pair $(\mu^{*,T}, W^{*,T})$. Write $h_t^* := \rho_t^{*,T}/\overline{\rho} - 1$, and let $U(t,s)\colon L^2_0(\overline{\mu}) \to L^2_0(\overline{\mu})$, $\tau_{\mathrm{reg}} \le s \le t \le T$, denote the evolution family of $\partial_t v = \mathsf{A}_{h_t^*, W_t^{*,T}} v$, which is well defined in the energy class by Lemma A.41 (applied in the additive variable, with $b_t = P_x^{\perp}(\chi_A[\mu_t^{*,T}] + W_t^{*,T}\sigma)$, whose $L^2(I;C^1)$-norm is finite by Lemma A.36 and (A.73), and $m = \rho^{*,T} \in L^\infty(I;L^2(\omega))$ by Lemma A.38).

Proposition A.44 (uniform stability of the linearized evolution)**.** *There exist $\eta_{\mathrm{lin}} \in (0, \eta_{\mathrm{sep}})$ and $C_{\mathrm{evo}} < \infty$, independent of T, such that for every sufficiently large T and every optimal pair,*

$$\|U(t,s)v\|_{\mathsf{H}} \le C_{\mathrm{evo}}\, \mathrm{e}^{-\frac{\kappa_{\mathrm{lin}}}{4}(t-s)}\|v\|_{\mathsf{H}}, \qquad \tau_{\mathrm{reg}} \le s \le t \le T, \quad v \in L^2_0(\overline{\mu}). \tag{A.105}$$

Proof. By Lemma A.38 and $\min_{\mathbb{S}^{d-1}} \overline{\rho} > 0$,

$$G_{\mathrm{reg}} := \sup_{t \in [\tau_{\mathrm{reg}}, T]} \|h_t^*\|_{L^2(\overline{\mu})} \le \left(\frac{C_{\mathrm{reg}}^2}{\min_{\mathbb{S}^{d-1}} \overline{\rho}} + 1\right)^{1/2} < \infty,$$

uniformly in T. Choose

$$0 < \eta_{\mathrm{lin}} \le \min\left\{\eta_{\mathrm{chart}},\ \frac{c_1\, c_H\, \kappa_{\mathrm{lin}}}{16\, C_A^2}\right\}, \tag{A.106}$$

with c_1 from (A.59) and η_{chart} from Lemma A.34, and split $[\tau_{\mathrm{reg}}, T] = G_T \cup B_T$, where $G_T := \{t : \mu_t^{*,T} \in \mathcal{C}_+(\eta_{\mathrm{lin}})\}$. On G_T: by Lemma A.40 every low-energy time lies in the component of the selected sign; by Lemma A.34 the state lies in the density chart; hence (A.59) gives $\|h_t^*\|_{L^2(\overline{\mu})}^2 \le \Delta_{\varepsilon,A}(\mu_t^{*,T})/c_1 \le \eta_{\mathrm{lin}}/c_1$, so that, by (A.106), the h^*-term in (A.103) is at most $\kappa_{\mathrm{lin}}/2$. On B_T: integrating (A.65) and using (A.75) and the slope gap (A.78) gives

$$|B_T| \le M_{\mathrm{lin}} := \frac{\mathfrak{D}_{\max}}{d_{\mathrm{set}}(\eta_{\mathrm{lin}})},$$

and the h^*-term is bounded by $(8C_A^2/c_H)G_{\mathrm{reg}}^2$ there. Integrating (A.103) over $[s,t]$, splitting into G_T and B_T, and using the control budget (A.73) yields

$$\|v_t\|_{\mathsf{H}}^2 \le \|v_s\|_{\mathsf{H}}^2 \exp\left\{-\frac{\kappa_{\mathrm{lin}}}{2}|G_T \cap [s,t]| + \frac{8C_A^2 G_{\mathrm{reg}}^2}{c_H} M_{\mathrm{lin}} + \frac{2\|\sigma\|_{L^\infty}^2}{c_H} R_{\mathrm{opt}}\right\}.$$

Since $|G_T \cap [s,t]| \ge (t-s) - M_{\mathrm{lin}}$, the estimate (A.105) follows with $C_{\mathrm{evo}}^2 = \exp\{(\frac{\kappa_{\mathrm{lin}}}{2} + \frac{8C_A^2 G_{\mathrm{reg}}^2}{c_H})M_{\mathrm{lin}} + \frac{2\|\sigma\|_{L^\infty}^2}{c_H} R_{\mathrm{opt}}\}$, after decreasing the exponent from $\kappa_{\mathrm{lin}}/2$ to $\kappa_{\mathrm{lin}}/4$ to absorb the constant on unit time scales. ∎

A.3.10 The optimality system after the regularization time

We first record that the optimal trajectory itself belongs to the energy class after the regularization time. Recall that for every $W \in L^2(0,T)$ and every initial distribution, (2.2) has a unique measure-valued weak solution: the drift $(x,\mu) \mapsto P_x^\perp(\chi_A[\mu](x) + W_t\sigma(x))$ is Lipschitz in x and in μ with respect to the Wasserstein distance, with time-integrable Lipschitz constants, so existence and uniqueness follow from the classical McKean–Vlasov theory; see, e.g., (Chaintron and Diez 2022).

Lemma A.45 (the optimal trajectory is an energy solution)**.** *Let $W \in L^2(0,T)$ and let $(\mu_t)_t$ be the weak solution of (2.2) whose density at time $\tau \geq \tau_{\text{reg}}$ satisfies $\rho_\tau \in C^\alpha(\mathbb{S}^{d-1})$. Then $(\rho_t)_{t\in[\tau,T]}$ coincides with the unique energy-class solution of the corresponding Fokker–Planck equation with datum ρ_τ.*

Proof. Freeze the drift $b_t := P_x^\perp(\chi_A[\mu_t] + W_t\sigma)$ along the given solution; by Lemma A.36, $\|b_\cdot\|_{C^2(\mathbb{S}^{d-1})} \in L^2(\tau,T)$. Both $(\mu_t)_t$ and the energy solution supplied by Lemma A.41 (with $m=0$ and this b) are distributional solutions of the same *linear* Fokker–Planck equation with the same initial datum. For such drifts the linear equation has at most one solution in the class of finite measures: given $\varphi \in C^\infty(\mathbb{S}^{d-1})$, the backward equation $-\partial_t\psi = \varepsilon\triangle\psi + b_t\cdot\nabla\psi$ has a solution $\psi \in C([\tau,t_1];C^2)$ for every terminal datum $\psi(t_1)=\varphi$, and $t \mapsto \int \psi_t \,\mathrm{d}\mu_t$ is constant along any distributional solution, which determines μ_{t_1}; see (Bogachev et al. 2015) for this duality argument. Hence the two solutions coincide. ■

By Assumption 3.8 and Lemma A.38, the set of probability densities ρ with $\|\rho\|_{C^\alpha} \leq C_{\text{reg}}$ is compact in $L^2(\omega)$ and contained in the open set on which ℓ is twice differentiable. Writing, in analogy with g_ℓ of Lemma 3.4,

$$g_T^* := \Pi_0^{\overline{\mu}} \frac{\delta\ell}{\delta\mu}\Big[\mu_T^{*,T}\Big] \in L_0^2(\overline{\mu}), \qquad L_\ell := \sup\left\{ \left\| \Pi_0^{\overline{\mu}} \frac{\delta\ell}{\delta\mu}[\rho\omega] \right\|_{\mathsf{H}^{-1}} : \|\rho\|_{C^\alpha} \leq C_{\text{reg}} \right\} < \infty, \tag{A.107}$$

the bound $L_\ell < \infty$ follows from this compactness and the continuity of the derivative; no additional quantitative assumption on ℓ is used.

Proposition A.46 (differentiability of the control-to-state map)**.** *Fix $\tau = \tau_{\text{reg}}$ and a nonnegative unit-mass datum $\rho_\tau \in L^2(\omega)$. For $W \in L^2(\tau,T)$ let ρ^W be the energy solution of (2.2) on $[\tau,T]$ with $\rho^W(\tau) = \rho_\tau$. Then the map $W \mapsto \rho^W$ is Fréchet differentiable from $L^2(\tau,T)$ to the energy class, on every bounded set of controls: the derivative $q = q[V]$ in the direction V is the energy solution of*

$$\partial_t q = \mathcal{A}_{\rho_t^W, W_t} q - \operatorname{div}\big[\rho_t^W\, P_x^\perp(V_t\sigma)\big], \qquad q(\tau) = 0,$$

and, for $\|V\|_{L^2(\tau,T)}$ small,

$$\sup_{t\in[\tau,T]} \big\|\rho_t^{W+V} - \rho_t^W - q_t[V]\big\|_{L^2(\omega)} \leq C_{\text{diff}} \|V\|_{L^2(\tau,T)}^2, \tag{A.108}$$

where C_{diff} depends on $T-\tau$ and the control bound but not otherwise on (W,V).

Proof. Testing (2.2) by ρ and using that $\big\|P_x^\perp\chi_A[\mu]\big\|_{W^{1,\infty}}$ is bounded uniformly over $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ gives $\sup_t \big\|\rho_t^W\big\|_{L^2} \leq M_R$ on every control ball. The difference $\delta\rho := \rho^{W+V} - \rho^W$ solves (A.98) with $b_t = P_x^\perp(\chi_A[\rho_t^{W+V}\omega] + (W_t+V_t)\sigma)$, $m = \rho^W$, $q_\tau = 0$, and source $f = -\operatorname{div}(\rho^W P_x^\perp(V\sigma))$, whose $L_t^2 H^{-1}$-norm is bounded by $CM_R\|V\|_{L^2}$; hence $\|\delta\rho\| \leq C_{\text{diff}}\|V\|_{L^2}$ in the energy norms by Lemma A.41. The remainder $r := \delta\rho - q$ solves the same linear equation with source $-\operatorname{div}\big[\delta\rho\, P_x^\perp\chi_A[\delta\rho\,\omega]\big] - \operatorname{div}\big[\delta\rho\, P_x^\perp(V\sigma)\big]$, whose $L_t^2H^{-1}$-norm is bounded by $C(\|\delta\rho\|_{L^\infty L^2}\|\delta\rho\|_{L^2L^2} + \|\delta\rho\|_{L^\infty L^2}\|V\|_{L^2}) \leq C_{\text{diff}}\|V\|_{L^2}^2$. Applying Lemma A.41 once more proves (A.108). ■

Proposition A.47 (adjoint equation and stationarity)**.** *Let $(\mu^{*,T}, W^{*,T})$ be an optimal pair. There is a unique adjoint state ϕ in the energy class on $[\tau_{\text{reg}},T]$ solving, in the weak sense,*

$$-\partial_t\phi = \Big(\mathsf{A}_{h_t^*, W_t^{*,T}}\Big)^\dagger \phi, \qquad \phi(T) = g_T^*, \tag{A.109}$$

where $\dagger$ denotes the $L^2(\overline{\mu})$-adjoint as in (A.30), and the optimal control satisfies the stationarity condition

$$\lambda_{\mathrm{reg}} W_t^{*,T} + \int_{\mathbb{S}^{d-1}} \rho_t^{*,T}(x)\, \nabla\phi_t(x)\, \sigma(x)^\top \,\mathrm{d}\omega(x) = 0 \qquad \text{for a.e. } t \in (\tau_{\mathrm{reg}}, T). \tag{A.110}$$

Proof. After reversing time, (A.109) is a linear parabolic equation whose bilinear form satisfies the same boundedness and Gårding bounds as in Lemma A.41: the only term requiring care is the control drift, for which an integration by parts moves the derivative off the test function at the cost of the factor $\|u_{W_t}\|_{L^\infty} + \|\mathrm{div}(\overline{\rho}\, u_{W_t})/\overline{\rho}\|_{L^\infty} \le C\|W_t\|_{\mathrm{F}} \in L^2(\tau_{\mathrm{reg}}, T)$. Hence the time-reversed Galerkin construction of Lemma A.41 applies verbatim and yields the unique backward energy solution ϕ.

For the stationarity condition, take $V \in L^2(\tau_{\mathrm{reg}}, T)$, extended by zero to $(0, \tau_{\mathrm{reg}})$, so that the perturbed trajectory agrees with the optimal one up to τ_{reg}; by Lemma A.45, on $[\tau_{\mathrm{reg}}, T]$ the states are the energy solutions of Proposition A.46. For small perturbations the terminal density stays in the open set of Assumption 3.8, so the chain rule, (A.108), and the substitution $q = \overline{\rho} v$ of Lemma A.42 give

$$DJ(W^{*,T})[V] = \int_{\mathbb{S}^{d-1}} g_T^*\, v_T \,\mathrm{d}\overline{\mu} + \lambda_{\mathrm{reg}} \int_{\tau_{\mathrm{reg}}}^T W_t^{*,T} : V_t \,\mathrm{d}t.$$

The Lions–Magenes product rule shows that $t \mapsto \int v_t \phi_t \,\mathrm{d}\overline{\mu}$ is absolutely continuous with derivative $\langle \mathsf{B}_{h_t^*} V_t, \phi_t \rangle_{L^2(\overline{\mu})} = \left(\int_{\mathbb{S}^{d-1}} \rho_t^{*,T} \nabla\phi_t\, \sigma^\top \,\mathrm{d}\omega\right) : V_t$, whence $\int g_T^* v_T \,\mathrm{d}\overline{\mu} = \int_{\tau_{\mathrm{reg}}}^T \left(\int_{\mathbb{S}^{d-1}} \rho_t^{*,T} \nabla\phi_t\, \sigma^\top \,\mathrm{d}\omega\right) : V_t \,\mathrm{d}t$. Since $DJ(W^{*,T})[V] = 0$ for all such V, the fundamental lemma of the calculus of variations gives (A.110). ∎

A.3.11 Adjoint decay and proof of Theorem 3.9

Proposition A.48 (backward decay of the adjoint state)**.** *With ϕ as in Proposition A.47,*

$$\|\phi_t\|_{\mathsf{H}^{-1}} \le C_{\mathrm{evo}}\, L_\ell\, \mathrm{e}^{-\frac{\kappa_{\mathrm{lin}}}{4}(T-t)}, \qquad \|\phi_t\|_{L^2(\overline{\mu})} \le (\varepsilon/c_H)^{1/2}\, C_{\mathrm{evo}}\, L_\ell\, \mathrm{e}^{-\frac{\kappa_{\mathrm{lin}}}{4}(T-t)}, \qquad t \in [\tau_{\mathrm{reg}}, T].$$

Proof. By duality, $\phi_t = U(T,t)^\dagger g_T^*$, and the Hilbert dual of $(L_0^2(\overline{\mu}), \|\cdot\|_{\mathsf{H}})$ under the $L^2(\overline{\mu})$ pairing is $(L_0^2(\overline{\mu}), \|\cdot\|_{\mathsf{H}^{-1}})$. The first bound follows from Proposition A.44 and (A.107); the second from the norm equivalence (A.95). ∎

Lemma A.49 (unit-window gradient estimate)**.** *There is $C_\nabla < \infty$, independent of T, such that for $s \in [\tau_{\mathrm{reg}}, T-1]$,*

$$\int_s^{s+1} \|\nabla\phi_t\|_{L^2(\overline{\mu})}^2 \,\mathrm{d}t \le C_\nabla \sup_{t\in[s,s+1]} \|\phi_t\|_{L^2(\overline{\mu})}^2.$$

Proof. Testing (A.109) by ϕ gives, by the Gårding bound of the adjoint form,

$$\frac{\varepsilon}{2} \int_s^{s+1} \|\nabla\phi_t\|_{L^2(\overline{\mu})}^2 \,\mathrm{d}t \le \frac{1}{2}\|\phi_{s+1}\|_{L^2(\overline{\mu})}^2 + C \int_s^{s+1} \left(1 + \left\|W_t^{*,T}\right\|_{\mathrm{F}}^2\right) \|\phi_t\|_{L^2(\overline{\mu})}^2 \,\mathrm{d}t,$$

and the control budget (A.73) bounds the time integral of the coefficient uniformly in T. ∎

Corollary A.50 (exponential smallness of the optimal control)**.** *There is $C_{\mathrm{tail}} < \infty$, independent of T, such that*

$$\int_s^{s+1} \left\|W_t^{*,T}\right\|_{\mathrm{F}}^2 \,\mathrm{d}t \le C_{\mathrm{tail}}\, \mathrm{e}^{-\frac{\kappa_{\mathrm{lin}}}{2}(T-s)}, \qquad s \in [\tau_{\mathrm{reg}}, T-1], \tag{A.111}$$

and hence, for every $L \ge 1$, $\int_{\tau_{\mathrm{reg}}}^{T-L} \left\|W_t^{*,T}\right\|_{\mathrm{F}}^2 \,\mathrm{d}t \le C'_{\mathrm{tail}}\, \mathrm{e}^{-\frac{\kappa_{\mathrm{lin}}}{2}L}$.

Proof. By (A.110), the Cauchy–Schwarz inequality, the uniform bound $\left\|\rho_t^{*,T}\right\|_{L^\infty} \le C_{\mathrm{reg}}$, and $\min_{\mathbb{S}^{d-1}} \overline{\rho} > 0$, $\lambda_{\mathrm{reg}}^2 \left\|W_t^{*,T}\right\|_{\mathrm{F}}^2 \le C\|\nabla\phi_t\|_{L^2(\overline{\mu})}^2$. Integrate over the unit window and apply Lemma A.49 and Proposition A.48. Summing the window bound over unit intervals gives the tail bound. ∎

Lemma A.51 (persistence in the chart away from the terminal layer)**.** *There is $L_{\text{end}} \geq 1$, independent of T, such that for all sufficiently large T, the optimal trajectory satisfies $\mu_t^{*,T} \in \mathcal{C}_+(\eta_{\text{chart}})$ and lies in the density chart for every $t \in [s_0^-, T - L_{\text{end}}]$, where $s_0^- \leq L$ is the entry time of Lemma A.39.*

Proof. Let a_{loc} and C_σ denote the constants a and C_σ in (A.65). Choose L_{end} so large that $\eta_{\text{lin}} + C_\sigma C'_{\text{tail}}\, \mathrm{e}^{-\frac{\kappa_{\text{lin}}}{2} L_{\text{end}}} < \eta_{\text{chart}}$. On the maximal interval on which the trajectory stays in the chart, the variation-of-constants form of (A.65) and Corollary A.50 give $\Delta_{\varepsilon,A}(\mu_t^{*,T}) \leq \eta_{\text{lin}} + C_\sigma \int_{s_0^-}^{T-L_{\text{end}}} \|W_r^{*,T}\|_{\text{F}}^2 \,\mathrm{d}r < \eta_{\text{chart}}$ for $t \leq T - L_{\text{end}}$. Since every such time has a regularizing window of length τ_{reg} with control budget (A.73), Lemma A.34 keeps the trajectory in the chart, and Proposition A.27 and Lemma A.40 exclude a switch of component. Hence the maximal interval extends to $T - L_{\text{end}}$. ∎

Proof of Theorem 3.9. Let $s_0^- \leq L$ be the entry time of Lemma A.39 and L_{end} the constant of Lemma A.51. For $t \in [s_0^-, T - L_{\text{end}}]$, the variation-of-constants form of (A.65) on the chart gives

$$\Delta_{\varepsilon,A}(\mu_t^{*,T}) \leq \mathrm{e}^{-a_{\text{loc}}(t-s_0^-)}\, \eta_{\text{lin}} + C_\sigma \int_{s_0^-}^{t} \mathrm{e}^{-a_{\text{loc}}(t-r)} \|W_r^{*,T}\|_{\text{F}}^2 \,\mathrm{d}r.$$

Splitting the convolution into unit windows and applying (A.111) on each yields $\int_{s_0^-}^{t} \mathrm{e}^{-a_{\text{loc}}(t-r)} \|W_r^{*,T}\|_{\text{F}}^2 \,\mathrm{d}r \leq C\, \mathrm{e}^{-\frac{1}{2}\min\{a_{\text{loc}}, \kappa_{\text{lin}}/2\}(T-t)} + C\, \mathrm{e}^{-\frac{a_{\text{loc}}}{2}(t-s_0^-)}$, by summing two geometric series. Since $s_0^- \leq L$ is bounded and $\Delta_{\varepsilon,A} \leq \mathfrak{E}_{\max}$ on the two boundary layers $[0, s_0^-]$ and $[T - L_{\text{end}}, T]$ of T-independent length, the bound of Theorem 3.9 follows with $a := \frac{1}{2}\min\{a_{\text{loc}}, \kappa_{\text{lin}}/2\}$ and a constant C depending only on $\mathfrak{E}_{\max}$, a_{loc}, κ_{lin}, C_σ, C_{tail}, L, and L_{end}. ∎

Remark A.52 (explicit size of the turnpike constants). All constants above are independent of the horizon T. The exponential rate is $a = \frac{1}{2}\min\{a_{\text{loc}}, \kappa_{\text{lin}}/2\}$, where a_{loc} is the local PL rate of (A.65) and, by (A.104), $\kappa_{\text{lin}} \geq \frac{c_{\text{1w}}(A,d)}{2}(1 - \lambda_2/\lambda_1)$ uniformly as $\varepsilon \downarrow 0$; the A-dependence of both quantities is governed by the spectral-gap factor $1 - \lambda_2/\lambda_1$ and by the one-well scale $c_{\text{1w}}(A,d) \gtrsim \lambda_1 \mathrm{e}^{-\lambda_1}$ (Proposition A.16 and Remark A.23). The prefactor is controlled by C_{evo}, that is, by the bad-time measure M_{lin}, the budgets (A.73)–(A.75), and the terminal-gradient bound L_ℓ of (A.107).

B Supplementary numerical experiments

B.1 Implementation details

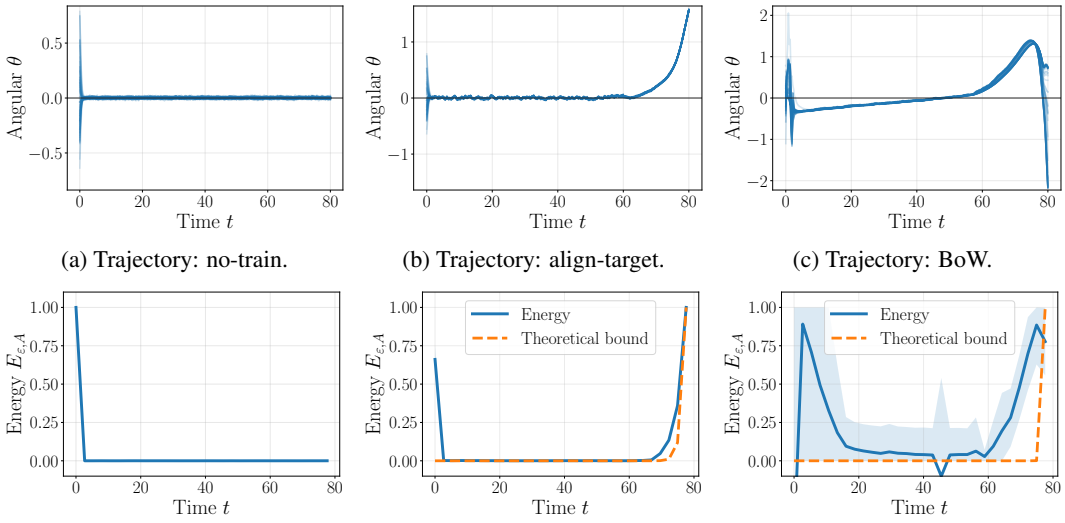
We solve the particle dynamics in (2.1) with Diffrax (Kidger 2022) and optimize the control parameters with Adam (Kingma and Ba 2015). Unless otherwise stated, we use horizon $T = 80$, inverse temperature $\beta = 1$, and $N = 64$ particles in the projected $d = 2$ setting with $A = \mathrm{diag}(1, \lambda_2)$. The control input u_W is piecewise constant with bin width $\Delta t_{\text{control}} = 0.25$, trained for 600 steps with learning rate 0.01. For the numerical optimization runs reported here we set $\lambda_{\text{reg}} = 0$; this choice is used only for the exploratory particle simulations and is separate from the positive-regularization assumptions in the turnpike theorem.

The initial particles are sampled by first drawing independent standard normal vectors ξ_i, then projecting

$$X_i(0) = \frac{e_1 + s\xi_i}{\|e_1 + s\xi_i\|}$$

onto the sphere, with $s = 0.3$. Thus, the initial empirical law is concentrated in a cap around e_1. For trajectory diagnostics on $\mathbb{S}^1$, we choose the stationary reference direction of each run and represent each particle by its signed angle $\theta_i(t)$ relative to that direction. The energy plots use $E_{\varepsilon,A}$ from (3.1).

For the rate comparison panels, we fit the late-time energy lift to $C\exp(-a(T-t))$. The theoretical rate is computed by solving the stationary problem on an $\mathbb{S}^1$ grid, linearizing the dynamics around the



(a) Trajectory: no-train. (b) Trajectory: align-target. (c) Trajectory: BoW.

(d) Energy: no-train. (e) Energy: align-target. (f) Energy: BoW.

Figure 6: Particle trajectories and energy time evolution. Panels (a)–(c) show all-particle angular trajectories in stationary-reference coordinates for no-train, align-target, and BoW runs. Panel (d) shows the corresponding no-train energy trajectory. Panels (e) and (f) show the normalized observed energy against the theoretical bound for representative align-target and BoW runs.

stationary density, and evaluating the rate diagnostic from Proposition 3.5. For numerical stability, the rate-comparison runs and the theoretical rate calculation use $\varepsilon = 0.1$, while the trajectory and energy-overlay panels use $\varepsilon = 10^{-4}$ unless otherwise stated. All experiments were run on an Apple M4 Max MacBook with 128GB RAM.

B.2 Experiment 0: Learning induces a turnpike

We verify that the turnpike profile is a learned effect rather than a property of the uncontrolled dynamics. Panels (a)–(c) of Figure 6 plot every particle trajectory on $\mathbb{S}^1$ in the signed angular coordinate θ, measured relative to a stationary reference angle. Panel (a) shows the untrained setting, while panels (b) and (c) show the trained counterparts: (b) the align-target trajectory and (c) the BoW trajectory. All three panels share the common parameters $\lambda_2 = 0.65$ and $\varepsilon = 10^{-4}$. The BoW run additionally uses $\kappa_{\text{cand}} = 2$. In the no-train case, the particles start from a scattered initial configuration, rapidly concentrate to a single location, and subsequently remain near $\theta = 0$. In contrast, the align-target trajectory shows particles that, after concentrating, remain clustered until the late layers before undergoing a coherent displacement at the terminal time. The BoW trajectory likewise shows an initial concentration followed by a dispersion near the terminal time. Panels (d)–(f) display the corresponding energy $E_{\varepsilon,A}$ in (3.1). Panel (d) shows the no-train energy trajectory, while panels (e) and (f) compare the observed energy curve with the theoretical bound in Proposition 3.5. Here, to ensure numerical stability, the lower bound in Proposition 3.5 is computed at $\varepsilon = 0.1$. The observed energy is evaluated with entropy coefficient $\varepsilon = 10^{-2}$ in panel (d), $\varepsilon = 10^{-4}$ in panel (e), and $\varepsilon = 10^{-1}$ in panel (f), while all underlying trajectories are simulated at $\varepsilon = 10^{-4}$. For readability, we display normalized values for both the observed energy and the theoretical bound. These results demonstrate that training gives rise to the turnpike profile.

B.3 Robustness protocol and full tables

This appendix records the protocol behind the robustness summary in Section 4. The sweeps are summarized in Table 2; the per-phase counts and the fitted quantities follow. At the reference setting, 20/20 runs satisfy the criterion, with fitted terminal rate 0.246 ± 0.007 and terminal angular displacement 1.554 ± 0.010 (mean $\pm$ SD). At $\varepsilon = 10^{-1}$ the clustering and the plateau degrade because the equilibrium cap width $\sigma_\varepsilon = \sqrt{\varepsilon/(e^{\lambda_1}\lambda_1)} \approx 0.19$ becomes comparable to the initial spread ≈ 0.3, while the terminal response persists in 20/20 runs. Across horizons the plateau duration

grows ($32.0 \to 71.2 \to 109.7$) while the initial and terminal widths vary mildly, consistent with boundary layers of T-independent width; across targets the plateau is robust whereas the strength and timing of the escape are target-dependent; under a global random initialization, the profile requires early entry into a single low-energy component, and regimes with additional metastable states lie outside our theorems.

Table 2: Robustness of the three-phase profile (20 runs per configuration; the fixed criterion is stated in Appendix B.3).

Sweep	Range	Full profile
Seeds (reference setting)	20 seeds	20/20
Diffusion ε	$\{0,\ 10^{-4},\ 10^{-3},\ 10^{-2}\}$	79/80
	10^{-1}	0/20 (terminal response 20/20)
Horizon T	$\{40, 80, 120\}$	60/60
Particles N	$\{32, 64, 128\}$	60/60
Regularization λ_{reg}	0.01	20/20
Targets κ_{cand}	$\{2, 3, 5, 7, 8, 9, 10\}$	138/140 (plateau 140/140)

Fixed three-phase criterion. Let $E(t)$ denote the energy $E_{\varepsilon,A}$ reconstructed along a run, $R_E := \max_t E(t) - \min_t E(t)$, and $E_{\mathrm{pl}} := \mathrm{median}_{t\in[0.4T,0.6T]}\, E(t)$. A run satisfies the criterion when

$$\frac{E(0) - E_{\mathrm{pl}}}{R_E} \geq 0.1, \qquad \frac{\mathrm{SD}_{t\in[0.4T,0.6T]}\, E(t)}{R_E} \leq 0.05, \qquad \frac{E(T) - E_{\mathrm{pl}}}{R_E} \geq 0.1,$$

where the standard deviation is the sample standard deviation on the interior window; no smoothing is applied, and the criterion is used without retuning across all sweeps. The initial width is the first time in $[0, 0.5T]$ at which 90% of the initial-to-plateau decrease has occurred; the terminal width is $T - t_{\mathrm{esc}}$, where t_{esc} is the first time in $[0.5T, T]$ at which 10% of the plateau-to-terminal increase is reached; the plateau duration is the interval between these crossings. The terminal rate a is fitted on $[0.7T, T]$ from $E(t) - E_{\mathrm{pl}} \approx C\mathrm{e}^{-a(T-t)}$ whenever the terminal lift is positive. Since ε enters $E_{\varepsilon,A}$, only within-run normalized quantities are compared across different ε.

Full tables. Diffusion sweep (20 runs per value; per-phase counts):

ε	initial clustering	plateau	terminal response	full profile
0	20/20	20/20	20/20	20/20
10^{-4}	20/20	20/20	20/20	20/20
10^{-3}	20/20	20/20	20/20	20/20
10^{-2}	20/20	19/20	20/20	19/20
10^{-1}	0/20	13/20	20/20	0/20

Horizon sweep (20 runs per value, all satisfying the full criterion):

T	plateau duration	initial width	terminal width
40	32.0	≈ 2.67	5.33
80	71.2	≈ 2.67	6.15
120	109.7	≈ 2.67	7.62

Target sweep: for $\kappa_{\mathrm{cand}} \in \{2, 3, 5, 7, 8, 9, 10\}$ (20 runs each), the interior-plateau criterion holds in $140/140$ runs and the full criterion in $138/140$, while the mean terminal angular displacement decreases from 0.763 ± 0.343 at $\kappa_{\mathrm{cand}} = 2$ to 0.169 ± 0.006 at $\kappa_{\mathrm{cand}} = 10$. Basin sensitivity: replacing the cap initialization by a global random initialization and classifying plateaus by the plateau-center norm (single well: ≥ 0.9; mixed well: ≤ 0.5; intermediate otherwise), 13 of 20 runs entered a single-well plateau, 6 remained mixed, and 1 was intermediate; the full criterion held in $11/13$ single-well and $0/6$ mixed-well runs. Supplementary checks: $N \in \{32, 64, 128\}$ give $20/20$ each; $\lambda_{\mathrm{reg}} = 0.01$ gives $20/20$ with terminal displacement 1.560 ± 0.008 (versus 1.554 ± 0.010 unregularized); an exploratory particle-model screening over $d \in \{2, 3, 8, 16\}$ gives $5/5$ at each dimension.

NeurIPS Paper Checklist

(i) **Claims**

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

Answer: [Yes]

Justification: The abstract and Section 1 summarize the contributions of this paper, citing section names within the paper.

Guidelines:

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

(ii) **Limitations**

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

Answer: [Yes]

Justification: See Section 5.

Guidelines:

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

(iii) **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions, Assumptions 3.1, 3.3, 3.6, and 3.8, are placed before each theorem, and the proof is provided in Appendix A.

Guidelines:

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

(iv) **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The numerical setup is specified in Section 4, and all source code will be included with a single script that regenerates the data and plots used in the reported experiments.

Guidelines:

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

(v) **Open access to data and code**

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

Answer:[Yes]

Justification: All source code will be included with the submission, and a single reproduction script regenerates the numerical data and the plots reported in the paper.

Guidelines:

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

(vi) **Experimental setting/details**

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

Answer: [Yes]

Justification: Section 4 reports the particle model, optimizer, horizon, control discretization, learning rate, noise levels, initialization, and swept parameters used in the numerical experiments.

Guidelines:

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

(vii) **Experiment statistical significance**

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

Answer: [Yes]

Justification: Section 4 reports parameter sweeps and fixed-seed representative runs supporting the numerical claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

(viii) **Experiments compute resources**

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

Answer: [Yes]

Justification: Section 4 reports the hardware used for the experiments, including the processor and memory.

Guidelines:

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

(ix) **Code of ethics**

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

Answer: [Yes]

Justification: The work is theoretical and uses only synthetic numerical simulations, with no human subjects, personal data, high-risk datasets, or deployed systems.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

(x) **Broader impacts**

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

Answer: [N/A]

Justification: It is a theoretical paper.

Guidelines:

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

(xi) **Safeguards**

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

Answer: [N/A]

Justification: It is a theoretical paper.

Guidelines:

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

(xii) **Licenses for existing assets**

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

Answer: [N/A]

Justification: It is a theoretical paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

(xiii) **New assets**

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

Answer: [Yes]

Justification: The source code for the numerical experiments will be submitted together with documentation describing how to reproduce the reported results.

Guidelines:

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

(xiv) **Crowdsourcing and research with human subjects**

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

(xvi) **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core proof strategy, theorem statements, and mathematical verification were developed by the authors. Any LLM usage was only auxiliary and did not affect the core methodology, scientific rigor, or originality of the work.

Guidelines:

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